Universitat Politècnica de València

Tesis Doctoral

Author Profiling en Social Media: Identificación de Edad, Sexo y Variedad del Lenguaje

Autor:
Francisco Manuel Rangel Pardo

Director:
Dr. Paolo Rosso

Junio 2016

Agradecimientos

A mi familia, por enseñarme de la vida
A Autoritas, por la libertad
A Socialancer, por la inspiración
A Paolo, por su amistad
A mi princesita, por ser mi mundo
A mi muñequita, por ser mi alegría
A mi pincesa, por serlo todo

Este trabajo ha sido parcialmente financiado por Autoritas Consulting SA (http://www.autoritas.net) y por el Ministerio de Economía y Competitividad de España bajo los códigos ECOPORTUNITY IPT-2012-1220-430000. La tarea de Author Profiling del PAN ha sido organizada en el marco del proyecto WIQ-EI IRSES (Grant No. 269180) dentro del 7FP Marie Curie People Framework of the European Commission.

Resumen

La posibilidad de conocer rasgos de una persona a partir únicamente de los textos que escribe se ha convertido en un área de gran interés denominada *author profiling*. Ser capaz de inferir de un usuario su sexo, edad, idioma nativo o los rasgos de su personalidad, simplemente analizando sus textos, abre todo un abanico de posibilidades desde el punto de vista forense, de la seguridad o del marketing.

Además, la proliferación de los medios sociales, que favorece nuevos modelos de comunicación y relación humana, potencia este abanico de posibilidades hasta cotas nunca antes vistas. La idiosincrasia inherente a estos medios sociales hace de ellos un entorno de comunicación especial, donde la libertad de expresión, la informalidad y la generación espontánea de temáticas y tendencias propician el acercamiento a la realidad diaria de las personas en su uso de la lengua. Sin embargo, esa misma idiosincrasia hace que en muchas ocasiones la aplicación de técnicas lingüísticas de análisis no sea posible, o sea extremadamente costoso.

En este trabajo hemos propuesto EmoGraph, una representación basada en grafos con el objetivo de modelar el modo en que los usuarios expresan sus emociones, y el modo en que las articulan en el marco de su discurso, teniendo en consideración no sólo su frecuencia, sino también su posición y relación con y respecto a los elementos del mismo. Nuestra hipótesis de partida es que los usuarios se expresan y expresan sus emociones de manera diferente dependiendo de su edad y sexo, y además, pensamos que esto es así independientemente de su idioma y del medio donde escriban. Hemos colaborado en la creación de un marco común de evaluación en el laboratorio PAN del CLEF, generando recursos que nos han permitido verificar nuestra hipótesis y conseguir resultados comparables y competitivos con los mejores resultados obtenidos por los investigadores del área.

Además, hemos querido investigar si la expresión de emociones permitiría diferenciar entre hablantes de diferentes variedades de una misma lengua, por ejemplo españoles, mexicanos o argentinos, o portugueses y brasileños. Nuestra hipótesis es que la variación entre lenguas se basa más en aspectos léxicos, y así lo hemos corroborado tras comparar EmoGraph con representaciones basadas en patrones, representaciones distribuidas y una representación que toma en consideración el vocabulario completo, pero reduciendo su dimensionalidad a únicamente 6 características por clase y que se erige idónea para su aplicación en entornos *big data* como los medios sociales.

Resum

La possibilitat de conèixer trets d'una persona únicament a partir dels textos que escriu s'ha convertit en una àrea de gran interès anomenada *author profiling*. Ser capaç d'inferir d'un usuari el sexe, l'edat, l'idioma nadiu o els trets de la seua personalitat tan sols analitzant els seus textos, obre tot un ventall de possibilitats des del punt de vista forense, de la seguretat o del màrketing.

A més, la proliferació dels mitjans socials, que afavoreix nous models de comunicació i de relació humana, potencia aquest ventall de possibilitats fins a cotes que no s'han vist fins ara. La idiosincràsia inherent a aquests mitjans socials en fa d'ells un entorn de comunicació especial, on la llibertat d'expressió, la informalitat i la generació espontània de temàtiques i tendències propicien l'aproximació a la realitat diària de les persones en l'ús que fan de la llengua. Tanmateix, aquesta idiosincràsia fa que en moltes ocasions no es puguin aplicar tècniques lingüístiques d'anàlisi, o que fer-ho resulti extremadament costós.

En aquest treball hem proposat EmoGraph, una representació basada en grafs que té l'objectiu de modelar la manera en què els usaris expressen les seves emocions, i la manera com les articulen en el marc de llur discurs, considerant-ne no només la freqüència sinó també la posició i la relació amb i respecte als elements del discurs. La nostra hipòtesi de partida és que els usuaris s'expressen i expressen llurs emocions de manera diferent depenent de l'edat i el sexe, i a més, pensem que això és així independentment de l'idioma i del mitjà en què escriguin. Hem col·laborat en la creació d'un marc comú d'avaluació al laboratori PAN del CLEF, generant recursos que ens han permès verificar la nostra hipòtesi i aconseguir resultats comparables i competitius amb els millors resultats obtinguts pels investigadors de l'àrea.

A més, hem volgut investigar si l'expressió d'emocions permetria establir diferències enre parlants de diferents varietats d'una mateixa llengua, per exemple espanyols, mexicans o argentins, o portuguesos i brasilers. La nostra hipòtesi és que la variació entre llengües es basa més en aspectes lèxics, i així ho hem corroborat després de comparar EmoGraph amb representacions basades en patrons, representacions distribuïdes i una representació que considera el vocabulari complet, però reduint-ne la dimensionalitat només a 6 característiques per classe i que s'erigeix de manera idònia per a aplicar-la en entorns *big data* com els mitjans socials.

Abstract

The possibility of knowing people traits on the basis of what they write is a field of growing interest named *author profiling*. To infer a user's gender, age, native language or personality traits, simply by analysing her texts, opens a wide range of possibilities from the point of view of forensics, security and marketing.

Furthermore, social media proliferation, which allows for new communication models and human relations, strengthens this wide range of possibilities to bounds never seen before. Idiosyncrasy inherent to social media makes them a special environment of communication, where freedom of expression, informality and spontaneous generation of topics and trends, enhances the knowledge of the daily reality of people in their use of language. However, the same idiosyncrasy makes difficult, or extremely costly, the application of linguistic techniques.

In this work we have proposed EmoGraph, a graph-based approach with the aim at modelling the way that users express their emotions, and the way they include them in their discourse, bearing in mind not only their frequency of occurrence, but also their position and relationship with other elements in the discourse. Our starting hypothesis is that users express themselves and their emotions differently depending on their age and gender, and besides, we think that this is independent on their language and social media where they write. We have collaborated in the creation of a common framework of evaluation at the PAN Lab of CLEF, generating resources that allowed us to verify our hypothesis achieving comparable and competitive results with the best ones obtained by other researchers on the field.

In addition, we have investigated whether the expression of emotions would help to differentiate among users of different varieties of the same language, for example, Spanish from Spain, Mexican and Argentinian, or Portuguese from Portugal and Brazil. Our hypothesis is that the variation among languages is based more on lexical aspects, and we have corroborated it after comparing EmoGraph with representations based on word patterns, distributed representations and a representation that uses the whole vocabulary, but reducing its dimensionality to only 6 features per class, what is suitable for its application to *big data* environments such as social media.

Tabla de contenidos

Lista de figuras

Lista de tablas

Todo por mis princesas...

Capítulo 1

Introducción

Vivimos una época de grandes cambios socioculturales a nivel global, facilitados por una revolución tecnológica que ha dado lugar, entre otras muchas, a la aparición y proliferación de los medios sociales de Internet, los cuales están propiciando nuevos modelos de relación y comunicación entre las personas. La teoría de las multitudes inteligentes de Rheingold [1] defiende el poder que entraña en la actualidad el hecho de que miles de personas puedan ponerse de acuerdo en cuestiones concretas para realizar acciones coordinadas, gracias y a través de estos medios sociales. Esta situación viene en parte dada por la continua necesidad del ser humano de expresarse y ser escuchado y valorado por los demás [2], una necesidad que llega en ocasiones a derivar en adicciones y problemas psicológicos [3]. Esta combinación de necesidad y capacidad propicia la generación de una suerte de inteligencia colectiva que fue definida por Pierre Levy [4] en 1997 como "una inteligencia repartida en todas partes, valorizada constantemente, coordinada en tiempo real, que conduce a una movilización efectiva de las competencias". Es innegable pues el abanico de oportunidades, y riesgos, que traen consigo los nuevos medios sociales.

Pero en esta suerte de inteligencia colectiva, en esta maraña de conversaciones y relaciones en que se han convertido los medios sociales, y que recibe comunmente el apelativo de *big data* [5], es el propio individuo el que queda difuminado entre la masa, incluso a pesar de sus esfuerzos, en muchas ocasiones, por destacar e influir [6]. En este nuevo entorno, millones de personas comparten información y se relacionan a partir de su identidad digital, avatar creado para tal fin en los medios, y que no siempre tiene que ser real ni coincidir con su identidad real; ciertos usuarios, en ciertas ocasiones y por diversos motivos, pueden querer tratar de ocultar su identidad, omitir cierta información sobre sí mismos o incluso querer resaltar ciertos aspectos para hacerse pasar por otros. Por ejemplo, un usuario que intenta mantener su anonimato para cometer un ciberdelito, otro usuario que intenta hacerse pasar por quién no es para cometer (ciber)acoso sexual a un menor [7, 8], o realizar *opinion spam* [9] sobre un producto, o simplemente un usuario que no ha informado ciertos datos de su perfil por dejadez, o porque el propio medio no se lo permite[1]. En definitiva, datos de la persona que serían útiles, pero que en ocasiones están ausentes o son falsos.

Sin embargo una cosa siempre está presente: el texto que escribe el usuario. Y aquí procede citar la Biblia, uno de los libros más distribuidos de la historia con cinco mil millones de copias[2], concretamente

[1] Por ejemplo, en Twitter no puedes informar ni tu edad ni tu sexo, y en Facebook cierta información del usuario permanece privada por defecto.

[2] http://www.guinnessworldrecords.com/world-records/best-selling-book-of-non-fiction

la contribución inicial de Juan 1:1[3] que dice: "*En el principio ya existía la Palabra. La Palabra estaba con Dios, y Dios mismo era la Palabra*". Y es que desde antiguo se ha considerado que la palabra está arraigada a la esencia de ser humano, lo que corroboran las teorías psicolingüísticas [10] actuales que postulan que el estilo discursivo refleja, en cierto modo, el perfil de su autor, quien decide, generalmente de manera inconsciente, qué palabras utilizar y cómo combinarlas. De este modo nace el *author profiling*, que trata de dar respuesta a la siguiente pregunta: *dado un texto, ¿qué podemos saber del autor que lo escribe?*.

El *author profiling* estudia aspectos psicolingüísticos y sociológicos (cómo se usa el lenguaje y qué rasgos son compartidos por grupos similares) para tratar de determinar, a partir de los textos, aspectos personales de su autor como la edad, el sexo, el idioma nativo o los rasgos de su personalidad. El interés es evidente desde perspectivas como la forense[4], donde ser capaz de conocer el perfil lingüístico de un mensaje de texto sospechoso (lenguaje utilizado por cierto tipo de personas) e identificar características (lenguaje como evidencia), ciertamente podría ayudar a atribuir su autoría [11, 12]; o desde el punto de vista de la mercadotecnia[5], lo que proporcionaría a las compañías la capacidad de segmentar su mercado en base, por ejemplo, al sexo, la edad o la región a la que pertenecen los usuarios que opinan de sus productos.

La generación de tecnología capaz de inferir este tipo de información sobre los usuarios a partir de lo que escriben, no sólo supondría un gran avance científico e industrial, sino también una gran responsabilidad que debería asumir la persona encargada de utilizarla. No olvidemos que la decisión final no debería realizarse sin la aplicación de la inteligencia, y la ética, humanas [13].

1.1 Motivación y objetivos

La mayoría de aproximaciones propuestas por los investigadores para abordar las diversas tareas de *author profiling*, se basan en la frecuencia de uso de determinadas características (e.g. categorías gramaticales, palabras vacías o signos de puntuación), o en modelos que emplean n-gramas. Nuestra hipótesis, especialmente cuando hablamos de medios sociales donde no hay censura y prima la libertad de expresión, es que los usuarios expresan sus emociones de manera diferente dependiendo de ciertos rasgos de su persona. Nuestro objetivo es profundizar en el modo en que los usuarios expresan dichas emociones en el marco de su discurso, no sólo tomando en consideración su frecuencia relativa de aparición, sino también su posición con y en relación con el resto de elementos del discurso, y analizar cómo puede esto ayudar a determinar su edad y sexo, independientemente del medio social y del idioma. Para hacerlo, hemos propuesto una representación basada en grafos, debido a su capacidad para modelar y analizar estructuras complejas como el lenguaje, que debido a la idiosincrasia propia de los medios sociales, hace compleja la aplicación de técnicas elaboradas de análisis sintáctico. Además, hemos querido investigar si, partiendo de esta supuesta independencia con el idioma, la expresión de las emociones sería un rasgo

[3] https://www.biblegateway.com/passage/?search=Juan+1%3A1&version=RVC

[4] La primera tarea internacional en *author profiling* ha sido patrocinada por el Forensiclab de la Universitat Pompeu Fabra de Barcelona http://pan.webis.de/clef13/pan13-web/author-profiling.html

[5] La segunda http://pan.webis.de/clef14/pan14-web/author-profiling.html y tercera http://pan.webis.de/clef15/pan15-web/author-profiling.html edición de la tarea en author profiling ha sido patrocinada respectivamente por las empresas Corex http://www.corex.es y Meaning Cloud http://www.meaningcloud.com/, todas ellas con la empresa Autoritas http://www.autoritas.es/ involucrada en su organización, con colaboraciones como Llorente y Cuenca http://www.llorenteycuenca.com/, y con empresas participantes como Xerox http://www.xerox.com/ o Daedalus http://www.sngularmeaning.team/.

distintivo entre usuarios de variedades diferentes de una misma lengua, por ejemplo, si serviría para diferenciar entre un español y un mexicano, o entre un portugués y un brasileño, o si por el contrario, como así postulamos nuestra segunda hipótesis, dichas variaciones se encuentran en mayor medida a nivel léxico. Todo ello nos lleva a plantear las siguientes cuestiones:

1.1.1 Cuestiones de investigación

- ¿Existe algún tipo de relación entre la forma en la que articulamos nuestro discurso y expresamos nuestras emociones, con la edad y el sexo que tenemos? Y en tal caso, ¿es independiente del medio social donde escribimos y del idioma en el que lo hacemos?
- Si la cuestión relativa al idioma es afirmativa, ¿implicaría en cierto modo que la expresión de emociones no difiere entre usuarios de lenguas similares o incluso variedades de una misma lengua? Y si es así, ¿en qué se diferencian principalmente dichas lenguas y variedades?
- ¿Disponemos de los recursos adecuados para investigar todas estas cuestiones, incluso en idiomas diferentes al inglés?

1.1.2 Objetivos

Para intentar dar respuesta a las anteriores cuestiones proponemos cumplir cada uno de los siguientes objetivos:

1. Proponer una representación que permita:
   (a) modelar la estructura del discurso y la expresión de las emociones en el mismo, tomando en consideración no sólo su frecuencia de aparición sino su posición con y en relación al resto de elementos del discurso;
   (b) verificar la hipótesis de que el modo en que el usuario articula su discurso y expresa en él sus emociones, sirve para determinar su edad y sexo;
   (c) comprobar la independencia de la hipótesis con respecto al medio social; y
   (d) con respecto al idioma.
2. Investigar si la expresión de las emociones es un rasgo diferenciador entre lenguas similares o variedades de una misma lengua, o si por el contrario, variaciones léxicas aportan más información a la tarea;
   (a) comprobar la adecuación de las representaciones distribuidas a la tarea;
   (b) proponer una representación que reduzca la dimensionalidad frente a las comunmente utilizadas basadas en n-gramas.
3. Crear los recursos necesarios para investigar las cuestiones planteadas y un marco de evaluación común que permita comparar las propuestas de diferentes investigadores construyendo así un estado del arte homogéneo, comparable y reproducible.

1.2 Aportaciones de la tesis

A continuación exponemos las contribuciones principales de la tesis, tanto desde la perspectiva científica como desde la perspectiva técnica relativa a la generación de recursos que permitan a otros investigadores profundizar en las tareas de *author profiling* abordadas:

1. Hemos propuesto la representación EmoGraph para modelar el estilo discursivo y la expresión de las emociones en textos, y la hemos aplicado a la identificación de edad y sexo. Además, hemos verificado su aplicabilidad y robustez a diferentes medios sociales e idiomas. De este modo hemos verificado nuestra primera hipótesis (Capítulo 5).
2. Hemos investigado la aplicabilidad de la representación EmoGraph en la tarea de identificación de variedades de una misma lengua, comprobando que la expresión de emociones y el estilo discursivo no varía de modo discriminativo. Para verificar nuestra hipótesis de que las variaciones se producen principalmente a nivel léxico, hemos analizado varias representaciones: una basada en patrones (IG-WP), dos basadas en representaciones distribuidas sobre el conocido modelo de Skip-gramas continuos, y la representación de baja dimensionalidad (LDR) que hemos propuesto y que permite trabajar de manera eficiente en entornos *big data*. De esta manera hemos verificado nuestra segunda hipótesis (Capítulo 7).
3. Hemos creado los recursos necesarios para llevar a cabo la investigación, concretamente:
   (a) Con respecto a la identificación de edad y sexo, hemos colaborado en la organización y creación de un marco de evaluación en la tarea de identificación de edad y sexo del PAN en el CLEF (Capítulo 3), lo que ha permitido crear un conjunto de corpus recopilados de diferentes medios sociales (Twitter, blogs, revisiones online, redes sociales) y en diferentes idiomas (inglés, español, holandés e italiano), etiquetados con edad y sexo (Capítulo 3, apartados 3.1.1, 3.2.1 y 3.3.1).
   (b) Con respecto a la identificación de emociones en medios sociales, y concretamente en Twitter, hemos compilado el corpus Barcenas con tuits tratando un caso de corrupción política ocurrido en España entre el 9 de julio y el 2 de octubre de 2013, con un total de 4.397.023 tuits en español (Capítulo 4, apartado 4.2.1).
   (c) Con respecto a la identificación de emociones en medios sociales y su relación con el sexo, así como con el uso de la ironía, hemos generado el corpus EmIroGeFB con comentarios de Facebook anotados con las seis emociones básicas de Ekman, la presencia/ausencia de ironía, y el sexo de los autores de los comentarios. El corpus se enmarca dentro de tres temáticas (política, fútbol, famosos) y consta de 1.200 comentarios en español (Capítulo 4, apartado 4.3).
   (d) Por último, con respecto a la identificación de la variedad de lenguaje, hemos construido el corpus HispaBlogs con posts escritos en blogs personales en cinco variedades del español: Argentina, Chile, España, México y Panamá. El corpus consta de dos particiones de 2.400 y 1.000 autores por partición (Capítulo 7, apartado 7.1).

1.3 Organización de la tesis

La presente tesis consta de 8 capítulos y 2 apéndices, donde se trata de responder a las preguntas de investigación formuladas en el apartado 1.1. Concretamente:

Capítulo 1. Introducción. En este capítulo introducimos la oportunidad que brindan los nuevos medios sociales y la necesidad de ser capaces de obtener información sobre las personas que participan en ellos, introduciendo así el concepto de *author profiling*.

Capítulo 2. Identificación de edad y sexo. En este capítulo efectuamos una revisión exhaustiva al estado del arte en identificación de edad y sexo, describiendo las representaciones propuestas, los corpus disponibles, las medidas de evaluación utilizadas y los resultados alcanzados. Además, sobre la base de las teorías psicolingüísticas actuales, realizamos un estudio estadístico relativo al uso de las categorías gramaticales por medio social y por sexo.

Capítulo 3. Author profiling en el PAN. En este capítulo realizamos una descripción detallada de los tres años en los que organizamos la tarea de identificación de edad y sexo en el PAN[6]. La organización del PAN ha propiciado la creación de recursos como corpus etiquetados con edad y sexo en idiomas diferentes al inglés, la definición de un marco común de evaluación y la generación de un estado del arte consistente, reproducible y útil para la comparación.

Capítulo 4. Identificación de emociones en medios sociales. En este capítulo abordamos la tarea de identificación de emociones en medios sociales e investigamos su relación con el *author profiling*. Nuestra hipótesis central es que la expresión de las emociones tiene una fuerte correlación con nuestro sexo y edad, algo que en el estado del arte no ha sido abordado. Comenzamos así con una revisión del estado del arte en procesamiento afectivo, desde la perspectiva de la generación de recursos y desde la perspectiva de la identificación automática de emociones en texto, para posteriormente analizar la utilidad de la expresión de las emociones para la identificación de tendencias, o su relación con la ironía, la edad y el sexo. En este capítulo presentamos nuestra investigación en identificación de edad y sexo a partir del mismo conjunto de características que utilizamos para la identificación de las emociones, sentando las bases de la hipótesis central de esta tesis.

Capítulo 5. EmoGraph: Una aproximación basada en grafos. En este capítulo investigamos con mayor profundidad cómo el modo en que los usuarios expresan las emociones sirve para conocer su edad y su sexo. Para ello, tratamos de modelar cómo los usuarios estructuran su discurso y cómo las emociones se enmarcan en el mismo, utilizando grafos debido a su potencia para representar y analizar estructuras complejas, como en este caso el lenguaje. Tras una revisión del estado del arte en uso de grafos para diversas tareas de procesamiento del lenguaje natural, decidimos aprovechar los grafos para extraer el conocimiento relacional entre las partes del discurso y las emociones, y obtener un esquema de asignación de pesos para el aprendizaje automático de los modelos. Para finalizar el capítulo investigamos en la robustez del método ante diversos medios sociales e idiomas.

Capítulo 6. Identificación del lenguaje nativo y de las variedades del lenguaje. En este capítulo se proporciona una visión detallada del estado del arte relativo dos tareas relacionadas: la identificación del idioma nativo de un usuario que escribe en una segunda lengua, y la discriminación entre variedades de una misma lengua. El objetivo del capítulo es presentar el transfondo de una tarea que tiene la doble vertiente de la clasificación de textos y el *author profiling*, y sobre la que deseamos contrastar una segunda hipótesis: la variación entre lenguas similares o variedades de una misma lengua, se debe más a cambios léxicos que al modo en que sus usuarios expresan las emociones.

[6] No se describe la del 2016 por estar en curso en el momento de escritura de esta tesis.

Capítulo 7. Aproximaciones para la identificación de variedades del lenguaje. En este capítulo investigamos la adecuación de la representación propuesta para identificar la edad y el sexo a partir del modo en que los usuarios articulan su discurso y expresan las emociones a la tarea de discriminar entre usuarios que hablan variedades de una misma lengua, o lenguas muy similares. Así mismo proponemos representaciones alternativas que permiten contrastar nuestra hipótesis de que en este caso, el léxico tiene un componente discriminativo mayor.

Capítulo 8. Conclusiones y trabajo futuro. En este capítulo presentamos las conclusiones al trabajo que hemos llevado a cabo en el marco de nuestro doctorado, subrayando los principales descrubrimientos que soportan nuestras hipótesis, las principales contribuciones al estado del arte y marcando las directrices de trabajos futuros que pueden derivarse del mismo.

Apéndice I. Author profiling en PAN 2014. En este apéndice se muestran las tablas de significación estadística en una comparación pareada de los sistemas participantes de la tarea del PAN 2014, además de las distancias entre la edad predecida y la edad real de los autores.

Apéndice II. Author profiling en PAN 2015. En este apéndice se muestran las tablas de significación estadística en una comparación pareada de los sistemas participantes en la tarea del PAN 2015, así como una comparativa de resultados entre idioma.

Capítulo 2

Identificación de edad y sexo

El estudio del modo en que ciertas características lingüísticas varían en función del perfil del autor que las utiliza es un tema de interés desde múltiples disciplinas como la psicología, la lingüística y más recientemente el procesamiento del lenguaje natural. Dentro de esta disciplina se enmarca el primero de los objetivos de este trabajo: conocer el sexo y la edad del autor de un texto anónimo.

Para comenzar su estudio hemos realizado una revisión del estado del arte, de modo que se obtenga una visión panorámica de las diferentes aproximaciones utilizadas y sobre los tipos de texto que se han aplicado. Esta visión se complementa con mediciones estilométricas realizadas sobre diferentes tipologías de textos, con el objetivo de verificar hasta qué punto los usuarios se expresan de manera diferente dependiendo del medio en el que lo hacen. Por último, y para un subconjunto de comentarios de Facebook, hemos realizado un análisis del uso de categorías gramaticales por sexo.

2.1 Trabajos previos

Pennebaker *et al.* [10] han investigado la conexión entre el uso del lenguaje y diferentes rasgos de la persona, estudiando cómo la variación de las características lingüísticas en un texto puede proporcionar información, entre otras, sobre la edad y el sexo de su autor. Algunas de las conclusiones de sus estudios para el inglés nos indican que las mujeres utilizan la primera persona del singular en mayor medida que los hombres debido a que son más conscientes de sí mismas y autocentradas, mientras que ambos sexos usan la primera persona del plural en similar medida, que los hombres utilizan los artículos determinados e indeterminados de manera más frecuente que las mujeres pues hablan más de cosas concretas (e.g. el carburador averiado o una parrilla para la barbacoa), mientras que ellas utilizan más palabras cognitivas (e.g. entender, pensar, saber, etc.) y sociales que los hombres, puesto que piensan y hablan más de las personas y sus relaciones. Respecto a la edad también ofrecen ciertas conclusiones, como que de jóvenes se tiende a usar más las primeras personas o a hablar de periodos temporales, principalmente en pasado, mientras que con la edad se tiende a usar más palabras cognitivas, hablando más del futuro, con palabras más grandilocuentes y usando más artículos, nombres y preposiciones. Estos estudios son la base de la construcción de uno de los recursos más utilizados en diferentes tareas relacionadas con *author profiling*: Linguistic Inquiry and Word Count. LIWC[1] es un software desarrollado por Pennebaker *et al.* [14] que permite obtener, a partir de un texto, hasta un total de 70 dimensiones psycolingüísticas referentes a

[1] http://liwc.wpengine.com/

frecuencias de uso de palabras que reflejan diferentes emociones, estilos de pensamiento, preocupaciones sociales, autoreferencias, palabras causales o partes del discurso. Estos estudios son la base e inspiración de la mayoría de aproximaciones utilizadas en procesamiento del lenguaje natural para abordar la tarea de identificación de edad y sexo.

Son pioneros los trabajos de Argamon *et al.* [15] que analizan 604 documentos formales extraídos del British National Corpus (BNC)[2], equilibrando por sexo a los autores de diferentes géneros literarios (e.g. ficción, no ficción, ciencias aplicadas, etcétera). Los autores utilizan un total de 1.081 características consistentes en la combinación de palabras de función[3] con partes del discurso y obteniendo aproximadamente un 80% de *accuracy*[4] en la identificación de sexo. En un estudio posterior, Koppel *et al.* [16] investigan la identificación de sexo a partir de un subconjunto equilibrado de 566 documentos extraídos del mismo corpus BNC. Los autores utilizan como características una lista de 405 palabras de función, n-gramas de 76 etiquetas referentes a partes del discurso y signos de puntuación. Los autores obtienen un 73,7% de *accuracy* usando las palabras de función, un 70,5% usando las partes del discurso y un 77,3% con la combinación completa de características. Los anteriores resultados ponen de manifiesto, en línea con los descubrimientos previos, la aportación de las palabras de función y las partes del discurso en la tarea de identificar el sexo.

En un dominio menos formal, Estival *et al.* [17] tratan, entre otras, la identificación de sexo y edad en 9.836 mensajes de correo electrónico en inglés con al menos 200 palabras cada uno. Los autores experimentan con diferentes algoritmos de clasificación y muestran la superioridad de las máquinas de vectores soporte (SVM de sus siglas en inglés) [18], reportando *accuracies* de 69,26% y 56,46% respectivamente. Para representar los documentos, los autores extraen un total de 689 características correspondientes a frecuencias de signos de puntuación, longitud de las palabras, letras mayúsculas/minúsculas, palabras de función, partes del discurso, saltos de párrafo o presencia de ciertas etiquetas HTML. En un trabajo posterior [19] los autores incorporan el árabe en la identificación y obtienen *accuracies* del 81,15% y 72,10% respectivamente.

Con el emerger de las redes sociales el foco se ha movido hacia otro tipo de escritos, más coloquiales, menos formales y estructurados, tales como blogs y foros. Schler *et al.* [20] estudian el efecto de la edad y el sexo del autor en el estilo de escritura en aproximadamente 70.000 blogs[5] obtenidos de Blogspot[6]. Los autores utilizan un conjunto de características de estilo como palabras fuera de diccionario, partes del discurso, palabras de función o hiperenlaces, combinadas a su vez con características de contenido como los unigramas de palabras con mayor ganancia de información. Los resultados obtenidos son del 80% de *accuracy* en la identificación de sexo y del 75% en la identificación de edad. Así mismo, demuestran que ciertas características del lenguaje tienen una alta correlación con la edad, como por ejemplo el uso de preposiciones y determinantes, en línea con los trabajos previos de Pennebaker. Por su parte, Goswami *et al.* [21] han investigado sobre el mismo corpus la adición de nuevas características como palabras de jerga o la longitud media de las oraciones, mejorando así la *accuracy* hasta el 80,3% en la detección de edad y hasta el 89,2% en la de sexo. Sobre un subconjunto de 19.000 blogs del mismo corpus, Argamon *et al.* [22] investigan la combinación de características léxicas y sintácticas, y obtienen *accuracies* de 77,7% y 76,1% respectivamente para edad y sexo. Los autores muestran ciertas correlaciones en línea con los estudios pioneros de Pennebaker, como por ejemplo que las mujeres utilizan

[2] http://www.natcorp.ox.ac.uk/

[3] Se considera palabras de función a aquellas que no tienen un significado concreto pero sirven para construir el discurso expresando relaciones con otras palabras. Son palabras de función las preposiciones, los pronombres y los verbos auxiliares

[4] https://en.wikipedia.org/wiki/Accuracy_and_precision

[5] http://www.cs.biu.ac.il/~koppel/BlogCorpus.htm

[6] http://blogspot.com

más los pronombres y los hombres los determinantes y preposiciones, estas últimas utilizadas más por los adultos en contraposición con el mayor uso de contracciones por los más jóvenes.

Yan y Yan [23] aplican Naive Bayes a la clasificación del sexo de 3.000 autores de Xanga[7] que han escrito un total de 75.000 posts. Para ello utilizan una representación basada en bolsas de palabras combinada con elementos extraídos del HTML como el color del fondo, la fuente utilizada o la tipografía. También tienen en cuenta características estilísticas como el uso de signos de puntuación o de emoticonos, estos últimos especialmente ligados a emociones. La *accuracy* reportada es del 73,11%.

Mukherjee y Liu [24] proponen un método para obtener patrones de secuencias de partes del discurso con una longitud variable calculada a partir de los datos de entrenamiento, lo que les ha permitido obtener una *accuracy* del 88,56%. Para realizar sus experimentos, han recopilado un total de 3.100 blogs que han etiquetado manualmente con el sexo de los autores.

Sarawgi *et al.* [25] aproximan la tarea de identificación de sexo en blogs y artículos científicos. La asunción inicial sería que la identificación en artículos científicos es más difícil, pero aún así obtienen un 61% de *accuracy* frente al 71,3% que obtienen en blogs. Para aproximar la tarea, utilizan gramáticas probabilísticas libres de contexto [26] para capturar regularidades sintácticas superficiales en patrones de n-gramas, mostrando que los n-gramas de caracteres funcionan mejor que los de palabras.

Soler-Company y Wanner [27] recuperan blogs de la sección correspondiente de periódicos conocidos y etiquetados a mano con el sexo de los autores para el inglés, español, francés, alemán, italiano y catalán. Los autores utilizan una combinación de 27 características que incluyen ratios de uso de comas, puntos, exclamaciones, frecuencia de interjecciones, palabras de afirmación y negación, primeras personas de singular y plural, palabras vacías, nombres propios, frases por post, media de palabras por frase o determinadas características sintácticas obtenidas del árbol de dependencias. La *accuracy* media obtenida es del 71,01% con la distribución por idioma mostrada en la Tabla 2.1.

Inglés	**Español**	**Alemán**	**Francés**	**Catalán**	**Italiano**
80%	88%	77%	83%	88%	86%

Tabla 2.1: Resultados por idioma de la aproximación de Soler-Company.

En el caso de Peersman *et al.* [28], los autores recopilan un total de 1.537.283 posts de los blogs de Netlog[8] en holandés, con una media de 12,2 tokens por posts. Los autores aplican máquinas de vectores soporte a una combinación de unigramas, bigramas y trigramas de palabras, junto con bigramas, trigramas y tetragramas de caracteres, seleccionando aquellos con mayor X^2. Entre diferentes experimentos de clasificación binaria entre pares de edades (menores de 16 vs. mayores de 16; menores de 16 vs. mayores de 18; menores de 16 vs. mayores de 25), destacan los resultados de la identificación conjunta de edad y sexo (hombres menores de 16 vs. mujeres menores de 16 vs. hombres mayores de 25 vs. mujeres mayores de 25), obteniendo un 88,7% de *accuracy* y mostrando cómo la identificación previa del sexo mejora la clasificación de ciertos grupos de edad.

La evolución de las redes sociales ha llevado a los investigadores a realizar estudios en otros medios diferentes a los blogs, como por ejemplo Schwartz *et al.* [29] que utiliza n-gramas y tópicos obtenidos con *Latent Dirichlet Allocation (LDA)*[9] como características para identificar, entre otras, la edad y el sexo

[7] http://xanga.com/
[8] http://www.netlog.com
[9] http://www.wikiwand.com/es/Latent_Dirichlet_Allocation

de usuarios de Facebook[10]. Los autores entrenan su sistema con 700 millones de palabras y obtienen una *accuracy* de 91,9% en la identificación de sexo, y un coeficiente de regresión lineal de 0,84 en la identificación de la edad como variable continua. También en Facebook, Sap *et al.* [30] obtienen un subconjunto en inglés de 75.394 posts del proyecto MyPersonality[11] donde los usuarios hayan etiquetado su edad, sexo y al menos tengan 1.000 palabras. Para la identificación de la edad han realizado 12 divisiones en rangos de 4 años desde los 13 hasta los 60. Los autores aproximan la tarea mediante la creación de un léxico predictivo donde calculan pesos para las palabras de cada clase usando *ridge regression* [31] en el caso de la edad, y clasificación con máquinas de vectores soporte [32] en el caso del sexo. Los resultados presentados son de 91,9% de *accuracy* en identificación de sexo y de 0,831 de coeficiente de correlación de Pearson[12] [33] en el caso de identificación de edad. Otterbacher [34] se ha centrado en la identificación de sexo de autores de revisiones de películas. En su estudio, no sólo tiene en cuenta características lingüísticas de las revisiones, como el uso de pronombres, medidas de riqueza y complejidad del vocabulario, uso de diferentes personas y tiempos verbales, o palabras relacionadas con violencia, familia o relaciones, entre otras, sino también metadatos referentes a las fechas de revisión, la valoración o el número de revisiones realizadas por el autor. Utilizando un clasificador de regresión logística obtiene 73,71% de *accuracy*, e indica que los mejores resultados se obtienen con la combinación de todas las anteriores características, descubriendo que la *utilidad* asignada a la revisión es un buen indicador ya que está asociado en mayor medida al sexo masculino.

Una de las necesidades que surgen cuando se investiga en autor profiling en social media, es la de obtener datos etiquetados con información demográfica de los usuarios, como en el caso que nos ocupa, su edad y sexo. Los estudios realizados en literatura clásica trabajaban con pequeños conjuntos de autores bien conocidos, donde el etiquetado manual podía realizarse de manera sencilla y fiable. Sin embargo y debido a la dimensión de los datos de las actuales redes sociales, este etiquetado manual es inviable, por lo que debe ser automatizado. En esta línea, algunos investigadores han etiquetado manualmente la colección de datos, como en el caso de Nguyen *et al.* [35], no sin cierto riesgo de desviación o parcialidad en el mismo. Sin embargo, en la mayoría de los casos citados anteriormente los investigadores han aprovechado la información proporcionada por los propios autores. Por ejemplo, en algunas plataformas de blogging los autores podían indicar ellos mismos este tipo de información. Este es el caso de los datos utilizados por Peersman *et al.* [28] que recuperan su colección desde Netlog, donde los autores de los posts pueden informar su sexo y edad exacta. Similar es el caso de Schler *et al.* [20] que recolectan el dataset desde Blogspot en un periodo en el que los usuarios podían reportar esta información. En ambos casos se debe tener en cuenta un problema común y es el uso de estos medios (especialmente los blogs) para promocionar posiciones webs en los motores de búsqueda mediante el uso de perfiles falsos (*black hat seo*[13]). Esto es similar a incorporar ruido al corpus de evaluación, pero por otro lado también refleja de manera realista el estado actual de los datos disponibles.

Otro aspecto a tener en cuenta, especialmente en social media, es el efecto que puede tener el tamaño de los textos en el rendimiento de los algoritmos de aprendizaje. En este sentido, Zhang y Zhang [36] experimentan con pequeños segmentos de posts de blog, concretamente 10.000 segmentos con 15 tokens por segmento. Los autores utilizan bolsas de palabras, medidas como la longitud de palabras y frases, partes del discurso o palabras con cierto contexto relativo a la familia, el hogar o la conversación, y obtienen un 72,1% de *accuracy* en identificación de sexo en comparación con más del 80% de *accuracy* reportado en anteriores trabajos. En esta línea, en los últimos tiempos muchas investigaciones se han

[10] https://www.facebook.com
[11] http://mypersonality.org/wiki/doku.php
[12] http://www.wikiwand.com/en/Pearson_product-moment_correlation_coefficient
[13] https://es.wikipedia.org/wiki/Black_hat_SEO

centrado en medios como Twitter[14], donde los mensajes son realmente cortos con una longitud media inferior a 10 términos por tuit (máximo 140 caracteres). Nguyen *et al.* [35] estudian el uso del lenguaje y la edad entre usuarios de esta red en holandés. Los autores modelan la edad como una variable continua, de manera similar a como hacen en estudios previos [37], y usan una aproximación basada en regresión logística. Además miden el efecto del sexo en el rendimiento de la detección de edad, considerando ambas variables como interdependientes, y obteniendo correlaciones superiores a 0,74 y errores medios absolutos entre 4,1 y 6,8 años. Los autores combinan como característica el sexo junto con n-gramas, partes del discurso y clases de palabras obtenidas con LIWC, mostrando que el sexo es una buena característica para determinar con mayor *accuracy* la edad en los más jóvenes.

Una preocupación recurrente de los investigadores es la fiabilidad a la hora de evaluar y comparar los resultados entre diferentes aproximaciones, de manera que se pueda construir una imagen fiel del estado del arte en un determinado campo. Así pues, por ejemplo Mukherjee y Liu [24] argumentan una mejora del 10% sobre los resultados de Schler *et al.* [20] ó Argamon *et al.* [22], entre otros, algo que no es del todo afirmable ya que se han utilizado colecciones de datos diferentes con conjuntos de autores diferentes, algo que puede afectar al resultado del aprendizaje y la generalización de los modelos. De ahí la importancia de las tareas de evaluación organizadas en el marco de las *Conference and Labs of the Evaluation Forum* (CLEF)[15], como la de autor profiling en el laboratorio *Plagiarism, Authorship and Social Software Misuse* (PAN)[16]. Es en 2013 cuando organizamos por primera vez una tarea internacional en *author profiling* [38–40] con el objetivo en la identificación de edad y sexo de autores de social media, y en otros idiomas además del inglés. En esta tarea, la mayoría de participantes han combinado características de estilo como frecuencias de signos de puntuación, el uso de mayúsculas y minúsculas, comillas, o partes del discurso con características basadas en el contenido como Latent Semantic Analysis (LSA), bolsas de palabras, *term frequency-inverse document frequency* (*tf-idf*), palabras de diccionario, palabras de tópico, etcétera. Independientemente de los buenos resultados de las características basadas en n-gramas reportados por Houvardas y Stamatatos [41] y Peersman *et al.* [28], es importante mencionar el efecto de características más elaboradas. Por ejemplo, la representación de segundo orden propuesta por el equipo ganador de las tres ediciones del PAN [42–44] y que relaciona documentos con perfiles (e.g. hombres, mujeres, adolescentes, etcétera) y subperfiles de autores (e.g. videojugadores, estudiantes, amas de casa, etcétera). López-Monroy *et al.* [45] proponen explotar información específica de los subperfiles dentro de cada perfil. Por lo tanto, en lugar de estudiar cada perfil asumiendo un grupo homogéneo de autores (e.g. mujeres, hombres, adolescentes, etcétera), buscan por subperfiles asumiendo heterogeneidad en los anteriores grupos (e.g. mujeres estudiantes, mujeres amas de casa, hombres adolescentes videojugadores, etcétera). De este modo se aprende información más específica de los subperfiles que componen cada posible perfil. Esta información intraperfil puede ser explotada de múltiples maneras, por ejemplo reduciendo la dimensionalidad para representar los documentos, o puede utilizarse para detectar perfiles de usuarios que tengan influencia en la reputación de una marca o sector [46]. Los autores han testado su método con la colección de Schler [20] obteniendo 82,01% y 77,68% de *accuracy* respectivamente en la identificación de sexo y edad, y mejorando así de manera significativa los resultados previos. Además, han testeado su método con los datasets PAN-AP-2013 (social media en inglés y español), y PAN-AP-2014 (social media, blogs y Twitter en inglés y español y revisiones de noticias en inglés), obteniendo los resultados mostrados en la Tabla 2.2. También es destacable el uso de colocaciones en el caso de Meina *et al.* [47], equipo que obtuvo el mejor resultado para el inglés en la tarea del 2013 con una *accuracy* de 59,21% y 64,91% en identificación de sexo y edad respectivamente. Usando la partición de inglés de este mismo dataset, Weren *et al.* [48] muestran la contribución de características relacionadas

[14] https://twitter.com
[15] http://www.clef-initiative.eu/
[16] http://pan.webis.de

con la recuperación de información en la identificación de edad y sexo, obteniendo *accuracies* de 62,1% y 68,2% respectivamente en la identificación de sexo y edad. O la investigación de Maharjan *et al.* [49] que aproxima la tarea con 3 millones de características procesadas en una configuración MapReduce y que le permite obtener resultados competitivos (valores de *accuracy* mayores de 61% para la identificación de sexo y edad, en inglés y español). Una revisión detallada de la tarea de *author profiling* en el PAN se dará en el Capítulo 3.

	Social Media		**Blogs**		**Twitter**		**Revisiones noticias**	
	Sexo	**Edad**	**Sexo**	**Edad**	**Sexo**	**Edad**	**Sexo**	**Edad**
Inglés	0,5539	0,3703	0,7823	0,4897	0,7169	0,4836	0,6927	0,3449
Español	0,6635	0,4276	0,6477	0,5302	0,6685	0,4819	-	-

TABLA 2.2: Resultados de López-Monroy en PAN-AP-2014 (validación cruzada).

Por último, el interés en diferentes aspectos del autor profiling es evidente también en la plataforma Kaggle[17] donde compañías y departamentos de investigación comparten sus necesidades e investigadores independientes aceptan el reto. Por ejemplo, el reto ICDAR2013 - Gender Prediction from Handwriting[18] proporciona a los participantes un dataset con 475 autores que han escrito un total de 4 documentos, cada uno de ellos en árabe e inglés. Esta competición se organiza en el marco del Twelfth International Conference on Document Analysis and Recognition ICDAR2013[19]. También la competición privada CME 250 Age Prediction Competition[20] donde los participantes deben identificar la edad del resto de los compañeros a partir de una serie de datos personales, aunque en este caso no incluyen ningún dato relacionado con el lenguaje.

En la Tabla 2.3 se resumen las diferentes investigaciones con el año en que fueron publicadas, el corpus utilizado, el idioma, las características utilizadas y los resultados obtenidos para la identificación de edad y sexo.

[17] http://www.kaggle.com/
[18] https://www.kaggle.com/c/icdar2013-gender-prediction-from-handwriting
[19] http://www.icdar2013.org/
[20] https://inclass.kaggle.com/c/cme-250-age-prediction

Autor	**Año**	**Corpus**	**Idioma**	**Características**	**Sexo**	**Edad***
Argamon	2003	BNC	Inglés	Palabras de función, partes del discurso	80%	-
Koppel	2003	BNC	Inglés	Palabras de función, *n*-gramas de partes del discurso, signos de puntuación	77,3%	-
Schler	2006	Blogs	Inglés	Palabras fuera de diccionario, partes del discurso, palabras de función, hiperenlaces, bolsas de palabras	80%	75%
Yan	2006	Blogs	Inglés	BOW, color fondo, tipografía, emoticonos	73,11%	-
Estival	2007	eMails	Inglés	Signos de puntuación, longitud de palabras, letras mayúsculas/minúsculas, palabras de función, partes del discurso, saltos de párrafo, presencia HTML	56,46%	69,26%
Estival	2008	eMails	Árabe	Las mismas que en inglés	81,15%	72,10%
Goswami	2009	Blogs	Inglés	Palabras de jerga, longitud media de las oraciones	89,2%	80,3%
Argamon	2009	Blogs	Inglés	Léxicas y sintácticas	76,1%	77,7%
Mukherjee	2010	Blogs	Inglés	Patrones de secuencias de partes del discurso	88,56%	-
Zhang	2010	Blogs	Inglés	BOW, longitud de palabras y frases, partes del discurso, palabras contextuales	72,1%	-
Otterbacher	2010	Rev. películas	Inglés	Uso de pronombres, medidas de riqueza y complejidad del vocabulario, uso de diferentes personas y tiempos verbales, palabras relacionadas con violencia, familia, relaciones, etc. + fechas de revisión, valoración, número de revisiones	73,71%	-
Sarawgi	2011	Blogs	Inglés	Gramáticas libres de contexto para generar patrones de *n*-gramas	71,3%	-
		Papers			61,5% %	-
Peersman	2011	Netlog	Holandés	*n*-gramas de palabras y caracteres	-	88,7%
Nguyen	2011/13	Twitter	Holandés	Sexo, *n*-gramas, partes del discurso, LIWC	-	4,1 - 6,8
Schartz	2013	Facebook	Inglés	*n*-gramas, LDA	91,9%	0,84
Sap	2014	Facebook	Inglés	Léxico ponderado por clase	91,9%	0.83
Weren	2014	Social Media	Inglés	Basadas en recuperación de información	62,1%	68,2%
Maharjan	2014	Social Media	Inglés	*n*-gramas + MapReduce	61,66%	65,62%
			Español		64,63%	61,73%
López-Monroy	2015	Blogs	Inglés	Representación de segundo orden	82,01%	77,68%
		Social Media	Inglés		55,39%	37,03%
			Español		66,35%	42,76%
		Blogs	Inglés		78,23%	48,97%
			Español		64,77%	53,02%
		Twitter	Inglés		71,69%	48,36%
			Español		66,85%	48,19%
		Revisiones hotel	Inglés		69,27%	34,49%
Soler-Company	2015	Blogs	Inglés	Comas, puntos, personas verbales, frases por post, dependencias	80%	-
			Español		88%	-
			Alemán		77%	-
			Francés		83%	-
			Catalán		88%	-
			Italiano		86%	-

* Todos los resultados de edad están dados en *accuracy* excepto Schartz que reporta el coeficiente de regresión lineal, Sap que reporta el coeficiente de correlación de Pearson o Nguyen que reporta el rango de error en edades.

Tabla 2.3: Estado del arte en identificación de edad y sexo.

2.2 Lenguaje y social media

La mayoría de estudios en *author profiling*, tal y como refleja el apartado anterior, han sido realizados para el idioma inglés y limitados inicialmente a textos formales, y más recientemente blogs y social media. Surge entonces la pregunta, ¿se utiliza el lenguaje de igual modo en los diferentes medios sociales de Internet?, es decir, ¿podemos aplicar estudios de lingüística computacional como los descritos en el apartado anterior de igual manera en Facebook que en Twitter, en la Wikipedia que en un foro, en un blog que en un artículo de prensa? Y además, ¿podemos aplicar tales estudios al español? Esta cuestión es crucial si se piensa por ejemplo en el diferente uso de los pronombres personales en inglés frente al español, siendo los pronombres de las palabras consideradas de función que más interés suscitan en la investigación [50–52]. En este apartado se presenta un estudio sobre el uso de los diferentes tipos de palabras en diferentes medios sociales de Internet, y en idioma español [53].

2.2.1 Metodología

Se ha determinado un conjunto de medios sociales de Internet para comparar el uso del lenguaje en cada uno de ellos, concretamente:

- *Wikipedia*[21]*:* Enciclopedia abierta y colaborativa dónde todo el mundo, salvo excepciones, puede editar su contenido. Es por lo tanto un medio de información formal, aunque escrito por gran cantidad de personas diferentes con estilos discursivos propios. Se ha descargado el fichero oficial de la Wikipedia en español de fecha 27/12/2012 de dónde se ha extraído aquellos ítems referentes a páginas. No se han tenido en cuenta ni revisiones, ni el histórico de las páginas, sólo el contenido de las páginas actualmente en uso.
- *Prensa:* Publicaciones formales emitidas por los medios de comunicación con el objetivo de informar. Son escritas por los reporteros y siguen un proceso de revisión y filtrado antes de su publicación final, siguiendo las pautas editoriales del periódico que las publica. Se ha recopilado durante un periodo de seis meses todo lo emitido en la prensa de 6.581 diarios online de España, Argentina, México, Chile y Panamá en idioma español.
- *Blogs:* Sitios web que se actualizan de manera periódica con contenidos publicados por uno o más autores según su propio criterio. Abarcan un amplio abanico de posibilidades, desde blogs corporativos hasta personales, pasando por medios de comunicación similares a la prensa. Se ha recopilado durante un periodo de seis meses todo lo emitido en un total de 78.289 blogs de España, Argentina, México, Chile y Panamá en idioma español.
- *Foros:* Sitios web que permiten a los usuarios discutir y compartir información relevante a un tema. Es un sitio de discusión libre e informal, a veces moderada en cuanto a la posibilidad de publicación o no, pero dejando libertad a los autores para expresarse en su propio estilo. Se ha recopilado durante un periodo de seis meses todo lo emitido en un total de 900 foros de España, Argentina, México, Chile y Panamá en idioma español.
- *Twitter:* Es un servicio de microblogging dónde los usuarios pueden compartir lo que hacen o están pensando con la limitación de un máximo de 140 caracteres por actualización (o tuit). Se utilizan desde perfiles personales hasta perfiles corporativos, incluyendo bots automáticos de generación de contenido. La restricción en el número de caracteres y la velocidad de generación de contenidos hace que sea un medio especialmente dinámico donde prima la síntesis y el ahorro

[21] https://www.wikipedia.org/

ortográfico. Se ha recopilado durante un periodo de dos años un stream de Twitter de diversas ciudades de España, todo en español.

- *Facebook*[22]*:* Es una red social compuesta por usuarios, páginas, grupos y eventos y la relación entre todos ellos. Un usuario de esta red social puede tener un perfil personal y conectarlo mediante amistad a otros perfiles personales. También puede crear una página o hacer "Me gusta" en otras páginas. Puede crear un grupo o pertenecer a grupos creados por otros usuarios. Y puede generar un evento o apuntarse a eventos generados por otros usuarios. El contenido es dinámico, libre y totalmente personal. El usuario decide qué comparte y qué no comparte con los demás. El uso de Facebook es tanto personal como corporativo. Se ha recopilado de un total de más de 250 proyectos de escucha activa de organizaciones e instituciones de España, Argentina, México, Chile y Panamá todo lo mencionado en páginas, grupos, muros y comentarios relativo a dichos proyectos, en español.

En la Tabla 2.4 se muestran estadísticas relativas al número de documentos recuperados y términos contenidos por medio.

	Wikipedia	**Prensa**	**Blogs**	**Foros**	**Twitter**	**Facebook**
Documentos	3.987.179	5.191.694	1.083.709	673.664	23.873.371	576.723
Términos	267.465.810	499.477.658	122.509.753	21.026.388	163.188.448	28.974.716

TABLA 2.4: Número de documentos y términos recuperados por medio.

A partir de los documentos recuperados se han obtenido una serie de indicadores sobre el uso de las palabras. Posteriormente se ha obtenido la categoría gramatical correspondiente a cada término con Freeling[23] [54, 55]. Las categorías gramaticales que se han tenido en consideración se presentan en la Tabla 2.5.

Adjetivo	Adverbio	Conjunción
Cuantitativo	Determinante	Interjección
Marcador del discurso	Preposición	Pronombre
Sustantivo	Verbo	

TABLA 2.5: Categorías gramaticales consideradas en el análisis estadístico.

Para los pronombres personales se han obtenido los rasgos morfológicos de persona (primera, segunda y tercera) y número (singular y plural). Debido a que en español se elude generalmente el uso de pronombres personales antes de los verbos (sujeto elíptico), se ha procedido a obtener para estos últimos también la persona y el número. Es preciso tener en cuenta que no se ha efectuado ningún tipo de corrección ortográfica ni reducción de términos deformados (e.g. hoooola, ke, t, kiero).

2.2.2 Riqueza léxica

En la Tabla 2.6 se muestran estadísticas relativas a la riqueza léxica de los diferentes medios sociales. La primera fila (términos únicos) muestra el número de palabras únicas del texto y que hayan sido

[22] https://www.facebook.com
[23] http://nlp.lsi.upc.edu/freeling/

identificadas como una de las categorías gramaticales propuestas. Teniendo en cuenta que el español parte de un lemario de aproximadamente 85.918 lemas según la 22° edición del Diccionario de la Real Academia Española, podemos obtener un ratio a partir de las cifras anteriores y el número de lemas del español, permitiendo una comparación relativa a la riqueza de la lengua, como se presenta en la segunda fila. La tercera fila presenta el ratio entre el número de términos totales y el número de documentos por medio, lo que permite obtener la longitud media en palabras de los textos en cada uno de los medios.

	Wikipedia	**Prensa**	**Blogs**	**Foros**	**Twitter**	**Facebook**
Términos únicos	162.357	157.457	162.412	93.145	128.147	110.040
Ratio léxico	1,89	1,83	1,89	1,08	1,49	1,28
Media de palabras	67	96	113	31	7	50

Tabla 2.6: Indicadores de riqueza léxica por medio.

Si podemos valorar la riqueza léxica de una comunicación a partir del número de palabras diferentes empleadas, podemos decir que en este sentido los medios correspondientes a Wikipedia (162.357), Prensa (157.457) y Blogs (162.412) son los que más riqueza léxica tienen, algo inherente a la función principalmente informativa de estos. De los tres anteriores, es la Prensa la que menos variedad léxica presenta, quizás por el esfuerzo extra de los autores de blogs y los colaboradores en Wikipedia por comunicar de una manera más elaborada frente a un estilo de comunicación más objetivo y directo de la Prensa. Es preciso notar el número de términos diferentes utilizados en Twitter (128.147), teniendo en cuenta su limitación a 140 caracteres (7 palabras de media como se ve en la tercera fila). Por último comentar el resultado en Foros (93.145), medio extremamente informal donde el objetivo se persigue mediante un uso del lenguaje directo, en muchas ocasiones basado en preguntas / respuestas sobre un tema, y que en cierta medida viene reflejado en la menor variedad léxica utilizada.

El ratio entre el número de términos totales y el número de documentos por medio permite obtener la longitud media en palabras de los textos en cada uno de los medios. De igual manera que los medios relativos a la Wikipedia (67 palabras), la Prensa (96 palabras) y los Blogs (113 palabras) disponían de una mayor variedad léxica, también disponen de una longitud media por documento mayor, siendo el caso de los Blogs el que supera a los demás. Estos tres medios se erigen como los que más información pretenden aportar. Twitter es el medio que tiene una longitud significativamente menor, de 7 palabras de media por documento, debido a la limitación de los 140 caracteres. Esta limitación del número de caracteres implica una necesidad de síntesis mayor para la transmisión de información. De nuevo los Foros (31 palabras) se ajustan a su característica como medio de información concisa y directa acerca de un tema. Facebook (50 palabras) por su parte se muestra como un medio intermedio entre los Foros y los medios de carácter más informativos como la Wikipedia, la Prensa y los Blogs.

2.2.3 Distribución gramatical

En la Tabla 2.7 se presentan los porcentajes de palabras etiquetadas en cada una de las categorías gramaticales (filas) para cada uno de los medios (columnas). Para los adjetivos podemos observar que el medio Twitter (6,62%) es el que los utiliza en menor proporción, reduciendo su uso a prácticamente la mitad del resto de medios, excepto Foros (9,27%), con un valor intermedio. El adjetivo es la categoría gramatical que acompaña al sustantivo determinándolo, calificándolo o indicando quién lo posee. Ayudan por lo tanto a describir. Según la tabla, su uso es inferior en Twitter y Foros, medios dónde menos descripción detallada se realiza de las cosas.

Categoría	Wikipedia	Prensa	Blogs	Foros	Twitter	Facebook
Adjetivo	13,57	12,50	13,67	9,27	6,62	12,06
Adverbio	2,78	3,46	3,87	4,74	6,30	3,49
Conjunción	1,52	2,10	1,80	4,18	7,00	2,65
Cuantitativo	3,34	4,47	4,15	5,34	5,53	4,29
Determinante	2,88	3,48	2,78	4,18	6,40	4,02
Interjección	0,35	0,04	0,06	0,42	0,38	0,07
Marcador	0,01	0,03	0,02	0,00	0,00	0,00
Preposición	4,00	5,49	5,07	8,94	13,81	6,15
Pronombre	0,65	0,92	1,12	2,22	3,32	1,39
Nombre	50,33	47,05	46,59	42,63	34,08	47,04
Verbo	20,55	20,47	20,88	18,08	16,56	18,83

TABLA 2.7: Distribución (%) de uso de categorías gramaticales por medio.

Para los adverbios observamos el efecto contrario, es en Twitter (6,30%), y en Foros (4,74%) en menor medida, donde se observa un uso mayor en comparación con el resto de medios. La función principal del adverbio es la de acompañar y modificar el significado del verbo, aunque también en ocasiones a otros adjetivos y adverbios. La modificación del verbo se produce en cuanto a tiempo, lugar, modo, intensidad... lo que indica, por su mayor uso en estos medios, que se procura aportar mayor información sobre el contexto de la acción: dónde sucede, cómo sucede, cuándo sucede algo. Ligado a lo anterior tiene que ver el uso de las preposiciones, significativamente superior de nuevo en los medios Twitter (13,81%) y Foros (8,94%). Las preposiciones tienen como principal función permitir una categorización jerárquica y espacial, complemento natural de la modificación adverbial realizada en estos medios.

En cuanto a los sustantivos es interesante notar el mayor uso de los mismos realizado en la Wikipedia (50,33%), algo que se espera por tratarse de una enciclopedia donde se describen objetos, lugares y personas, aunque a un ratio similar al resto de medios, excepto Twitter (34,08%) que es significativamente inferior, lo que muestra un interés menor en hablar de cosas en este último medio. De igual modo y referido a los verbos se detecta un uso significativamente menor en el medio Twitter (16,56%), y en los medios Foros (18,08%) y Facebook (18,83%) frente al resto de medios. El uso de los verbos implica acción, presión o estado, pero también es preciso para la formación de oraciones. Lo primero puede explicar el mayor uso de los verbos en los medios que efectúan descripciones de cosas (Wikipedia, Blogs) y acciones o sucesos (Prensa). Así mismo, el menor uso de los verbos en el medio Twitter se puede deber a la necesidad de reducción y síntesis de lo que se comenta, debido a su limitación en espacio, primando frases cortas a oraciones elaboradas.

Es importante resaltar el mínimo uso de los pronombres en todos los medios, pero siendo significativo su mayor uso en el medio Twitter (3,32%) frente al resto. Para el caso de los pronombres (Tabla 2.8) y los verbos (Tabla 2.9) se han incorporado los ragos morfológicos de persona y número. En el caso de los pronombres se ha obtenido la persona y el número para los personales, y se ha añadido una fila más (Otros) con los porcentajes de los otros pronombres sin persona y/o número (pe. relativos).

Persona	Número	Wikipedia	Prensa	Blogs	Foros	Twitter	Facebook
1	Singular	13,61	14,58	18,85	54,47	65,81	22,30
	Plural	0,00	0,00	0,00	0,00	0,00	0,00
2	Singular	4,58	1,18	2,23	1,54	3,53	3,95
	Plural	1,92	1,75	5,31	4,61	5,62	3,49
3	Singular	55,06	50,75	39,26	24,08	12,70	34,68
	Plural	13,42	18,22	16,93	8,91	3,35	17,14
Otros		11,41	13,52	17,42	6,39	8,99	18,44

TABLA 2.8: Distribución (%) por medio del uso de personas y números en pronombres.

Se aprecia un incremento significativo en el uso de la primera persona del singular en los medios Foros (54,47%) y Twitter (65,81%), describiendo a estos medios como los más egocéntricos en el sentido que emiten más información desde el YO. Es interesante cómo en todos los medios el uso de la primera persona del plural queda completamente en desuso. Así mismo, es importante notar el uso de la tercera persona del singular cuando se trata de los medios Wikipedia (55,06%) y Prensa (50,75%) reafirmando de nuevo a estos medios como descriptores de lo que pasa a las cosas o a terceros. Nótese el decremento significativo del uso de la tercera persona del plural en el caso de Twitter (3,35%), así como el menor uso de otros pronombres en Foros (6,39%) y Twitter (8,99%), reafirmando de nuevo el uso directo y personal de estos medios.

En español el uso del pronombre personal queda reducido en muchas ocasiones por la elipsis del mismo, así que se hace imprescindible el análisis de la persona desde el análisis de las formas verbales. En la Tabla 2.9 se muestra el porcentaje, para cada medio, de uso de las personas y los números en los verbos.

Persona	**Número**	**Wikipedia**	**Prensa**	**Blogs**	**Foros**	**Twitter**	**Facebook**
1	Singular	19,95	17,41	17,50	28,94	24,00	16,61
	Plural	2,10	2,42	4,19	2,68	4,68	4,89
2	Singular	6,02	1,55	3,58	3,55	6,77	2,95
	Plural	0,46	0,42	0,69	0,98	1,65	0,76
3	Singular	31,40	34,00	29,92	28,80	31,21	31,21
	Plural	40,07	44,20	45,11	35,05	31,69	43,59

TABLA 2.9: Distribución (%) por medio del uso de personas y números en verbos.

En este caso es el medio Foros (28,94%) (e.g. *tengo una pregunta*; *¿cómo puedo...?*) el que mayor uso hace de la primera persona del singular, seguido de Twitter (24,00%), en ambos casos significativamente superior al resto de medios. También notar el uso de la primera persona del plural que frente al no uso del pronombre en este tiempo, muestra que en todos los medios ésta es la persona y el número en el que en mayor medida se produce la elipsis.

En la Tabla 2.10 se muestra el detalle de las 15 palabras más frecuentemente utilizadas por cada uno de los medios. En ella se corrobora la afirmación de Pennebaker sobre la alta frecuencia de uso de las palabras de función, pues prácticamente ocupan todas las posiciones.

Wikipedia	**Prensa**	**Blogs**	**Foros**	**Twitter**	**Facebook**
de	de	a	de	de	de
en	la	de	y	que	la
la	el	la	que	a	el
y	en	en	a	la	en
el	a	el	la	el	y
por	que	y	el	y	a
un	y	que	en	en	que
una	del	del	un	no	los
que	los	los	no	me	del
a	por	un	pregunta	un	por
los	un	por	es	es	para
del	se	se	por	se	un
es	con	con	abierta	lo	con
las	las	para	se	con	se
con	para	las	para	por	no

TABLA 2.10: Top 15 palabras más frecuentes por medio.

Es de notar que el pronombre reflexivo *se* aparece en prácticamente todos los medios, en casos como la prensa para sustituir el uso de una persona en la configuración del discurso. De igual modo el verbo *ser* en su tercera persona *es* aparece con cierta frecuencia en todos los medios. Un primer punto de máximo interés es la aparición de dos sustantivos en el medio Foros, medio que por su parte tenía el menor uso de tal categoría gramatical (42,63%) junto con Twitter, y que el primero de ellos (*pregunta*) demuestra el uso principal de este medio como mediador entre usuarios que necesitan información y usuarios que la proporcionan. Un segundo punto también de máximo interés es la aparición de dos pronombres personales en el medio Twitter, uno de primera persona (*me*) y otro de tercera (*lo*), lo que posiciona este medio como el más personal y egocéntrico.

En un análisis por medio se aprecia la similitud entre los medios de la Wikipedia, la Prensa y los Blogs en cuanto al uso de categorías gramaticales como los sustantivos, los verbos y los adjetivos para la descripción de objetos, personas, lugares y situaciones. Es interesante notar la similitud entre los medios Facebook y Prensa, lo que viene determinado por el uso que se realiza del primero de ellos, en una frecuencia muy elevada, de compartir las noticias y comentarlas. Además, el medio Facebook no destaca significativamente en ningún otro aspecto, lo que de nuevo posiciona a este medio como el gran desconocido. El medio Twitter destaca por el uso de los pronombres, especialmente en primera persona al igual que los verbos, los adverbios y las preposiciones, lo que conforma y hace honor a lo que viene a representar, "en qué estás pensando", "qué estás haciendo" o "qué está sucediendo". El medio Foros destaca por su estilo directo (baja variedad léxica) involucrado en un proceso de obtención de información guiado por preguntas/respuestas.

2.2.4 Uso del lenguaje por sexo

A partir del corpus de Facebook anterior, hemos tomado una muestra de 1.200 comentarios donde el perfil del usuario aportaba información suficiente para la anotación de su sexo[24]. En la Tabla 2.11 se muestra la distribución por sexo del uso de categorías gramaticales [56].

Categoría	**Total**	**Mujeres**	**Hombres**
Adjetivo	6,49	6,45	6,53
Adverbio	3,93	3,91	3,94
Conjunción	9,51	9,46	9,55
Cuantificador	5,46	5,12	5,76
Determinante	7,25	7,74	6,81
Interjección	0,23	0,30	0,18
Marcador	0,00	0,00	0,00
Preposición	6,06	5,85	6,25
Pronombre	2,45	2,67	2,24
Nombre	31,89	31,53	32,21
Verbo	15,38	15,32	15,44

TABLA 2.11: Distribución (%) de categorías gramaticales por sexo en Facebook.

Se pueden apreciar algunas variaciones en línea con las apuntadas por Pennebaker para el inglés, como que los hombres usan más preposiciones que las mujeres (+6,84%), quizás porque ellos intentan categorizar jerárquicamente los objetos de su entorno, o que las mujeres usan más los pronombres (+19,20%),

[24]El dataset aquí utilizado se describe en el Apartado 4.3

determinantes (+13,66%) e interjecciones (+66,67%) que los hombres, quizás porque ellas están más interesadas en las relaciones sociales.

2.2.5 Discusión

En el presente estudio se han puesto de manifiesto las variaciones en el uso del lenguaje según el medio de Internet donde se está efectuando la comunicación. Esto implica que el propio medio acentúa el uso de determinadas categorías gramaticales, por ejemplo los pronombres en primera persona en Twitter, y que en la literatura son rasgos identificativos para diferentes tareas de autor profiling. Por ello, se debe tener en cuenta la incorporación de ciertos agentes correctivos en los estudios de *author profiling* basados en texto cuando se realicen sobre diferentes medios sociales de Internet, especialmente en representaciones frecuentistas como los n-gramas.

Por último, hemos analizado la variación de categorías gramaticales por sexo apreciando significativas diferencias. Por ejemplo, que los hombres usan más preposiciones que las mujeres, o que las mujeres utilizan más los pronombres, determinantes e interjecciones, confirmando en este medio lo previamente descubierto por Pennebaker.

2.3 Conclusiones

El estado del arte nos muestra que son diversas las aproximaciones aplicadas al problema de identificar edad y sexo de autores de textos anónimos, y se aprecia que los resultados en muchas ocasiones dependen del tipo de texto sobre el que se aplican las diferentes técnicas. Se observa que la mayoría de trabajos se centran en el inglés, inicialmente en textos formales y cada vez más en textos de social media.

Es interesante notar que los modelos del lenguaje basados en n-gramas son una de las características más utilizadas, conjuntamente con toda una serie de medidas estilísticas tales como frecuencias relativas al uso de signos de puntuación, longitudes de palabras, frases, y similares, combinadas con información morfosintática como las partes del discurso. En esta línea se han desarrollado diversos recursos entre los que destaca por su uso LIWC, el cual permite obtener a partir de un texto todo un conjunto de características relativas a su estilometría, categorías gramaticales o las emociones y sentimientos relacionados con las palabras utilizadas.

Del estudio sobre el uso del lenguaje en diferentes medios se desprenden hechos a tener en consideración. Así como en el estado del arte se postulan hechos como el diferente uso de categorías gramaticales por sexo, o de diferentes tiempos verbales por edad, al analizar su distribución por medios se aprecia que su uso también varía en función de los mismos. Esto es importante y se debe tener en cuenta a la hora de utilizar este tipo de características dependiendo de la tipología de los textos, ya que puede variar del resultado esperado.

Capítulo 3

Author profiling en el PAN

Los trabajos en el área se han enfocado inicialmente en textos formales en inglés, evolucionando con el tiempo hacia las redes sociales y otros idiomas como el español. Aún así, no existe un marco común donde se disponga de corpus y medidas de evaluación normalizadas que permitan obtener, para diferentes medios y en diferentes idiomas, un estado del arte reproducible y comparable.

La creación de un marco de evaluación común es la *alma mater* del *CLEF*, donde a través de la organización de laboratorios se pilotan iniciativas innovadoras de evaluación de tareas relacionadas con el acceso a la información, especialmente multilingüe y multimedia. El laboratorio *PAN* tiene como objetivo, en un sentido amplio, proveer de un espacio de evaluación relacionada con tareas de autoría, originalidad y credibilidad en la red. Dentro de su objetivo de autoría, en 2013 comenzamos a organizar la tarea de *author profiling*.

Uno de los elementos diferenciadores del laboratorio PAN es la utilización de la plataforma TIRA [57, 58] donde los usuarios, en lugar de enviar el resultado de sus ejecuciones para su evaluación, ejecutan directamente su software en la plataforma. Esto proporciona las siguientes ventajas:

- Todos los participantes tienen la misma capacidad computacional.
- Incrementa el compromiso de los participantes.
- Incrementa la sostenibilidad, replicabilidad y reproducibilidad de la tarea.
- Permite la evaluación a través de varios años.

En los siguientes apartados se describe la tarea de identificación de edad y sexo en el laboratorio PAN sobre *author profiling* en los años 2013, 2014 y 2015. El interés de la tarea desde el ámbito forense y del marketing se pone de manifiesto dados los patrocinadores de la misma, respectivamente Forensiclab de la Universitat Pompeu Fabra de Barcelona[1], Corex, Building Knowledge Solutions[2] y Meaning Cloud[3].

[1] http://www.iula.upf.edu/forensiclab/fpresuk.htm
http://pan.webis.de/clef13/pan13-web/author-profiling.html

[2] http://www.corex.es/
http://pan.webis.de/clef14/pan14-web/author-profiling.html

[3] http://www.meaningcloud.com/
http://pan.webis.de/clef15/pan15-web/author-profiling.html

3.1 Author profiling en PAN 2013

En 2013 organizamos la primera tarea internacional en *author profiling* [38] con el objetivo de identificar automáticamente la edad y el sexo de usuarios de medio sociales. A diferencia de la mayoría de investigaciones en el área [10, 15, 16, 20–22, 50, 59–61], nuestro interés se centra también en idiomas diferentes al inglés, por lo que se introduce además el español. Puesto que la tarea se enmarca en un entorno big data como es el de los medios sociales, se pretende que los participantes hagan frente a retos inherentes a las mismas, como por ejemplo:

- Gran variedad de temas, lo que proporciona un mayor espectro de tópicos sobre los que realizar una tarea más realista. Esta gran variedad de temas permite investigar clichés como por ejemplo que los hombres hablan mucho sobre cerveza y fútbol mientras que las mujeres lo hacen sobre sus uñas o las compras. De este modo, se podría reforzar o romper estos tópicos cuando de medios sociales se trata.
- Concurrencia de usuarios falsos (*fakes*), como los bots que se utilizan para la generación automática de contenido, por ejemplo, para mejorar el posicionamiento en buscadores.
- Los medios sociales se usan para hablar de sexo, pero en algunas ocasiones ciertas personas cruzan la línea y las usan para acosar sexualmente a menores. En la edición del 2013 se ha querido comprobar la robustez de los métodos de autor profiling para la identificación de de edad y sexo en depredadores sexuales. Para ello se han incluido algunos textos del corpus de la tarea sobre identificación de depredadores sexuales del PAN 2012 [7].
- En ocasiones los usuarios de los medios sociales las utilizan sólo para saludar o felicitar a sus contactos, por lo que puede resultar en textos extremadamente cortos sin apenas estructura ni contenido, lo que puede dificultar la tarea de identificación.

3.1.1 Corpus

Para la construcción del corpus se ha buscado repositorios abiertos y públicos, como la red social Netlog[4], donde sus usuarios se etiquetan con información demográfica como la edad y el sexo, y cuyos contenidos a modo de posts están también abiertos y públicos. De esta manera se ha construido un corpus de grandes dimensiones de manera automática. Estos posts se han agrupado por autor, seleccionando aquellos autores que tuvieran al menos un post, y descartando al resto. Se ha dividido en más de un fichero aquellos autores que contuvieran más de mil palabras en sus posts. Para mantener un entorno de evaluación realista, se han mantenido también a autores con muy pocos posts y posiblemente cortos (por ejemplo saludando o felicitando).

El corpus se ha dividido aleatoriamente en tres partes: entrenamiento, preevaluación y evaluación. La partición de preevaluación se proporciona a mitad de la tarea para que los usuarios puedan verificar el funcionamiento de sus propuestas y ajustarlas para la evaluación final. La partición de evaluación incorpora a la partición de preevaluación más un 20% de nuevos posts. Se ha asegurado que un mismo autor no estuviera presente en más de uno de los anteriores conjuntos, evitando así posibles sobreajustes. El corpus se ha equilibrado por sexo, pero se ha mantenido desequilibrado por edad, siguiendo una distribución realista del uso de las redes sociales[5]. Los grupos de edad considerados son los siguientes: 10s (de los 13 a los 17 años), 20s (de los 23 a los 27 años) y 30s (de los 33 a los 47 años). Así como se ha

[4] http://www.netlog.com
[5] http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/

comentado en el capítulo anterior, esta división responde a los trabajos pioneros de Koppel *et al.* [16]. Las estadísticas correspondientes al número de autores del corpus se muestran en la Tabla 3.1.

Edad	Sexo	Entrenamiento		Preevaluación		Evaluación	
		Inglés	Español	Inglés	Español	Inglés	Español
10s	hombre	8.600	1.250	740	120	888	144
	mujer	8.600	1.250	740	120	888	144
20s	hombre	(72) 42.828	21.300	3.840	1.920	(32) 4.576	2.304
	mujer	(25) 42.875	21.030	3.840	1.920	(10) 4.598	2.304
30s	hombre	(92) 66.708	15.400	6.020	1.360	(40) 7.184	1.632
	mujer	66.800	15.400	6.020	1.360	7.224	1.632
Σ		236.600	75.900	21.200	6.800	25.440	8.160

TABLA 3.1: Distribución del número de autores del corpus.

Continuando con el intento de mantener un marco realista[6], se ha incorporado un pequeño número de conversaciones sobre depredadores sexuales [7] junto con ejemplos de conversaciones sobre sexo entre adultos. En el corpus en inglés se muestran entre paréntesis las cifras correspondientes a depredadores sexuales para hombres del grupo 20s y 30s, y para conversaciones sobre sexo entre mujeres del grupo 20s. Se libera el corpus bajo el nombre de PAN-AP-13[7].

3.1.2 Aproximaciones

El número de equipos participantes ha sido de 21. A continuación se muestra el análisis de los sistemas descritos por los 18 equipos que enviaron los artículos describiendo sus aproximaciones. En esta sección presentamos un resumen de sus aproximación en términos de preprocesamiento, características utilizadas y algoritmos de clasificación.

Preprocesado. Pocos equipos han preprocesado los datos, por ejemplo sólo *patra13* [62], *moureau13* [63], *meina13* [47], *weren13* [64] ó *aditya13* [65] han limpiado el HTML para obtener texto plano. El equipo de *flekova13* [66] ha borrado documentos que contenían al menos 0.1% de palabras de spam y el equipo de *yong13* [67] ha usado análisis de componentes principales para reducir linealmente la dimensionalidad. Durante la fase de entrenamiento, equipos como el de *caurceldiaz13* [68], *flekova13* [66], *moureau13* [63], *farias13* [69] ó *sapkota13* [70] han seleccionado un subconjunto del corpus para reducir la dimensionalidad. Sólo el equipo de *meina13* [47] ha intentado discriminar entre humanos y robots (*spam* ó *chatbots*).

Características. Muchos de los equipos participantes, como el de *yong13* [67], *haro13* [71], *aditya13* [65], *patra13* [62], *jimenez13* [72], *meina13* [47], *flekova13* [66], *ayala13* [73] ó *santosh13* [74], han utilizado características de estilo como la frecuencia de signos de puntuación, mayúsculas, comillas, etcétera. En algunos casos como en *yong13* [67], *meina13* [47], *ayala13* [73], *haro13* [71] ó *santosh13* [74] han combinado además el uso de partes del discurso (POS, por su nombre en inglés *Part of Speech*). Los equipos

[6] E.g. Las estadísticas existentes muestran que aproximadamente se producen unos 200 tuits por hora en inglés generados por depredadores sexuales (http://www.mirror.co.uk/news/uk-news/paedophiles-using-twitter-to-find-victims-1253833). Twitter ha generado aproximadamente 200 millones de tuits por día en 2011 (https://blog.twitter.com/2011/200- million-tweets-day), llegando a 400 millones de tuits por día en 2013 (http://www.webpronews.com/twitter-turns-7-boasts-400m-tweets-per-day-2013-03). Esto implica que los tuits de los depredadores representen aproximadamente el 0,0012% del total.

[7] https://github.com/autoritas/RD-Lab/tree/master/data/PAN-AP13

de *santosh13* [74], *ramirez13* [70] y *meina13* [47] también han usado características propias del HTML como el uso de imágenes o enlaces a otras páginas. Características de legibilidad han sido ampliamente utilizadas por equipos como el de *patra13* [62], *yong13* [67], *meina13* [47], *flekova13* [66], *ayala13* [73], *weren13* [64] ó *gillam13* [75]. Este último ha utilizado únicamente este tipo de características. Dos equipos, los de *ayala13* [73] y *farias13* [69] han utilizado emoticonos, característica que fue descartada por *ramirez13* [70] por empeorar sus resultados.

Diferentes características de estilo como LSA, bolsas de palabras, *tf-idf*, palabras basadas en diccionario, palabras basadas en tópicos o palabras con alta entropía, han sido usadas por los equipos de *ramirez13* [70], *patra13* [62], *yong13* [67], *mechti13* [76], *hidalgo13* [68], *flekova13* [66], *meina13* [47], *haro13* [71], *santosh13* [74], *aditya13* [65] y *farias13* [69]. El equipo de *flekova13* [66] ha utilizado entidades nombradas, el de *patra13* [62] palabras de sentimiento y los equipos de *meina13* [47], *flekova13* [66] y *farias13* [69] han usado palabras de emoción. Además, los equipos de *flekova13* [66], *diazcaurcel13* [68], *ayala13* [73] y *farias13* [69] han utilizado palabras de jerga, contracciones y palabras con *character flooding*.

Una aproximación diferente basada en recuperación de información ha sido presentada por el equipo de *weren13* [64]. En ella, el texto del autor a ser identificado se utiliza como consulta a un motor de búsqueda. El equipo de *jimenez13* [72] ha utilizado una gran variedad de estadísticas del corpus para construir de manera no supervisada su representación. Modelos basados en n-gramas han sido usados por los equipos de *meina13* [47], *jankowska13* [77], *moreau13* [63] y *ramirez13* [70]. Finalmente, el equipo de *meina13* [47] ha utilizado características lingüísticas más avanzadas como las colocaciones y el equipo de *pastor13* [42], así como se introdujo en el capítulo anterior, ha usado una representación de segundo orden basada en la relación entre los documentos y los perfiles de identificación (edad y sexo).

Métodos de clasificación. Todos los equipos han utilizado métodos de aprendizaje supervisado. La mayoría de ellos, como *santosh13* [74], *patra13* [62], *mechti13* [76], *gillam13* [75] ó *weren13* [64], han usado árboles de decisión. Los equipos de *yong13* [67], *haro13* [71] y *ramirez13* [70] han usado máquinas de vectores soporte, los de *jimenez13* [72] y *flekova* [66] regresión logística y el resto, Naïve Bayes en el caso de *meina13* [47], máxima entropía en el de *pavan13* [65], Stochastic Gradient Descent en el de *hidalgo13* [68] y random forest en el de *ayala* [73].

3.1.3 Resultados

La medida de evaluación utilizada ha sido *accuracy*, calculada como el porcentaje de casos acertados frente al total de casos. Esta evaluación se hace de manera independiente por idioma. Además, se evalúa la identificación conjunta de edad y sexo como el porcentaje de casos acertados para ambos en cada autor frente al número total de autores. Para obtener un resultado global, se hace media de esta identificación conjunta para inglés y español. Mediante esta evaluación, el equipo de *pastor13* ha obtenido el mejor rendimiento general con una precisión de 0.3986. En la Tabla 3.2 se muestran los resultados de todos los participantes [38].

Inglés			
Equipo	Total	Sexo	Edad
meina13	38,94	59,21	64,91
pastor13	38,13	56,90	65,72
mechti13	36,77	58,16	58,97
santosh13	35,08	56,52	64,08
yong13	34,88	56,71	60,98
ladra13	34,20	56,08	61,18
ayala13	32,92	55,22	59,23
gillam13	32,68	54,10	60,31
kern13	31,15	52,67	56,90
haro13	31,14	54,56	59,66
aditya13	28,43	50,00	60,55
hidalgo13	28,40	50,00	56,79
farias13	28,16	56,71	50,61
jankowska13	28,14	53,81	47,38
flekova13	27,85	53,43	52,87
weren13	25,64	50,44	50,99
ramirez13	24,71	47,81	54,15
jimenez13	24,50	49,98	48,85
moreau13	23,95	49,41	48,24
baseline	16,50	50,00	33,33
patra13	15,74	56,83	28,95
cagnina13	7,41	50,40	12,34

Español			
Equipo	Total	Sexo	Edad
santosh13	42,08	64,73	64,30
pastor13	41,58	62,99	65,58
haro13	38,97	61,65	62,19
flekova13	36,83	61,03	59,66
ladra13	35,23	61,38	57,27
jimenez13	31,45	56,27	54,29
kern13	31,34	57,06	53,75
yong13	31,20	54,68	57,05
ramirez13	29,34	51,16	56,51
aditya13	28,24	50,00	56,43
jankowska13	25,92	58,46	42,76
meina13	25,49	52,87	49,30
gillam13	25,43	47,84	,53,77
moreau13	25,39	49,67	50,49
weren13	24,63	53,62	46,15
cagnina13	23,39	55,16	41,48
hidalgo13	20,00	50,00	40,00
farias13	17,57	49,82	35,54
baseline	16,50	50,00	33,33
ayala13	16,38	55,26	29,15
mechti13	2,87	54,55	5,12
patra13	–	–	–

TABLA 3.2: Resultados de la evaluación en términos de *accuracy* para textos en inglés (izquierda) y español (derecha).

Es complicado establecer cierta correlación entre las aproximaciones de los participantes y los resultados obtenidos, desde el momento que gran cantidad de características son compartidas por la mayoría de los equipos. Es destacable el uso de la representación de segundo orden basada en relaciones entre documentos y perfiles que ha permitido a *pastor13* obtener el mejor resultado global, así como el uso de colocaciones por parte del equpo de *meina13* que ha obtenido el mejor resultado en inglés. Sin embargo, las colocaciones no parecen haber dado el mismo resultado en español, aunque podría deberse a la mayor dificultad de obtenerlas en este idioma. Las características de estilo y contenido han sido de las más utilizadas por la mayoría de los equipos, y sus resultados aparecen por diferentes posiciones del ranking, por lo que resulta complejo su análisis. Sin embargo, las POS han sido utilizadas por diferentes equipos, entre ellos el de *meina13* que obtiene el mejor resultado en inglés y el de *santosh13* que lo obtiene en español. En todos los casos los resultados obtenidos con aproximaciones que utilizan POS están bajo la mediana del ranking. Esto nos indica que generalmente mejoran el rendimiento en la tarea tal y como se postulaba en los estudios de Pennebaker [10, 50]. Las características de legibilidad son también ampliamente utilizadas por la mayoría de equipos, pero en este caso sí que podemos comparar su rendimiento ya que el equipo de *gillam13* ha utilizado sólo este tipo de características y ha obtenido la octava posición en inglés y la treceava en español. El uso de palabras de sentimiento y emoción no parecen mejorar el resultado, de la misma manera que las de jerga, aunque es difícil de afirmar ya que se usan conjuntamente con gran variedad de características.

3.1.4 Discusión

En esta primera tarea de *author profiling* organizada en el PAN el objetivo se ha fijado en la identificación de edad y sexo en medios sociales en inglés y en español. El corpus proporcionado proviene de una red social que permite a sus usuarios la anotación de su edad y su sexo, lo que ha propiciado la generación automática de un corpus de gran tamaño, con gran variedad temática, pero que adolece de ruido posiblemente por la presencia de robots o de anotaciones fraudulentas.

Los resultados no son concluyentes con respecto a su relación con las aproximaciones utilizadas, ya que la mayoría de equipos participantes comparten combinaciones de características. Sin embargo, destacan *i*) la representación de segundo orden que relaciona perfiles con autores; *ii*) el uso de colocaciones, especialmente en los resultados en inglés; y *iii*) el uso de partes del discurso, en ambos idiomas.

Los valores de *accuracy* obtenidos muestran la dificultad de la tarea, especialmente en el caso de la identificación conjunta de la edad y el sexo para el mismo usuario, propiciado posiblemente por el entorno *big data* de evaluación, donde se analizan textos de usuarios de medios sociales, con las dificultades que introducíamos al principio del capítulo.

3.2 Author profiling en PAN 2014

En 2014 organizamos la segunda tarea internacional en *author profiling* [39] con el objetivo de identificar automáticamente la edad y el sexo tanto en inglés como en español, pero introduciendo varias novedades respecto a la primera edición de la tarea:

- El foco se centra en investigar como las diferentes aproximaciones de *author profiling* funcionan aplicadas a autores de diferentes medios. Para ello se incorpora, además de un subconjunto refinado de los datos de social media del PAN 2013 [38], datos obtenidos de Twitter, blogs y revisiones de hoteles de TripAdvisor.
- Se colabora estrechamente con el laboratorio RepLab[8] con el objetivo de dotar a la tarea de una segunda perspectiva, la reputación. La colaboración se ha llevado a cabo en dos ámbitos, por un lado la construcción del dataset de Twitter se ha realizado conjuntamente y por otro se ha ofrecido a RepLab la utilización de la plataforma TIRA para su uso en la evaluación de su tarea.

3.2.1 Corpus

Con el objetivo marcado de evaluar las técnicas de *author profiling* aplicadas a diferentes medios, se ha construido el corpus a partir de los siguientes medios: *i*) social media; *ii*) blogs; *iii*) Twitter y; *iv*) revisiones de hotel. Los respectivos subcorpus se han proporcionado tanto el inglés como el español, excepto en el caso de las revisiones de hotel que sólo se proporcionan en inglés. Al igual que en PAN 2013, los contenidos se han agrupado por autor y el autor se ha etiquetado con sexo y edad. En el caso de la edad, en lugar de los tres grupos utilizados en 2013 y que eran *i*) 10s (13-17); *ii*) 20s (23-27); *iii*) 30s (33-47), este año se ha realizado un etiquetado más detallado optando por las siguientes clases: *i*) 18-24; *ii*) 25-34; *iii*) 35-49; *iv*) 50-64; *v*) 65+ . Al igual que en la edición previa, se ha dividido el corpus en tres partes: entrenamiento, preevaluación y evaluación respectivamente. A continuación se

[8] http://nlp.uned.es/replab2014

presentan en detalle los diferentes subcorpus y cómo han sido construidos, y se liberan bajo el nombre de PAN-AP-14[9].

Social Media El subcorpus de social media se ha obtenido como una parte del corpus del PAN 2013 donde se han seleccionado aquellos autores que de media tuvieran un número de palabras superior a 100. Además se han revisado los documentos manualmente para eliminar autores que a priori pudieran parecer falsos como los bots. Para ello se han eliminado los autores que tuvieran repetidos varias veces el mismo post y que además tratara por ejemplo de venta de productos móviles, publicidad o similares, así como autores con un alto contenido de texto reutilizado, como por ejemplo adolescentes compartiendo poesías o citas históricas. La distribución de autores se muestra en la Tabla 3.3. El número de autores está equilibrado por sexo, por lo que la cifra final es el doble de la mostrada.

	Entrenamiento		Preevaluación		Evaluación	
	Inglés	Español	Inglés	Español	Inglés	Español
18-24	1.550	330	140	30	680	150
25-34	2.098	426	180	36	900	180
35-49	2.246	324	200	28	980	138
50-64	1.838	160	160	14	790	70
65+	14	32	12	14	26	28
Σ	7.746	1.272	692	122	3.376	566

TABLA 3.3: Distribución del número de autores del subcorpus de social media.

Blogs Para construir el subcorpus de blogs, se han seleccionado y etiquetado manualmente todos los documentos. En primer lugar, se ha buscado por perfiles públicos en LinkedIn[10] que tuvieran compartido un enlace al blog personal del perfil. Se ha verificado que ese blog existe, que está escrito en alguno de los idiomas en los que estamos interesados (inglés o español) y que además sólo lo escribe una persona y esta persona es identificable con el perfil en cuestión. Hemos descartado blogs corporativos cuando no hemos tenido certeza de que dicho blog lo escribe la persona del perfil analizado. En segundo lugar, hemos buscado información sobre la edad del perfil. En algunos casos el usuario publica su fecha de nacimiento en el perfil de LinkedIn. Si no publica su fecha de nacimiento, hemos tenido en cuenta las fechas de inicio y fin de grado, diplomatura o licenciatura (y sus equivalentes técnicas) para inferir su edad. Para ello hemos usado la información que se muestra en la Tabla 3.4. Hemos descartado usuarios cuyas fechas en el apartado de educación no resultaban claras. En tercer lugar, y si hemos podido inferir la edad, se ha identificado el sexo del usuario a partir de su fotografía y su nombre. De nuevo, en los casos en los que no estuviera clara la información de sexo, descartamo al usuario. Finalmente, este proceso ha sido llevado a cabo por dos anotadores independientes, y en caso de desacuerdo, se ha pedido la opinión a un tercero. Para cada blog, se ha proporcionado un máximo de 25 posts. El texto proporcionado se ha obtenido de los RSS del blog, asegurando la calidad del texto, pero permitiendo al participante descargar el texto completo a partir del enlace permanente del post.

La distribución del número de autores se muestra en la Tabla 3.5. De nuevo, este subcorpus está equilibrado por sexo por lo que se muestra la mitad del número de autores.

[9] https://github.com/autoritas/RD-Lab/tree/master/data/PAN-AP14
[10] https://www.linkedin.com

Fecha de inicio de grado	Grupo de edad
2006-. . .	18-24
1997-2006	25-34
1982-1996	35-49
1967-1981	50-64
. . . -1966	+65

TABLA 3.4: Rangos de edad a partir de la sección educación de LinkedIn (grado, diplomatura o licenciatura y sus equivalentes técnicas).

	Entrenamiento		Preevaluación		Evaluación	
	Inglés	Español	Inglés	Español	Inglés	Español
18-24	6	4	4	2	10	4
25-34	60	26	6	4	24	12
35-49	54	42	8	4	32	26
50-64	23	12	4	2	10	10
65+	4	4	2	2	2	2
Σ	147	88	24	14	78	56

TABLA 3.5: Distribución del número de autores del subcorpus de blogs.

Twitter El subcorpus de Twitter se ha obtenido seleccionando y etiquetando manualmente los documentos siguiendo la misma metodología que el subcorpus de blogs. Como se ha comentado anteriormente, hemos construido este subcorpus en colaboración con RepLab. Desde el punto de vista de la monitorización de la reputación en Twitter, el principal objetivo del *author profiling* es determinar como de influyente es un usuario en el dominio de estudio. Esto implica determinar el tipo de autor (v.g. periodista, stakeholder, profesional) y su grado de influencia en las opiniones del sector. En el caso de la tarea de identificación de edad y sexo, se han etiquetado con edad y sexo un total de 131 usuarios de Twitter de diferentes dominios (energía, medio ambiente, banca, automoción y responsabilidad social corporativa). Los perfiles se han seleccionado del corpus de la tarea de RepLab 2013 [78] y de una lista de autores influyentes proporcionada por la división online de la consultora Llorente & Cuenca[11]. Estos perfiles están además etiquetados manualmente por expertos en reputación respecto a *i*) tipo de autor; *ii*) generador de opinión (influyente, no influyente, indecidible). Para más información específica de este subcorpus, refiérase el lector al overview de la tarea [79]. Debido a los términos del servicio de Twitter[12], se ha proporcionado únicamente la url de los tuits de manera que los participantes se los puedan descargar. Para cada usuario se proporciona hasta un máximo de 1000 tuits. La distribución del número de autores para este subcorpus se puede ver en la Tabla 3.6. El subcorpus de Twitter está equilibrado por sexo por lo que las cifras mostradas se corresponden con la mitad del total.

[11] http://www.llorenteycuenca.com/
[12] https://twitter.com/tos

	Entrenamiento		Preevaluación		Evaluación	
	Inglés	Español	Inglés	Español	Inglés	Español
18-24	20	12	2	2	12	4
25-34	88	42	6	4	56	26
35-49	130	86	16	12	58	46
50-64	60	32	4	6	26	12
65+	8	6	2	2	2	2
Σ	306	178	30	26	154	90

TABLA 3.6: Distribución del número de autores del subcorpus de Twitter.

Revisiones de hotel Para estudiar la aplicabilidad de las técnicas de *author profiling* al sexo de las revisiones de hoteles se ha compilado el corpus Webis-TripAd-13. Este corpus se ha derivado de otro que originalmente fue utilizado para la tarea de *aspect-level rating prediction* [61].[13] El orpus original se recuperó del sitio TripAdvisor[14] en el periodo de un mes desde mitad de febrero hasta mitad de marzo de 2009. Contiene 235 793 revisiones sobre un total de 1 850 hoteles diferentes. Cada revisión comprende el nombre del autor, el texto de la revisión y la fecha de la misma. Además, hay siete posibles valoraciones y una puntuación global asignada por el usuario, lo que sirve como ground-truth para aspect-level rating prediction o análisis del sentimiento. Sin embargo, el dataset original no disponía de anotaciones para la edad y el sexo.

Para realizar la anotación de edad y sexo asegurando su calidad se han seguido los siguientes cuatro pasos. En primer lugar, se han eliminado revisiones cortas con menos de 10 palabras. En segundo lugar, se han eliminado revisiones cuyos texto no se identifiquen como escritos en inglés según un detector automático de idioma. En tercer lugar, ya que el dataset original no incorporaba información sobre sexo y edad, se ha compilado una lista de nombres de usuarios que enviaron las revisiones y se ha recuperado el perfil de dichos usuarios de la web de TripAdvisor. En cuarto lugar, dada esta información se ha descartado aquellas revisiones escritas por autores cuya edad y sexo no están presentes en el perfil o si el perfil está inactivo. Además, para asegurar la calidad de los datos, se han revisado manualmente los perfiles y las revisiones para asegurar que la información proporcionada tiene sentido. El corpus final contiene 58 101 revisiones. La distribución de autores y revisiones se muestra en la Tabla 3.7.[15]

	Entrenamiento	Evaluación
18-24	360	148
25-34	1.000	400
35-49	1.000	400
50-64	1.000	400
65+	800	294
Σ	4.160	1.642

TABLA 3.7: Distribución del número de autores del subcorpus de revisiones de hotel.

[13] http://times.cs.uiuc.edu/~wang296/Data
[14] http://www.tripadvisor.com
[15] Esta versión del corpus se ha liberado en: http://www.webis.de/research/corpora

3.2.2 Aproximaciones

El número de equipos participantes en esta edición de la tarea ha sido de 10, de los cuales 8 han enviado un artículo explicativo de sus aproximaciones, *liau4* nos ha proporcionado una descripción y *castillojuarez14* nos ha comentado que ha participado con su sistema de la edición anterior *ayala13* [73]. En esta sección presentamos un resumen de sus aproximaciones en términos de preprocesamiento, características utilizadas y algoritmos de clasificación.

Preprocesado. Equipos como los de *shrestha14* [80], *marquardt14* [81], *baker14* [82], *ashok14* [83] ó *weren14* [84] han limpiado el HTML y XML para obtener texto plano, así como *ashok14* [83] ha eliminado metadatos como las urls, las menciones a otros usuarios o los hashtags. El equipo de *baker14* [82] ha convertido a minúsculas los textos y ha eliminado caracteres extraños de manera similar a *weren14* [84] , así como han eliminado múltiples espacios consecutivos. Un preprocesado más elaborado ha sido realizado por los equipos de *villenaroman14* [85] y *weren14* [84] que han tokenizado el texto, y además éste último ha estudiado el efecto de la selección de características. Por último el equipo de *marquardt14* [81] ha intentado eliminar tuits de spam o creados por bots mediante la heurística de eliminar contenidos con una alta frecuencia de aparición del caracter %.

Características. Diferentes equipos son los que han utilizado características de estilo, como por ejemplo *ashok14* [83], *mechti14* [86], *baker14* [82], *shrestha14* [80] y *weren14* [84] que han usado frecuencias de diferentes tipos de signos de puntuación, longitud de frases y palabras que aparecen una o dos veces, *mechti14* [86] que ha tenido en cuenta el uso de la flexión verbal o *weren14* [84] que ha considerado el número de caracteres, palabras y frases. El equipo de *marquardt14* [81] ha obtenido el número de posts por usuario y la frecuencia de caracteres y palabras en mayúscuas, mientras que el equipo de *weren14* [84] ha medido la corrección, claridad y diversidad de los textos. Tan sólo los equipos de *weren14* [84] y *marquardt14* [81] han aprovechado la información proporcionada por el HTML como los elementos *img*, *href* ó *br*. También han sido ampliamente explotadas las características de legibilidad por equipos como los de *mechti14* [86], *marquardt14* [81], *ashok14* [83], *baker14* [82] y *weren14* [84]. Por ejemplo, los equipos de *marquardt14* [81] y *ashok14* [83] han calculado el Automated Readability Index, el Coleman-Liau Index y el Rix Readability Index. Independientemente, *ashok14* [83] ha calculado el Gunning Fog Index y *weren14* [84] el Flesch-Kinkaid Coleman-Liau. Los equipos de *mechti14* [86] y *ashok14* [83] han realizado un análisis léxico, obteniendo las partes del discurso como características y combinándolas con la identificación de nombres propios o el uso de palabras con *character flooding* (e.g. hellooooo). Los equipos de *shrestha14* [80], *marquardt14* [81] y *liau14* han utilizado la aparición de emoticonos.

Con respecto a características de contenido, equipos como los de *villenaroman14* [85], *shrestha14* [80] ó *liau14* han modelado el lenguaje con n-gramas o bolsas de palabras. El equipo de *mechti* [86] ha extraído palabras tópico como *money*, *home*, *smartphone*, *games*, *sports*, *job*, *marketing*, etcétera. Por su parte, el equipo de *marquardt* [81] ha utilizado recursos como MRC[16] y LIWC para obtener características como las frecuencias de palabras relacionadas con diferentes conceptos sociolingüísticos como *familiarity*, *concreteness*, *imagery*, *motion*, *emotion*, *religion*, entre otras. Otros equipos han usado diccionarios para diferenciar palabras por subcorpus y clases, como en el caso de *baker14* [82], palabras foráneas en el caso de *ashok14* [83] o en el caso de *marquardt14* [81] identificar errores léxicos o frases específicas como *mi marido* o *mi mujer*.

El equipo de *weren14* [84] ha utilizado características específicas del campo de la recuperación de información, como por ejemplo la medida de similaridad por coseno o el índice Okapi BM24. Finalmente, el

[16] http://www.psych.rl.ac.uk/

equipo de *marquardt14* [81] ha estimado el sentimiento de las frases y el equipo de *lopezmonroy14* [43] ha utilizado una representación de segundo orden que relaciona términos, documentos, perfiles y subperfiles.

Métodos de clasificación. Todos los participantes han utilizado métodos de aprendizaje supervisado. Por ejemplo, *shrestha14* y *liau14* han utilizado regresión logística, como *weren14* que además ha utilizado diferentes algoritmos por subcorpus, como logic boost, rotation forest, clasificadores multiclase, perceptrón multicapa o logística simple. El equipo de *villenaroman14* ha utilizado Naïve Bayes multinomial, el de *lopezmonroy14* [43] libLINEAR, el de *ashok14* [83] random forest, el de *marquardt14* máquinas de vectores soporte y el de *metchi14* [86] tablas de decisión. Por último, el equipo de *baker14* [82] ha implementado su propia función de predicción basada en frecuencias.

3.2.3 Resultados

A continuación se muestran y discuten los resultados obtenidos en cada idioma y para cada uno de los subcorpus. Junto con los resultados de los participantes, se ha considerado una baseline consistente en los 1 000 trigramas de caracteres más frecuentes.

Los mejores resultados se han obtenido para Twitter, donde la identificación de sexo ha obtenido mejores resultados en inglés a diferencia de la identificación de edad y la identificación conjunta que los ha obtenido en español. En cuanto a los blogs, los mejores resultados en identificación de sexo se obtienen en inglés y en identificación de edad en español. Aunque los resultados de la identificación conjunta son similares para ambos idiomas, en inglés hay más participantes con precisiones más elevadas. La menor *accuracy* se ha obtenido en la identificación de blogs en español, con valores muy cercanos a la baseline estadística. En caso de identificación de edad, los resultados obtenidos en blogs y social media son comparables. Los peores resultados en la indentificación conjunta se han obtenido en social media en inglés y en revisiones de hotel. En este último además se han obtenido los peores resultados en identificación de edad. Los peores resultados en identificación de sexo se han obtenido en blogs en inglés, de nuevo con valores muy próximos a la baseline estadística. Por el contrario, los mejores resultados para la identificación de sexo se han obtenido en Twitter y en revisiones de hotel. Es destacable la posición en el ranking de la baseline en revisiones de hotel con valores para la identificación de sexo de 66,26 y con una posición media en la identificación conjunta.

El valor de *accuracy* más alto se ha obtenido por *liau14* en la identificación de sexo en Twitter en inglés (73,38), por *shrestha14* [80] en la identificación de edad en Twitter en español (61,11) y en la identificación conjunta en Twitter en español (43,33). Es difícil de establecer una correlación entre aproximaciones y resultados, pero analizando el top tres por subcorpus y subtarea podemos apreciar que características simples como las bolsas de palabras o los n-gramas de palabras obtienen los mejores resultados. En este sentido, las bolsas de palabras utilizadas por *liau14*, los n-gramas de palabras usados por *shrestha14* y los modelos de vectores de términos utilizados por *villenaroman14* han obtenido los mejores resultados en la mayoría de los subcorpus. También es remarcable la contribución de las características basadas en recuperación de información utilizadas por *weren14* en todas las subtareas en blogs en inglés, en la identificación conjunta de social media en inglés, en la identificación de edad en Twitter en español, social media en español y revisiones de hotel, en la identificación de sexo en blogs en español y en la identificación conjunta de social media en inglés. La mezcla de características de contenido y estilo de *marquardt14* ha obtenido buenos resultados en la identificación de sexo en Twitter en español y en todas las subtareas en blogs en español. La segunda posición en la identificación de sexo en social media en español la obtiene la baseline, pero los bajos valores de esta baseline en el resto de subcorpus

demuestra que el uso de n-gramas de caracteres no parece ser una aproximación demasiado buena en *author profiling*.

Inglés			
Equipo	Conjunta	Sexo	Edad
shrestha14	20,62	53,82	36,52
liau14	19,52	53,85	36,05
weren14	19,14	53,61	34,89
villenaroman14	19,05	54,21	35,81
lopezmonroy14	19,02	52,37	35,52
castillojuarez14	14,45	50,53	28,55
marquardt14	14,28	52,16	27,01
ashok14	13,18	51,98	25,15
baker14	12,77	50,12	24,94
mechti14	012,44	51,98	23,55
baseline	9,30	50,74	19,25

Español			
Equipo	Conjunta	Sexo	Edad
liau14	33,57	68,37	48,94
shrestha14	28,45	64,49	42,76
lopezmonroy14	28,09	64,31	45,23
weren14	27,92	63,07	43,82
marquardt14	21,02	64,31	34,45
villenaroman14	19,61	57,24	36,22
baseline	18,20	65,55	28,62
baker14	16,78	50,00	34,45
castillojuarez14	12,54	49,82	25,09
mechti14	10,60	59,19	21,91
ashok14	-	-	-

Tabla 3.8: Resultados en social media en términos de *accuracy*.

Inglés			
Equipo	Conjunta	Sexo	Edad
lopezmonroy14	30,77	67,95	39,74
villenaroman14	30,77	64,10	39,74
weren14	29,49	64,10	46,15
liau14	26,92	65,38	34,62
shrestha14	23,08	57,69	38,46
castillojuarez14	17,95	51,28	33,33
ashok14	12,82	42,31	25,64
baker14	12,82	50,00	29,49
marquardt14	12,82	46,15	26,92
baseline	8,97	57,69	14,10
mechti14	8,97	58,97	17,95

Español			
Equipo	Conjunta	Sexo	Edad
lopezmonroy14	32,14	58,93	48,21
marquardt14	26,79	51,79	48,21
shrestha14	25,00	42,86	46,43
baker14	23,21	50,00	44,64
liau14	23,21	50,00	44,64
villenaroman14	23,21	51,79	46,43
mechti14	17,86	50,00	28,57
weren14	17,86	53,57	25,00
castillojuarez14	8,93	44,64	26,79
baseline	5,36	53,57	16,07
ashok14	-	-	-

Tabla 3.9: Resultados en blogs en términos de *accuracy*.

Inglés			
Equipo	Conjunta	Sexo	Edad
lopezmonroy14	35,71	72,08	49,35
liau14	35,06	73,38	50,65
shrestha14	30,52	66,88	44,16
villenaroman14	20,78	51,30	41,56
weren14	20,13	57,14	33,12
ashok14	19,48	50,00	38,96
marquardt14	19,48	52,60	37,66
baker14	16,88	50,65	33,77
baseline	14,94	59,74	27,92
mechti14	5,84	53,90	11,04
castilloJuarez14	-	-	-

Español			
Equipo	Conjunta	Sexo	Edad
shrestha14	43,33	65,56	61,11
lopezmonroy14	34,44	60,00	53,33
liau14	32,22	63,33	50,00
marquardt14	31,11	61,11	52,22
weren14	27,78	53,33	52,22
villenaroman14	26,67	54,44	50,00
baseline	23,33	47,78	46,67
baker14	21,11	50,00	48,89
mechti14	14,44	51,11	22,22
ashok14	-	-	-
castillojuarez14	-	-	-

TABLA 3.10: Resultados en Twitter en términos de *accuracy*.

Inglés			
Equipo	Conjunta	Sexo	Edad
liau14	25,64	72,59	35,02
lopezmonroy14	22,47	68,09	33,37
shrestha14	22,23	66,87	33,31
weren14	22,11	67,78	33,43
villenaroman14	21,99	68,45	31,43
baseline	18,21	66,26	27,53
marquardt14	14,37	57,00	24,36
baker14	13,82	52,92	25,94
ashok14	12,91	51,89	24,54
castillojuarez14	12,36	50,91	24,18
mechti14	4,51	50,12	9,01

TABLA 3.11: Resultados en revisiones de hotel en términos de *accuracy*.

En la Tabla 3.12 se muestra la *accuracy* de las diferentes aproximaciones en la tarea de identificación conjunta por subcorpus y la media global. Se puede observar que el mejor rendimiento general lo obtiene el equipo de *lopezmonroy14* con su representación de segundo orden. De esta tabla podemos inferir que: *i*) los mejores resultados se han obtenido en Twitter quizás debido al elevado número de documentos (tuits) por autor en comparación con el resto de medios y muy probablemente a la espontaneidad con la que sus usuarios se expresan en este medio; *ii*) el peor resultado se ha obtenido en social media en inglés y en revisiones de hotel, debido a los peores resultados en la identificación de sexo en el primer caso y en la identificación de edad en el segundo.

Ranking	Equipo	Media	Social Media		Blogs		Twitter		Revisiones
			EN	ES	EN	ES	EN	ES	EN
1	lopezmonroy14	28,95	19,02	28,09	30,77	32,14	35,71	34,44	22,47
2	liau14	28,02	19,52	33,57	26,92	23,21	35,06	32,22	25,64
3	shrestha14	27,60	20,62	28,45	23,08	25,00	30,52	43,33	22,23
4	weren14	23,49	19,14	27,92	29,49	17,86	20,13	27,78	22,11
5	villenaroman14	23,15	19,05	19,61	30,77	23,21	20,78	26,67	21,99
6	marquardt14	19,98	14,28	21,02	12,82	26,79	19,48	31,11	14,37
7	baker14	16,77	12,77	16,78	12,82	23,21	16,88	21,11	13,82
8	baseline	14,04	9,30	18,20	8,97	5,36	14,94	23,33	18,21
9	mechti14	10,67	12,44	10,60	8,97	17,86	5,84	14,44	4,51
10	castillojuarez14	9,46	14,45	12,54	17,95	8,93	-	-	12,36
11	ashok14	8,34	13,18	-	12,82	-	19,48	-	12,91

TABLA 3.12: Resultados de la identificación conjunta en términos de *accuracy*.

En la Figura 3.1 se muestra la media y la desviación típida de las distancias entre las edades predecidas y las reales. La mayor distancia de media se produce en las revisiones de hotel con un valor de 1.69. Las menores distancias tanto en media como desviación típica se producen para Twitter. La lista completa de distancias entre participantes y para cada subcorpus se muestra en el Apéndice B.

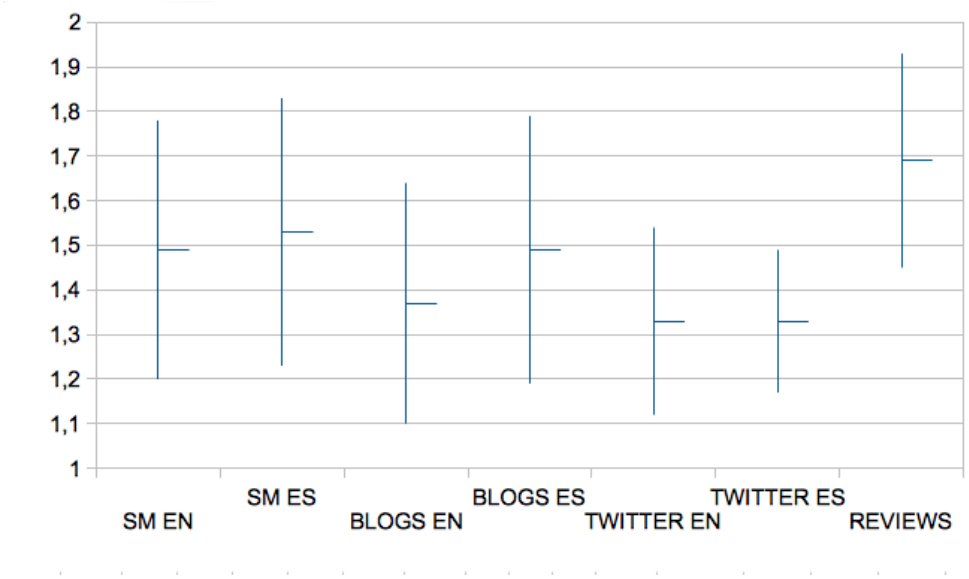

FIGURA 3.1: Distancias entre edades predecidas y reales por subcorpus.

En el Apéndice B se muestran los valores de significación estadística par a par entre todos los sistemas participantes. Como se puede apreciar en la Tabla A17, aunque *lopezmonroy14* obtiene la primera posición en el ranking, estadísticamente su sistema no es significativamente diferente del de *shrestha14*, *villenaroman14* y *weren14*. Todos los sistemas son significativamente diferentes de la baseline, aunque los de *weren14*, *villenaroman14* y *marquardt14* forman un grupo muy cercano a la baseline. Es destacable que la mayoría de los sistemas son estadísticamente indistinguibles en cuanto a social media en inglés, Twitter en español y blogs en ambos idiomas.

Respecto a la identificación de edad, todos los sistemas son significativamente diferentes de la baseline excepto *ashok14*. Sistemas como el de *lopezmonroy14* y *liau14* ó *weren14* y *villenaroman14* no son significativamente diferentes. En blogs la mayoría de los sistemas son indistinguibles entre ellos pero significativamente diferentes de la baseline. En el resto de medios la mayoría de sistemas son además diferentes de la baseline. Analizando las precisiones se aprecia que la mayoría de los sistemas obtienen resultados significativamente mejores que la baseline en la identificación de edad.

Respecto a la identificación de sexo, todos los sistemas son significativamente diferentes de la baseline, pero los sistemas de *lopezmonroy14*, *marquardt14*, *shrestha14*, *villenaroman14* y *weren14* forman un grupo cerrado. En social media en inglés, en blogs en inglés y español y en Twitter en español, la mayoría de los sistemas no son significativamente diferentes. Aunque todos los sistemas son significativamente diferentes de la baseline, la mayoría de ellos son estadísticamente indistinguibles. Sin embargo, no podemos concluir que los sistemas funcionan mejor o peor que la baseline de manera significativa en la identificación de sexo. Por ejemplo, en social media en inglés todos los sistemas que son diferentes de la baseline funcionan mejor en la identificación de sexo, al igual que en Twitter, pero en social media en español ocurre lo contrario y todos los sistemas funcionan peor. Lo mismo ocurre en las revisiones de hotel, donde la mayoría de los sistemas funcionan peor.

3.2.4 Resultados entre años

Se ha ejecutado el software de los participantes de la tarea del PAN 2013 en los documentos del subcorpus de social media del PAN 2014. La comparación para la identificación de edad no ha sido posible realizarla por el cambio en los rangos utilizados, por lo que se ha centrado la comparación en la identificación de sexo. Algunos softwares fallaron por lo que sólo mostramos los que pudieron ejecutarse en la plataforma TIRA.

Inglés	
Equipo	Sexo
lopezmonroy13	54,38
villenaroman14	54,21
liau14	53,85
shrestha14	53,82
weren14	53,61
cagnina13	52,87
lopezmonroy14	52,37
marquardt14	52,16
ashok14	51,98
mecthi14	51,98
baseline	50,74
castillojuarez14	50,53
haro13	50,36
baker14	50,12
ramirez13	49,82
jimenez13	49,67
patra13	49,17

Español	
Equipo	Sexo
cagnina13	69,43
haro13	68,55
liau14	68,37
baseline	65,55
shrestha14	64,49
lopezmonroy14	64,31
marquardt14	64,31
lopezmonroy13	63,36
weren14	63,07
jimenez13	62,37
mechti14	59,19
villenaroman14	57,24
ramirez13	54,59
baker14	50,00
castillojuarez14	49,82

TABLA 3.13: Aproximaciones del PAN 2013 evaluadas en el subcorpus de social media del PAN 2014 para la identificación de sexo.

El único participante con resultados para ambos años es *lopezmonroy.*[17] La comparación de resultados se muestra en la Tabla 3.13. En inglés, aunque el mejor resultado fue obtenido por *lopezmonroy13* [42],

[17]El equipo de lopezmonroy se identificaba como pastor en PAN 2013 (equipo que obtiene el mejor resultado en ambas ediciones)

la mayoría de las aproximaciones del PAN 2014 obtienen mejores resultados que las del PAN 2013. En español los resultados están más equilibrados entre los equipos de ambos años, aunque los dos mejores resultados se obtuvieron respectivamente por los equipos de *cagnina13* y *haro13* [71]. Es destacable el alto número de aproximaciones por debajo de la baseline en español, así como que los resultados obtenidos en español son superiores a los obtenidos para el inglés (siendo el español un idioma marcado por género). Respecto a los participantes de ambos años, *lopezmonroy13* obtuvo mejores resultados que *lopezmonroy14* en inglés pero no en español.

3.2.5 Discusión

En esta segunda tarea de *author profiling* organizada en el marco del PAN 2014 en el CLEF 2014 el objetivo se ha mantenido en la identificación de edad y sexo en cuatro diferentes medios: social media, blogs, Twitter y revisiones de hotel. Todos ellos se han realizado en inglés y español, excepto las revisiones de hoteles que se proporcionaron sólo en inglés. Al igual que en la anterior edición, los equipos participantes han utilizado gran cantidad de características diferentes, donde características sencillas como las bolsas de palabras, los n-gramas de palabras o los vectores de términos han permitido obtener buenos resultados. Los n-gramas de caracteres han demostrado no ser demasiado buenos para la tarea. También en esta edición el mejor resultado global lo ha obtenido la representación de segundo orden que relaciona términos, documentos, perfiles y subperfiles propuesta por el equipo de *lopezmoroy14*. Con respecto a los diferentes medios, podemos concluir lo siguiente: *i*) los mejores resultados se han obtenido en Twitter. Pensamos que esto se debe al hecho de que aunque los tuits son cortos, se dispone de gran cantidad de ellos por cada autor, y que además son un medio muy espontáneo de comunicación; *ii*) los peores resultados se han obtenido en inglés en social media y revisiones de hoteles, debido principalmente a los peores resultados en identificación de edad y sexo respectivamente; *iii*) la mayor distancia entre la edad predecida y la real se produce en las revisiones de hoteles. Sería necesario un análisis más detallado para determinar por ejemplo si se debe a la existencia de casos de opiniones falsas.

3.3 Author profiling en PAN 2015

En 2015 organizamos la tercera edición de la tarea en *author profiling* [40] con el objetivo de continuar con la identificación de sexo y edad, además de las siguientes novedades:

- Se introduce una nueva tarea de identificación de rasgos de personalidad en base a la teoría de los Big Five [87–89];
- Se incorporan nuevos idiomas además del inglés y el español, concretamente el italiano y el holandés, aunque sólo para la identificación de edad.

Puesto que la tesis se centra en identificación de edad y sexo, se describirá en este apartado atendiendo únicamente a estas dimensiones y refiriendo al lector al overview de la tarea [40] para cuestiones relacionadas con el reconocimiento de la personalidad.

3.3.1 Corpus

El corpus del PAN 2015 lo hemos recopilado de Twitter teniendo en consideración cuatro idiomas diferentes: inglés, español, italiano y holandés. El corpus ha sido anotado con sexo y en el caso del

español y el inglés con grupos de edad. La información sobre edad y sexo la ha proporcionado el propio usuario de Twitter en el test proporcionado online[18]. Para los casos del italiano y del holandés, no se ha conseguido una participación suficiente de los usuarios en el test y no se pudo compilar información sobre la edad. Para etiquetar la edad hemos considerado las siguientes clases: *a)* 18-24, *b)* 25-34, *c)* 35-49, and *d)* 50+. El corpus se ha dividido aleatoriamente en tres partes: entrenamiento, preevaluación y evaluación, asegurando que un mismo usuario no forme parte de más de una parte a la vez. El corpus está equilibrado por sexo pero desequilibrado por edad. La distribución de número de autores se muestra en la Figura 3.14. El corpus se libera bajo el nombre de PAN-AP-15[19].

	Entrenamiento				Preevaluación				Evaluación			
	EN	ES	IT	DU	EN	ES	IT	DU	EN	ES	IT	DU
18-24	58	22			16	6			56	18		
25-34	60	56			16	14			58	44		
35-49	22	22			6	6			20	18		
50+	12	10			4	4			8	8		
Hombre	76	55	19	17	21	15	6	5	71	44	18	16
Mujer	76	55	19	17	21	15	6	5	71	44	18	16
Σ	152	110	38	34	42	30	12	10	142	88	36	32

TABLA 3.14: Distribución del número de usuarios de Twitter del corpus.

3.3.2 Aproximaciones

El número de equipos participantes en esta edición ha sido de 22, siendo 21 los equipos que han presentado un artículo describiendo su trabajo. En esta sección presentamos un resumen de sus aproximación en términos de preprocesamiento, características utilizadas y algoritmos de clasificación.

Preprocesado. La mayoría de los equipos han realizado algún tipo de preprocesamiento, especialmente la limpieza del HTML que pudiera aparecer en los tuits como ha sido en el caso de los equipos de *arroju15* [90], *grivas15* [91], *cheema15* [92] y *ashraf15* [93]. Equipos como los de *arroju15* [90], *gonzalezgallardo15* [94], *grivas15* [91], *maharjan15* [95] ó *nowson15* [96] han hecho uso de hashtags, urls y menciones. Por ejemplo, el equipo de *gonzalezgallardo15* [94] ha sustituido menciones, urls y hashtags por tokens específicos. De manera similar el equipo de *maharjan15* [95] ha reemplazado urls por el token URL. Aunque antes de liberar el dataset se ha procurado eliminar elementos como los RTs, algunos equipos como el de *bartoli15* [97] o *poulston15* [98] han preprocesado el dataset para eliminar los posibles RTs y tuits compartidos todavía existentes. El equipo de *weren15* [99] ha pasado a minúsculas el texto y eliminado números y palabras vacías, así como ha aplicado un proceso de stem. El equipo de *nowson15* [96] ha eliminado todas las secuencias de caracteres que representan emojis y el equipo de *markov15* [100] ha eliminado tuits con menos de cinco palabras.

Características. De manera similar a ediciones anteriores, los equipos han utilizado múltiples combinaciones de características basadas en estilo y basadas en contenido, así como diferentes combinaciones de modelos de n-gramas como los equipos de *kiprov15* [101] y *poulston15* [98]. Por ejemplo, los equipos de *gonzalezgallardo15* [94], *maharjan15* [95] ó *sulea15* [102] han utilizado n-gramas de caracteres, los

[18] http://your-personality-test.com/
[19] https://github.com/autoritas/RD-Lab/tree/master/data/PAN-AP15

de *arroju15* [90], *gimenez15* [103], *cheema15* [92], *nowson15* [96] y *mezaruiz15* [104] *n*-gramas de palabras, los de *gimenez15* [103], *grivas15* [91], *mezaruiz15* [104] y *sulea15* [102] *n*-gramas ponderados con *tf-idf* y los de *gonzalezgallardo15* [94] y *mezaruiz15* [104] *n*-gramas de partes del discurso. El equipo de *markov15* [100] ha combinado 10 tipos diferentes de *n*-gramas obtenidos del árbol de dependencias sintácticas, como por ejemplo de lemas, palabras, relaciones, partes del discurso, etcétera.

Los participantes también han combinado gran variedad de características estilísticas. Por ejemplos, los equipos de *miculicich15* [105], *mezaruiz15* [104] y *ameer15* [106] han tenido en consideración los signos de puntuación, *nowson15* [96], *mezaruiz15* [104], *markov15* [100] ó *teisseyre15* [107] los emoticonos, grivas15 [91] la longitud de las palabras, *ameer15* [106] la longitud de frases, *gimenez15* [103], *kiprov15* [101] y *nowson15* [96] el uso de *character flooding*, *sulea15* [102] diversos índices de verbosidad, y *gimenez15* [103], *grivas15* [91] y *kiprov15* [101] el uso de mayúsculas y minúsculas. El uso de interrogación ha sido tenido en cuenta por el equipo de *maharjan15* [95] y el de frases interrogativas por el de *ameer15* [106]. Otros participantes han aprovechado elementos específicos de Twitter como links, hashtags o menciones, como ha sido el caso de los equipos de *gimenez15* [103], *grivas15* [91], *kiprov15* [101], *miculicich15* [105], *nowson15* [96], *mezaruiz15* [104] y *markov15* [100].

Con respecto a características de contenido, los equipos de *maharjan15* [95], *mccollister15* [108], *miculicich15* [105], *poulston15* [98] y *ashraf15* [93] han usado modelado de tópicos con LDA. El equipo de *kocher15* [109] ha obtenido los 200 términos más frecuentes y el equipo de *mezaruiz15* [104] ha combinado bolsas de palabras. Además, el equipo de *maharjan15* [95] ha considerado el uso de términos referentes a la familia como mi mujer/mi marido, mi novio/mi novia, etcétera, o el equipo de *cheema15* [92] que ha obtenido listas de las palabras más discriminantes por clase. Finalmente, el equipo de *nowson15* [96] ha utilizado la ocurrencia de entidades nombradas.

Equipos como los de *arroju15* [90], *bartoli15* [97] ó *miculicich15* [105] han utilizado recursos psycolingüísticos como LIWC, o el recurso NRC[20] que ha sido usado por los equipos de *gimenez15* [103] y *kiprov15* [101]. En ambos casos han sido utilizados entre otras cosas para obtener palabras de polaridad y emoción. También se han usado otros diccionarios, algunos de ellos compilados manualmente, como por ejemplo el de palabras irónicas y el de palabras tabú compilados por *mezaruiz15* [104] o el de palabras informativas utilizado por *bartoli15* [97].

El equipo de *weren15* [99] ha utilizado características específicas del campo de la recuperación de información tales como la similaridad por coseno o el modelo Okapi BM25. Finalmente es destacable la combinación de LSA con la representación de segundo orden que relaciona términos, documentos, perfiles y subperfiles realizada por el equipo de *alvarezcarmona15* [44], obteniendo con ella los mejores resultados globales.

Métodos de clasificación. Todos los participantes han aproximado la tarea como un problema de aprendizaje automático, concretamente un problema de clasificación superviada. La mayoría de participantes han utilizado máquinas de vectores soporte, como por ejemplo los equipos de *alvarezcarmona15* [44], *gonzalezgallardo15* [94], *cheema15* [92], *markov15* [100], *grivas15* [91], *kiprov15* [101], *nowson15* [96] y *poulston15* [98]. Otros equipos han utilizado árboles de decisión, como los de *mezaruiz15* [104] y *ashraf15* [93] que han usado *random forest*, o el equipo de *mccollister15* [108] que ha usado *rotation forest*. Otros algoritmos como *logistic regression* han sido usados por *maharjan15* [95] o aproximaciones basadas en distancias por *mirco15* [109], *ameer15* [106] o *teisseyre15* [107].

[20] urlhttp://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

3.3.3 Resultados

En esta sección se muestra un resumen de los resultados obtenidos por los 22 participantes. En la Tabla 3.15 se muestran los resultados por equipo para las tareas de identificación de edad, sexo y conjunta en inglés y español, y de sexo en italiano y holandés. Se puede observar que el equipo de *alvarezcarmona15*, que ha usado la representación de segundo orden que le permitió obtener los mejores resultados en las dos ediciones anteriores del PAN, obtiene los mejores resultados tanto en inglés (72,54) como en español (77,27), la segunda posición en holandés tras *grivas15* (96,88) y empatando con *gonzalezgallardo15* (93,75), que a su vez obtiene el mejor resultado en italiano (86,11) y resultados top en el resto de tareas e idiomas. El equipo de *gonzalezgallardo15* ha utilizado combinaciones de *n*-gramas de caracteres y partes del discurso y el equipo de *grivas15* ha combinado *n*-gramas ponderados por *tf-idf* y diferentes características de estilo.

	Inglés			**Español**			**Italiano**	**Holandés**
Equipo	*Conjunta*	Sexo	Edad	*Conjunta*	Sexo	Edad	Sexo	Sexo
alvarezcarmona15	72,54	85,92	83,80	77,27	96,59	79,55	72,22	93,75
ameer15	50,70	69,01	71,83	42,05	69,32	53,41	58,33	59,38
arroju15	57,04	76,76	70,42	48,86	75,00	69,32	58,33	53,13
ashraf15	39,44	55,63	69,72	-	-	-	-	-
bartoli15	47,18	64,79	74,65	32,95	85,23	42,05	50,00	71,88
bayot15	24,65	50,00	59,15	36,36	61,36	56,82	52,78	56,25
cheema15	42,25	59,15	66,90	45,45	84,09	56,82	52,78	46,88
gimenez15	38,73	63,38	59,86	42,05	62,50	56,82	69,44	71,88
gonzalesgallardo15	69,72	85,21	78,17	70,45	89,77	72,73	86,11	93,75
grivas15	66,90	85,92	74,65	68,18	94,32	69,32	83,33	96,88
kiprov15	59,15	84,51	72,54	72,73	90,91	78,41	-	-
kocher15	55,63	71,13	71,13	67,05	81,82	73,86	77,78	81,25
maharjan15	56,34	74,65	69,01	57,95	79,55	62,50	69,44	78,13
markov15	36,62	59,15	58,45	38,64	65,91	51,14	-	53,13
mccollister15	51,41	72,54	71,83	31,82	68,18	50,00	55,56	81,25
mezaruiz15	21,83	50,00	40,85	40,91	82,95	51,14	50,00	50,00
miculicich15	57,04	78,87	69,01	62,50	92,05	68,18	63,89	81,25
nowson15	37,32	77,46	49,30	48,86	77,27	67,05	80,56	78,13
poulston15	52,11	69,01	73,94	54,55	84,09	59,09	75,00	50,00
sulea15	61,97	76,76	78,87	65,91	87,50	75,00	63,89	84,38
teisseyre15	64,79	83,10	75,35	21,59	55,68	36,36	41,67	59,38
weren15	55,63	76,06	70,42	69,32	84,09	77,27	58,33	65,63

TABLA 3.15: Resultados de la evaluación en términos de *accuracy* para las tareas de identificación de edad y sexo por idioma.

Respecto al inglés se puede apreciar los elevados valores obtenidos por *alvarezcarmona15*, *gonzalezgallardo15* y *grivas15* en todas las subtareas. En español los mejores resultados en identificación de sexo han sido obtenidos por *alvarezcarmona15* (96,59), *grivas15* (94,32) y *kiprov15* (90,91). Estos resultados son significativamente superiores a los del estado del arte descritos en el apartado 2. Además, los resultados para la identificación de edad son muy elevados, con valores sobre el 70%, como por ejemplo *alvarezcarmona15* (83,80), *gonzalesgallardo15* (78,17), *kiprov15* (74,65), *kocher15* (71,13), *sulea15* (78,87) y *weren15* (70,42). Recordando las aproximaciones, *kiprov15* ha combinado diferentes modelos de *n*-gramas con características como la *character flooding* o elementos de Twitter como hashtags, urls y urls, *kocher15* que ha utilizado las palabras más frecuentes por perfil, *sulea15* ha combinado *n*-gramas de caracteres y *tf-idf* con características de estilo como el ratio de verbosidad y *weren15* ha utilizado diferentes características del campo de la recuperación de información. En italiano los resultados para la identificación de sexo son superiores al 80% en algunos casos. Son interesantes los obtenidos por

gonzalesgallardo15 (86,11), *grivas15* (83,33) y *nowson15* (80,56). El último de ellos, ha aplicado combinaciones de n-gramas con partes del discurso, entidades nombradas, *character flooding*, emoticonos y hashtags. Respecto al holandés, los valores más significativos para la identificación de sexo son obtenidos por *alvarezcarmona*(93,75), *gonzalezgallardo15* (0.9375) y *grivas15* (96,88).

Equipos como los de *ashraf15*, *kiprov15* y *markov15* no han participado en todos los idiomas. Son destacables los resultados de *kiprov15* en español (90,91; 78,41 y 72,73 respectivamente en identificación de sexo, edad y conjunta), obteniendo la segunda posición usando un gran conjunto de características que incluyen n-gramas, partes del discurso, *character flooding*, longitud media de las frases, diccionarios como NRC, Hashtag Emotion Lexicon [110], lexicón de palabras malsonantes[21], World Well-Being Project Personality Lexicon [29] o características específicas de Twitter como hashtags, urls, menciones y retuits.

En la Tabla 3.16 se muestra un resumen con los mejores resultados por idioma y tarea. Se puede apreciar que los mejores resultados en identificación del sexo se producen en español y holandés, con valores superiores al 95%. Respecto a la edad, en inglés (83,80) se obtienen resultados ligeramente superiores al español (79,55), no siendo suficientes como para compensar en la identificación conjunta (77,27 del español frente a 72,54 del inglés). En comparación con ediciones previas de la tarea, las aproximaciones de la edición del 2015 han obtenido resultados significativamente superiores tanto para la identificación de edad y sexo, como para la identificación conjunta. Esto sugiere que, a pesar de la limitación en el número de caracteres de Twitter y del uso de un registro más informal por parte de sus usuarios, el número de tuits por autor es suficiente para predecir su edad y sexo con cierta precisión.

Language	*Conjunta*	Sexo	Edad
Inglés	72,54	85,92	83,80
Español	77,27	96,59	79,55
Italiano	-	86,11	-
Holandés	-	96,88	-

Tabla 3.16: Mejores resultados por idioma y tarea.

Si analizamos la distribución de resultados de las diferentes subtareas por separado, podemos apreciar por ejemplo que en la identificación de sexo mostrada en la Figura 3.2, la mayor dispersión en los resultados se produce en holandés, donde además la desviación del mejor resultado al resto es mucho mayor que en el resto de idiomas. Así mismo, excepto en italiano, la mayoría de los sistemas se encuentran bajo la mediana, lo que indica que hay unos pocos sistemas que sobresalen de los demás. Esto sucede especialmente en el caso del español.

Respecto a la distribución de resultados en la identificación de edad mostrada en la Figura 3.3 el caso del inglés destaca por la existencia de dos resultados atípicamente inferiores correspondientes a los sistemas de *mezaruiz15* (40,85) y *nowson15* (49,30). Sin tener en cuenta estos resultados se aprecia que la distribución de los resultados para el inglés, no sólo es ligeramente superior al español, sino que es más compacta en torno a la media y uniformemente distribuida en torno a la mediana. En español por su parte hay una ligera tendencia hacia valores extremos por su rango inferior y una ligera existencia de mayor número de resultados por encima de la mediana, debido principalmente a dichos valores extremos.

Finalmente, la distribución de resultados en la distribución conjunta mostrada en la Figura 3.4 muestra una dispersión mayor en los resultados en español, con valores superiores para unos pocos sistemas

[21] Generado a partir de la combinación de una compilación manual de términos y el diccionario de Google "what do you love" https://opennlp.apache.org

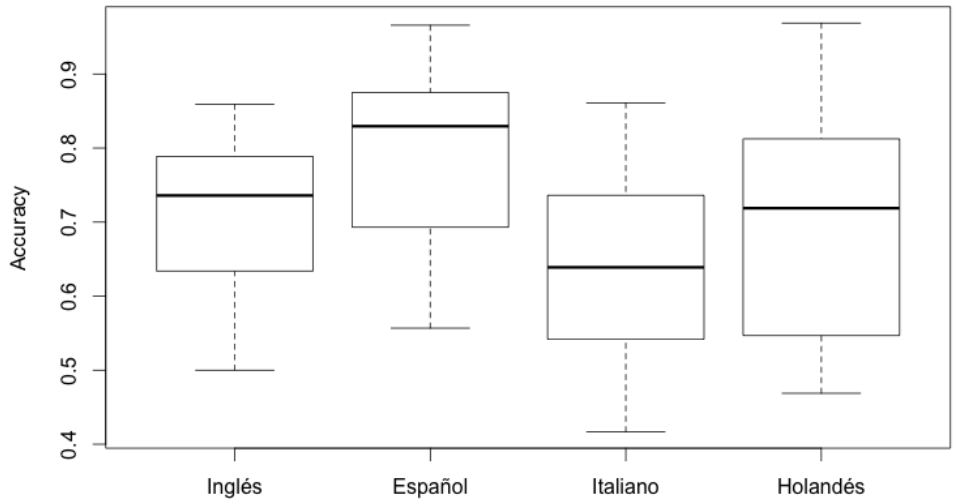

FIGURA 3.2: Distribución de resultados en identificación de sexo por idioma.

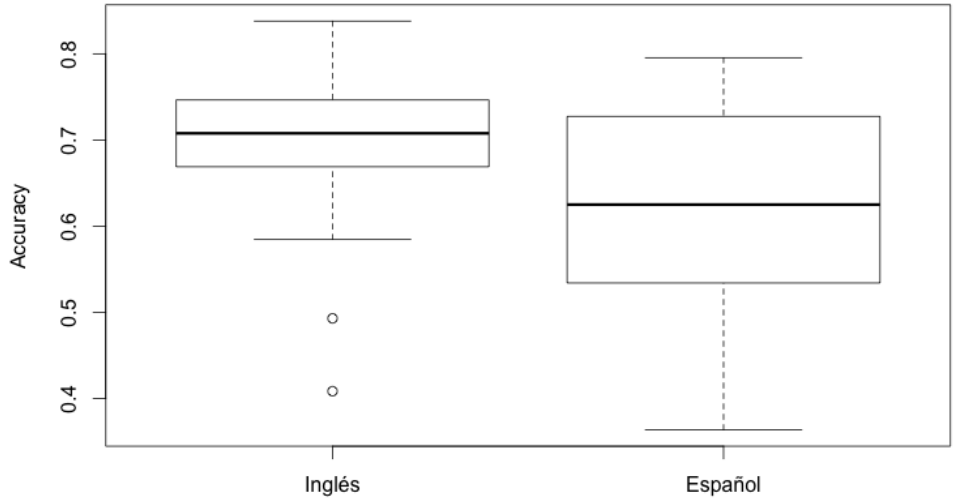

FIGURA 3.3: Distribución de resultados en identificación de edad por idioma.

donde la mayoría se encuentran por debajo de la mediana a diferencia del caso del inglés, donde la mayoría se encuentran por encima de la mediana y cuya dispersión es ligeramente inferior. En ambos casos la tendencia al extremo se produce por abajo donde el peor resultado es similar en ambos casos y correspondiente a *mezaruiz15* (21,83) y *teisseyre15* (21,59) en inglés y español respectivamente.

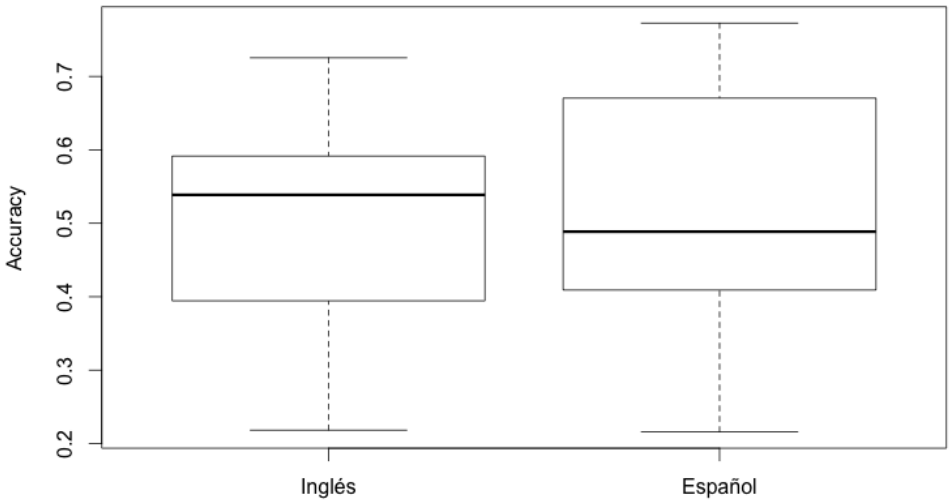

FIGURA 3.4: Distribución de resultados en identificación conjunta por idioma.

3.3.4 Discusión

En esta tercera edición de la tarea de *author profiling* organizada en el marco del PAN 2015 en el CLEF, además de novedades como el reconocimiento de personalidad, el objetivo se ha mantenido en la identificación de sexo y edad, el primero no sólo en español e inglés sino también en italiano y holandés, de usuarios de la red social Twitter. La tarea ha contado con 22 participantes que han utilizado combinaciones de características de contenido (bolsas de palabras, n-gramas de palabras, vectores de términos, n-gramas ponderados con *tf-idf*, entidades nombradas, diccionarios de palabras, palabras de jerga, palabras irónicas, palabras de sentimiento y palabras de emoción, entre otras) y características de estilo (frecuencias, signos de puntuación, partes del discurso e indicadores de verbosidad, entre otros, así como elementos específicos de Twitter como las menciones, los hashtags o el uso de urls).

Las valores superiores de *accuracy* en la identificación de sexo se han producido en holandés y español. En comparación con ediciones anteriores del PAN, los sistemas ha obtenido resultados significativamente superiores tanto en identificación de edad como sexo. Esto sugiere que independientemente de la limitación del número de caracteres de los tuits individuales y el registro informal que se usa, el número de tuits por autor parece ser suficiente para perfilar tanto la edad como el sexo de los mismos con gran precisión.

Respecto a las características que mejores resultados han producido es difícil de determinar debido al elevado número de las mismas que se han usado y combinado en las diferentes aproximaciones. Una vez más en esta edición la representación de segundo orden propuesta por *alvarezcarmona15* ha obtenido los mejores resultados. Sin embargo, representaciones basadas en n-gramas como la propuesta por *gonzalezgallardo15* o *grivas15* han obtenido posiciones en el top tres del ranking para cada uno de los diversos idiomas y subtareas.

3.4 Conclusiones

En este capítulo se ha realizado una revisión de la tarea de *author profiling* organizada en el laboratorio PAN, cuyo objetivo principal en sus tres ediciones ha sido la identificación de la edad y el sexo de autores de textos de social media y en diferentes idiomas. Concretamente, *i*) en 2013 la tarea se ha enfocado en un entorno big data de textos extraídos de un medio social tanto en inglés como en español; *ii*) en 2014 se ha extendido el número de medios además a Twitter, blogs y revisiones de noticias, todos ellos en inglés y español, excepto el último que se ha realizado sólo en inglés. Además, se le ha dotado de una segunda perspectiva relativa a la reputación mediante la colaboración con Replab; y *iii*) en 2015 se ha complementado la tarea con la identificación de personalidad, incorporando el italiano y el holandés, estos últimos sólo para la identificación de sexo.

El número total de participantes ha sido elevado, contribuyendo a la variedad de aproximaciones propuestas. Básicamente tales aproximaciones se pueden catalogar en: *i*) basadas en estilometría, tales como frecuencias de uso de diferentes elementos como las comas, puntos, interrogaciones, emoticonos, longitudes de palabras, frases, y similares; *ii*) basadas en contenido, tales como bolsas de palabras o LDA; *iii*) modelos del lenguaje basados en diferentes tipos de n-gramas, por ejemplo de caracteres, palabras, partes del discurso ó *tf-idf*; y *iv*) modelos de representación novedosos elaborados por sus autores.

La combinación de varias de las anteriores características y la variación en el ranking de los resultados no permite afirmar nada concluyente, aunque sí que algunas representaciones propias como la del sistema que ha obtenido los mejores resultados en las tres ediciones de la tarea, han permitido obtener ciertas ventajas a la hora de capturar la diferencia entre sexos y grupos de edad. Es destacable el bajo uso de características relacionadas con las emociones, limitándose en la mayoría de casos a la identificación de emoticonos o algunos conteos de frecuencias basadas en diccionarios como LIWC.

Capítulo 4

Identificación de emociones en medios sociales

En el Capítulo 2 hemos realizado una revisión exhaustiva del estado del arte en identificación de edad y sexo. Siguiendo los estudios pioneros de Pennebaker [50], hemos investigado cómo se distribuye el uso de rasgos morfosintácticos, como las categorías gramaticales o los tiempos verbales, en diferentes medios sociales de internet. Hemos terminado el capítulo analizando, para un medio social en concreto, cómo usan los autores dichos rasgos en función de su sexo.

En el Capítulo 3 hemos presentado la tarea de *author profiling* organizada en el marco del PAN, cuyo objetivo principal ha sido la identificación de edad y sexo, y que ha permitido generar un marco de evaluación común mediante la generación de los recursos necesarios así como la definición de una medida de evaluación común. En dicho capítulo se aprecia que, pese a la variedad de aproximaciones presentadas, las emociones no han jugado un papel relevante en ellas.

Visto el limitado impacto de las emociones en la identificación de edad y sexo, en este capítulo investigamos más en detalle lo relativo al procesamiento afectivo: cómo se puede modelar y qué puede aportar a la identificación de tendencias en medios sociales, a la identificación del sexo y edad, e incluso de manera colateral al uso de la ironía.

4.1 Trabajos previos

Se debe realizar una clasificación de los trabajos previos bajo dos perspectivas: *i*) la generación de recursos afectivos; y *ii*) los métodos de procesamiento afectivo.

4.1.1 Generación de recursos afectivos

Uno de los recursos más comunes son los diccionarios que incorporan dimensiones afectivas. Son pioneros el Lasswel Value Dictionary [111] donde cada palabra se anota con respecto a dimensiones como riqueza (*wealth*), poder (*power*), rectitud (*rectitude*), respeto (*respect*), capacidad de ilustración (*enlightenment*), habilidad (*skill*), afecto (*affection*) o bienestar (*wellbeing*), y el General Inquirer [112], donde cada

palabra se anota con la existencia de dimensiones como activo (*active*), pasivo (*passive*), fuerte (*strong*), débil (*weak*), placer (*pleasure*), dolor (*pain*), sentimiento (*feeling*), excitación (*arousal*), virtud (*virtue*), vicio (*vice*), sobreestimado (*overstated*) o subestimado (*understated*). Ambos diccionarios sólo utilizan etiquetas binarias sin considerar el grado de aparición.

Al igual que los anteriores, que se basan en obtener la existencia de ciertas dimensiones emotivas, aparece el Clairvoyance Affect Lexicon [113], que además de etiquetar categorías como ira, felicidad y miedo, añade las dimensiones de centralidad e intensidad para complementar la relación de la palabra con la clase afectiva etiquetada.

En la línea de identificar las dimensiones emocionales aparece el Dictionary of Affect in Language (DAL) [114], que consiste en un total de 8.742 palabras marcadas en función de su activación e imágenes (capacidad de imaginar la emoción), y el Affective Norms for English Words (ANEW) [115], cuyo objetivo es tener medidas el mayor número de palabras del inglés en términos de activación, evaluación y control. De manera similar, se desarrolla WordNetAffect [116] como subdominio de WordNet donde cada palabra se etiqueta con su categoría emocional, evaluación y activación.

En [117] los autores usan Mechanical Turk[1] para crear un léxicon de emociones de alta calidad y un tamaño moderado de acerca de 2.000 términos. Los autores muestran cómo términos relacionados con emociones están entre los unigramas y bigramas más comunes, y además identifican qué emociones se tienden a evocar simultáneamente por el mismo término. Para detectar y desechar anotaciones erróneas, los autores generan opciones de palabras de manera automática.

LIWC es un recurso lingüístico, desarrollado por Pennebaker et. al. [118], que mediante análisis de texto proporciona hasta 70 dimensiones tales como el grado de emociones positivas y negativas, auto-referncias, palabras causales, etcétera.

Es destacable que apenas existen recursos en español, siendo destacable la adaptación al español de ANEW [119]. Los autores, con la ayuda de 720 participantes, etiquetan la traducción de 1.034 palabras de ANEW en las dimensiones de polaridad, activación y control. Además, el Spanish Emotion Lexicon (SEL) [120] consistente en 2.036 palabras asociadas con la medida "Probability Factor of Affective use" (PFA) referida a cada una de las seis emociones básicas de Ekman [121]: alegría (*joy*), disgusto (*disgust*), enfado (*anger*), miedo (*fear*), tristeza (*sadness*) y sorpresa (*surprise*). Para calcular el PFA, en base a lo indicado por 19 anotadores referente a cuatro posibles grados de relación entre cada palabra y cada emoción básica (nulo, bajo, medio, alto), se obtiene como media de los porcentajes asignados a cada grado.

Aún no siendo un recurso conviene citar el estudio de Osherenko [122] sobre la necesidad de usar diccionarios afectivos en el análisis de emociones, dónde tratan de contestar a cuestiones como si realmente mejoran la identificación o si pueden ser sustituibles por diccionarios de propósito general.

4.1.2 Métodos de procesamiento afectivo

El procesamiento automático de la afectividad se ha enfocado en gran medida hacia el análisis de sentimiento, donde se trabaja sobre una de las dimensiones del sentimiento: la clasificación de la polaridad. Sin embargo, existe un conjunto de métodos orientados a etiquetar los documentos dentro de la categoría emocional que expresan, por lo general, basándose en las seis emociones básicas de Eckman.

[1] https://www.mturk.com/mturk/welcome

Son destacables tres métodos presentados en SemEval 2007[2] *i*) UPAR7; *ii*) UA; y *iii*) SWAT, donde se incorpora la tarea de anotación de emociones a partir de 1.000 cabeceras de noticias. UPAR7 [123] utiliza el analizador sintáctico de Stanford para identificar qué se está diciendo del tópico principal, estimando cada palabra con ayuda de los recursos SentiWordnet y WordnetAffect, y obteniendo la calificación global de manera incremental. UA [124] utiliza tres motores de búsqueda sobre los que se lanzan como consulta todas las palabras de la cabecera de la noticia y cada una de las emociones, y se calcula el Pointwise Mutual Information según el número de documentos devueltos. SWAT [125] es un sistema supervisado basado en unigramas entrenado a partir de otras 1.000 noticias que anotaron manualmente sus autores y que utiliza el tesauro de Roget[3] para expandir sinónimos y construir las características.

En [126] se presenta el detalle de los resultados de la tarea, que a su vez se comparan con cinco propuestas que ellos mismos realizan: *i*) WN-AFFECT PRESENCE, que anota las emociones basándose en la presencia de palabras de WordnetAffect; *ii*) LSA SINGLE WORD, que calcula la similitud LSA entre cada texto y cada emoción, tomando ciertas palabras como *joy* como representativas de la emoción; *iii*) LSA EMOTION SYNSET, que añade sinónimos de Wordnet; *iv*) LSA ALL EMOTIONS, que además incorpora todas las palabras anotadas en WordnetAffect como contenedoras de emoción; y *v*) NB TRAINED ON BLOGS, basado en un clasificador Naïve Bayes sobre un corpus de blogs.

Método	r	Precisión	*Recall*	F1
WN-AFFECT PRESENCE	9,54	38,28	1,54	4,00
LSA SINGLE WORD	12,36	9,88	66,72	16,37
LSA EMOTION SYNSET	12,50	9,20	77,71	13,38
LSA ALL EMOTION WORDS	9,06	9,77	90,22	17,57
NB TRAINED ON BLOGS	10,81	12,04	18,01	13,22
SWAT	25,41	19,46	8,61	11,57
UA	14,15	17,94	11,26	9,51
UPAR7	28,38	27,60	5,68	8,71

TABLA 4.1: Resultados de SemEval 2007 en identificación de emociones.

El mejor rendimiento global medido por F_1[4] lo obtiene LSA ALL EMOTION WORDS con un 17.57%, consiguiendo por contra los métodos presentados en SemEval un r[5] superior.

Otros trabajos relacionados con la anotación emocional se basan en la detección de palabras clave [127], en la afinidad léxica según la probabilidad de ciertas palabras de ser afines a una determinada emoción [128], o en bases de conocimientos como OMCS2 [129].

En cualquier caso, los métodos anteriores utilizan, de una u otra forma, características y enfoques basados en el contenido semántico del texto analizado. En los trabajos de Dhaliwal *et al.* [130] se introducen características de estilo como la identificación de oraciones imperativas, el uso de signos de exclamación o las mayúsculas, o el uso del presente y el futuro, para la identificación de la polaridad y la categoría emocional.

Intentando desvincular el método del idioma, inglés en todos los casos vistos hasta el momento, García *et al.* [131] describen una arquitectura que incorpora módulos como el de desambiguación semántica (específico por idioma) o el uso de diccionarios afectivos como ANEW, sistema que ha sido aplicado al español.

[2] http://nlp.cs.swarthmore.edu/semeval
[3] http://www.roget.org/
[4] Media armónica entre *precision* y *recall*: https://es.wikipedia.org/wiki/Valor-F
[5] Kappa de Pearson [33] para medir la correlación entre el resultado obtenido y el azar.

Siguiendo la línea de características de estilo y para un idioma diferente al inglés, se tiene el trabajo de Sugimoto y Yoneyama [132], que para identificar la emoción consideran los nombres, adjetivos y verbos mediante la identificación de palabras clave y los tipos de oraciones del japonés.

En español se presenta un método basado en el diccionario SEL [133] junto con la anotación de historias cortas. En el trabajo se comparan diferentes métodos de aprendizaje, demostrando una mejora sobre la baseline. En [133] se presenta un método basado en el Spanish Emotion Lexicon (SEL) para la identificación de emociones en cuentos cortos en español.

Un paso hacia el enlace entre el análisis de emociones y las dimensiones de la personalidad se da en [134], donde se incorpora el análisis de emociones por sexo, analizando emails de tres tipos: cartas de amor, emails de odio y notas de suicidio.

4.2 Emociones y tendencias en Twitter

El uso de los medios sociales y especialmente de Twitter está creciendo rápidamente, permitiendo a los usuarios compartir información en tiempo real y generando una atención creciente en dominios como el de las campañas políticas o la comunicación de desastres. Uno de los factores principales de difusión de esta red es la posibilidad de transmitir información desde la red de usuarios a los que sigues (*friends*) hacia la red de usuarios que te siguen (*followers*), lo que propicia la generación de tendencias. Pero también, tal y como se puso de manifiesto en los capítulos anteriores, que Twitter sea el medio social donde de manera más espontánea sus usuarios pueden compartir sus pensamientos, libres de la censura y/o las pautas editoriales de otros medios, puede propiciar que la libre expresión de sus emociones, tenga un impacto significativo en la generación de dichas tendencias.

Los investigadores han prestado especial atención a la detección de tendencias (*trending topics*) [135], aunque su estimación es una tarea compleja debida al ruido y los efectos cíclicos. La mayoría de aproximaciones se basan en una base temporal absoluta para analizar el comportamiento dinámico de las discusiones en Twitter, por ejemplo, la evolución de cambios porque es lunes por la mañana [136, 137]. En [137] se utilizan modelos estadísticos para predecir eventos futuros, y los autores en [136] usan normalización sobre al menos dos semanas de datos para detectar anormalidades y eventos. Nuestra propuesta [138] se basa en modelar la evolución de la emotividad de ciertos tópicos en Twitter, con ayuda del diccionario SEL, y utilizar una aproximación basada en estado de espacios (*state space approach* [139]) con el doble objetivo de: *i*) mitigar el efecto del ruido y los efectos cíclicos; y *ii*) reducir la dimensionalidad de la representación.

4.2.1 Metodología

Hemos recuperado una colección de Twitter consistente en las discusiones políticas durante el escándalo de corrupción del caso Bárcenas al final del verano del 2013 ocurrido en España. Concretamente, los datos comprenden 4.397.023 tuits en español recuperados entre el 9 de julio y el 2 de octubre de 2013. Hemos llamado a este corpus *Barcenas* y lo hemos liberado a la comunidad[6].

Para cada término de cada tuit se obtiene la probabilidad de uso afectivo para cada una de las seis emociones básicas de Ekman con ayuda del diccionario SEL. Además, se incorporan características

[6] https://github.com/autoritas/RD-Lab/tree/master/data/Barcenas

específicas de Twitter como el número de seguidores, de seguidos, de retuits, el número único de tuits, el número total de tuits y el número de usuarios únicos.

Un modelo basado en estados expresa ciertas características como dependientes del estado previo, por ejemplo, el número de tuits para la hora actual depende del número de tuits de la hora previa. Usamos una aproximación basada en dos pasos para estimar el modelo de espacio de estados global S, dado por la Definición 1 acorde a [140]. En el primer paso se utiliza un modelo predefinido de tendencia lineal local con componentes estacionales para descomponer los datos univariantes y estimar su tendencia. En el segundo paso, las tendencias estimadas se combinan en un vector de características multivariantes y se estima el modelo de estados global.

Definición 1. Un sistema de espacio de estados S se define como:

$$S : \begin{cases} \mathbf{x}_{t+1} & = \mathbf{F}\mathbf{x}_t + \mathbf{W}_t, \quad \mathbf{W}_t \sim MND(0, \mathbf{Q}_t) \\ \mathbf{y}_t & = \mathbf{G}\mathbf{x}_t + \mathbf{V}_t, \quad \mathbf{V}_t \sim MND(0, \mathbf{R}_t) \end{cases} \text{ t=1,2,...}$$

donde $\mathbf{x}_t \in \mathbb{R}^n$, $\mathbf{y}_t \in \mathbb{R}^q$, $\mathbf{W}_t \in \mathbb{R}^n$ y $\mathbf{V}_t \in \mathbb{R}^q$.

Un sistema S basado en estados consiste en dos ecuaciones: una ecuación de observación y_t y una ecuación de estado x_{t+1}. La ecuación de observación describe los datos observables del sistema, en este caso las características extraídas de una discusión, y la ecuación de estado denota el comportamiento interno del sistema. Adicionalmente, los estados y las observaciones son cubiertas por una distribución multinomial con ruido (MND). Los estados suelen estar ocultos y no son completamente observables, por lo que el filtrado de Kalman [141] se utiliza para calcular las estimaciones lineales con el menor error cuadrático medio del vector de estado x_t en términos de las observaciones $y_1, ..., y_n$. Además, los parámetros F, Q_t y R_t del sistema son estimados a partir de la estimación de máxima verosimilitud.

4.2.2 Resultados

En la Figura 4.1 se muestra la evolución del número de tuits por hora para un periodo de 22 días. Se puede apreciar una fuerte periodicidad diaria, con fuertes y periódicos picos ariba y abajo en la evolución de los datos. Este efecto ha sido considerado también por [136] y se debe al horario diario de publicación. Sobre las 2 h la actividad en Twitter decrece, mientras que a partir de las 7 h vuelve a incrementarse. El mayor número de tuits se suele observar durante el mediodía.

Obviamente hay picos mayores que sugieren ciertos eventos, por ejemplo, el pico alrededor del 15 de julio se debe a la publicación de una conversación privada entre el presidente Rajoy y el tesorero del partido Bárcenas, conversación que días antes Rajoy había negado mantener. Estos picos causan un crecimiento exponencial de la conversación, que se puede suavizar tomando logaritmos tal y como se muestra en la sección (a) de la Figura 4.2. Aplicando el filtro Kalman se desomponen los datos en el nivel lineal local mostrado en la sección (b), y en el componente de tendencia mostrado en la sección (c), ésta última como pendiente de la gráfica mostrada en (a). Los eventos cíclicos de los datos se han eliminado y modelado en los restantes estados del modelo de espacio de estados.

Cada característica se descompone de acuerdo a la Figura 4.2, normalizando a máximos para visualizar las diferencias entre los datos. El resultado es una tendencia de variación lenta para cada característica donde los componentes cíclicos y el ruido se han separado. Después de aplicar el filtro Kalman, las tendencias se muestran mucho más aparentes, así como las interacciones entre características. En comparación con [136], las anormalidades en la evolución de la discusión pueden ser visualizadas incluso

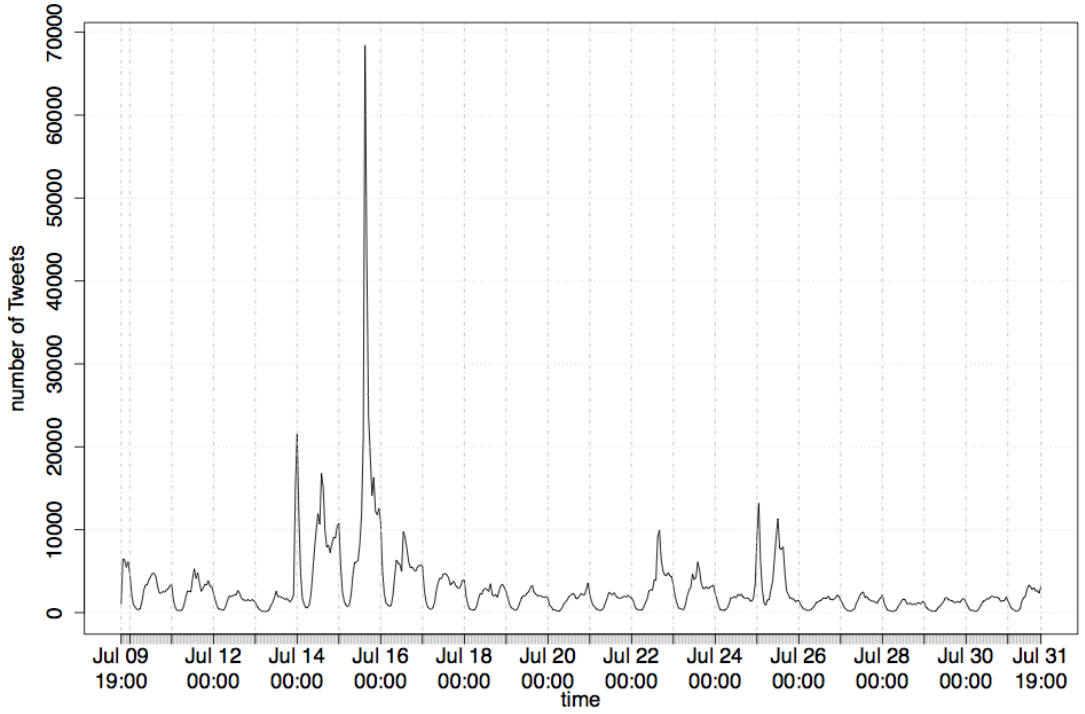

FIGURA 4.1: Número de tuits por hora.

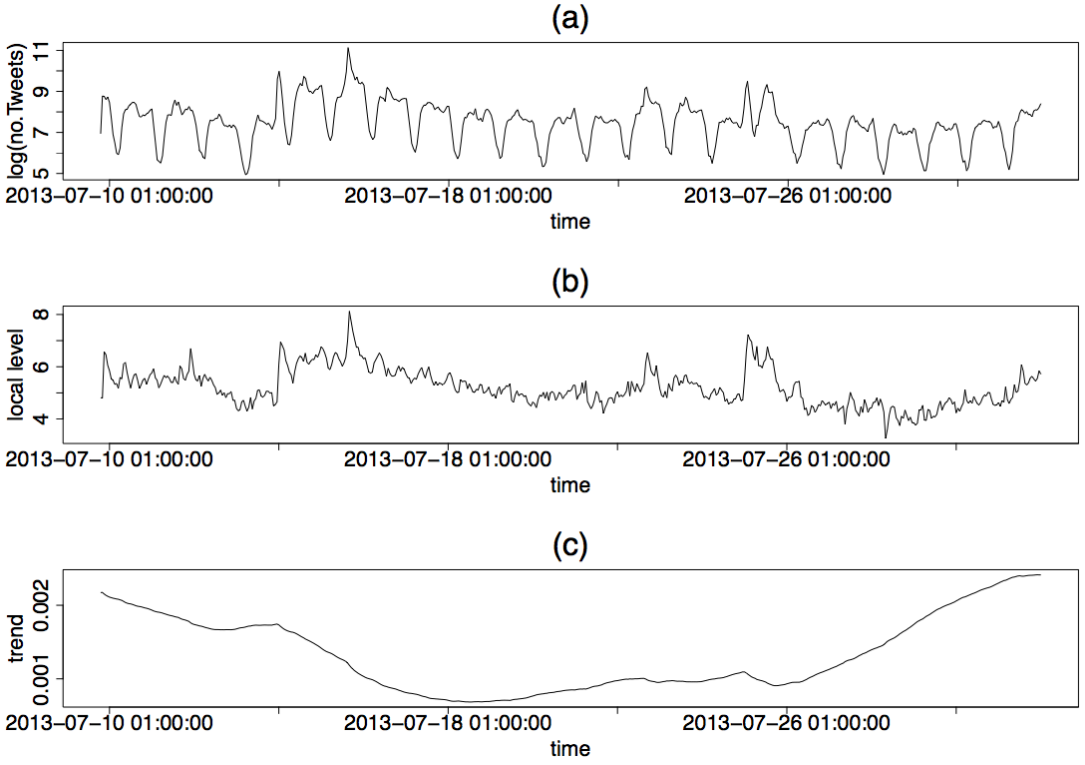

FIGURA 4.2: Descomposición del número de tuits por hora: (a) datos logarítmicos; (b) componente de tendencia lineal local; (c) componente de tendencia.

con cortos periodos de tiempo. De este modo, a pesar de que el periodo de observación tenga un efecto en la aproximación de aprendizaje automático utilizada, no hay una restricción en la cantidad de los

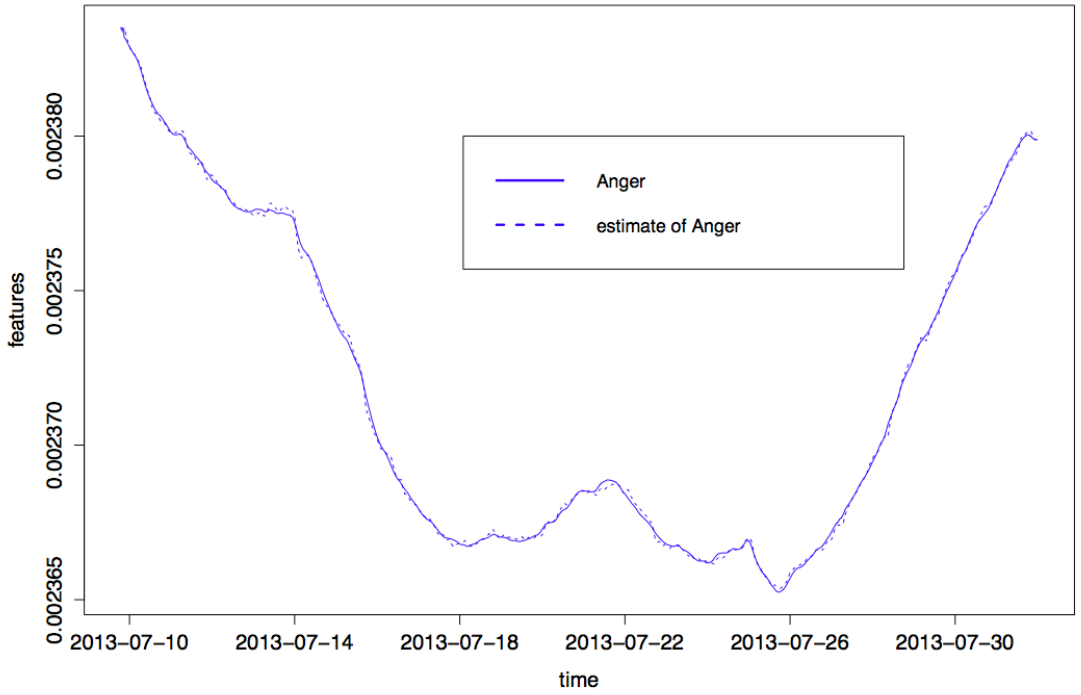

FIGURA 4.3: Tendencia y estimación para la emoción enfado (Anger).

datos. En la Figura 4.3 se muestra un ejemplo de estimación de estado del modelo aprendido. En este ejemplo, la línea sólida denota la tendencia para la emoción *enfado* (*anger*) y la línea punteada denota la tendencia estimada usando el filtro Kalman. El filtro Kalman estima los datos observados con ligeras diferencias y la estimación de la tendencia sigue de cerca la tendencia original, a pesar de que los valores sean muy pequeños.

4.2.3 Discusión

En este apartado hemos realizado un análisis de la evolución de la conversación producida en Twitter en torno a un tema de gran polémica como el caso Bárcenas, desde la perspectiva de la emotividad de los usuarios, con ayuda del diccionario SEL. Hemos creado un modelo de estados que ha permitido: *i*) reducir la dimensionalidad de la representación; *ii*) utilizar de este modo el conjunto completo de datos, a diferencia de la mayoría de los trabajos existentes que se limitan a cierta cantidad de datos; *iii*) separar los componentes estacionales y cíclicos, sin acudir a una base temporal absoluta; y *iv*) poner de manifiesto y hacer más evidentes los cambios ocultos en la percepción emotiva de los usuarios, especialmente la relativa al enfado (*anger*); *v*) verificar que la expresión de las emociones permite reflejar la evolución de las tendencias en los medios sociales (el siguiente paso sería predecirlas).

4.3 Emociones y sexo en Facebook

La mayoría de las investigaciones en identificación de emociones se han basado en obtener características representativas de la semántica de los documentos, esto es, intentan analizar el contenido de los textos, lo que puede implicar sobreajuste y dependencia con el dominio, el contexto o la temática. En base a los estudios realizados para el inglés por autores como Pennebaker [50], nuestro objetivo es analizar

el estilo de escritura de un autor para identificar las emociones que expresa, y ver si ello ayuda en la identificación de su sexo [142].

4.3.1 Características de estilo

Con el objetivo de modelar el estilo de escritura, hemos considerado un conjunto de características relativas a fecuencias de uso de diferentes elementos léxicos y signos de puntuación, así como el uso de los emoticonos, lo que proporciona tanto un aspecto estilístico como emocional. Para complementar en el modelado del estilo, se han tenido en consideración las categorías gramaticales junto con toda su información morfosintáctica asociada y obtenida con Freeling [54, 55]. Por último, se ha utilizado el Spanish Emotion Lexicon para detectar palabras con carga afectiva relativa a cada una de las seis emociones básicas de Ekman. Cada elemento en las listas se constituye como una característica independiente en el modelo, formando un modelo de espacio vectorial compuesto por 59 características.

Frecuencias	Ratio entre el número de palabras únicas y el número total de palabras (*unique-words*), palabras comenzando por mayúsculas (*capital-words*), palabras en mayúsculas (*upper-words*), longitud de las palabras (*words-length*), número de letras en mayúsculas (*upper-chars*) y número de palabras con *character flooding* (e.g. Hoooolaaaaa) (*running-words*).
Signos de puntuación	Frecuencia de uso de signos de puntuación (*punctuation*), de puntos (*punctuation-dots*), comas (*punctuation-comma*), dos puntos (*punctuation-colon*), punto y coma (*punctuation-semicolon*), exclamaciones (*punctuation-exclamation*), preguntas (*punctuation-question*) y comillas (*punctuation-quotes*).
Categorías gramaticales	Frecuencia de uso de cada categoría gramatical (*Adj, Adv, Conj, Quant, Det, Intj, MD, Prep, Pron, Noun, Verb*), número y persona de verbos (*Verb1S, Verb2S, Verb3S, Verb1P, Verb2P, Verb3P*) y pronombres (*Pron1S, Pron2S, Pron3S, Pron1P, Pron2P, Pron3P*), modo de los verbos (*VerbIndicative, VerbSubjuntive, VerbConditional, VerbImperative, VerbInfinitive, VerbParticipe*), nombres propios (*Proper*) y palabras no reconocidas (*NotRecognized*).
Emoticonos	Ratio entre el número de emoticonos y el número total de palabras (*emoticons*), número de tipos diferentes de emoticonos: alegría (*emoticon-happy*), tristeza (*emoticon-sad*), disgusto (*emoticon-disgust*), enfado (*emoticon-angry*), sorpresa (*emoticon-surprise*), burla (*emoticon-wink*) o tontería (*emoticon-dumb*).
Lexicón de emociones en español	Para cada palabra se obtiene su lema. Para el lema, se busca en SEL su factor de probabilidad de uso afectivo. Si el lema no tiene una entrada en el diccionario, se obtienen sus sinónimos. Para cada sinónimo, se busca el citado factor. Para cada emoción se suman todos los valores obtenidos, creando una característica por emoción (*sel-enjoyment, sel-surprise, sel-anger, sel-disgust, sel-sadness, sel-fear*).

Tabla 4.2: Conjunto de características para representar los textos.

4.3.2 Metodología

Hemos elegido los comentarios de Facebook en español como la fuente de datos para la realización de nuestros primeros experimentos, principalmente por su orientación social donde, frente a métodos tradicionales como las noticias, se posee la libertad de expresión (y estilo) sin pautas editoriales, y frente a otros medios como los blogs, se posee la espontaneidad en el uso del lenguaje. Por otro lado, el uso masivo de Facebook por la población y la carga afectiva en dicho medio, nos permite tener una fuente actual adecuada para nuestros objetivos. Facebook además nos permite obtener rasgos demográficos como el sexo, por lo que podemos investigar de manera conjunta la expresión de emociones y el sexo de los que las expresan.

Facebook se compone de una jerarquía de objetos. El primer nivel de la jerarquía lo componen las *páginas*, así como los *perfiles*, los *eventos* y los *grupos*. Cada *página* tiene un propietario que publica *posts*, que son el segundo nivel de objetos de la jerarquía. Los *posts* son escritos por el propietario de la *página* y por lo tanto siguen sus propias pautas y temáticas. Pero en ocasiones, los *posts* permiten a otros usuarios participar en la conversación mediante los *comentarios*. Los *comentarios* son el tercer nivel de objetos de la jerarquía. La gente puede expresar mediante *comentarios* lo que piensan acerca del tema del *post*, pero sin las pautas editoriales del propietario de la *página*. Para construir el corpus, nos hemos centrado en los *comentarios*.

Hemos seleccionado tres temáticas de actualidad, con gran volumen de participación[7] de los usuarios, y además susceptibles de comentarios emotivos. Las temáticas con *política*, *fútbol* y *famosos*. Para cada una de las temáticas, se han seleccionado cuatro de las *páginas* más conocidas en esa temática en España. Información detallada de las mismas se muestra en la Tabla 4.3.

Hemos recuperado al menos 1.000 posts de cada página, y todos los comentarios escritos en cada post. Hemos seleccionado los comentarios para cuyo autor se disponía de la información sobre su sexo[8]. Aleatoriamente hemos seleccionado 200 comentarios para cada temática y cada sexo, equilibrando los datos como se muestra en la Tabla 4.4. No se ha efectuado más selección y limpieza de los datos que aquella orientada a asegurar en la medida de lo posible que los comentarios estén escritos en español y contengan texto (no sean meras comparticiones de enlaces).

Tema	Número comentarios
Política	200/200
Fútbol	200/200
Famosos	200/200

Tabla 4.4: Corpus de comentarios de Facebook en español equilibrado por sexo.

Etiquetado de las emociones

Tres anotadores independientes[9] han etiquetado el total de 1.200 documentos con respecto a las seis emociones básicas de Ekman. Hay diferentes formas de anotar emociones en textos, basándose en:

- La emoción expresada por el autor;

[7] http://ow.ly/XWFhc

[8] Aunque la anotación de sexo es proporcionada a Facebook por el propio usuario, nos hemos asegurado de que sea correcta mediante la verificación manual del nombre y las fotos del usuario.

[9] Dos mujeres y un hombre.

POLÍTICA: Cuatro páginas oficiales de partidos políticos de España
Partido Popular
https://www.facebook.com/pp
Partido Socialista Obrero Español
https://www.facebook.com/psoe
Izquierda Unida
https://www.facebook.com/izquierda.unida
Union por el Progreso y la Democracia
https://www.facebook.com/Union.Progreso.y.Democracia
FÚTBOL: Cuatro páginas oficiales de clubs de fútbol de España
Real Madrid CF
https://www.facebook.com/RealMadrid
FC Barcelona
http://www.facebook.com/fcbarcelona
Valencia Club de Futbol
http://www.facebook.com/vcf1919
Atletico Bilbao
http://www.facebook.com/pages/ATLETICO-BILBAO/103997686354572
FAMOSOS: Cuatro páginas oficiales del famoseo en España
Belen Esteban
http://www.facebook.com/BelenEstebanM
Kiko Hernandez
http://www.facebook.com/ElConfesionariodeKiko
David Bisbal
http://www.facebook.com/davidbisbal
Santiago Segura
http://www.facebook.com/pages/Santiago-Segura-Silva/12459228767

TABLA 4.3: Páginas oficiales seleccionadas para recuperar los datos para cada temática.

- La emoción producida en el lector;
- La emoción que se describe o expresa en el texto.

Hemos instado a los anotadores a usar la última aproximación, tratando así de reducir al máximo los juicios de valor. Como ayuda para el etiquetado, se les ha proporcionado la información de la Tabla 4.5 obtenida de Greenberg [143], que muestra los estados emocionales, o emociones secundarias, más cercanos a cada una de las seis emociones básicas. Es de resaltar que algunas de estas emociones secundarias están compartidas por más de una de las emociones básicas; por ejemplo, *indignación* es compartida por *enfado* y *disgusto*, y *fascinación* es compartida por *alegría* y *sorpresa*. Este hecho obstaculiza la identificación única de tales emociones, como ha sido evidenciado por Ortony y Turner [144]. Además, se permite la identificación de emociones múltiples en cada texto, así como la ausencia de emoción alguna.

Hemos calculado la concordancia entre anotadores siguiendo el método Kappa_DS propuesto por Díaz [133], que permite múltiples evaluadores (tres en nuestro caso: A1, A2 y A3) y variables multinomiales (en nuestro caso, seis no mútuamente excluyentes, las seis emociones básicas). Se muestran los resultados en la Tabla 4.6.

ALEGRÍA	ENFADO	MIEDO	REPULSIÓN	SORPRESA	TRISTEZA
Agradecido	Agresivo	Acomplejado	Aborrecimiento	Extrañeza	Abatido
Alegre	Colérico	Alarmado	Desagrado	Sobresalto	Agobiado
Animado	Crispado	Angustiado	Grima	Susto	Apenado
Calmado	Descontento	Ansioso	Repulsión	Consternación	Confuso
Confiado	Enfadado	Atemorizado	Antipatía	Pasmo	Decepcionado
Contento	Enojado	Aterrado	Aversión	Desconcierto	Deprimido
Dichoso	Excitado	Avergonzado	Repugnancia	Estupor	Desalentado
Encantado	Fastidiado	Confuso	Disgusto	Asombro	Desanimado
Entusiasmado	Furioso	Desesperado	Repudia	**Fascinación**	Desdichado
Eufórica	Insatisfecho	Desorientado	Repulsa	Admiración	Desmoralizado
Esperanzado	Irascible	Horrorizado	Odio	Confusión	Frustrado
Feliz	Malhumorado	Inquieto	Manía	Chasco	Nostálgico
Gozoso	Molesto	Inseguro	Rabia	Impresión	Soledad
Satisfecho	Nervioso	Intranquilo	Animadversión	Exclamación	Triste
Tranquilo	Rabioso	Pánico	Nauseabundo	Conmoción	Infeliz
Complacido	Tenso	Preocupado	**Indignación**	Estupefacción	Desconsolado
Libre	Violento	Temeroso	Enfado		Afligido
Fascinado	Irritado	Tenso	Desprecio		Amargado
Seguro	**Indignado**	Indeciso	Distanciamiento		Impotente
		Impotencia			

TABLA 4.5: Emociones secundarias relacionadas con las seis emociones básicas.

	A1	A2	A3	REST
A1	-	0,0587	0,2738	0,1662
A2	0,0587	-	0,1042	0,0814
A3	0,2738	0,1042	-	0,1890
TOT		0,1455		

TABLA 4.6: Kappa_DS: Concordancia entre anotadores.

Un valor medio de Kappa igual a 0,1455 muestra una baja concordancia de acuerdo a las recomendaciones originales de Landis y Koch [145]. Aunque como indica Díaz, tenemos que tener en cuenta la cantidad de variables interviniendo en la evaluación para hacer una correcta interpretación del índice, haciendo que no sea comparable con la recomendación original. Por otro lado, hemos agrupado aquellas emociones básicas que comparten emociones secundarias, como se mostraba en la Tabla 4.5: *alegría/sorpresa* y *disgusto/enfado*. Los resultados de concordancia se muestran en la Tabla 4.7. En este caso, el índice muestra un valor mayor de concordancia (0,6016), lo que sugiere que tenemos que tener en cuenta esta discordancia entre anotadores cuando analicemos los resultados, especialmente con respecto a estas parejas de emociones.

	A1	A2	A3	REST
A1	-	0,6618	0,5656	0,6137
A2	0,6618	-	0,5773	0,6196
A3	0,5656	0,5773	-	0,5715
TOT		0,6016		

TABLA 4.7: Kappa DS: Concordancia entre anotadores con emociones agrupadas: *alegría/sorpresa* y *disgusto/enfado*.

La selección final de etiquetas emocionales para cada documento se basa en la concordancia de al menos dos de los tres anotadores, con las distribución mostrada en la Tabla 4.8. El bajo número de documentos etiquetados con la emoción *miedo* no nos ha permitido experimentar con dicha emoción.

	TOTAL	%
Alegría	338	28,17
Enfado	151	12,58
Miedo	3	0,25
Disgusto	129	10,75
Sorpresa	390	32,50
Tristeza	76	6,33
Neutro	262	21,83

TABLA 4.8: Documentos por emoción.

Hemos llamado a este corpus *EmIroGeFB* [146] y lo hemos liberado públicamente a la comunidad[10].

Algoritmo de aprendizaje

Se propone la generación de clasificadores binarios que determinen si un texto contiene o no cada una de las emociones. Cada clasificador se entrena con textos etiquetados con esa emoción como ejemplos positivos y con el resto como negativos. La evaluación se realiza mediante validación cruzada de 10 etapas. Se toman dos tipos de medidas de evaluación, siguiendo lo realizado en SemEval 2007. En primer lugar, se obtiene el índice de correlación Kappa de Pearson [33] entre el resultado obtenido y el propio azar (r). En segundo lugar, se obtienen la precisión, el *recall* y la medida F1 [147]. Se testean cuatro algoritmos de aprendizaje diferentes implementados en Weka[11], usando sus parámetros por defecto: árboles de decisión J48, Naïve Bayes, BayesNet y máquinas de vectores soporte. Para la identificación del sexo, se ha entrenado una máquina de vectores soporte, parametrizándola con un núcleo gausiano con g=0,01 y c=3.500 tras probar diversas opciones.

4.3.3 Resultados

A continuación se presentan los resultados para la identificación de emociones en posts de Facebook, tanto para las emociones básicas como para las agrupadas, y para la identificación del sexo de sus autores.

Identificación de emociones

En la Tabla 4.9 se presentan los resultados para la identificación de emociones en Facebook, con los mejores resultados para cada métrica resaltados en negrita. Se puede apreciar que diferentes métodos de aprendizaje tienen diferentes fortalezas. Por ejemplo, J48 obtiene la precisión más elevada en la mayoría de los casos, pero al coste de reducir el *recall*. De manera similar, BayesNet obtiene mejor *recall*, pero a costa de la precisión, aunque es el mejor método en términos de F1. Los valores de r parecen tener menor correlación con el método utilizado. Sin embargo, en la mayoría de casos los mejores resultados son obtenidos por los métodos estadísticos como Naïve Bayes y BayesNet.

Con respecto a las emociones, los resultados para *alegría* y *sorpresa* son los mejores, principalmente para la medida F1, algo que puede tener relación con el número de muestras para dichas emociones en

[10] https://github.com/autoritas/RD-Lab/tree/master/data/EmIroGeFB
[11] http://www.cs.waikato.ac.nz/ml/weka

la partición de entrenamiento. Los resultados para *tristeza* son menores que para el resto, probablemente debido a que se tienen menos documentos etiquetados para esta emoción que para el resto (ver Tabla 4.8). Esto implica que hay cierta dependencia del algoritmo de aprendizaje con el número de ejemplos utilizados en el entrenamiento.

Es necesario resaltar el bajo resultado de las máquinas de vectores soporte en algunos experimentos, debido al desequilibrio entre clases y al pequeño número de ejemplos de entrenamiento, mostrándose este método más sensible a ambos factores.

En cualquier caso, los resultados obtenidos con las características propuestas, todas ellas independientes de la temática y basadas en el estilo de escritura, obtienen resultados comparables a los del estado del arte, en este caso en el dominio de los medios sociales en español.

Emoción	Algoritmo	r	Precisión	*Recall*	F1
Alegría	J48	27,1	49,7	43,2	46,2
	NB	**27,9**	45,4	56,8	50,5
	BN	25,6	40,9	**73,7**	**52,6**
	SVM	24,9	**56,9**	30,5	39,7
Disgusto	J48	21,7	**36,1**	23,3	28,3
	NB	15,7	19,7	55,8	29,1
	BN	**24,9**	25,5	**64,3**	**36,5**
	SVM	6,2	11,7	5,4	7,4
Enfado	J48	16,6	**32,3**	19,9	24,6
	NB	**22,6**	25,9	60,3	**36,3**
	BN	22,2	25,6	**60,9**	36,0
	SVM	10,8	25,8	15,2	19,2
Sorpresa	J48	**25,8**	**50,4**	48,7	49,5
	NB	20,6	42,7	**67,2**	**52,2**
	BN	20,7	43,0	64,6	51,6
	SVM	17,2	49,4	30,5	37,7
Tristeza	J48	12,1	**20,0**	14,5	16,8
	NB	6,1	9,8	35,5	15,4
	BN	**16,7**	16,3	**51,3**	**24,7**
	SVM	8,2	17,9	0,92	12,2
Resultados medios					
	J48	20,7	**37,7**	29,9	33,1
	NB	18,6	28,7	55,1	36,7
	BN	**22,0**	30,3	**63,0**	**40,3**
	SVM	13,5	32,3	16,5	23,2

Tabla 4.9: Resultados para la identificación de emociones en Facebook.

Identificación de emociones agrupadas

Algunas categorías de emociones son difícilmente identificables de manera unívoca, incluso para anotadores humanos, por compartir emociones secundarias muy cercanas. Para comprobar este efecto, se ha realizado una clasificación conjunta agrupando las categorías *alegría/sorpresa*, y *enfado/disgusto*, y se presentan los resultados en la siguiente tabla:

Emoción	Algoritmo	r	Precisión	*Recall*	F1
Alegría + Sorpresa	J48	38,8	69,5	69,6	69,5
	NB	42,1	71,3	71,1	71,1
	BN	40,1	70,5	70,4	70,2
	SVM	**44,5**	**72,9**	**72,1**	**72,1**
Enfado + Disgusto	J48	26,0	76,2	**77,3**	**76,7**
	NB	**33,9**	**80,2**	73,0	75,2
	BN	33,0	79,3	73,8	75,7
	SVM	18,9	74,2	**77,3**	75,3

TABLA 4.10: Resultados para la identificación de emociones agrupadas: *alegría/sorpresa* y *enfado/disgusto*.

Se puede apreciar un incremento significativo de los resultados, especialmente en el valor de r, pasando de valores de 27,9 y 25,8 *alegría* y *sorpresa* respectivamente a 44,5 para la combinación de ambas, y de 22,6 y 24,9 para *enfado* y *disgusto* respectivamente a 33,9 para su combinación. Para las medidas de precisión y *recall*, o su representación media en F1, se aprecia no sólo un incremento sino también unos valores más equilibrados. Esto sugiere que los clasificadores realmente se han encontrado con dificultades a la hora de discernir entre estos dos pares de emociones, seguramente producido por la propia dificultad al anotar.

Identificación de sexo

Con el objetivo de enlazar las emociones con la demografía hemos llevado a cabo un experimento consistente en, usando las mismas características para la identificación de las primeras, aprender un modelo que nos permita identificar el sexo de los autores de los comentarios de Facebook. Los resultados se muestran en la Tabla 4.11.

Sexo	Acc	r
Mujer / Hombre	59	18

TABLA 4.11: Resultados para la identificación de sexo en términos de *accuracy* y coeficiente de Pearson r.

Un valor de r igual al 18% significa que el clasificador trabaja sobre el azar y sugiere que las características propuestas proporcionan cierta información sobre el sexo, tal y como Koppel *et al.* [16] muestran para el inglés.

4.3.4 Discusión

Hemos construido el corpus EmIroGeFB de comentarios hechos en páginas de Facebook en español y lo hemos etiquetado manualmente con las seis emociones básicas de Ekman. Al calcular los índices de concordancia entre anotadores, se observa un bajo valor del mismo, debido principalmente a la dificultad de discernir entre los pares *alegría/sorpresa* y *disgusto/enfado*. Esto ha dado lugar a resultados dispares en la clasificación de dichas emociones, que se corrige cuando se toman en consideración de manera conjunta.

El hecho de que las características utilizadas para identificar emociones, basadas en una combinación de categorías gramaticales, frecuencias estilísticas y palabras de emoción del diccionario SEL, nos hayan

permitido identificar el sexo con cierta precisión, sugiere que hay cierta correlación entre la expresión de emociones en textos y el sexo de sus autores. Una *accuracy* del 59% nos permite pensar que el método es competitivo para tal tarea, por ejemplo teniendo en cuenta los resultados obtenidos por los participantes de la tarea del PAN, aunque con distintos corpus. En el apartado 4.5 compararemos nuestra aproximación basada en el lexicón de emociones en español SEL con los demás sistemas que participaron en la tarea de identificación de edad y sexo en la edición del 2013 del PAN.

4.4 Emociones, sexo e ironía en Facebook

La ironía es un modo de comunicación exclusivamente humano por el cual el emisor dice algo diferente a su intención [148]. Así pues, Grice [149] y Attardo [150] consideran la ironía como una violación *intencional* de las máximas conversacionales. Debido al uso creciente de la ironía en los medios sociales[12][153–155], su relación tanto con la expresión de afectividad como con el estilo particular de su autor, y al interés de utilizarla como característica en la tarea de identificación de edad y sexo [104], en este apartado hemos querido investigar la relación entre la ironía, las diferentes emociones y el sexo de los autores [146].

4.4.1 Metodología

Tomando el corpus EmIroGeFB descrito en el apartado anterior y obtenido a partir de comentarios realizados por usuarios de Facebook, se le ha solicitado a los mismos anotadores que indicasen cuando consideraban que un comentario es irónico. No se les ha dado mayor indicación de lo que es o no la ironía, solicitándoles que realizasen el etiquetado en base a su propio concepto de ironía. Hemos calculado el índice Kappa de Fleiss [156] de concordancia entre anotadores. Este método permite múltiples anotadores (tres en nuestro caso: A1, A2 y A3) y una variable binaria (irónico/no irónico). El resultado obtenido es de 0,0989, valor concordancia extremadamente bajo. Pensamos que este valor tan bajo se debe a la propia tarea: el concepto de ironía es extremadamente subjetivo y depende completamente la persona que lo percibe, de su humor, contexto cultural y lingüístico, etc. para ser correctamente entendido; así como información contextual que en nuestro caso, no se ha proporcionado. Por ejemplo, no se ha proporcionado una definición general de ironía para construir un marco común de características.

Con el objetivo de conocer el índice de concordancia entre anotadores con respecto a las emociones etiquetadas en comentarios irónicos, se ha calculado el índice Kappa_DS teniendo en cuenta sólo el subconjunto de comentarios identificados como irónicos. Los resultados se muestran en la Tabla 4.12.

	A1	A2	A3	Rest
A1	-	-0,0854	0,0001	-0,0426
A2	-0,0854	-	-0,1128	-0,0991
A3	0,0001	-0,1128	-	-0,0563
Total		-0,0660		

TABLA 4.12: Kappa_DS: Concordancia entre anotadores en el etiquetado de contenidos emocionales con ironía.

[12]Se ha organizado una tarea piloto en análisis del sentimiento e ironía (en italiano) en Evalita-2014 [151], así como otra tarea en inglés se ha organizado en SemEval-2015 [152] sobre análisis de sentimiento en lenguaje figurado en Twitter.

Se obtiene un valor negativo de -0,0660, lo que significa que no hay ninguna concordancia entre anotadores cuando se tiene en cuenta conjuntamente emociones e ironía. En este aspecto, no se ha realizado más análisis, sólo se ha obtenido este índice de concordancia. Sin embargo, pensamos que hay una estrecha relación entre ironía y emociones, especialmente, la clase de emoción lanzada por una declaración irónica. Esto se proyecta analizar en el futuro y fuera del marco de esta tesis.

4.4.2 Emociones

El bajo índice de concordancia entre anotadores mostrado en el apartado anterior (0,1455) sugería la dificultad de la tarea de etiquetar emociones, algo que aún queda más patente con algunos de los ejemplos que se muestran a continuación:

guuapaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa ers la mejorrr

El anterior comentario es claramente un comentario de ánimo a un famoso. Pero ¿cuál es la emoción básica? Usando la información de la Tabla 4.5, ¿qué palabra describe mejor el humor del texto: *euforia*, *fascinación*, *admiración*? Dependiendo de la elección, los anotadores pueden dudar entre *alegría* y *sorpresa*. En este caso, el anotador A1 seleccionó *alegría* y *sorpresa*, el anotador A2 *alegría* y el anotador A3 *sorpresa*.

Algo similar ocurre con comentarios de repugna como el siguiente, donde el autor critica algunas entidades por ser financiadas por el gobierno pese a ser privadas; comentarios como éste pueden ser etiquetados de manera diferente dependiendo del humor del anotador o de su propia visión del mundo. Por ejemplo, los anotadores A1 y A2 lo etiquetaron tanto con *disgusto* como con *enfado*, mientras que el anotador A3 sólo utilizó *enfado*.

Es una vergüenza, que se financien ellos, que para algo son privadas

En el comentario siguiente el autor recomienda a un famosillo que guarde algo en una cámara acorazada para guardarla de los ladrones. Este comentario es tremendamente ambiguo de modo que la emoción etiquetada por cada anotador es diferente: el anotador A1 indicó *ausencia de emoción*, el A2 reportó *miedo* y el A3 *sorpresa*.

guarda semejante alhaja en una camara acorazada por si os la roban

En la Tabla 4.13 se muestra el número de los comentarios etiquetados en cada emoción por cada anotador. Como se ha mencionado previamente, la diferencia entre anotadores es elevada en casos como *alegría*, *sorpresa*,*disgusto* y *enfado*. Parece ser que algunos anotadores (A1 y A3) perciben como *sorpresa* lo que otros (A2) lo hacen como *alegría*, de manera similar a *enfado* (A2) y *disgusto* (A1).

	A1	A2	A3
Alegría	255	756	215
Disgusto	255	78	166
Enfado	96	265	148
Miedo	19	6	7
Sorpresa	626	140	460
Tristeza	165	72	83
Ninguna	97	42	160

TABLA 4.13: Número de comentarios etiquetados por cada anotador en cada emoción.

Finalmente se seleccionan aquellas anotaciones en las que al menos dos de los tres anotadores coinciden. En la Tabla 4.14 se muestra el número y porcentaje de comentarios anotados con cada emoción. Por ejemplo, *alegría* tiene un valor superior (388) que la obtenida por dos de los anotadores (A1=255 y A3=215). Algo similar sucede con otras emociones. Esto significa que la percepción de *alegría* o *sorpresa* es muy subjetiva.

	Total	%
Alegría	338	28,17
Disgusto	129	10,75
Enfado	151	12,58
Miedo	3	0,25
Sorpresa	390	32,50
Tristeza	76	6,33
Ninguna	262	21,83

TABLA 4.14: Número y porcentaje de comentarios por emoción.

En la Tabla 4.15 se muestra la distribución de emociones etiquetadas para cada sexo. Los resultados se muestran bastante equilibrados, independientemente de que haya menos comentarios con emoción para el caso de las mujeres (18 vs. 37), o más emociones positivas/neutras como *alegría* (194 vs. 144) o *sorpresa* (215 vs. 175).

	Mujeres	Hombres
Alegria***	194	144
Disgusto	63	66
Enfado	72	79
Miedo	1	2
Sorpresa***	215	175
Tristeza	39	37
Ninguna***	18	37

TABLA 4.15: Emociones por sexo.

En la Tabla 4.16 se muestra la distribución de emociones por temática. Como era de esperar, la *política* se percibe con mayor negatividad y por tanto mayor uso de las emociones *enfado*, *disgusto* y *tristeza*. Así mismo, es la temática con mejor número de comentarios etiquetados con la *ausencia de emoción*. Tanto *fútbol* como *famosos* tienen valores similares para *alegría* y *sorpresa*, pero *famosos* tiene valores mayores para *disgusto*. Quizás esto se debe a que la gente escribe en las páginas de dichos famosos para apoyarles o criticarles, dependiendo de la afinidad con los mismos.

	Política	Fútbol	Famoseo
Alegría	50	153	135
Disgusto	79	7	43
Enfado	114	10	27
Miedo	2	1	0
Sorpresa	53	180	157
Tristeza	52	9	15
Ninguna	9	23	23

Tabla 4.16: Emociones por temática.

4.4.3 Ironía

A continuación se muestran algunos ejemplos de comentarios irónicos y una explicación de los mismos, aportando el contexto cultural necesario para su entendimiento y mostrando así la dificultad de la tarea.

Pitbul es cultura, ¿ no ves que te enseña a contar? aunque sea sólo hasta 3

En el anterior comentario el autor critica al cantante por incluir en sus canciones siempre la coletilla "uno, dos, tres...". Por ello dice que esto es "cultura" porque escuchando a este cantante, cualquiera puede aprender a contar, aunque sea hasta tres. El autor expresa un comentario positivo, recalcando un matiz, con el objetivo de enfatizar su opinión negativa acerca de este cantante. En este comentario, dos de los tres anotadores coincidieron.

Que viva, pero muy lejos!

En este comentario el autor expresa su intención de vivir bien lejos de alguien, utilizando inicialmente un deseo positivo pero mostrando finalmente su intención real. En este comentario los tres anotadores concuerdan.

Pobres, en el fondo producis ternura...que triste tiene que ser haber votado al PP.

En este comentario, el autor expresa lástima frente a la gente que ha votado al partido político PP. El autor utiliza este matiz para mostrar su desprecio por el juicio de los que han votado al partido político ganador. El autor utiliza un comentario negativo para mostrar su verdadera intención. En este comentario concuerdan dos de los tres anotadores.

Eres muy injusto y quiero que sepas que la infanta cuando se fue a vivir a su nueva vivienda recien reformada y a pesar de ser mucho mas pequeña que la zarzuela se mudo convencida de que era una VPO o no?.....

En el comentario anterior el autor dice que la infanta se traslada a una nueva residencia y que la infanta estaba convencida de que era una vivienda de protección oficial porque era mucho más pequeña que el Palacio de la Zarzuela, residencia de la familia real. El autor expresa una observación positiva sobre el juicio de alguien incluyendo una comparación para enfatizar así el sentido irónico del comentario. En este comentario, todos los anotadores concuerdan.

Yo soy presunta ciudadana española y digo esto porque no estoy segura de si realmente lo soy o si vivo en una realidad paralela donde nuestro presi es más inútil que una neurona de Paris Hilton.

En este último comentario, los autores aluden a la posibilidad de vivir en una realidad paralela porque el país está siendo gobernado por alguien tan inútil como una neurona de Paris Hilton. El autor compara dos observaciones en el mismo comentario con el objetivo de enfatizar su intención real de rechazar al actual gobierno del pais. En este comentario los tres anotadores concuerdan.

En la Tabla 4.17 se muestra el número de comentarios irónicos identificados por cada anotador. El porcentaje de comentarios etiquetados como irónicos es muy bajo, aunque uno de los anotadores etiquetó como irónicos un número de comentarios muy superior al resto.

Anotador	Comentarios	%
A1	52	4,33
A2	189	15,75
A3	48	4,00

Tabla 4.17: Número de comentarios irónicos etiquetados por anotador.

Finalmente, se marcan como irónicos aquellos comentarios donde al menos dos de los tres anotadores concuerdan. Como se puede apreciar en la Tabla 4.18 sólo 42 comentarios cumplen este criterio.

	Total	%
Irónicos	42	3,62
No irónicos	1158	96,37

Tabla 4.18: Número y porcentaje de comentarios irónicos y no irónicos.

En la Tabla 4.19 se muestra el número de comentarios irónicos por sexo y temática. Se puede apreciar que en este corpus los hombres usan la ironía en mayor medida que las mujeres y que en política es donde más comentarios irónicos se generan.

	Mujeres	Hombres	Total
Famosos*	3	8	12
Fútbol	1	3	4
Política	11	16	27
Total	15	27	42

Tabla 4.19: Número de comentarios irónicos por sexo y temática.

Finalmente, en la Tabla 4.20 se muestra el número de comentarios irónicos por emoción.

Emoción	Comentarios irónicos
Alegría	8
Disgusto	6
Enfado	4
Miedo	0
Sorpresa	6
Tristeza	0
Ninguna	3

TABLA 4.20: Número de comentarios irónicos por emoción.

4.4.4 Discusión

Partiendo del corpus EmIroGeFB descrito en el apartado anterior se ha realizado un análisis estadístico relacionando las emociones etiquetadas con la anotación de ironía, teniendo en cuenta tanto el sexo de los autores como las temáticas de las páginas donde comenta.

Así como en el capítulo anterior se mostraba la dificultad de etiquetar las emociones, apreciando valores del índice Kappa_DS muy bajos, en este apartado se aprecia que la tarea de etiquetar ironía es incluso más compleja, dando indicadores de Kappa negativos, indicando que no hay concordancia alguna entre los anotadores. La principal razón es la alta subjetividad cuando se anota este tipo de lenguaje figurado. Además, este resultado negativo puede suponer que la gente expresa ironía independientemente de las emociones que sienten.

Las estadísticas muestran, al menos en este corpus: i) que las mujeres tienden a usar muchas más palabras relacionadas con emociones que los hombres, principalmente emociones positivas; ii) que los hombres tienden a ser más irónicos que las mujeres, posíblemente debido también a las temáticas de estudio, siendo una de ellas el fútbol; y iii) que el tema de la política es en el que más emociones negativas e ironía se expresan.

4.5 Emociones, sexo y edad en PAN

En apartados anteriores hemos mostrado una serie de características basadas en el diccionario SEL y en la variación del uso de categorías gramaticales y elementos estilísticos que ha resultado competitiva en la identificación de emociones en comentarios efectuados en Facebook, así como del sexo de sus autores. En este último apartado, hemos utilizado las características propuestas para identificar la edad y el sexo de manera comparativa a los participantes de la tarea PAN del 2013 [56].

4.5.1 Metodología

Para realizar los experimentos hemos utilizado la partición en español del dataset presentado en la tarea PAN del 2013 y descrito en detalle en el apartado 3.1.1. Este dataset consiste en un gran número de autores anónimos etiquetados con sexo y edad. Para la edad, se agrupan los autores siguiendo los trabajos de Koppel *et al.* [16] en las siguientes clases: *i*) 10s (13-17); *ii*) 20s (23-27); y *iii*) 30s (33-47). Los datos están equilibrados por sexo pero no lo están por edad. Cada autor puede contener desde un post hasta diez. La distribución del número de autores por rango de edad se muestra en la Tabla 4.21.

Rango edad	**Training**	**Test**
10s	2.500	240
20s	42.600	3.840
30s	30.800	2.720
Total	75.900	6.800

TABLA 4.21: Distribución de autores por rango de edad (PAN-AP-2013).

Hemos utilizado máquinas de vectores soporte, ajustando sus parámetros hasta obtener los mejores resultados con un núcleo gausiano con *g=0,01* y *c=2.000*. Para poder comparar los resultados con los de la tarea del PAN 2013 hemos utilizado la medida de *accuracy*, concretamente calculando el ratio entre el número de autores correctamente predecidos y el número total de autores.

4.5.2 Resultados

En la Tabla 4.22 se muestran los resultados obtenidos de manera comparada con los oficiales de la tarea. Como se puede apreciar, obtenemos la séptima posición en la predicción de sexo y la tercera en la predicción de edad.

Posición	**Sexo**		**Edad**	
1	Santosh	64,73	Pastor	65,58
2	Pastor	62,99	Santosh	64,30
3	Haro	61,65	**Rangel**	63,50
4	Ladra	61,38	Haro	62,19
5	Flekova	61,03	Flekova	59,66
6	Jankowska	58,46	Ladra	57,27
7	**Rangel**	57,13	Yong	57,05
8	Kern	57,06	Ramirez	56,51
9	Jimenez	56,27	Aditya	56,43
10	Ayala	55,26	Jimenez	54,29
11	Cagnina	55,16	Gillam	53,77
12	Yong	54,68	Kern	53,75
13	Mechti	54,55	Moreau	50,49
14	Weren	53,62	Meina	49,30
15	Meina	52,87	Weren	46,15
16	Ramirez	51,16	Jankowska	42,76
17	*Baseline*	50,00	Cagnina	41,48
18	Aditya	50,00	Hidalgo	40,00
19	Hidalgo	50,00	Farias	35,54
20	Farias	49,82	*Baseline*	33,33
21	Moreau	49,67	Ayala	29,15
22	Gillam	47,84	Mechti	5,12

TABLA 4.22: Resultado comparado con la tarea PAN 2013 en español.

En el caso de la predicción de edad, no hay diferencia significativa entre nuestra aproximación (63,50%) y los equipos de Santosh (64,30%) y Pastor (65,58%) al 95% y 99% de significación respectivamente en un test t-Student [157], lo que sugiere que las características propuestas para representar los textos (véase Tabla 4.3.1) han permitido obtener resultados competitivos, sobre todo para la identificación de edad. Quizás esto se debe a que el estilo de escritura depende más de la edad que del sexo, confirmando lo que afirman Goswami *et al.* [21] sobre la correlación entre edad y uso del lenguaje. En cualquier caso, la tarea de identificación de edad parece ser menos difícil que la de sexo, tal y como se puede apreciar por los bajos resultados cercanos a la predicción aleatoria (50%) en este último caso.

4.5.3 Discusión

Sobre la base de las características descritas al principio del capítulo, cuyo objetivo ha sido modelar el modo en que los usuarios expresan sus emociones, hemos experimentado en la identificación de edad y sexo, poniendo los resultados en perspectiva con los obtenidos por los participantes de la tarea PAN 2013. Los resultados obtenidos muestran que la representación es competitiva, lo que implica en cierto sentido que la expresión de las emociones está íntimamente ligada a aspectos de la persona como su sexo, y por la significación de los resultados, especialmente a su edad.

4.6 Conclusiones

Hemos investigado el impacto de las emociones en diferentes tareas tales como el seguimiento de tendencias y la identificación de sexo y edad. Para lo primero, hemos generado un corpus con textos publicados en la red social Twitter en torno a un tema de elevada temperatura política: el caso Bárcenas[13]. Se ha procesado el texto obteniendo un conjunto de características basadas en parámetros cuantitativos de la red social, como el número de tuits o el número de autores en un momento dado, y se ha combinado con las emociones expresadas en cada momento. La construcción de un modelo basado en estados a partir de la combinación propuesta de características ha permitido hacer más aparentes las tendencias con respecto a la temática dada, poniendo de manifiesto el papel de las emociones en la evolución de las tendencias en los medios sociales.

Un segundo aspecto que hemos investigado es el relativo al propio procesamiento afectivo en relación a la identificación de las seis emociones básicas y su relación con otros usos del lenguaje como la ironía, o con aspectos demográficos del emisor de los mensajes tales como su sexo. Para ello hemos construido un corpus, EmIroGeFB, y lo hemos etiquetado manualmente respecto a las seis emociones básicas, y a la presencia o ausencia de ironía. La construcción del corpus ha sido necesaria debido a la inexistencia de recursos etiquetados con estas variables en español y en social media. A partir de este corpus, presentamos una serie de estadísticas que muestran la dificultad de la tarea de anotación, incluso para un ser humano, y especialmente para el caso de la ironía.

Hemos propuesto un conjunto de características para tratar de modelar los aspectos estilísticos y emotivos del discurso del usuario, por ejemplo rasgos morfosintácticos, frecuencias de aparición de signos de puntuación o la afectividad de las palabras con ayuda del diccionario SEL. A partir de estas características, hemos generado un modelo para identificar en los textos las seis emociones básicas de Ekman, obteniendo resultados similares a los del estado del arte. Con este mismo conjunto de características, hemos identificado el sexo de los autores de los comentarios de Facebook de EmIroGeFB, así como el sexo y la edad de los autores de medios sociales de la tarea del PAN 2013. Los resultados obtenidos muestran que la representación es competitiva para ambos cometidos (identificación de emociones e identificación de edad y sexo) y que por lo tanto parece existir cierta relación entre la expresión de las emociones y los ragos de edad y sexo del autor que las expresa.

En el próximo capítulo, partiendo de esta premisa que relaciona emociones y rasgos de edad y sexo, profundizamos en el análisis del discurso con el objetivo de capturar mejor el modo en que dichas emociones son expresadas y cómo se articulan y combinan con el resto de elementos participantes en el discurso, verificando además su aplicabilidad a la tarea de identificación de edad y sexo de autores de textos de medios sociales.

[13] https://es.wikipedia.org/wiki/Caso_B%C3%A1rcenas

Capítulo 5

EmoGraph: Una aproximación basada en grafos

Nuestro objetivo es modelar el modo en que los usuarios construyen sus textos y el papel que tienen las emociones con respecto a los temas que en ellos se describen. Nuestra hipótesis de partida presume que el modo en que los usuarios expresan sus emociones sobre los diferentes temas tiene un elevado nivel de dependencia con su edad y sexo. Tal hipótesis se ve soportada por diferentes investigaciones así como por nuestra propia experimentación, presentada en el capítulo previo. Por ello, nuestro objetivo es modelar el modo en que tales emociones son expresadas, y obtener información sobre su importancia relativa no sólo por su frecuencia de aparición, sino por su posición con y en relación con el resto de elementos del discurso [158].

En estudios previos como los de Pennebaker [50] se postula que los hombres usan más preposiciones que las mujeres porque tienden a describir con mayor detalle su entorno, o que las mujeres tienden a hablar más sobre relaciones sociales. Sobre esta base se puede conjeturar que los hombres utilizarán más sintagmas preposicionales que las mujeres, o que éstas expresarán de modo diferente sus emociones con respecto a determinados temas, lo que puede dotar de cierta importancia a determinadas secuencias como:

preposición + determinante + nombre + adjetivo

Tal importancia no será capturada por modelos basados en conteos de frecuencias y puede resultar compleja de modelar mediante aproximaciones basadas en n-gramas, por la propia elección del valor de n. Por otro lado, la idiosincrasia propia de los textos objeto, de alta informalidad por su expresión en todo tipo de medios sociales, hacen de difícil y costosa aplicación técnicas elaboradas de análisis sintáctico. Por ello utilizamos grafos, dada su capacidad de modelar y analizar estructuras complejas, y que como veremos a continuación, permiten modelar diferentes situaciones relativas al uso del lenguaje, en diferentes tareas de procesamiento automático del mismo.

5.1 Grafos en procesamiento del lenguaje natural

Los grafos han sido ampliamente utilizados en recuperación de información y procesamiento del lenguaje natural. Por ejemplo, en [159] se describe en detalle cómo aprender modelos basados en grafos para tareas como recuperación y clasificación de documentos, filtrado colaborativo, análisis unificado de enlaces o recuperación de imágenes. En [160] se presenta un nuevo modelo basado en grafos para la representación de documentos independientemente del idioma junto con una medida de similitud para comparar documentos en diferentes idiomas. En [161] los autores representan los documentos con grafos considerando múltiples niveles lingüísticos (léxico, morfológico, sintáctico y semántico), y usan el camino más corto para extraer patrones de texto útiles. En [162] se presenta una revisión completa de teoría de grafos, redes y algoritmos para recuperación de información y diferentes tareas de procesamiento del lenguaje natural.

Con relación al procesamiento afectivo, la mayoría de las aproximaciones basadas en grafos se han centrado en análisis de sentimiento. Una de las primeras propuestas aproxima el problema de identificación automática de la polaridad de adjetivos [163]. En [164] se investiga la relación entre significados de palabras y subjetividad, mientras que en [165] los autores modelan frases como grafos de coocurrencia de palabras, obviando los signos de puntuación y palabras vacías. Los autores obtienen propiedades del grafo y utilizan aprendizaje automático para distinguir entre opiniones positivas y negativas. Finalmente, en [166] los autores construyen grafos de opiniones a nivel de discurso para realizar análisis de opinión.

Hasta donde conocemos, este es el primer caso en el que se aplica un grafo enriquecido con emociones para modelar la edad y el sexo de autores de textos anónimos.

5.2 EmoGraph

En este apartado se propone la construcción de un grafo con las diferentes categorías morfosintácticas del texto, enriquecido con información semántica como los tópicos de los que trata, los tipos de verbo que se utilizan, y las emociones y sentimientos que se expresan. Con el texto modelado, se obtienen medidas basadas en teoría de grafos que proporcionan los pesos relativos a la importancia de cada una de las características utilizadas según su posición en el discurso y su relación con las demás.

5.2.1 Construcción y enriquecimiento del grafo

Para cada autor, se construye un grafo con todos sus textos, en lugar de un grafo por frase, lo que permite modelar como el autor enlaza una frase con la siguiente, e incluso un párrafo con el siguiente. El grafo construido es dirigido, pues se pretende modelar la secuencia del discurso. Para cada texto, se realiza un análisis morfológico con Freeling[1] [54, 55], obteniendo las partes del discurso y los lemas de

[1] http://nlp.lsi.upc.edu/freeling/

cada palabra. Freeling describe cada parte del discurso con una etiqueta Eagle[2] [167, 168]. Se modela cada parte del discurso como un nodo N del grafo G y cada arco E define la secuencia entre partes del discurso en el texto, modelada como enlaces dirigidos entre la parte del discurso precedente y la actual. Por ejemplo, consideremos un ejemplo simple como el siguiente:

El gato come pescado y bebe agua.

Que genera la siguiente secuencia de etiquetas Eagle, y que se modela como el grafo mostrado en la Figura 5.1. Debido a que el enlace VMIP3S0 -> NCMS000 se produce dos veces, el peso de ese enlace es el doble que el resto.

DA0MS0->NCMS000->VMIP3S0->NCMS000->CC->VMIP3S0->NCMS000->Fp

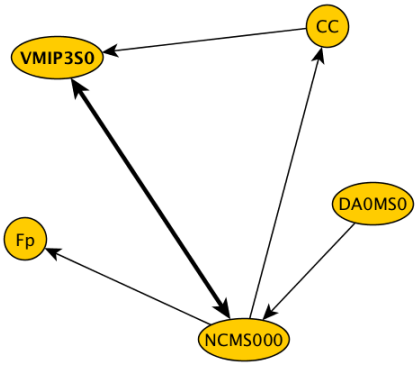

FIGURA 5.1: Grafo de partes del discurso de "El gato come pescado y bebe agua."

El siguiente paso es el enriquecimiento del grafo mediante información semántica y afectiva. Para cada palabra en el texto, se hace uso de ciertos recursos lingüísticos para obtener la siguiente información:

- **Dominios:** Si la palabra es un nombre común, un adjetivo o un verbo, se busca por el dominio principal asociado a su lema. Para ello se utiliza Wordnet Domains[3], enlazándolo con la versión en español de Euro Wordnet[4]. Si la palabra tiene uno o más tópicos relacionados, se crea un nuevo nodo, si previamente no existe, por cada uno de dichos dominios y un nuevo arco desde la etiqueta Eagle actual a cada uno de ellos. En el ejemplo anterior, *gato* está relacionado con ambos *biology* y *animals*, por lo que se crearán dos nodos con un enlace a cada uno de ellos desde la etiqueta NCMS000 correspondiente a gato: NCMS000 -> biology y NCMS000 -> animals.

[2]El grupo Eagles (http://www.ilc.cnr.it/EAGLES96/intro.html) propone una serie de recomendaciones para la anotación morfosintáctica de corpus. En este apartado se ha utilizado la versión en español (http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html), donde por ejemplo: en la frase "El gato come pescado y bebe agua, la palabra "gato" se devuelve como NCMS000 donde NC significa nombre común, M significa masculino, S significa singular y 000 es un relleno hasta los 7 caracteres; o la palabra "come" que se devuelve como VMIP3S0 donde V significa verbo, M significa verbo principal (en este caso, no auxiliar), I significa modo indicativo, P significa tiempo presente, 3 significa tercera persona, S significa singular, y 0 es el relleno hasta los 7 caracteres.

[3]http://wndomains.fbk.eu/

[4]http://www.illc.uva.nl/EuroWordNet/

- **Clasificación semántica de los verbos:** Si la palabra es un verbo se busca por la clasificación semántica de su lema. Sobre la base de lo investigado en [169], hemos anotado manualmente un total de 158 verbos en una de las siguientes categorías semánticas: *i*) *percepción (perception)*: ver, escuchar, oler...; *ii*) *comprensión (understanding)*: saber, conocer, entender, pensar...; *iii*) *duda (doubt)*: dudar, ignorar...; *iv*) *lenguaje (language)*: decir, declarar, hablar...; *v*) *emoción (emotion)*: sentir, querer, amar...; *vi*) and *permiso (will)*: deber, prohibir, permitir.... Se crea un nodo con la etiqueta semántica correspondiente, si aún no existe, y se añade un arco desde la etiqueta Eagle actual a la nueva. Por ejemplo, si el verbo es de percepción, se crea un nodo llamado *perception* y se enlaza con el nodo VMIP3S0: VMIP3S0 -> percepcion.

- **Polaridad:** Si la palabra es nombre común, adjetivo, adverbio o verbo, se busca por su polaridad en un lexicón de sentimiento[5]. Por ejemplo, considerando la siguiente frase:

  Ella es una amiga increíble.

  Se modela con la siguiente secuencia de etiquetas Eagles:

  PP3FS000->VSIP3S0->DI0FS0->NCFS000->AQ0CS0(->positive & negative)->Fp

  El nodo adjetivo AQ0CS0 correspondiente a *increíble* tiene enlaces a ambas etiquetas *positive* y *negative* ya que *increíble* puede tener ambas polaridades dependiendo del contexto, por lo que desde el punto de vista del sentimiento es una palabra ambigua. Sin embargo, desde el punto de vista del *author profiling* nos proporciona un par de nodos de afectividad con sus respectivos arcos.

- **Emociones:** Si la palabra es nombre propio, adjetivo, adverbio o verbo, se busca su posible relación con cada una de las emociones mediante el empleo del lexicón SEL. Se crea un nodo para cada una de las emociones encontradas y se asigna un enlace a cada uno de ellos. Sirva de ejemplo la siguiente frase:

  He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

  El EmoGraph de la sentencia anterior se muestra en la Figura 5.2. La secuencia se debe seguir empezando por el nodo VAIP1S0 correspondiente a *He*. El tamaño de los nodos se ha representado proporcional a su valor *eigenvector* [171] y su color en base a la *modularidad* [172]. Por último, se enlaza el último elemento de cada frase (e.g. Fp) con el primero de la siguiente frase, ya que como se ha comentado anteriormente, estamos interesados también en el modo en que los autores enlazan sus frases, señal inequívoca de cómo utilizan el lenguaje, y cómo viene representado por ciertos signos de puntuación (e.g. . ; :).

[5]Con la ayuda de expertos en análisis de opiniones en Autoritas (http://www.autoritas.net), hemos traducido manualmente y adaptado al español el lexicón descrito en [170]

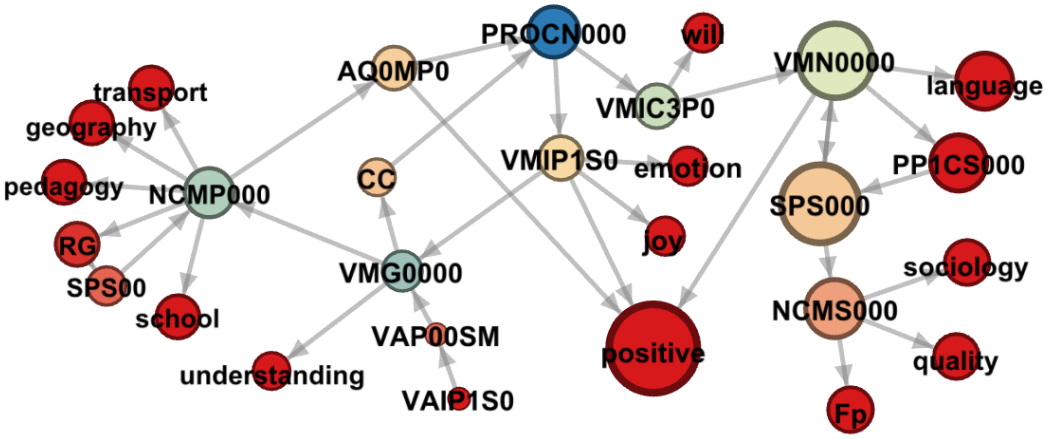

FIGURA 5.2: EmoGraph de "He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público".

En la Figura 5.3 se puede apreciar un ejemplo de EmoGraph para un texto completo de un autor.

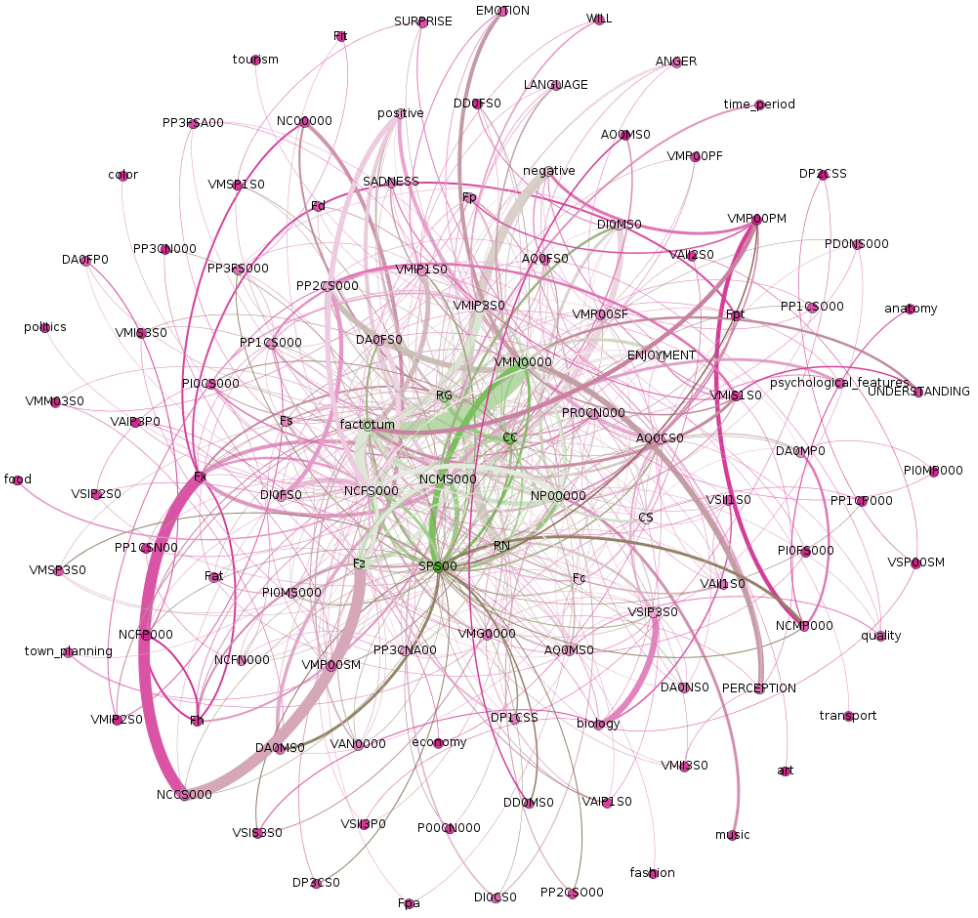

FIGURA 5.3: EmoGraph de un texto completo de un autor.

5.2.2 Características

Una vez construido el grafo se obtiene un conjunto de características para construir un modelo de espacio vectorial que permita aplicar técnicas de aprendizaje automático para la clasificación de los textos. Se obtienen tales características en base al análisis de grafos, teniendo en cuenta dos grupos: *i*) propiedades generales del grafo, que proporcionan aspectos de la *estructura* global de los textos; y *ii*) propiedades específicas de sus *nodos* y de cómo se relacionan entre ellos, que proporcionan propiedades específicas sobre cómo los usuarios usan el lenguaje.

5.2.2.1 Características generales de la estructura del grafo

En esta sección se describe el modo de obtener ocho de las características del espacio vectorial a partir de propiedades generales del grafo. Concretamente:

- **Ratio nodos-arcos**. Se calcula el ratio entre el número total de nodos N y el número de arcos E del grafo G={N,E}. El número máximo posible de nodos (429) viene dado por la suma de: *i*) el número total de etiquetas Eagles (247); *ii*) el número total de tópicos de Wordnet Domains (168); *iii*) el número total de clases semánticas de los verbos (6); *iv*) el número total de emociones (6); y *v*) el número total de polaridades de sentimiento (2). El número máximo posible de arcos (183.612) en un grafo dirigido se calcula teóricamente como:

  $$max(E) = N * (N - 1)$$

  donde N es el número total de nodos. El ratio entre el número de nodos y arcos nos proporciona un indicador de cómo de conectado está el grafo, o en el caso propuesto, cómo de complicada es la estructura del discurso del usuario.

- **Grado medio** del grafo, que indica cómo de interconectado está. El grado de un nodo es el número de sus vecinos; en nuestro caso, esto viene dado por el número de otras categorías gramaticales o información semántica que precede o sucede a cada nodo. El grado medio se calcula como media de los grados de cada nodo.

- **Grado medio ponderado** del grafo, que se calcula como el grado medio pero dividiendo el grado de cada nodo por el máximo número de arcos que un nodo puede tener (N-1). Por lo tanto, el resultado se transforma al rango [0, 1]. El significado es el mismo que en el caso anterior, pero variando la escala.

- **Diámetro** del grafo, que indica la mayor distancia entre cualquier par de nodos. Se obtiene mediante el cálculo de todos los caminos más cortos entre cada par de nodos del grafo y se selecciona el camino de mayor distancia entre todos ellos. Esto es:

  $$d = max_{n \in N} \varepsilon(N)$$

  donde $\varepsilon(n)$ representa la eccentricidad o distancia geodésica [173] mayor entre n y cualquier otro nodo. En nuestro caso, mide cómo de lejos una categoría gramatical, o información semántica asociada, está de todas las demás, por ejemplo cómo de lejos se encuentra un tópico de una emoción.

- **Densidad** del grafo, que mide cómo de cerca está el grafo de ser completo [174], o en nuestro caso, cuan denso es el texto en el sentido de cómo cada categoría gramatical se usa en combinación

al resto. Dado un grafo G=(N,E), mide cuántos arcos hay en el conjunto E comparado con el máximo número posible de arcos entre los nodos del conjunto N. Entonces, la densidad se calcula como:

$$D = \frac{2*|E|}{(|N|*(|N|-1))}$$

- **Modularidad** del grafo, que mide la fuerza de la división del grafo en módulos, grupos, clusters o comunidades. Una modularidad elevada indica que los nodos dentro de los módulos tienen conexiones densas mientras que tienen conexiones dispersas con nodos en otros módulos. En nuestro caso, puede indicar cómo se modela el discurso en diferentes unidades estructurales o estilísticas. La modularidad se calcula siguiendo el algoritmo descrito en [175].
- **Coeficiente de clustering**, que indica el grado de transitividad del grafo, esto es, si a está directamente enlazado a b y b está directamente enlazado a c, la probabilidad de que a esté también enlazado a c. Indica cómo están ensamblados los nodos en su vecindad, o en nuestro caso, cómo las diferentes categorías gramaticales, o información semántica como las emociones, están relacionadas unas con otras. Para cada nodo, el coeficiente de clustering ($cc1$) se calcula siguiendo la fórmula de Watts-Strogatzt [176]:

$$cc1 = \frac{\sum_{i=1}^{n} C(i)}{n}$$

donde cada $C(i)$ mide cómo de cerca están los vecinos de un nodo i de ser un grafo completo, y se calcula como sigue:

$$C(i) = \frac{|\{e_{jk} : n_j, n_k \in N_i, e_{jk} \in E\}|}{k_i(k_i-1)}$$

donde e_{jk} es el arco que conecta al nodo n_j con el nodo n_k y k_i es el número de vecinos del nodo i. Finalmente, se calcula el coeficiente de clustering global como la media de los coeficientes de todos los nodos, excluyendo aquellos con grados 0 ó 1, y siguiendo el algoritmo descrito en [177].

- **Longitud media de camino**, que es la distancia media entre todos los pares de nodos y que puede ser calculada siguiendo [178]. Proporciona un indicador de cuán lejos están algunos nodos de otros, o en nuestro caso, cómo de lejos algunas categorías gramaticales se encuentran entre sí y de las emociones que se expresan.

5.2.2.2 Características específicas de los nodos

Para cada nodo en el grafo, se calculan dos medidas de centralidad [179]: *betweenness* y *eigenvector*. Se usan ambos valores como pesos para dos características nombradas respectivamente BTW-xxx y EIGEN-xxx, donde *xxx* es el nombre del nodo (e.g. AQ0CS0, positive, joy, animal, etcétera).

- **Betweenness**: mide cómo de importante es un nodo por el número de caminos más cortos que lo atraviesan. La medida de centralidad *betweenness* de un nodo x es el ratio entre el total de caminos más cortos entre dos nodos del grafo que pasan a través de x. Se calcula como sigue:

$$BC(x) = \sum_{i,j \in N-\{n\}} \frac{\sigma_{i,j}(n)}{\sigma_{i,j}}$$

donde $\sigma_{i,j}$ es el número total de caminos más cortos entre el nodo i y el j, y $\sigma_{i,j}(n)$ es el número total de aquellos caminos que pasan por n. En nuestro caso, si un nodo tiene un alto valor de *betweenness* significa que es un elemento común usado para enlazar entre diferentes partes del discurso, por ejemplo: preposiciones, conjunciones, o incluso verbos y nombres. Esta medida nos proporciona un indicador de cuáles son los elementos de enlace más utilizados en las estructuras lingüísticas de un autor.

- **Eigenvector**: mide la influencia de un nodo en el grafo [180]. Dado un grafo y su matriz de adyacencia $A = a_{n,t}$ donde $a_{n,t}$ es 1 si un nodo n está enlazado a un nodo t, y 0 en caso contrario. Se calcula el valor de *eigenvector* como:

$$x_n = \frac{1}{\lambda} \sum_{t \in M(n)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{n,t} x_t$$

donde λ es una constante que representa el mayor *eigenvalue* asociado con la medida de centralidad, $M(n)$ es un conjunto de vecinos del nodo n y x_t representa cada nodo diferente a x_n en el grafo. Esta medida proporciona un indicador de cuáles son las categorías gramaticales con un uso más central en el discurso del autor, por ejemplo: nombres, verbos o adjetivos.

5.2.2.3 Construcción del espacio vectorial

A partir de los cálculos realizados sobre el grafo y descritos en las subsecciones previas, se obtiene el siguiente conjunto de características que constituyen el modelo de espacio vectorial que representa EmoGraph:

Estructura del grafo	8 características
ENRatio	Ratio nodos-arcos
Degree	Grado medio del grafo
WeightedDegree	Grado medio ponderado
Diameter	Diámetro del grafo
Density	Densidad del grafo
Modularity	Grado de modularidad
Clustering	Coeficiente de agrupamiento
PathLength	Longitud media del camino
Específicas de los nodos	944 características
BTW-*xx*	Valor de *betweenness* de cada nodo (*xx*)
EIGEN-*xx*	Valor *eigenvector* de cada nodo (*xx*)
Características de estilo	59 características
	Descritas en la Tabla 4.2

Tabla 5.1: Modelo de espacio vectorial de EmoGraph.

Cada *xx* del conjunto de características específicas de los nodos se corresponde con un nodo del grafo, que a su vez representa cada una de las 247 posibles etiquetas morfosintácticas, los 168 tópicos de Wordnet Domains, los 6 verbos, las 6 emociones y las 2 polaridades, sumando el total de las 944 características descritas como específicas de los nodos, que se combinan con las 8 características estructurales del grafo y las 58 características estilísticas descritas en el apartado 4.3.1, sumando un total de 1.011 características.

5.2.3 Metodología

Se ha utilizado el corpus proporcionado por la tarea de *author profiling* del PAN del 2013. Se han probado diferentes algoritmos de aprendizaje automático en su implementación en Weka. Los mejores resultados con la partición de entrenamiento se obtienen con: *i*) máquinas de vectores soporte con núcleo gausiano con g=0.20 y c=1 para la identificación de sexo; y *ii*) máquinas de vectores soporte con núcleo gausiano con g=0.08 y c=1 para la identificación de edad. Para que los resultados sean comparables a los de la tarea del PAN, se evalúa el modelo propuesto en la partición de pruebas mediante la medida de *accuracy*. Se efectúa un cálculo de significación mediante la prueba t-Student, asumiendo como hipótesis

nula que de media todos los clasificadores son iguales $H_0 : p_1 = p_2$, y se rechaza dicha hipótesis en caso contrario.

Se compara EmoGraph no sólo con los resultados de los participantes sino también con dos propuestas propias: *i*) Rangel-S, modelado con las características de estilo [142] descritas en el apartado 4.3.1; y *ii*) Rangel-nG consistente en n-gramas de partes del discurso, donde se obtiene el valor de n como aquél que mejores resultados proporciona al iterar su valor desde 1 hasta 10. Aunque los n-gramas de partes del discurso también capturarían la estructura sintáctica del discurso, y la representación basada en características de estilo capturaría otros aspectos como la emotividad, nuestra intuición es que EmoGraph no sólo captura dicha información sino que además captura información valiosa como su localización en el texto y su relación con otros aspectos del discurso.

5.2.4 Resultados

En este apartado se presentan los resultados para la identificación de sexo y edad, en términos de *accuracy*, evaluando en la partición de pruebas de la tarea PAN-AP-13, calculando y mostrando la significación estadística de los mismos en comparación con los mejores resultados obtenidos por los participantes de la tarea del PAN.

5.2.4.1 Identificación de sexo

En la Tabla 5.2 se muestran los resultados para la identificación de sexo. Se puede apreciar que EmoGraph mejora significativamente los resultados tanto de Rangel-S como de Rangel-nG. Además, sus resultados son competitivos con respecto a los participantes de la tarea, obteniendo una segunda posición en el ranking.

Ranking	Equipo	Accuracy
1	Santosh	64,73
2	**EmoGraph**	63,65
3	Pastor	62,99
4	Haro	61,65
5	Ladra	61,38
6	Flekova	61,03
7	**Rangel-nG**	60,16
8	Jankowska	58,46
9	**Rangel-S**	58,00
...	...	
19	Baseline	50,00
...	...	
24	Gillam	47,84

TABLA 5.2: Resultados en términos de *accuracy* para la identificación de sexo en la partición de pruebas en español del corpus PAN-AP-2013.

Se lleva a cabo la prueba t-Student para verificar la significación de los resultados. Comparando *Santosh* con EmoGraph ($z_{0.05} = 1.4389 < 1.960$) se aprecia que no hay diferencia signficativa entre ellos a una confianza del 95%. *Santosh* [74] aproximó la tarea combinando características basadas en el contenido (n-gramas de palabras), características basadas en el estilo (n-gramas de partes del discurso y medidas estilométricas) y características basadas en tópicos (LDA). La comparación de *Santosh* con *Pastor*($z_{0.05} = 2.3135 > 1.960$; $z_{0.01} = 2.3135 < 2.576$) muestra que no hay diferencia significativa

al 95% de confianza pero sí la hay al 99%. *Pastor* [42] aproximó la tarea con una representación de segundo orden basada en la relación entre documentos y perfiles. Comparando EmoGraph con *Pastor* ($z_{0.05} = 0.8748 < 1.960$), no se aprecia diferencia significativa al 95% de confianza. La conclusión es que, independientemente del ranking, las tres primeras aproximaciones obtienen resultados similares al 95% y 99% de confianza respectivamente. El resto de participantes aproximaron la tarea con características más estándar. Por ejemplo, *Haro* [71] y *Flekova* [66] usaron bolsas de palabras, y *Jankowska* [77] n-gramas de caracteres.

5.2.4.2 Identificación de edad

En la Tabla 5.3 se muestran los resultados para la identificación de edad. Se puede apreciar que EmoGraph mejora significativamente los resultados obtenidos por la representación basada en características de estilo o los n-gramas de partes del discurso. Además, es competitiva con los participantes de la tarea pues obtiene la primera posición en el ranking.

Ranking	Equipo	Accuracy
1	**EmoGraph**	66,24
2	Pastor	65,58
3	Santosh	64,30
4	**Rangel-S**	62,59
5	Haro	62,19
6	**Rangel-nG**	61,62
7	Flekova	59,66
...	...	
21	Baseline	33,33
...	...	
23	Mechti	5,12

TABLA 5.3: Resultados en términos de *accuracy* para la identificación de edad en la partición de pruebas en español del corpus PAN-AP-2013.

Se lleva a cabo la prueba t-Student mostrando que no hay diferencia significativa entre EmoGraph y *Pastor* ($z_{0.05} = 0.8894 < 1.960$), ni entre *Pastor* y *Santosh* ($z_{0.05} = 1.7139 < 1.960$) al 95% de confianza, pero sí una diferencia significativa entre EmoGraph y *Santosh* ($z_{0.05} = 2.6027 > 1.960$). En este caso hay una mayor diferencia entre las tres primeras propuestas, aunque estadísticamente no hay significación entre la primera (EmoGraph) y la segunda. De manera similar a la identificación de sexo, estos métodos más elaborados superan significativamente a aproximaciones basadas en características estándar menos elaboradas.

5.2.5 Discusión

En la sección anterior se ha mostrado la efectividad de EmoGraph en la identificación de edad y sexo. En esta sección se investiga la diferencia en el uso de las palabras, verbos y emociones por sexo y edad, se analiza cuáles son las características más discriminantes para ambas tareas, se efectúa un análisis del error así como de la complejidad computacional de la propuesta, y se termina analizando el impacto de las emociones, modeladas mediante EmoGraph, en la tarea de *author profiling*.

5.2.5.1 Diferencias en base a edad y sexo

Es de sentido común pensar que escribimos de manera diferente dependiendo de nuestra edad y sexo, pero en muchos casos existen clichés sin ningún transfondo científico. En este apartado se muestran diferencias en el uso del lenguaje dependiendo de dichas variables demográficas.[6]

FIGURA 5.4: Palabras más frecuentemente usadas por las mujeres en el corpus PAN-AP-13.

FIGURA 5.5: Palabras más frecuentemente usadas por los hombres en el corpus PAN-AP-13.

En las Figuras 5.4 y 5.5 se muestran las palabras más frecuentemente usadas por mujeres y hombres respectivamente. Se puede apreciar que no hay grandes diferencias entre sexos. Independientemente del sexo, las personas parecen preocuparse sobre la vida, el amor (*amor*, *quiero*) y la esperanza (*espero*).

Con la ayuda de Wordnet Domains se obtienen los tópicos más utilizados por los autores del corpus. Se eliminan los más frecuentes[7] ya que no son demasiado informativos al estar en el top de la jerarquía. Las nubes correspondientes a los tópicos se muestran en las Figuras 5.6, 5.7 y 5.8 para mujeres en cada uno de los grupos de edad (10s, 20s y 30s), y en las Figuras 5.9, 5.10 y 5.11 para hombres en los mismos grupos de edad.

Se aprecia que los más jóvenes tienden a escribir más sobre diferentes disciplinas como física (*physics*), lingüística (*linguistics*), literatura (*literature*), metrología (*metrology*), derecho (*law*), medicina (*medicite*), química (*chemistry*), etcétera, debido seguramente a que hablan sobre lo que están estudiando en el colegio, instituto o universidad. Las mujeres parece que hablan más de química o gastronomía, mientras que los hombres lo hacen sobre física o derecho. Ambos escriben sobre música (*music*) y entretenimiento (*play*). Por el contrario a lo que podríamos pensar, las chicas adolescentes (10s) hablan más de sexo y sexualidad (*sexuality*) mientras que los chicos no lo hacen, y sucede lo contrario con las compras (*shopping*). Según crecen, ambos grupos muestran más interés en lo relacionado con la vivienda (*buildings*) seguramente por la búsqueda de su propia independencia (comprar/alquilar vivienda), y respecto a lo relacionado con los animales (*animals*), la gastronomía (*gastronomy*), la medicina (*medicine*) o la religión (*religion*), aunque este último en mayor medida en el caso de los hombres.

[6] Las conjeturas enunciadas en este apartado se basan en los estudios realizados sobre el corpus de la tarea PAN-AP-13 y no hay que generalizarlas.

[7] e.g. biology, quality, features, psychological, economy, anatomy, period, person, transport, time y psychology

FIGURA 5.6: Dominios más frecuentes para mujeres en el grupo 10s en el corpus PAN-AP-13.

FIGURA 5.7: Dominios más frecuentes para mujeres en el grupo 20s en el corpus PAN-AP-13.

FIGURA 5.8: Dominios más frecuentes para mujeres en el grupo 30s en el corpus PAN-AP-13.

FIGURA 5.9: Dominios más frecuentes para hombres en el grupo 10s en el corpus PAN-AP-13.

FIGURA 5.10: Dominios más frecuentes para hombres en el grupo 20s en el corpus PAN-AP-13.

FIGURA 5.11: Dominios más frecuentes para hombres en el grupo 30s en el corpus PAN-AP-13.

En la Figura 5.12 se muestra la proporción de uso de palabras emocionales, donde no se aprecia una tasa de expresión de emociones diferente por sexo, aunque se puede observar que las mujeres parecen expresar más disgusto que los hombres, que expresan más la tristeza.

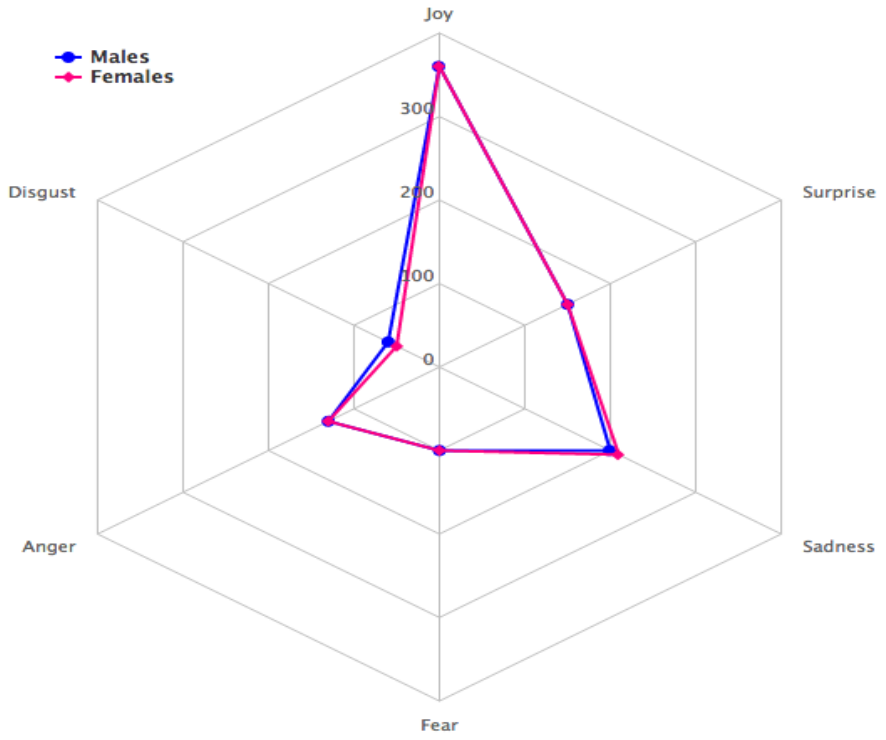

FIGURA 5.12: Palabras emocionales por sexo en el corpus PAN-AP-13.

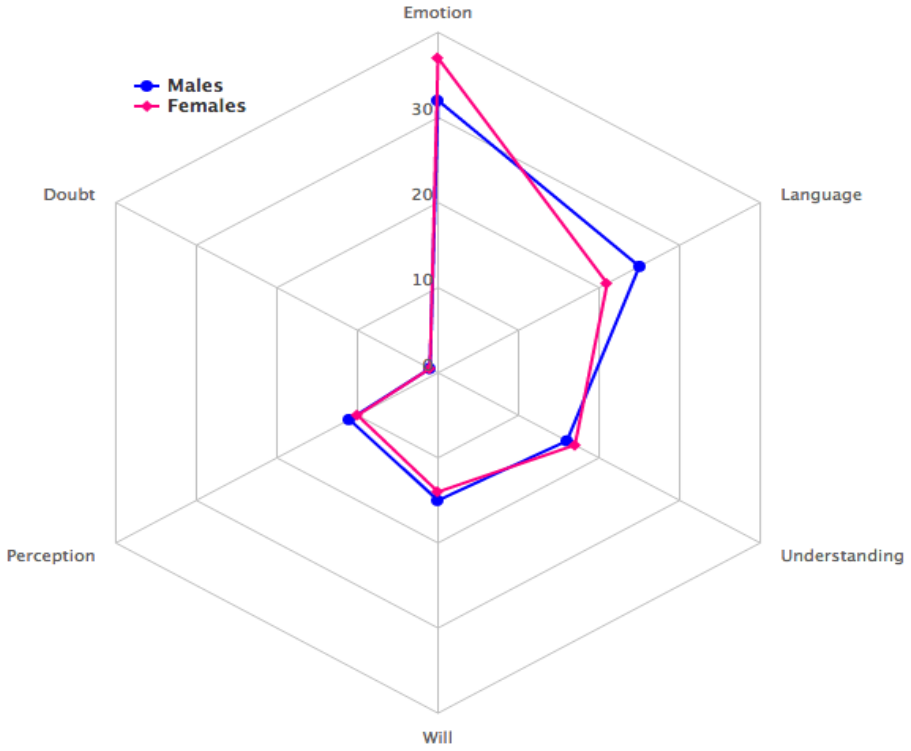

FIGURA 5.13: Uso de tipos de verbo por sexo en el corpus PAN-AP-13.

Con respecto al uso de tipos de verbos, el interés se centra en investigar qué clase de acciones (verbos) son más referidas por hombres y mujeres, y cómo esto cambia a lo largo del tiempo. En la Figura 5.13 se

ilustra que las mujeres usan más verbos de tipo *emocional* (e.g. sentir, querer, amar...) que los hombres, quienes usan más verbos del *lenguaje* (e.g. decir, hablar, explicar...). Éste es un resultado interesante porque, aunque en la Figura 5.12 se muestra que ambos hombres y mujeres usan palabras emocionales en una proporción similar, la Figura 5.13 muestra que las mujeres comunican con más verbos relativos a las emociones que los hombres.

Puesto que el uso de los tipos de verbo nos muestra ciertas diferencias entre sexos, se analiza su evolución a lo largo del tiempo. En las Figuras 5.14 y 5.15 se muestra la evolución a través de los rangos de edad 10s, 20s y 30s. El uso de verbos de *emoción* decrece a lo largo de los años, mientras que los verbos de *entendimiento* (e.g. saber, entender, pensar...) se incrementan para hombres y permanecen estables para mujeres, aunque hay que remarcar que las mujeres empiezan a utilizarlos a edades más tempranas a ratios similares al que lo hacen los hombres más tarde. De manera similar, los verbos de *permiso* (e.g. deber, permitir, prohibir...) se incrementa para ambos sexos aunque en mayor medida para hombres. Se puede afirmar que las mujeres usan más los verbos de *emoción* que los hombres en cualquier estadio de su vida, y lo contrario sucede con los verbos de *lenguaje*.

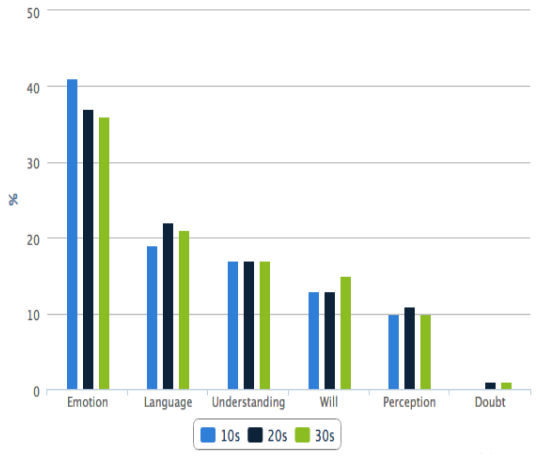

FIGURA 5.14: Evolución en el uso de tipos de verbos por las mujeres en el corpus PAN-AP-13.

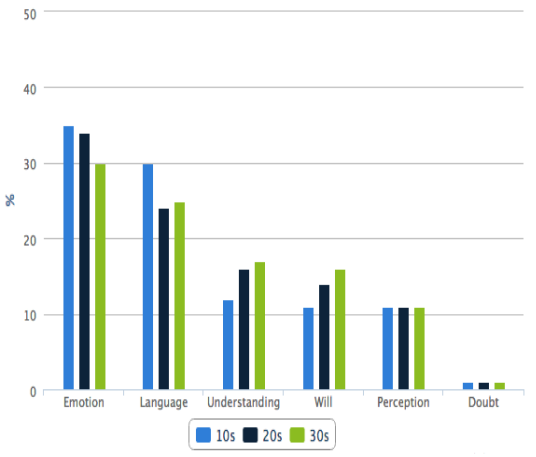

FIGURA 5.15: Evolución en el uso de tipos de verbos por hombres en el corpus PAN-AP-13.

5.2.5.2 Características más discriminantes

Se obtienen las características más discriminantes en función de su ganancia de información [181] en la clasificación. En la Tabla 5.4 se muestran las 20 más discriminantes del total de 1.011 características, donde se puede observar que las basadas en los nodos del grafo (Betweenness *BTW-xxx* y eigenvector *EIGEN-xxx*) se encuentran entre las más discriminantes[8]. Se puede identificar un número mayor de características *eigen* (principalmente verbos, nombres y adjetivos) en la identificación de sexo en comparación con un mayor número de características *betweenness* (principalmente preposiciones o signos de puntuación) en la identificación de edad. Esto significa que características que describen nodos importantes en el discurso proporcionan más información en la identificación de sexo, mientras que características que describen los enlaces más comunes en el discurso proporcionan más información en la identificación de edad. En otras palabras, la selección de la posición en el discurso de palabras como

[8]Referirse al apartado 5.2.2 donde se explica el nombre de las características y al sitio web de Eagles para conocer el significado de las anotaciones morfosintácticas: http://www.cs.upc.edu/~nlp/tools/parole-sp.html

nombres, verbos o adjetivos, que principalmente proporcionan el significado de las frases, es el mejor tipo de características para la identificación de sexo, mientras que la selección de conectores como preposiciones, signos de puntuación o interjecciones, proporciona las características más discriminantes para la identificación de edad. Es importante resaltar el elevado número de características relacionadas con las emociones (SEL-sadness, SEL-disgust, SEL-anger) para la identificación del sexo, y la presencia de determinadas categorías gramaticales (Pron, Intj, Verb) para la identificación de edad.

Ranking	Sexo	Edad	Ranking	Sexo	Edad
1	punctuation-semicolon	words-length	11	BTW-NC00000	EIGEN-SPS00
2	EIGEN-VMP00SM	Pron	12	BTW-Z	BTW-NC00000
3	EIGEN-Z	BTW-SPS00	13	EIGEN-DA0MS0	punctuation-exclamation
4	EIGEN-NCCP000	BTW-NCMS000	14	BTW-Fz	emoticon-happy
5	Pron	Intj	15	BTW-NCCP000	BTW-Fh
6	words-length	EIGEN-Fh	16	EIGEN-AQ0MS0	punctuation-colon
7	EIGEN-NC00000	BTW-PP1CS000	17	SEL-disgust	punctuation
8	EIGEN-administration	EIGEN-Fpt	18	EIGEN-DP3CP0	BTW-Fpt
9	Intj	EIGEN-NC00000	19	EIGEN-DP3CS0	EIGEN-DA0FS0
10	SEL-sadness	EIGEN-NCMS000	20	SEL-anger	Verb

TABLA 5.4: Características más discriminantes para la identificación de sexo y edad según su mayor ganancia de información.

5.2.5.3 Análisis del error

En este apartado se intenta arrojar cierta luz sobre la relación entre la tasa de error y el número de palabras por autor. Para analizarlo en más detalle, se ha calculado la tasa de error teniendo en cuenta particiones de 50 palabras. Los resultados se muestran en la Figura 5.16, donde claramente se ve que la mayoría de errores se han producido con autores para los que se tiene menos de 50 palabras.

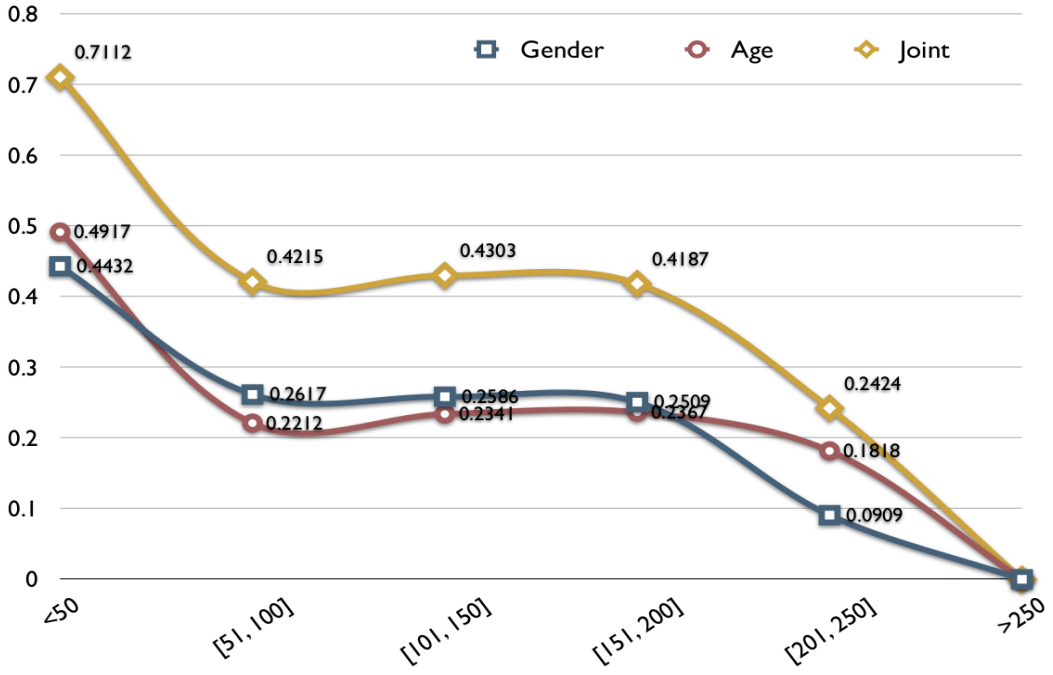

FIGURA 5.16: Tasa de error para particiones cada 50 palabras.

Analizando en detalle la tasa de error en función del número de palabras, en las Figuras 5.17, 5.18 y 5.19 se puede apreciar que tal ratio: *i*) parece ser mayor y más concentrado cuando hay menos de 50 palabras por autor; *ii*) tiende a ser cercano a 0 cuando hay más de 150 ó 200 palabras; y *iii*) mucho más disperso en casos intermedios.

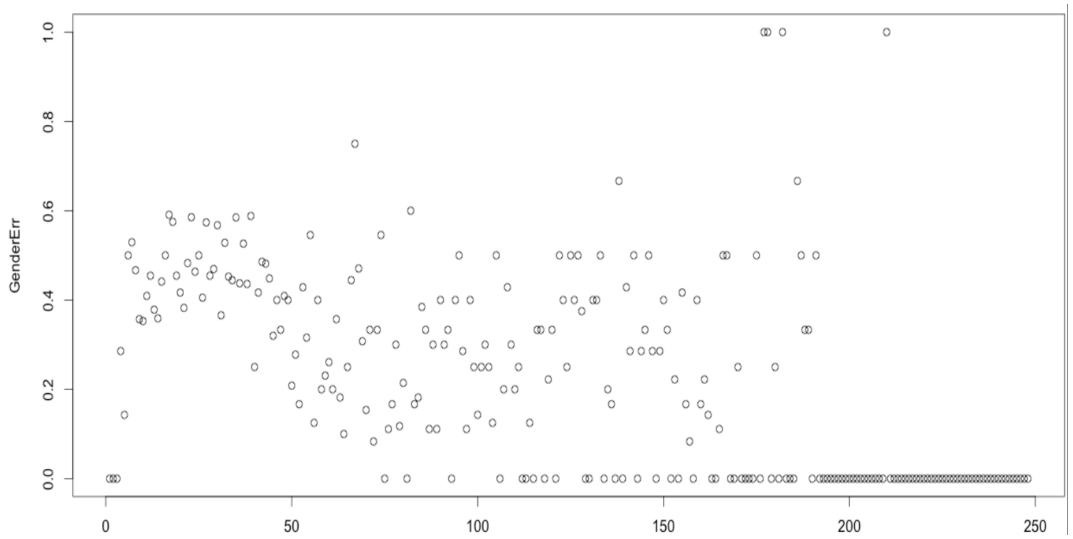

FIGURA 5.17: Tasa de error (eje Y) en función del número de palabras por autor para la identificación de sexo.

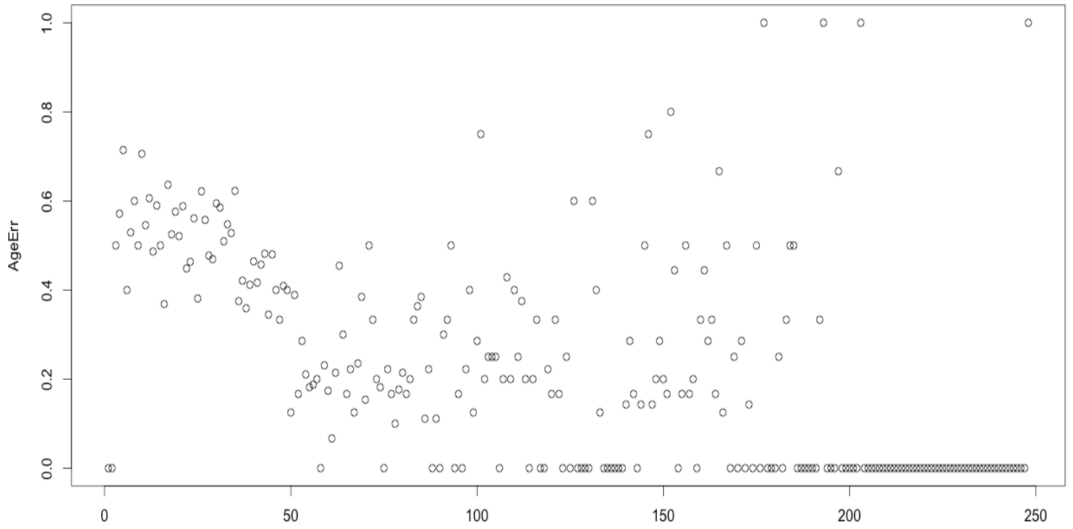

FIGURA 5.18: Tasa de error (eje Y) en función del número de palabras por autor para la identificación de edad.

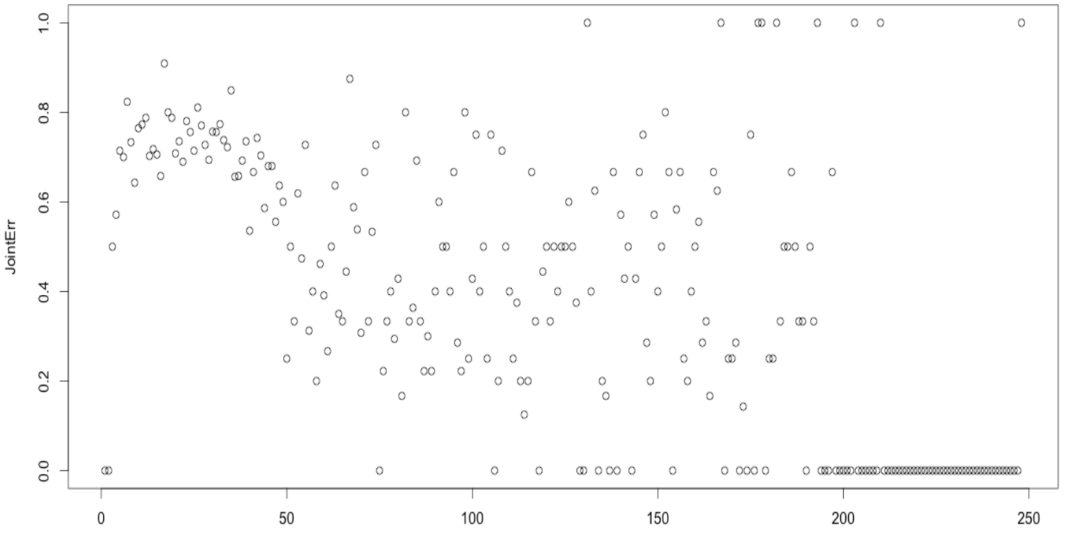

FIGURA 5.19: Tasa de error (eje Y) en función del número de palabras por autor para la identificación conjunta.

Atendiendo a los casos donde la predicción falla y que el autor tiene menos de 50 palabras, se pueden encontrar textos como los siguientes:

- *Bibliografia libros*
- *El amor es una mentira*
- *Hola a todos los cocolinos de corazon*

Como es esperable, en tales casos la predicción es cosa del azar. Por lo tanto, el rendimiento de EmoGraph en la tarea depende en cierta medida de un mínimo de palabras por autor. Por otro lado, tal y como se muestra en la Figura 5.16, la identificación de edad es mucho más dependiente del número mínimo de palabras que la de sexo. En general, debido al hecho de que hay tres categorías en la identificación de edad frente a dos en la identificación de sexo, y a que el número de errores es muy similar, se puede concluir que la identificación de sexo es una tarea incluso más difícil que la identificación de edad. Con respecto a la identificación conjunta, los resultados aún son más dependientes del número de palabras. Por ejemplo, cuando el número de palabras se incrementa, la tasa de error en la identificación conjunta decrece con respecto a las tareas individuales, al contrario que cuando hay pocas palabras donde la tasa de error para la identificación conjunta es aproximadamente el doble que para cada una de las tareas.

5.2.5.4 Análisis de la complejidad

La extracción de características se realiza en tres fases: *i*) la anotación morfosintáctica del texto con Freeling; *ii*) la construcción del grafo con las etiquetas morfosintácticas y su enriquecimiento con emociones, tópicos, sentimientos y tipos de verbos; *iii*) el cálculo de las diferentes medidas del grafo y sus nodos. El proceso se muestra en la Figura 5.20. A continuación se describe cada fase y se estima su complejidad.

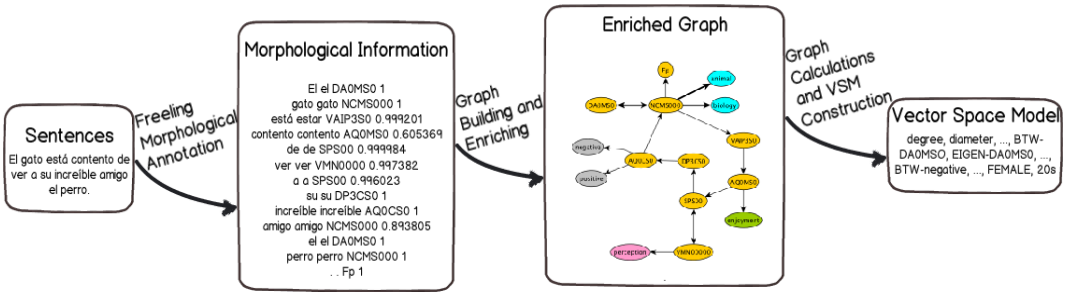

FIGURA 5.20: Extracción de características en tres fases.

Fase (i) Anotación morfosintáctica con Freeling

El anotador morfosintáctico de Freeling implementa un algoritmo Viterbi [182] estándar sobre un modelo oculto de Markov [183], por lo que su complejidad es $O(W \cdot S^2)$, donde W es el número de palabras de la frase y S es el número de etiquetas morfosintácticas (en este caso, la media de etiquetas posibles por palabra)[9].

Fase (ii) Construcción y enriquecimiento del grafo

Se construye el grafo del siguiente modo. Cada anotación morfosintáctica se constituye como un nodo y se crea una relación de precedencia entre pares de nodos. En el caso de que la palabra sea nombre común, adjetivo, adverbio o verbo, entonces se busca información relativa a su polaridad y emoción. Si la palabra es nombre común, adjetivo o verbo, se busca también su dominio. Además, si la palabra es verbo, entonces se busca su categoría semántica. La búsqueda se realiza en una tabla hash [184], por lo que su complejidad en el peor de los casos es $O(W)$, donde W es el número de palabras anotadas.

Fase (iii) Cálculos en el grafo

Una vez construido el grafo, la mayoría de los cálculos se pueden realizar en paralelo. Cálculos como el ratio entre nodos y arcos tienen complejidad constante $O(1)$, el grado medio y el grado medio ponderado, al igual que el coeficiente de clustering, tienen complejidad $O(N)$ donde N es el número de nodos. A continuación, se calcula la complejidad de las restantes características del grafo (*diámetro* y *modularidad*), y las correspondientes a las medidas de centralidad de los nodos (*betweenness* y *eigenvector*).

- *Diámetro*: Se calculan los caminos más cortos y se obtiene el diámetro como el más largo de todos ellos. Para calcular los caminos más cortos se utiliza el algoritmo de Bellman-Ford [185, 186] pues permite manejar grafos dirigidos. Su complejidad es $O(N \cdot E)$, donde N es el número de nodos y E el número de arcos.
- *Modularidad*: Se utiliza el método descrito en [175], consistente en inicializar cada nodo como una comunidad diferente y seguir los dos siguientes pasos: *i*) para cada nodo y sus vecinos, se calcula la ganancia en modularidad de acuerdo a una fórmula dada; *ii*) una vez se calcula el paso anterior para todos los nodos, cada comunidad se trata como un nodo y se repite de nuevo el paso anterior múltiples veces. Como el número de comunidades decrece a cada paso, la complejidad se puede estimar como $O(N \cdot logN)$, aunque los autores del algoritmo lo han llegado a estimar como lineal de media.

[9] http://nlp.lsi.upc.edu/freeling/index.php?Itemid=65&func=view&id=3998

- Centralidad *betweenness*: De acuerdo con la fórmula descrita en el apartado 5.2.2.2, para cada nodo n se debe calcular el ratio entre todos los caminos más cortos entre pares de nodos y todos aquellos que pasan a través de dicho nodo n. El cálculo de los caminos más cortos con el algoritmo de Bellman-Ford tiene una complejidad de $O(N \cdot E)$. La complejidad de la fórmula de *betweenness*, una vez calculados los caminos más cortos, es $O(N^2)$. En este punto se aprovechan los cálculos previamente realizados para los caminos más cortos en el cálculo del diámetro. Aunque se necesita almacenar todos los caminos más cortos, debido al máximo número posible de nodos en nuestro grafo (429), y debido a que se está modelando texto y por tanto el número posible de arcos es muy limitado, ya que las estructuras gramaticales imponen ciertas reglas, el coste espacial es despreciable.
- Centralidad *eigenvector*: Dada la matriz cuadrada A de adyacencia del grafo, un *eigenvector* es un vector sin ceros v que cuando se multiplica por A proporciona un *eigenvalue*, un escalar múltiplo de sí mismo llamado λ. La relación es $A \cdot v = \lambda \cdot v$. Se usa un algoritmo de iteración en potencia [187] para calcular el valor de λ con el mayor valor absoluto. Su coste es $O(N^2)$, donde N es el número de nodos.

Los dos cálculos con mayor complejidad son los relativos a la anotación morfosintáctica y las medidas de centralidad, con lo que se puede estimar la complejidad del proceso como: $O(W \cdot S^2) + [O(N \cdot E) + O(N^2)]$, donde W es el número de palabras de cada frase, S es el número de etiquetas morfosintácticas, N el número de nodos y E el número de arcos.

Aunque la complejidad global para la extracción de características es superior que la que se podría tener extrayendo n-gramas, EmoGraph permite una reducción considerable de la dimensionalidad a 1.011 características, aspecto de vital importancia en entornos big data.

5.2.5.5 Impacto de EmoGraph en Facebook

En el apartado 5.2 se ha mostrado la mejora de los resultados cuando se combina EmoGraph con el conjunto de características de estilo descritas. En este apartado se incide en el impacto de EmoGraph frente a las características basadas en estilo, para lo que se ha realizado una serie de experimentos con el corpus EmIroGeFB [146] descrito en el Capítulo 4.3. EmIroGeFB, utilizado también en [142], consiste en comentarios de Facebook etiquetados con sexo, emoción e ironía. Además se analiza el impacto de EmoGraph frente a otras variantes del grafo, concretamente:

- **Simple Graph**: Un grafo construido únicamente con las categorías gramaticales de las etiquetas Eagle (sólo el primer caracter de la etiqueta Eagle), esto es, verbo, nombre, adjetivo, etcétera.
- **Complete Graph**: Un grafo construido con la etiqueta Eagle completa (categoría gramatical + información morfológica), pero sin tópicos, verbos, polaridad ni emociones.
- **Semantic Graph**: Un grafo construido con toda las características descritas arriba (etiquetas Eagle, tópicos y clasificación de los verbos), pero sin las emociones.

Se combinan todas las características basadas en grafo con las basadas en estilo (S) (ver apartado 4.3.1): *i*) (F)recuencia; *ii*) Signos de (P)untuación; *iii*) (C)ategorías Gramaticales; *iv*) (E)moticonos; *v*) Spanish Emotion Lexicon (SEL). Los resultados se muestran en la Tabla 5.5.

Características	Accuracy
F + P	52,92
C	55,42
F + P + C	56,25
F + P + C + E + SEL	**59,09**
Simple Graph + S	50,83
Complete Graph + S	51,92
Semantic Graph + S	55,01
EmoGraph + S	**65,96**

TABLA 5.5: Resultados en términos de *accuracy* para la identificación de sexo; **c**orpus: EmIroGeFB (Facebook en español).

Se puede apreciar que EmoGraph obtiene mejoras significativas ($z_{0.05} = 3.4764 > 1.960$) en la tarea de identificación de sexo en comentarios de Facebook, lo que refuerza los resultados obtenidos previamente.

5.3 Robustez ante idiomas y social media

Con el objetivo de verificar la robustez del método EmoGraph para la identificación de edad y sexo ante diferentes idiomas y géneros, se propone utilizar el corpus de la tarea del PAN 2014, que incorpora textos de diferentes social media en inglés y español. Además, la utilización de este corpus permite comparar los resultados obtenidos con los de los participantes de la tarea del PAN [188].

5.3.1 Metodología

Así como se ha descrito en el Capítulo 2, el corpus PAN-AP-14 incorpora cuatro géneros diferentes: *i*) social media; *ii*) blogs; *iii*) Twitter; *iv*) y revisiones de hotel (TripAdvisor). Los respectivos subcorpus cubren el inglés y el español, a excepción de las revisiones de hotel, que se proporcionan sólo en inglés. Los autores de los textos están etiquetados con edad y sexo. Para la edad se consideran las siguientes clases: *i*) 18-24; *ii*) 25-34; *iii*) 35-49; *iv*) 50-64; *v*) y 65+ .

Se propone combinar EmoGraph con n-gramas de caracteres para paliar los efectos negativos de la identificación en textos excesivamente cortos como los descritos en el apartado 5.2.5.3. Se proponen n-gramas de caracteres por su sencillez y lo competitivo de sus resultados en la tarea en sus diferentes ediciones. Se seleccionan los 1.000 n-gramas más frecuentes, y se elige empíricamente 6 como valor para n a partir de iterar su valor desde 1 hasta 10 en validación cruzada con la partición de entrenamiento. Se prueban diferentes algoritmos de aprendizaje, seleccionando los mejores en cada caso: *i*) Simple logistic en identificación de sexo en Twitter en inglés; *ii*) máquinas de vectores soporte en blogs en español, en revisiones de hotel en inglés y en social media en inglés, tanto para identificación de sexo como edad, y en Twitter en español para identificación de edad; y *iii*) AdaBoost con Decision Stump para el resto de casos. Con esta configuración, se construye el modelo con la partición de entrenamiento y se evalúa con la partición de pruebas, manteniendo así la independencia que caracteriza a la tarea y que nos permite comparar nuestros resultados con los oficiales.

5.3.2 Resultados

Como se aprecia en la Figura 5.21, los resultados en español son superiores al inglés, excepto quizás en blogs. La causa se puede encontrar en la mayor variedad de rasgos morfosintácticos para español que se obtienen con la herramienta Freeling. El grupo Eagles[10] propuso una serie de recomendaciones para el etiquetado morfosintáctico de textos. Sobre la base de estas recomendaciones, Freeling obtiene 247 anotaciones diferentes para el español mientras que sólo obtiene 53 para el inglés. Por ejemplo, en la versión en español[11] la palabra *cursos* se etiqueta como NCMP000, donde NC significa nombre común, M significa masculino, P significa plural y 000 es un relleno hasta los 7 caracteres. Sin embargo, en la versión en inglés *courses* se anota como NNS, etiqueta mucho menos detallada.

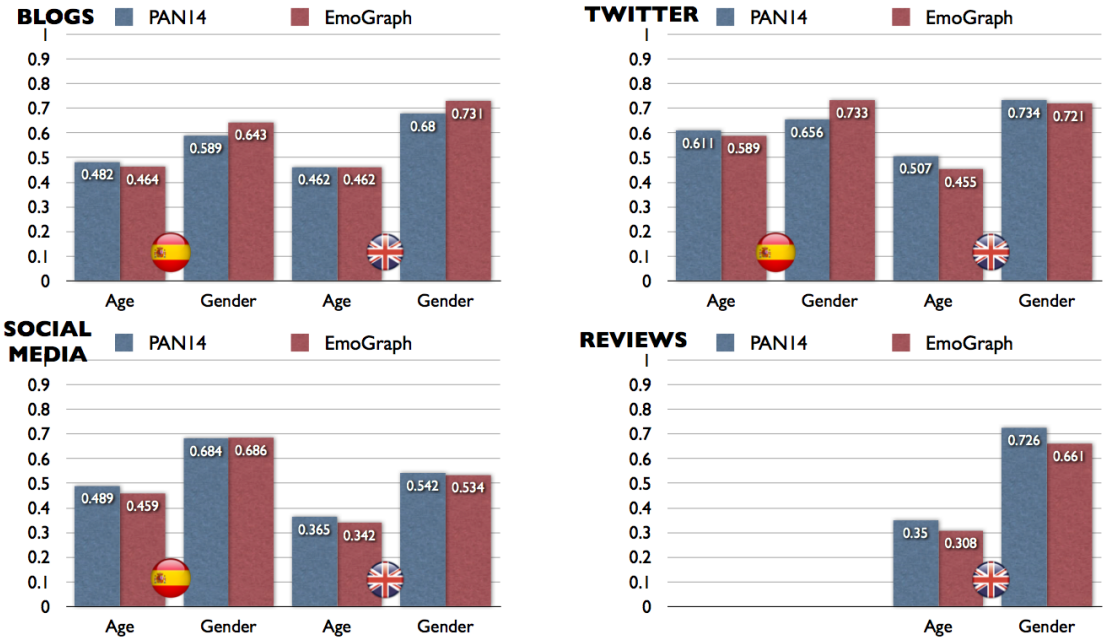

FIGURA 5.21: *Accuracy* de los mejores equipos del PAN 2014 vs. EmoGraph para los diferentes idiomas y géneros.

Al contrario de lo que obtuvimos con el corpus PAN-AP-13 para el español (ver apartado 5.2.4), los resultados para la identificación de sexo son mejores que para la de edad, hecho éste en línea con los resultados obtenidos por el resto de participantes de la tarea del 2014. Esto se puede deber al mayor número de clases para la edad; 5 clases continuas en 2014 frente a 3 clases en 2013. Los resultados para blogs y Twitter son mejores que para social media y revisiones de hotel, cuya causa puede ser que tanto blogs como Twitter han sido anotados manualmente, asegurando la veracidad de la información sobre sexo y edad de sus autores. Por el contrario, en social media y revisiones de hotel, por ejemplo en TripAdvisor, se asume como cierta la información que sus propios autores reportan, la cual no siempre es cierta. Además, en blogs hay suficiente texto por autor como para poder obtener un mejor perfil. De manera similar, aunque en Twitter cada tuit es muy corto (como mucho 140 caracteres), se dispone de cientos de tuits por autor, algo que permite disponer de suficiente texto como para mejorar el perfil de cada autor. Por otro lado, aunque la calidad de los textos de social media con respecto a la edición de la tarea del PAN del 2013, se ha mejorado mediante una revisión manual, se continúa con niveles de ruido superiores a blogs o Twitter. Respecto a las revisiones de hotel, donde EmoGraph obtiene los peores

[10] http://www.ilc.cnr.it/EAGLES96/intro.html
[11] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

resultados, cabe una mención especial. Además de textos cortos y de la posibilidad de información engañosa respecto a la edad y el sexo de los autores, las revisiones están enmarcadas en el dominio de los hoteles y a la expresión de dos tipos de emociones: *quejas* o *alabanzas*. Tal limitación en cuanto a dominios y emociones reduce la capacidad discriminativa de EmoGraph.

5.3.3 Contribución de EmoGraph

Para verificar la contribución de EmoGraph en la clasificación, en la Figura 5.22 comparamos los resultados obtenidos sólo con EmoGraph, sin combinar con n-gramas de caracteres, con los obtenidos con la combinación de ambos.

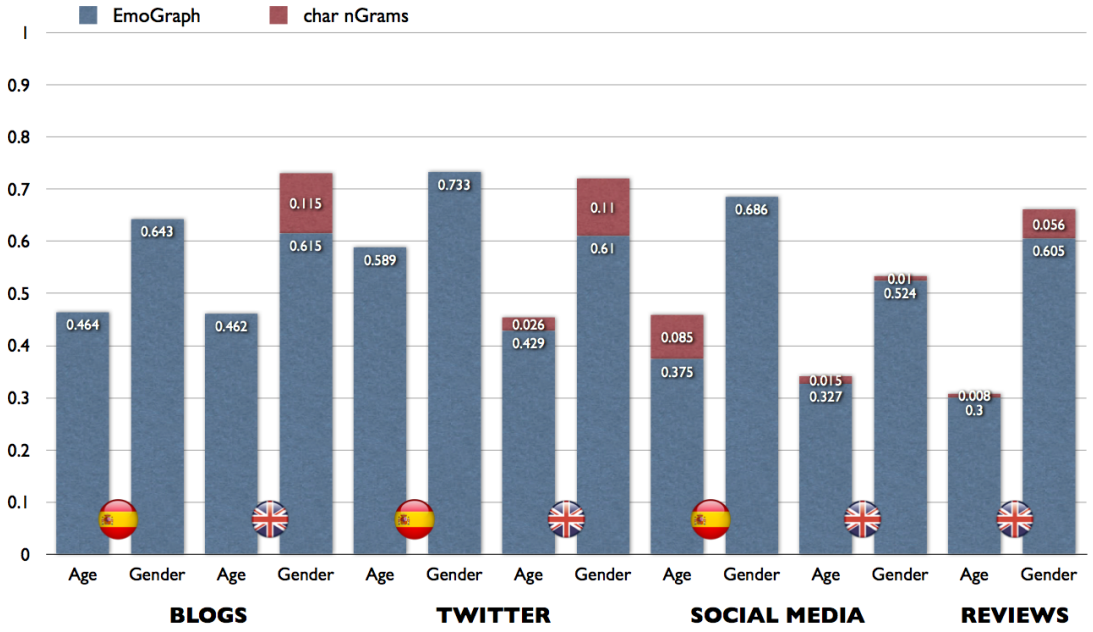

FIGURA 5.22: Contribución de EmoGraph a la *accuracy* por tarea, género e idioma.

Es destacable que en general EmoGraph en español obtiene los mejores resultados sin necesidad de ser combinado con n-gramas de caracteres. Esto se puede deber al menor número de etiquetas morfosintácticas en inglés, como se comentó previamente. Con respecto al desempeño en la identificación de sexo en blogs y Twitter en inglés, al contrario de los obtenidos en social media en español, los n-gramas de caracteres pueden ayudar a capturar información faltante.

5.3.4 Discusión

La aplicación de EmoGraph para la identificación de edad y sexo en diferentes géneros e idiomas se muestra robusta y competitiva con respecto a los mejores resultados obtenidos en la tarea. Se puede verificar que los mejores resultados se obtienen para el español, posiblemente debido a la más detallada anotación morfosintáctica para este idioma. Esto implica que, pese a lo robusto del método respecto al idioma de aplicación, se produce una fuerte dependencia con los recursos lingüísticos utilizados.

Los resultados para las revisiones de hotel nos aportan información esclarecedora respecto al rendimiento de EmoGraph cuando el medio está limitado en cuanto a temáticas y/o emociones expresadas, reduciendo

en estos casos su poder discriminativo. Esto está en línea con los mejores resultados obtenidos en Twitter y blogs, medios donde se dispone de mayor cantidad de textos por autor, lo que favorece al modelado de su estructura gramatical, y que además permiten una libre expresión de las más variadas emociones, enriqueciendo el grafo.

Hay que destacar que en casos como la identificación de edad en blogs en español, el resultado no es mucho mejor que el que se obtendría por un clasificador de clase mayoritaria. Esto se debe principalmente a la distribución sesgada del número de autores por edad en estos medios, algo que por otro lado es representativo del uso real de las diferentes redes sociales en los diferentes estadios de la vida[12].

5.4 Conclusiones

En este capítulo se ha propuesto una representación basada en grafos con el objetivo de capturar el modo en que la expresión de emociones se enmarca dentro de la estructura discursiva de un autor, y aprovechar dicha información para identificar rasgos demográficos como su edad o sexo. La motivación de utilizar grafos es doble, por un lado paliar la dificultad de realizar análisis lingüísticos complejos en textos donde la informalidad e incluso la incorrección son características inherentes, y por otro lado evitar la limitación de los modelos frecuentistas, basados en su mayoría en n-gramas, para capturar la importancia relativa de los diferentes elementos discursivos más allá de su frecuencia de uso y dentro de una ventana de aparición. Con ello, hemos definido una representación denominada EmoGraph.

EmoGraph aprovecha la potencia de los grafos para modelar estructuras complejas, como por ejemplo texto en lenguaje natural. Al comparar los resultados obtenidos por EmoGraph con los obtenidos por las características de estilo o los n-gramas de partes del discurso, hemos corroborado nuestra hipótesis de partida, ya que EmoGraph ha permitido obtener información sobre la importancia relativa de los elementos del discurso no sólo por su frecuencia de aparición, sino por su posición con y en relación a los demás elementos. Además, EmoGraph enriquece el modelo con la expresión de las emociones, algo inherentemente humano que determina qué decimos y cómo lo hacemos. Hemos utilizado la representación propuesta en el marco de evaluación del PAN, y los resultados obtenidos se muestran competitivos con los sistemas que obtuvieron las mejores prestaciones, así como extensibles y robustos frente a diferentes idiomas y medios sociales.

Del análisis de las características más discriminantes se desprenden conclusiones en línea con los estudios previos pero a un nivel superior de detalle. Por ejemplo, en la identificación de sexo predominan como más significativas las características de tipo *eigen*, lo que significa que determinadas categorías gramaticales como los *nombres*, *verbos* o *adjetivos*, junto con la *emotividad* que pueden expresar, ayudan a diferenciar entre sexos. Por otro lado, en la identificación de edad predominan como más discriminativas las características de tipo *betweenness*, lo que indica que el modo en que utilizamos los conectores discursivos proporciona información valiosa para determinar nuestra edad, conclusión fuertemente respaldada por el estado del arte donde se apunta a la mayor complejidad en la creación de estructuras lingüísticas como un rasgo distintivo de la misma. En ambos casos, la aparición de características emocionales entre las más discriminativas pone de manifiesto la importancia que tiene la expresión de *emociones* en la definición de los rasgos demográficos objeto de estudio.

Un análisis detallado del error muestra que el método EmoGraph requiere de un mínimo de palabras para comenzar a ser eficaz; se pasa de errores superiores al 50% para un número menor a 50 palabras a menos

[12] http://ow.ly/XWPIZ

de la mitad cuando el número de palabras supera esta cifra y siendo mínimo a partir de 150∼200 palabras. Los resultados obtenidos, junto con el análisis del error, muestran que la capacidad discriminativa de las características propuestas y modeladas mediante grafos, es superior a representaciones más simples siempre y cuando se tenga en consideración un mínimo de texto sobre el que aplicar el método, siendo en cualquier caso el acierto en caso contrario (menos de 50 palabras) fruto del azar.

Aunque el análisis de complejidad muestra puntos críticos en el proceso, como por ejemplo el análisis morfosintáctico de las oraciones, la reducción final de la dimensionalidad proporcionada por EmoGraph lo hacen adecuado para ser aplicado en entornos big data como los social media.

Capítulo 6

Identificación del lenguaje nativo y de las variedades del lenguaje

La tarea de identificar automáticamente el idioma en el que está escrito un texto ha sido ampliamente abordada desde los años 60 [189]. En un trabajo pionero en el área, Canvar y Trenkle [190] utilizan las frecuencias de aparición de n-gramas de caracteres para identificar noticias entre nueve idiomas diferentes, reportando un 99,8% de precisión. De manera similar, Dunning [191] combina n-gramas y modelos ocultos de Markov para discriminar entre textos escritos en inglés y en español, obteniendo una precisión superior al 99%. Realizando una selección de características basada en su propuesta de la medida *LD* (*Language-Domain*) [192] a un modelo de n-gramas, Lui y Baldwin [193] obtienen precisiones entre el 88-99% en la detección de hasta 97 idiomas. De manera similar, precisiones del 99,7% son obtenidas por Kruengkrai *et al.* [194] mediante el uso de string kernels en máquinas de vectores soporte en la clasificación de noticias en 17 idiomas diferentes. A partir de los resultados obtenidos, se empieza a pensar que la tarea se encuentra resuelta.

Pero con el auge de los medios sociales se plantean nuevos retos, como por ejemplo: *i*) la identificación a nivel de texto y no de documento completo, donde además el texto puede tener unas dimensiones considerablemente reducidas [195]; *ii*) la aparición de textos en idiomas desconocidos por el clasificador; *iii*) la existencia de textos escritos en un idioma diferente a la lengua nativa del que los escribe; o *iv*) la discriminación entre lenguas muy cercanas o variedades de una misma lengua.

Estos retos han dado lugar a nuevos campos de investigación que están cobrando auge en la actualidad, pero que tienen un matiz que va más allá de la clasificación de textos utilizada en la identificación de idiomas. En esta tarea, el propio texto tiene todos los indicadores necesarios para determinar su idioma; por ejemplo las palabras utilizadas o la frecuencia de las diferentes secuencias de caracteres. Sin embargo, en la identificación del idioma nativo de un autor que escribe en una segunda lengua o en la identificación de la variedad regional de un idioma, va a influir fuertemente la idiosincrasia del autor que los escribe, como por ejemplo su estilo discursivo, sus preferencias por el vocabulario o incluso las influencias socioculturales y educacionales a las que ha estado expuesto. De este modo, estas nuevas tareas se abordan desde la doble perspectiva de la clasificación de textos y el *author profiling*.

En los siguientes apartados realizamos un repaso a las investigaciones realizadas en cada una de estas tareas: *i*) identificación de idioma nativo; *ii*) discriminación entre idiomas similares; e *iii*) identificación de variedad del lenguaje.

6.1 Identificación del lenguaje nativo

La identificación del lenguaje nativo se formula como la tarea de identificar el idioma nativo (L1) de un autor en base a un ejemplo de este autor escribiendo en otro idioma (L2). Por ejemplo, un periodista español escribiendo noticias en árabe, un estudiante francés escribiendo ensayos en inglés, o un brasileño tuiteando en español.

En este apartado se describen: *i*) el conjunto de corpus disponibles para abordar la tarea; y *ii*) las aproximaciones más comunes utilizadas para abordarla.

6.1.1 Corpus

Principalmente a partir de escuelas de enseñanza y exámenes oficiales del inglés como segunda lengua, se han construido un conjunto de corpus que permiten abordar la tarea de identificación del idioma nativo de un autor. A continuación se proporciona una lista de ejemplos de corpus, donde la segunda lengua (L2) es el inglés:

- *International Corpus of Learner English* (ICLE) [196], consistente en 3.640 ensayos escritos por estudiantes universitarios de inglés de diferentes nacionalidades: búlgara, china, checa, holandesa, finlandesa, francesa, alemana, italiana, polaca, rusa, española y sueca.
- *International Corpus of Learner English versión 2* (ICLEv2) que aumenta el número de ensayos a 6.085, y el número de nacionalidades incorporando la japonesa, noruega, tsuana y turca.
- *First Certificate in English* (FCE) [197], que contiene 1.244 exámenes para obtener el título correspondiente de inglés, escritos por alumnos hablantes nativos de catalán, chino, holandés, francés, alemán, griego, italiano, japonés, coreano, polaco, portugués, ruso, español, sueco, tailandés y turco.
- *International Corpus Network of Asian Learners of English* (ICNALE) [198], que consiste en 5.600 ensayos de estudiantes universitarios de nacionalidades china, inglesa, filipina, hong kong, indonesia, japonesa, coreana, paquistaní, singapurea, taiwanesa y thai.
- *Test of English as a Foreign Language* (TOEF11) [199], con un conjunto de 12.100 ensayos escritos para exámenes de acceso a la universidad de alumnos con nacionalidades árabe, china, francesa, alemana, índica, italiana, japonesa, coreana, española, telugu y turca.
- *International Corpus of Crosslinguistic Interlanguage* (ICCI) [200], conteniendo 9.000 ensayos descriptivos y argumentativos escritos por jóvenes estudiantes de inglés de diferentes niveles y con nacionalidades como la austriaca, china, hong kong, israelita, japonesa, polaca, española y taiwanesa.
- *National University of Singapore Corpus of Learner English* (NUCLE) [201], compuesto por un conjunto de 1.375 ensayos de estudiantes de grado en el *Center for Language Communication*[1].
- *Corpus of English Essays by Asian University Student* CEEAUS [202], que lo compone un total de 1.008 ensayos restringidos a únicamente dos tópicos (*"It is important for college students to have a part-time job and Smoking should be completely banned at all restaurants in the country"*), y que han sido escritos por estudiantes de nacionalidades china, inglesa y japonesa.
- *BUiD Arab Learner Corpus* (BALC) [203], conjunto de 1.875 textos escritos por estudiantes universitarios de primer año o del último año de secundaria.

[1] http://www.nus.edu.sg/celc/

- Lang-8[2], servicio colaborativo en red donde los estudiantes de diferentes idiomas pueden escribir textos que son corregidos por nativos.

6.1.2 Aproximaciones

Como hemos introducido previamente, la identificación del lenguaje nativo se considera una tarea de *author profiling* ya que los textos escritos en una segunda lengua se ven influidos por elementos idiosincrásicos del autor que los escribe, como por ejemplo las influencias culturales recibidas. Koppel *et al.* [204] publican uno de los primeros trabajos en esta línea aprovechando la aparición de errores que son comunes a autores de diferentes nacionalidades cuando escriben en inglés. Para ello, los autores combinan una lista con 400 palabras de función y 200 n-gramas de caracteres, con 250 bigramas de partes del discurso considerados raros y 180 tipos de error. Concretamente: *i*) errores ortográficos, como letras repetidas (remmit vs. remit) o letras dobles que se omiten (comit vs. commit); *ii*) errores sintácticos, como falta de concordancia singular/plural, no coincidencia en tiempos verbales o confusión en el uso de that y which, o frases excesivamente largas; *iii*) uso de neologismos como fantabolous; o *iv*) bigramas de partes del discurso extrañas identificadas con ayuda del corpus Brown[3]. Mediante la aplicación de un multiclasificador lineal basado en máquinas de vectores soporte y en una configuración de validación cruzada en diez capas sobre el corpus ICLE, los autores reportan una *accuracy* de alrededor del 80%.

Una aproximación similar es seguida por Tofight *et al.* [205], que sobre un corpus de 600 textos recopilados de agencias de noticias y escritos por nativos en inglés, persa, turco y alemán, los autores extraen: *i*) hasta 64 características léxicas como n-gramas de caracteres, frecuencia de longitud de palabras o riqueza del vocabulario; *ii*) hasta 308 características sintácticas como signos de puntuación o palabras de función[4]; *iii*) hasta 13 características estructurales como la longitud de los párrafos o el uso de diferentes tipos de saludos; y *iv*) características específicas del contenido como los n-gramas con frecuencia superior a 10. Mediante el uso de máquinas de vectores soporte, los autores obtienen *accuracies* de 70-80% en validación cruzada a diez capas.

Dado el interés en el área, en 2013 Tetreault *et al.* [207] organizan la primera tarea internacional en identificación de idioma nativo en el taller BEA-8 en el NAACL-HT[5]. La tarea tiene una participación de 29 equipos, de los cuales un total de 24 envía artículos descriptivos de sus propuestas, y donde cada equipo ha podido presentar hasta en un total de 5 sistemas. La tarea se realiza con el corpus TOEF11, donde los participantes tienen que clasificar los textos entre los 11 idiomas proporcionados, mediante la participación en tres posibles modalidades de la tarea: *i*) *closed-training*, donde los participantes sólo pueden utilizar el corpus proporcionado para entrenar sus modelos; *ii*) *open-training 1*, donde los participantes pueden entrenar sus modelos con cualquier corpus, excepto el que ha sido proporcionado; y *iii*) *open-training 2*, donde los participantes pueden entrenar sus modelos con cualquier corpus, incluido el proporcionado. Las características más utilizadas han sido los n-gramas, de palabras, caracteres y partes del discurso, combinados con clasificadores basados en máquinas de vectores soporte, máxima entropía o métodos *ensemble*. Las *accuracies* reportadas son del 83,6% en el caso del *closed-training*, 56,5% en el caso de *open-training 1* y 83,5% en el caso de *open-training 2*.

[2] http://lang-8.com/

[3] http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/

[4] Las palabras de función son aquellas que tienen poco significado o un significado ambiguo, pero que en cambio sirven para expresar relaciones gramaticales con otras palabras dentro de una oración, o especificar la actitud o humor del hablante [206].

[5] https://sites.google.com/site/nlisharedtask2013/

Como ejemplo de sistema presentado en la tarea BEA-8, cabe mencionar el que presentan Brooke y Hirst [208]. Los autores utilizan los corpus ICLE, FCE, ICCI, ICANLE y Lang-8 para entrenar sus modelos, utilizando diferentes conjuntos de características como los n-gramas de palabras, y una mezcla de partes del discurso con n-gramas de palabras de función. Mediante el uso de máquinas de vectores soporte obtienen *accuracies* del 80,2% en la *closed-training*, 56,5% en la *open-training 1* y 81,6% en la *open-training 2*.

Es destacable el uso que hacen Bykh y Meurers [209] de gramáticas libres del contexto. Los autores construyen el corpus NT11 mediante la combinación y alineamiento de los corpus ICLE, FCE, BALC, ICNALE y TÜTEL-NLI, y lo utilizan para aprender un modelo de regresión logística que evalúan con el corpus TOEFL11, reportando una *accuracy* del 84,82%.

6.2 Identificación de lenguas similares y variedades del lenguaje

La discriminación entre lenguas similares, como bosnio, serbio o croata, o entre variedades de un mismo lenguaje, como el portugués brasileño frente al europeo, o el español peninsular frente al utilizado en los diferentes países latinoamericanos, suponen un reto añadido a la tarea clásica de identificación de idioma, aumentando su dificultad tanto por la mayor similitud léxica, sintáctica y semántica de los textos, como por la idiosincrasia histórica y cultural de los autores que los escriben.

El auge del interés en ambas tareas se pone de manifiesto en las tareas organizadas en los últimos años. Por ejemplo:

- *Workshop on Language Technology for Closely Related Languages and Language Variants* [210] organizada en 2014 en EMNLP[6];
- *Applying NLP Tools to Similar Languages, Varieties and Dialects - VarDial Workshop* [211] organizada en 2014 en COLING[7]; y
- *Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialect - LT4VarDial* [212] organizada en 2015 en RANLP[8].

En este apartado se describen: *i*) el conjunto de corpus disponibles para abordar ambas tareas; y *ii*) las aproximaciones más comunes utilizadas para abordarlas.

6.2.1 Corpus

En un área en la que, a diferencia de la presentada en identificación de idioma nativo, no se encuentran fácilmente conjuntos de datos sobre los que poder investigar. Uno de los productos de valor generados como consecuencia de la organización de las tareas descritas al inicio, son los recursos que se producen tales como los siguientes corpus compilados a partir de cabeceras de noticias:

[6] http://alt.qcri.org/LT4CloseLang/index.html
[7] http://corporavm.uni-koeln.de/vardial/sharedtask.html
[8] http://ttg.uni-saarland.de/lt4vardial2015/dsl.html

- DSLCC[9], que contiene 20.000 instancias por cada uno de los siguientes idiomas: bosnio, croata, serbio, indonesio, malayo, checo, eslovaco, portugués brasileño, portugués europeo, español peninsular, español argentino, inglés americano e inglés británico.
- DSLCC v2.0[10], que contiene 20.000 instancias por cada uno de los siguientes idiomas o variedades: búlgaro, macedonio, serbio, croata, bosnio, checo, eslovaco, español argentino, español peninsular, portugués brasileño, portugués europeo, malayo, indonesio y un grupo conteniendo textos escritos en otros idiomas, como por ejemplo el catalán.
- DSLCC v. 2.1, que contiene todas las instancias del DSLCC v.2.0 además de español mexicano y portugués macanés.

6.2.2 Aproximaciones

Respecto a la identificación de lenguas similares, son respresentativos los trabajos presentados a las diferentes ediciones de la tarea DSL. Concretamente, Goutte et al. [213] obtuvieron los mejores resultados en la edición del 2014 con una combinación de diferentes núcleos usando máquinas de vectores soporte [214] y n-gramas de caracteres y palabras. De manera similar, Malmasi y Dras [215] obtienen en 2015 uno de los resultados más competitivos a partir de la combinación de máquinas de vectores soporte y n-gramas de caracteres (con n entre 1 y 6), unigramas y bigramas de palabras.

En el caso de variedades de un mismo idioma como es el caso del portugués, Zampieri *et al.* [216] compilan un corpus de 1.000 documentos de prensa a partir del Folha do Säo Paulo[11] para portugués brasileño y del Diário de Notícias[12] para portugués europeo. Los autores utilizan una combinación de n-gramas de caracteres y de palabras tratando de capturar diferencias entre ambos a nivel: *i*) ortográfico, como signos gráficos (e.g. econômico vs. económico), o consonantes mudas (e.g. ator vs. actor); *ii*) sintáctico, como el uso de los pronombres (eu te amo vs. eu amo-te); *iii*) variaciones léxicas (e.g. multa vs. coima); o *iv*) los diferentes nombres propios utilizados (e.g. topónimos ó personas). Los autores utilizan distribuciones de probabilidad del lenguaje con funciones *log-likelihood* para la estimación de las probabilidades y reportan, mediante evaluación por partición al 50-50, *accuracies* del 99,6% con unigramas de palabras, 91,2% con bigramas de palabras y 99,8% con 4-gramas de caracteres.

Sadat *et al.* [217] investigan la identificación de 6 variedades regionales de árabe en social media. En concreto, variedades como: *i*) la egipcia; *ii*) la iraquí; *iii*) la del golfo, que incluye autores de Bahrein, Emiratos, Kuwait, Qatar, Oman y Arabia Saudí; *iv*) la magrebí, que incluye autores de Argelia, Túnez, Marruecos, Libia y Mauritania; *v*) la levantina, incluyendo Jordania, Líbano, Palestina y Siria; o *vi*) un conjunto de autores de Sudán. Los autores comparan el uso de n-gramas de caracteres con modelos del lenguaje de Markov, reportando una *accuracy* del 98% en una evaluación por partición al 50-50.

Con respecto al español, Maier y Gómez-Rodríguez [218] investigan la identificación de variedades de español de Argentina, Chile, Colombia, México y España en Twitter. Para ello, combinan cuatro tipos de características con un meta clasificador: *i*) perfiles frecuentes de n-gramas de caracteres; *ii*) modelos de lenguaje basados en n-gramas de caracteres; *iii*) método de compresión LZW; y *iv*) modelos del lenguaje basados en sílabas. Los autores reportan *accuracies* entre 60-70% en una evaluación cruzada.

[9] https://bitbucket.org/alvations/dslsharedtask2014
[10] https://github.com/Simdiva/DSL-Task
[11] http://www.folha.uol.com.br/
[12] http://www.dn.pt/

6.3 Conclusiones

En este capítulo hemos realizado una revisión a las investigaciones realizadas en tareas que matizan ciertos aspectos de la identificación del idioma, concretamente: *i*) la identificación del idioma nativo de un autor que lo hace en una segunda lengua (e.g. un español, italiano o frances escribiendo en inglés); *ii*) la discriminación entre idiomas similares (e.g. bosnio, serbio y croata; indonesio y malayo); y *iii*) la identificación de variedad dentro de una misma lengua (e.g. español de España, Argentina, México, etc.; portugués de Portugal vs. Brasil; inglés de UK vs. US).

Se puede apreciar, que a diferencia de para la tarea de identificación de idioma nativo, para la discriminación entre idiomas similares o variedades de una misma lengua apenas se dispone de corpus. La única excepción son los generados por la organización de la tarea DSL, aunque están centrados en cabeceras de noticias y por lo tanto no recopilan todo el potencial de los medios sociales.

En cuanto a las aproximaciones utilizadas, en su mayoría son combinaciones de diferentes algoritmos de aprendizaje con diferentes modelos basados en n-gramas. Destaca la variabilidad de los resultados obtenidos dependiendo del corpus y el modelo de evaluación, en su mayoría *10-fold cross-validation* o algún tipo de *split* entre entrenamiento y pruebas, aunque en ninguno de los casos se asegura que no existan textos de un mismo autor en más de una de las particiones.

Capítulo 7

Aproximaciones para la identificación de variedades del lenguaje

En capítulos anteriores nos hemos centrado en las tareas de identificación de edad y sexo, mostrando una descripción detallada del estado del arte y proponiendo EmoGraph como método basado en grafos para representar el estilo discursivo de los autores y cómo se enmarca en el mismo la expresión de las emociones, obteniendo resultados competitivos y robustos frente a diferentes géneros e idiomas. En este capítulo proponemos investigar la utilidad de EmoGraph en la identificación de variedad idiomática y la comparamos con los siguientes modelos de representación del contenido: *i*) dos modelos de representación distribuida sobre el modelo de Skip-gramas continuos: Skip-gramas y SenVec; y *ii*) un modelo de baja dimensionalidad que permita trabajar en entornos *big data*. Nuestro interés se centra en conocer qué clase de características capturan mejor las diferencias entre idiomas similares o variedades de un mismo idioma. Las características más usadas en el estado del arte, como veíamos en el Capítulo 6, son diferentes variaciones del modelo de n-gramas, bien de palabras, bien de caracteres. Ambas capturan diferencias en el contenido de los documentos, así como las de caracteres también capturan ciertas diferencias sintácticas. Nuestra intuición es que las variedades del lenguaje se diferencian más en el vocabulario usado que, por ejemplo, en la estructura gramatical de las frases. Por ejemplo, tal y como se ilustra en la Tabla 7.1, variedades del español como la de Argentina y México se parecen más entre sí en comparación con la de España. Pero tales similitudes/diferencias se producen a nivel léxico más que sintáctico.

ES-Argentina	Estaba haciendo **boludeces** con mi perro y **extravié** el **celular**.
ES-México	Estaba haciendo el **pendejo** con mi perro y **extravié** el **celular**.
ES-España	Estaba haciendo el **tonto** con mi perro y **perdí** el **móvil**.

Tabla 7.1: El mismo ejemplo en tres variedades del español (Argentina, México y España).

Algunas de las diferencias con respecto a trabajos previos para la identificación de variedades del lenguaje son las siguientes: *i*) nos centramos en textos de medios sociales porque estamos interesados en cómo

se expresan los usuarios en su día a día; *ii*) evaluamos nuestros modelos en particiones separadas de los datos para evitar el problema del sobreajuste; *iii*) algunos de los datasets utilizados en el estado del arte no son públicos: nosotros construimos y liberamos un dataset para uso por parte de la comunidad científica; *iv*) proponemos modelos de representación más elaborados que los típicamente usados basados en n-gramas. En el último apartado mostramos el rendimiento de las representaciones presentadas en la tarea *Discriminating Similar Languages* (DSL) en la cual participamos.

Es importante aquí matizar el punto *ii*) ya que es el que marca la diferencia entre considerar la identificación de variedad del lenguaje como un problema clásico de clasificación de textos a considerarlo un problema de *author profiling*. A diferencia de un problema de identificación de idioma, donde cada texto proporciona indicadores característicos del idioma en el que está escrito, como las palabras utilizadas o la frecuencia de las diferentes secuencias de caracteres, con cierta independencia del autor que los escriba, cuando se trata de textos en diferentes variedades de un mismo idioma, el estilo de sus autores, su forma de construir las oraciones o las preferencias por el vocabulario utilizado, pueden llegar a tener una influencia decisiva. Por este motivo, el problema clásico de clasificación de textos se ve extendido y complementado con la necesidad de identificar rasgos distintivos del el autor que los escribe. Es por ello que los métodos de evaluación que no tienen en cuenta este matiz, donde se pueden encontrar textos de un mismo autor en entrenamiento y pruebas, pueden incurrir en sobreajuste al estilo y preferencias de un autor, y aumentar artificialmente los resultados de la identificación.

7.1 HispaBlogs

Con la ayuda de expertos analistas ubicados en diferentes países de habla hispana, hemos generado un censo de blogeros conocidos pertenecientes a dichos países y que escriben en la variedad regional de los mismos. Concretamente, se han censado blogs de los siguientes cinco países: *i*) Argentina; *ii*) Chile; *iii*) México; *iv*) Perú; y *v*) España. A partir del censo, hemos seleccionado un total de 650 blogs por país, dividiendo el corpus en dos particiones, entrenamiento y pruebas, con un total de 450 y 200 blogs respectivamente. De cada blog se recupera hasta un total de 10 artículos, generando un corpus de 2.250 y 1.000 artículos respectivamente por país. Hemos llamado a este corpus HispaBlogs y lo hemos liberado a la comunidad.[1] En la Tabla 7.2 se muestran estadísticas detalladas.

Variedad	**Número de palabras**	
	Entrenamiento	**Pruebas**
AR - Argentina	1.408.103	590.583
CL - Chile	1.081.478	298.386
ES - España	1.376.478	620.778
MX - México	1.697.091	618.502
PE - Perú	1.602.195	373.262
TOTAL	7.164.935	2.501.511

Tabla 7.2: Número de palabras por variedad del lenguaje.

[1] https://github.com/autoritas/RD-Lab/tree/master/data/HispaBlogs

7.2 Representaciones distribuidas

El uso de modelos log-lineales ha sido propuesto por Mikolov [219] como un modo eficiente de generar representaciones distribuidas de palabras ya que permite reducir la complejidad de la capa oculta a la vez que mejorar la eficiencia. El modelo de bolsa de palabras continuo intenta maximizar la clasificación de una palabra mediante el uso de las palabras que la rodean, sin tener en cuenta el orden de la secuencia. Por el contrario, el modelo de Skip-gramas continuo utiliza el orden de las palabras para tomar como frecuencia de muestreo una proporción inversa a la distancia con la palabra en cuestión en el momento del entrenamiento.

Si se compara con propuestas tradicionales como los modelos del lenguaje *Feedforward Neural Net* [220] o *Recurent Neural Net* [221], se obtiene mejor rendimiento con un tiempo de entrenamiento considerablemente inferior en tareas de relaciones semánticas y sintáticas entre palabras. Los resultados experimentales además han demostrado que el modelo Skip-gramas ofrece mejor rendimiento de media, sobresaliendo especialmente en el nivel semántico. Por ello en este apartado se muestra su utilización para generar representaciones distribuidas aplicables a la tarea de identificación de variedad del lenguaje [222].

7.2.1 Modelo continuo de Skip-gramas

El modelo de Skip-gramas continuos [219, 223] parte de un algoritmo iterativo que intenta maximizar la clasificación del contexto alrededor de una palabra (ver Figura 7.1).

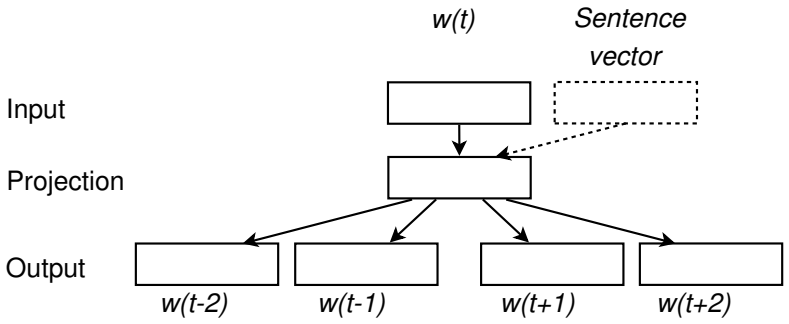

FIGURA 7.1: Arquitectura del modelo de Skip-gramas. El objetivo es predecir palabras dentro de cierto rango antes y después de la palabra actual. La parte punteada se usa sólo en lugar de $w(t)$ cuando se aprenden los vectores de sentencias.

Formalmente, dada una palabra $w(t)$ y las que la envuelven $w(t-c),\ w(t-c+1), ...,\ w(t+c)$ dentro de una ventana de tamaño $2c+1$, el objetivo del entrenamiento es maximizar la media de la log probabilidad.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \tag{7.1}$$

$$p(w_O|w_I) = \frac{\exp({v'_{w_O}}^T v_{w_I})}{\sum_{w=1}^{W}\exp({v'_w}^T v_{w_I})} \tag{7.2}$$

Aunque $p(w_{t+j}|w_t)$ se puede estimar usando la función softmax (Eq. 7.2) [224], su normalización depende del tamaño del vocabulario W y se hace impracticable para valores elevados de W. Por esta razón

se utilizan alternativas más eficientes en términos computacionales. Se propone el uso de Softmax jerárquico [225] para aproximar los resultados de la función softmax. Esta función se basa en un árbol binario con todas las palabras $w \in W$ como hojas, siendo cada nodo la probabilidad relativa de sus nodos hijos. Con este algoritmo sólo es necesario procesar $\log_2(W)$ palabras para cada estimación de probabilidad. Una alternativa se introduce en [223] mediante muestreo negativo. Esta función es una simplificación del *noise contrastive estimation* (NCE) [226, 227] que tiene relación con preservar la calidad del vector en el contexto de aprendizaje de Skip-gramas. La idea básica es la de utilizar regresión logística para distinguir la palabra objetivo W_O de una distribución de ruido $P_n(w)$, teniendo k ejemplos negativos para cada clase. Formalmente, las muestras negativas estiman $p(w_O|w_I)$ como sigue:

$$\log \sigma({v'_{w_O}}^T v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i} \sim P_n(w) \left[\log \sigma(-{v'_{w_i}}^T v_{w_I}) \right] \tag{7.3}$$

donde $\sigma(x) = 1/(1 + \exp(-x))$. Destacar que la complejidad computacional es lineal con el número de ejemplos negativos k. Los resultados experimentales en [223] muestran que esta función obtiene mejores resultados a nivel semántico que softmax jerárquica y NCE. Por lo tanto, en este trabajo hemos usado muestreo negativo en nuestros experimentos.

Aprendiendo vectores de frases El modelo de Skip-gramas lineales puede ser fácilmente adaptado para generar vectores representativos de frases (o documentos). Los vectores de frases (SenVec) [228] siguen la arquitectura de Skip-gramas para entrenar un vector especial sv que represente a la frase. Básicamente, antes de cada movimiento de la ventana de contexto, SenVec usa sv en lugar de $w(t)$ con el objetivo de maximizar la clasificación de las palabras que la envuelven (ver Figura 7.1).

Clasificación usando representaciones distribuidas Aunque SenVec se puede aplicar directamente como entrada de un clasificador, necesitamos combinar los vectores de palabras generados con el modelo de Skip-gramas para usarlos cuando se clasifican los documentos. El uso de *convolutional neural networks* con vectores de Skip-gramas de palabras como entrada fue propuesto en [229], obteniendo excelentes resultados para tareas de clasificación de frases. Sin embargo, debido a la complejidad computacional de estas redes, se emplea una solución más simple consistente en, dada una lista de vectores de palabras[2] $(w_1, w_2, ..., w_n)$ pertenecientes a un documento, se genera una representación vectorial v de su contenido mediante la estimación de la media de sus dimensiones: $v = n^{-1} \sum_{i=1}^{n} w_i$. En la evaluación nos referimos a este método como Skip-gramas.

7.2.2 Metodología

Estamos interesados en comparar el rendimiento de las representaciones distribuidas con respecto a tres representaciones alternativas que se describem a continuación: *i*) grafos *tf-idf*, donde se representa cada palabra como un nodo y su secuencia como arcos, ponderando los nodos por *tf-idf*; *ii*) Information Gain Word-Patterns (IG-WP); y *iii*) EmoGraph, que se describió en detalle en el Capítulo 5. Además se comparan los resultados con baselines utilizadas comunmente como n-gramas de palabras, de caracteres y ponderados por *tf-idf*. Para determinar el valor de n, hemos iterado desde 1 hasta 10, obteniendo los mejores resultados para valores de n igual a 1, 4 y 2 respectivamente. En todos ellos se seleccionan los 10.000 n-gramas más frecuentes.

[2] Se permite el uso de repeticiones de palabras.

Se ha medido la calidad de los modelos evaluando la *accuracy* de la clasificación en el conjunto de pruebas, asegurando la independencia con respecto al conjunto de entrenamiento. Se ha observado que durante la fase de prototipado, los vectores de sentencias y los vectores de medias de palabras ofrecían mejores resultados cuando eran estimados a partir de un número reducido de palabras. Aprovechando el dataset, se ha tratado cada post como una instancia independiente[3], y se ha determinado la variedad del lenguaje del blog en función de las probabilidades de clasificación de cada uno de sus posts: $class = argmax_{c \in C} \sum_{i=1}^{n} P(c|po_i)$, donde C es el número total de clases y $(po_1, ..., po_n)$ representa la lista de posts de un blog concreto.

Siguiendo el estado del arte [228], en la evaluación se utiliza un clasificador logístico[4] para ambas aproximaciones, SenVec y Skip-gramas. Se usan vectores de 300 dimensiones, ventanas de contexto de tamaño 10, y 20 palabras negativas para cada muestra. Se procesa el texto para convertir todos los caracteres a minúscula, tokenizar y eliminar palabras de un solo caracter, así como detección de frases mediante las herramientas word2vec[5].

Information Gain Word-Patterns Word-Patterns [230] es un método ascendente para la generación de patrones léxico-sintácticos capaces de representar el contenido de los documentos. Este método se basa en la hipótesis de construcción de patrones que establece que aquellos contextos que son relevantes para la definición de un grupo de palabras semánticamente relacionadas tienden a ser (parte de) construcciones léxico-sintácticas[6]. Este método consiste en las siguientes fases sucesivas. En primer lugar, el corpus de origen se anota morfológicamente con Freeling[7] [54, 55], obteniendo lemas y partes del discurso, y se anota con dependencias sintáticas usando Treeler[8]. En segundo lugar, se construye un modelo de espacio vectorial en forma de matriz en el que los contextos se modelan como relaciones de dependencia entre dos lemas. Para cada lema en cada fila de la matriz (lema origen), se define un contexto mediante una tupla de tres elementos: *i*) la dirección de la dependencia; *ii*) la etiqueta de la dependencia; y *iii*) el lema destino.

matriz_contexto = (dirección dependencia, etiqueta dependencia, lema contexto)

Un ejemplo de matriz sería:

$$contexto_1 = (<, subj, \textbf{\textit{robar}})$$
$$contexto_2 = (<, dobj, \textbf{\textit{peinar}})$$

donde *subj* hace referencia a sujeto y *dobj* a objeto directo, < indica la dirección de la dependencia, que en este caso significa que el lema en el contexto (*robar* y *peinar*) es el nodo padre del lema origen. En tercer lugar, se utiliza la herramienta CLUTO[9] para obtener los clusters de palabras semánticamente

[3] Aunque el método propuesto asegura que todos los contextos son tenidos en cuenta a la vez, una ventana deslizante podría ser usada como alternativa.

[4] Se han obtenido resultados similares pero con tiempos de entrenamiento superiores con otros clasificadores como las máquinas de vectores soporte.

[5] https://code.google.com/p/word2vec

[6] Una construcción es un patrón recurrente en el lenguaje.

[7] http://nlp.lsi.upc.edu/freeling

[8] Treeler es una librería abierta escrita en C++ de métodos de predicción enfocada en etiquetado y análisis. http://devel.cpl.upc.edu/treeler/svn/trunk

[9] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

Cluster: 25	
Lemas	barba, bigote, cabellera, cabello, cana, ceja, hebra, mecha, mechón, melena, pelo, peluca, pestaña, rizo, trenza

TABLA 7.3: Cluster 25 con los correspondientes lemas.

Cluster	**Contextos**	**Lemas**
643	<:* (2.2) <:subj (2.2) <:cd (1.3)	afeitar, ahuecar, alisar, cepillar, encrespar, enmarañar, erizar, mesar, ondear, peinar, rapar, rizar, sombrear, trenzar, tupir

TABLA 7.4: Cluster 643 relacionado con el 25

relacionadas que comparten los mismos contextos. A continuación, se establecen las relaciones entre clusters usando los contextos más descriptivos y discriminativos de cada cluster. Cada contexto consiste en una dirección de dependencia, una etiqueta de dependencia, un lema y una puntuación:

$$cluster\_contexto =$$
$$(dirección\ dependencia,\ etiqueta\ dependencia,\ lema\ contexto,\ puntuación)$$

Como resultado se obtiene un grafo de clusters relacionados, ejemplificado en la Tabla 7.3 y 7.4, donde el cluster número 25 se relaciona con el número 643 a partir de relaciones de sujeto y objeto directo. En la Tabla 7.3 se describen los lemas del cluster 25 y en la Tabla 7.4 se muestra uno de los clusters relacionados (el 643 en la primera columna) como resultado del proceso de enlazar clusters para el número 25. En la segunda columna se muestran los contextos que relacionan el cluster 25 con el 643, y en la tercera columna se mustran los lemas en el cluster relacionado. Todos los miembros (nombres) en el cluster 25 son buenos candidatos para ser sujetos y objetos directos de todos los miembros (verbos) del cluster 643.

Por último y para evitar relaciones espurias, se deriva un conjunto de los patrones léxico-sintácticos después de aplicar diferentes filtros. Los patrones léxico-sintácticos son tuplas involucrando dos lemas, relacionadas ambas por una dirección de dependencia y una etiqueta de dependencia:

$$pattern = (lemma_u,\ dep\text{-}dir,\ dep\text{-}lab,\ lemma_v)$$

Considerando los ejemplos del cluster 25 y 643, se generan todas las posibles combinaciones de cada lema del cluster 25 con cada lema del cluster 643. Ejemplos de patrones léxico-sintácticos derivados de la relación entre los clusters 25 y 643 son:

$$(\boldsymbol{bigote}_{c_{25}},\ <,\ dobj,\ \boldsymbol{afeitar}_{c_{643}}),$$
$$(\boldsymbol{peluca}_{c_{25}},\ <,\ dobj,\ \boldsymbol{peinar}_{c_{643}}),$$
$$(\boldsymbol{pelo}_{c_{25}},\ <,\ subj,\ \boldsymbol{encrespar}_{c_{643}})$$

En los experimentos llevados a cabo se seleccionan como características las 1.000 palabras obtenidas de los patrones que tienen la mayor ganancia de información. Los patrones se han generado utilizando el corpus Araknion [230].

7.2.3 Resultados

Se ha comparado el rendimiento de los modelos Skip-gramas y SenVec con bolsa de palabras (BOW), 4-gramas de caracteres. 2-gramas ponderados por *tf-idf*, grafos *tf-idf*, IG-WP y EmoGraph. Los resultados se muestran en la Tabla 7.5. Se resaltan en negrita los resultados con significación estadística resultantes de aplicar t-Student.

Method	**Accuracy**
Skip-gramas	72,20
SenVec	70,80
BOW	52,70
IG-WP	52,00
Char. 4-grams	51,50
EmoGraph	39,30
tf-idf 2-grams	32,20
Random baseline	20,00
tf-idf graphs	18,10

TABLA 7.5: Resultados en términos de precisión (*accuracy*) en la identificación de variedad del lenguaje.

Como se puede apreciar, el peor resultado se obtiene con grafos *tf-idf* (18,1%), incluso inferiores que la baseline aleatoria (20,0%). Teniendo en cuenta además los resultados obtenidos por 2-gramas ponderados por *tf-idf* (32,2%), pensamos que en esta tarea concreta los modelos basados en *tf-idf* no son capaces de capturar las diferencias entre variedades de lenguas. EmoGraph tampoco obtiene resultados competitivos (39,3%), aunque toma ventaja de información adicional como tópicos, verbos, sentimientos y emociones, mejorando ligeramente su rendimiento.

Las dos baselines, BOW (52,7%) y 4-gramas de caracteres (51,5%), son competitivas a pesar de su simplicidad. Los modelos de *n*-gramas de caracteres han probado ser capaces de extraer variaciones sintáticas (diferencias en vocabulario y en inflexiones verbales) entre diferentes variedades del lenguaje. El método IG-WP (52,0%) no parece obtener mejoras frente a BOW, pero demuestra la potencialidad de los patrones de palabras y tiene como ventaja la considerable reducción de la dimensionalidad, teniendo en cuenta además información lingüística.

Por último se aprecia que ambas representaciones distribuidas, Skip-gramas (72,2%) y SenVec (70,8%), obtienen resultados significativamente superiores al resto. El uso de la media de los vectores de palabras en el modelo Skip-gramas mejora ligeramente los resultados de SenVec, el cual infiere una única representación del documento y prueba ser una buena alternativa a aproximaciones más complejas.

7.2.4 Análisis del error

En la Tabla 7.6 se puede observar la dificultad en la clasificación de las diferentes variedades usando el modelo de Skip-gramas. La variedad de español de España es la más sencilla de detectar frente a la de Argentina, para la cual se obtienen los peores resultados. En general, las variedades latinoamericanas están más cercanas entre ellas y es más difícil diferenciarlas.

	Clasificado como				
Varidedad	AR	CL	ES	MX	PE
AR	58,5	8	8,5	11	14
CL	5	73,5	5	6	10,5
ES	3	3,5	85,5	4	4
MX	8	4,5	5	70	12,5
PE	6,5	6	4	10	73,5

TABLA 7.6: Matriz de confusión (en %) del modelo de Skip-gramas aplicado al conjunto de pruebas.

En la Figura 7.2 se muestra la capacidad de las representaciones distribuidas de modelar propiedades semánticas del lenguaje que sean extensibles más allá del corpus del que se obtienen. Para ello se ha comparado el rendimiento de todos los métodos presentados cuando se evalúan con validación cruzada en la misma partición de datos frente a su evaluación con una partición independiente. Se puede apreciar como en la mayoría de los casos, especialmente los basados en *tf-idf*, muestran cierto grado de sobreajuste observable en la mayor diferencia de rendimiento entre ambos métodos de evaluación.

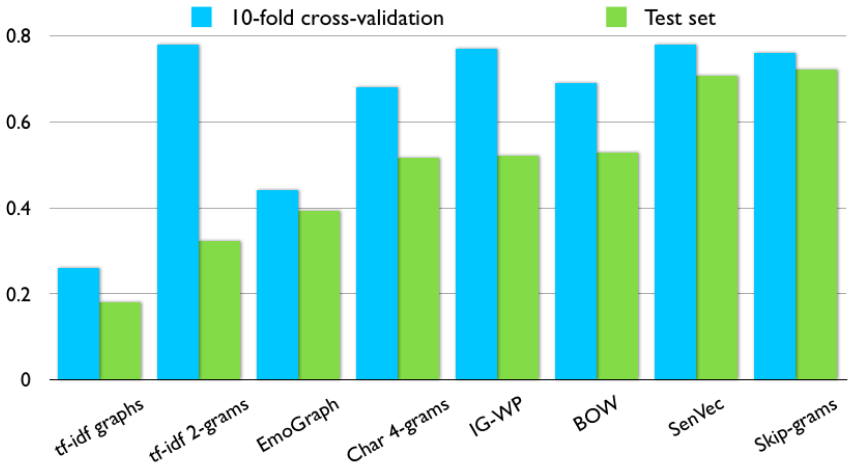

FIGURA 7.2: Análisis del sobreajuste de los modelos.

7.2.5 Discusión

Se observa que características basadas en contenido como BOW o IG-WP obtienen mejores resultados que las que toman información estructural como EmoGraph o grafos *tf-idf*. Esto puede corroborar la hipótesis de partida de que las diferencias entre variedades de un mismo lenguaje se deben en mayor medida al uso de las palabras que a la manera de estructurar el discurso, esto es, lo que se dice es más importante que como se dice. Es por ello que la representación EmoGraph, que captura la esencia de la expresión de emociones dentro del estilo discursivo del usuario, no ofrece resultados competitivos, demostrando así mismo que la expresión de las emociones no es un rasgo diferenciador entre variedades del lenguaje. En esta línea, las representaciones distribuidas analizadas, en concreto Skip-gramas y

SenVec, son capaces de diferenciar entre variedades de una misma lengua de manera significativamente superior a representaciones clásicas como n-gramas de palabras y/o caracteres.

7.3 Representación de baja dimensionalidad

Debido a que EmoGraph no parece adecuado para la identificación de variedades del lenguaje, poniendo de manifiesto que la expresión de las emociones no es un rasgo diferenciador entre variedades, y a partir de los resultados obtenidos previamente con representaciones distribuidas, donde se pone de manifiesto el poder discriminativo del léxico en esta tarea en concreto, en este apartado proponemos una aproximación que optimiza el procesamiento de textos en el entorno *big data* de los medios sociales. Cuando se trabaja en este tipo de medios, donde por ejemplo en Twitter se generan más de 500 millones de tuits por día[10], se hace imprescindible la aplicación de técnicas de reducción de la dimensionalidad y de representaciones que no resulten costosas de obtener. Por regla general los modelos basados en n-gramas son rápidos de obtener pero generan representaciones altamente dimensionadas. Para reducir su dimensionalidad se tiende a seleccionar los n-gramas más frecuentes, lo que puede dejar fuera características altamente discriminativas. En este apartado se propone un modelo de representación que permite utilizar el conjunto completo del vocabulario de un corpus a la par que reducir de manera drástica el número de características necesarias para modelar los textos. Esta representación de baja dimensionalidad (LDR de sus siglas en inglés) permite mantener la competitividad con el estado del arte en cuanto a precisión de la identificación de diferentes variedades del lenguaje, y su aplicación a entornos *big data* como los descritos [231].

7.3.1 Esquema de representación

El concepto clave es el de peso que representa la probabilidad de pertenencia de cada término a cada una de las variedades del lenguaje. La distribución de pesos para un documento debería estar más cercana a la de los documentos de su correspondiente variedad que al resto. Para obtener los pesos, se procede del siguiente modo. En primer lugar, se obtienen los términos de los documentos de la partición de entrenamiento D, lo que conforma el vocabulario disponible. Para cada término de este vocabulario se obtiene su peso *tf-idf*. Con los pesos calculados, se procede a construir la siguiente matriz:

$$\Delta = \begin{bmatrix} w_{11} & w_{12} & ... & w_{1m} & \delta(d_1) \\ w_{21} & w_{22} & ... & w_{2m} & \delta(d_2) \\ ... & ... & ... & ... & \\ w_{n1} & w_{n2} & ... & w_{nm} & \delta(d_n) \end{bmatrix} \tag{7.4}$$

Cada línea en la matriz Δ representa a un documento d y cada columna representa un término t del vocabulario. Cada celda contiene w_{ij} que es el peso *tf-idf* de cada término j en el documento i. Cada celda en la última columna $\delta(d_i)$ representa la clase c asignada al documento i, esto es, la variedad del lenguaje en la que está escrito el documento. Se calculan los pesos de cada término en cada variedad del lenguaje siguiendo la ecuación 7.5:

$$W(t,c) = \frac{\sum_{d \in D/c=\delta(d)} w_{dt}}{\sum_{d \in D} w_{dt}}, \forall d \in D, c \in C \tag{7.5}$$

[10] http://www.internetlivestats.com/twitter-statistics/

Donde d es cada uno de los documentos en el corpus representado por la matriz Δ, t representa cada término en el vocabulario T, c es una de las clases del conjunto de variedades del lenguaje C, $\delta(d)$ la clase c asignada al documento d, y w_{dt} el peso del término t en el documento d. El proceso es el siguiente: *i*) para cada término t representado como una columna y dada una variedad del lenguaje c, su peso *W(t,c)* es el ratio entre la suma de pesos para los documentos que pertenecen a dicha variedad del lenguaje y la suma de todos los pesos para ese término en todos los documentos, pertenezcan o no a la variedad en cuestión; *ii*) una vez calculados los pesos para cada término para cada variedad del lenguaje, el documento se representa en base a la ecuación 7.6:

$$d = \{F(c_1), F(c_2), ..., F(c_n)\} \sim \forall c \in C \tag{7.6}$$

donde cada $F(c_i)$ se compone del conjunto de características mostrado en la ecuación 7.7 y que se describe en la Tabla 7.7.

Es destacable que todo documento se representa por un total de características igual a 6 por el número de categorías, en nuestro caso tenemos 5 categorías (variedades del lenguaje) y por lo tanto un total de 30 características. Esto supone una reducción radical de la dimensionalidad pese a haber tenido en cuenta el vocabulario completo:

$$F(c_i) = \{avg, std, min, max, prob, prop\} \tag{7.7}$$

avg	La media de los pesos de un documento se calcula como la suma de todos los pesos *W(t,c)* de sus términos divivido por el número total de términos del vocabulario presentes en el documento.
std	La desviación estándar de los pesos se calcula como la raíz cuadrada del cuadrado de la suma de todos los pesos *W(t,c)* menos la media dividida por el número de elementos menos uno.
min	El peso mínimo es el menor peso *W(t,c)* encontrado para los términos de un documento.
max	El peso máximo es el mayor peso *W(t,c)* encontrado para los términos de un documento.
prob	La suma de los pesos *W(t,c)* de los términos del documento dividido entre el número total de términos del documento.
prop	Proporción entre el número de términos en el vocabulario que están presentes en el documento y el número total de términos del documento.

TABLA 7.7: Conjunto de características para cada categoría (variedad del lenguaje) usado en la ecuación 7.7.

Dado un nuevo documento, por ejemplo del conjunto de pruebas, se representaría utilizando los cálculos previos obtenidos para la partición de entrenamiento, de modo que se asegure la independencia de la representación. Para el documento en cuestión, se obtienen sus términos y se calculan los pesos para cada una de las variedades utilizando la ecuación 7.5. Con los pesos calculados, se obtiene cada una de las medidas $F(c_i)$ para cada variedad siguiendo la ecuación 7.7 y se representa el nuevo documento en función de la ecuación 7.6.

7.3.2 Metodología

Hemos experimentado con una clasificación a cinco clases, comparando el método propuesto con otros métodos usados en el estado del arte. Concretamente: *i*) bolsa de palabras con las 10.000 palabras más frecuentes; *ii*) 4-gramas de caracteres; *iii*) y los 10.000 2-gramas con mayor *tf-idf*. Además lo hemos comparado con los resultados obtenidos por EmoGraph y las representaciones Skip-gramas y SenVec presentadas en el apartado anterior.

Hemos probado diferentes algoritmos de aprendizaje, tal y como se muestra en el apartado 7.3.4.1, seleccionando el que mejores resultados ha proporcionado: clasificador multiclase con máquinas de vectores, seleccionando el método *exhaustive correction code* para convertir el problema multiclase en diferentes problemas de clasificación binaria.

7.3.3 Resultados

En la Tabla 7.8 se muestran los resultados en valores de *accuracy* para cada uno de los métodos descritos. Como se puede apreciar, la representación de baja dimensionalidad propuesta obtiene aproximadamente un 35% de mejora sobre el mejor de los métodos del estado del arte.

Método	**Accuracy**
Skip-gramas	72,20
LDR	**71,10**
SenVec	70,80
BOW	52,70
Char. 4-grams	51,50
EmoGraph	39,30
tf-idf 2-grams	32,20
Random baseline	20,00

TABLA 7.8: Resultados en *accuracy* para la identificación de variedad del lenguaje.

La bolsa de palabras obtiene ligeramente mejores resultados que los 4-gramas de caracteres, y ambos mejoran significativamente los 2-gramas ponderados por *tf-idf*. Esto último puede deberse a una distribución diferente de términos entre las particiones de entrenamiento y pruebas que pudieran hacer que la ponderación resultase ineficiente. La representación de baja dimensionalidad propuesta aprovecha las diferencias en el uso de las palabras como la bolsa de palabras o los n-gramas de caracteres, pero a diferencia de ellas, utiliza el vocabulario completo del corpus, dando mayor importancia a las palabras más discriminantes (que no tienen por qué ser las más frecuentes).

Se puede apreciar que los resultados de la representación LDR propuesta obtiene resultados similares a los métodos presentados en el apartado anterior basados en Skip-gramas y SenVec. Concretamente, aplicando un test t-Student se aprecia que las diferencias no son significativas ($z_{0,05} = 0,5457 < 1,960$ con respecto a Skip-gramas y $z_{0,05} = 0,7095 < 1,960$ con respecto a SenVec).

7.3.4 Discusión

Debido a que EmoGraph no obtiene resultados competitivos en la identificación de variedad del lenguaje, poniendo de manifiesto la hipótesis de partida de que las variedades léxicas de los textos permiten discriminar mejor entre variedades que las estructuras discursivas o la expresión de las emociones, hemos propuesto una representación de baja dimensionalidad que obtiene resultados superiores a los del estado del arte como las basadas en n-gramas de caracteres y palabras, y comparables a métodos en auge como las representaciones distribuidas. La principal contribución del método es la reducción drástica de la dimensionalidad, desde miles de características como en los casos anteriores a un total de 6 características por categoría. Esto permite aplicar la técnica en entornos *big data* como los medios sociales. Además, la representación tiene en cuenta el conjunto global del vocabulario de los documentos, lo que evita eliminar términos poco frecuentes pero que podrían contener un elevado poder discriminativo.

En este apartado analizamos en detalle diferentes aspectos del método propuesto, concretamente *i*) determinamos cuáles son los algoritmos de aprendizaje automático que mejor se ajustan a la tarea dada la representación propuesta; *ii*) comprobamos el efecto del preprocesamiento de los textos en el rendimiento de la representación; *iii*) analizamos el error con el objetivo de comprender mejor cuáles son las variedades del lenguaje más difíciles de identificar o cuáles son las que más se confunden entre sí; *iv*) obtenemos el ranking de las características más discriminativas en función de la ganancia de información en la tarea, experimentando con subconjuntos de las mismas para mejorar la comprensión sobre su aportación a la discriminación entre variedades; *v*) realizamos un análisis de costes desde la doble perspectiva de la complejidad computacional de la generación de la representación y de la reducción de la dimensionalidad conseguida; y por último *vi*) investigamos la aplicabilidad de la representación para la identificación de edad y sexo, poniéndola en perspectiva con EmoGraph y con las diferentes representaciones propuestas en la tarea del PAN.

7.3.4.1 Algoritmos de aprendizaje

Con el objetivo de determinar qué algoritmo se adapta mejor al problema de la identificación de variedad del lenguaje dada la representación LDR descrita, hemos experimentado con un conjunto de clasificadores. Como se puede apreciar en la Tabla 7.9, los mejores resultados se obtienen con el clasificador multiclase con máquinas de vectores soporte (*MulticlassClassifier*[11]). Se ha realizado un test t-Student para verificar la significación de los resultados del algoritmo con respecto a los dos que mejores rendimientos obtienen: SMO ($z_{0,05} = 0,880 < 1,960$) y LogitBoost ($z_{0,05} = 1,983 > 1,960$).

Algorithm	**Accuracy**	**Algorithm**	**Accuracy**	**Algorithm**	**Accuracy**
Multiclass Classifier	**71,10**	Rotation Forest	66,60	Multilayer Perceptron	62,50
SVM	69,30	Bagging	66,50	Simple Cart	61,90
LogitBoost	67,00	Random Forest	66,10	J48	59,30
Simple Logistic	66,80	Naive Bayes	64,10	BayesNet	52,20

Tabla 7.9: *Accuracy* de la representación LDR con diferentes algoritmos de aprendizaje.

[11] Se utilizan SVM con parámetros por defecto y *exhaustive correction code* para convertir el problema multiclase en respectivos problemas de clasificación binaria.

7.3.4.2 Efecto del preprocesamiento

La representación LDR utiliza todo el vocabulario para calcular sus pesos, pero en los medios sociales los textos pueden tener muchas faltas de ortografía e incluso palabras inventadas, lo que puede derivar en palabras que se usan unas pocas veces, incluso en ocasiones por un único autor, y que pueden introducir ruido en la representación. Con el objetivo de investigar el efecto de tales palabras en el rendimiento de LDR, hemos realizado un preprocesamiento de los textos para eliminar aquellas palabras que aparecen en menos de n autores, iterando el valor de n desde 1 hasta 100. En la Figura 7.3 se muestra el resultado de la identificación en términos de *accuracy* para diferentes valores de n. En la gráfica (a) se muestran los resultados para una iteración continua de n entre 1 y 10. En la gráfica (b) se muestran los resultados para diferentes valores de n entre 10 y 100, en una escala no continua.

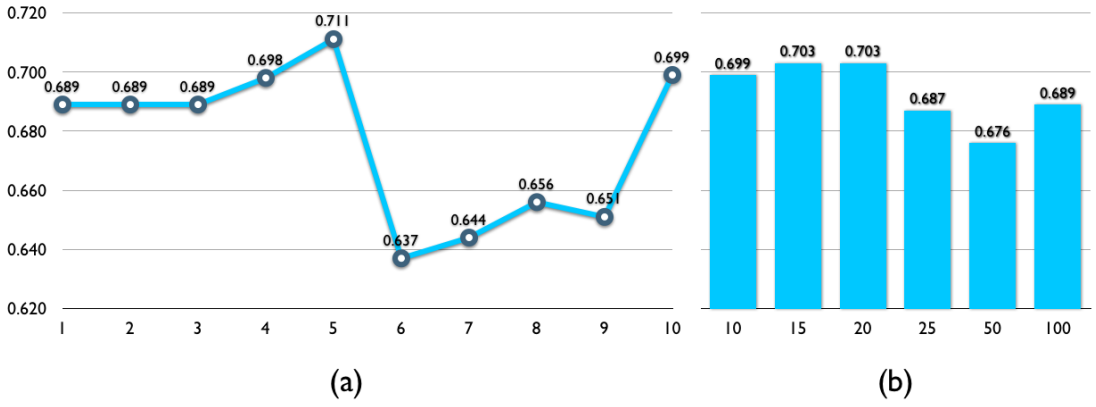

FIGURA 7.3: Accuracy obtenida después de eliminar palabras que aparecen n o menos veces en los autores.

Se puede apreciar que el mejor resultado se obtiene con n igual a 5, con una *accuracy* del 71,10%. Esto sugiere que aunque la representación propuesta tome ventaja de todo el vocabulario, es recomendable eliminar aquellas palabras que aparecen demasiado pocas veces entre los autores ya que pueden distorsionar los resultados. Cuando analizamos algunas de estas palabras, podemos ver casos como los mostrados en la Tabla 7.10.

# Apariciones = 1	# Apariciones = 2	# Apariciones = 3
aaaaaaaah	aaaaa	aaaa
aaaaaaaarrrgh	aaaayyy	aaaaaaaaae
aaaaaaggghhhhh	aaavyt	aaaaaaaacu
aaaah	aach	aantofagastina
aaaahhhh	aachen	abalanzaron
aaaauuuuuuuuuuuu	aackoc	wasp
aaaay	aang	ñam
aaayyyy	aarag	ñirripil

TABLA 7.10: Palabras que aparecen en pocos autores.

Relacionado con esto, es interesante analizar la evolución en el número de palabras a procesar por la representación en función del valor de n seleccionado. Como se puede ver en la Figura 7.4, se sigue una distribución potencial donde son muchas las palabras con frecuencias de aparición muy bajas. Esto puede ocasionar una distorsión elevada en los pesos calculados por LDR, al mismo tiempo que su eliminación permite reducir las necesidades de procesamiento y aumentar así la eficiencia en términos de coste

computacional. Por ejemplo, entre no eliminar ninguna palabra y eliminar las que aparecen menos de 5 veces en el corpus, hay una diferencia de 178.092 palabras.

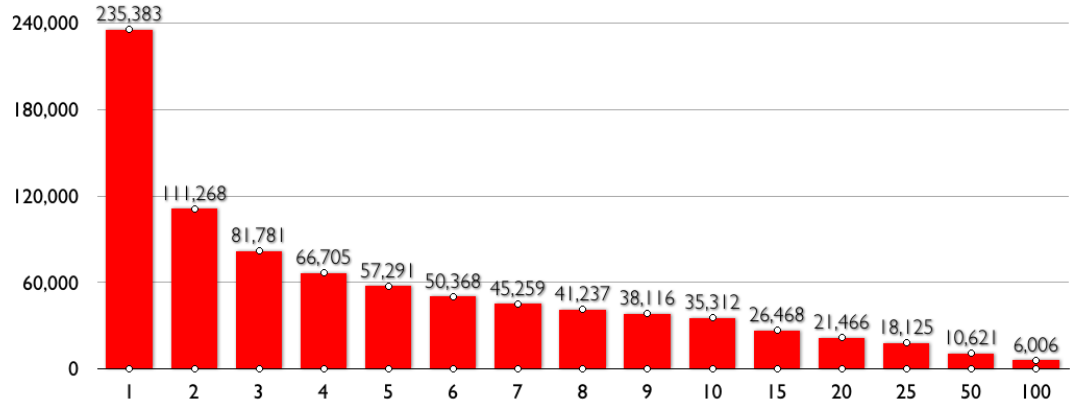

FIGURA 7.4: Evolución del número de palabras tras eliminar las de frecuencia de aparición igual o menor que n. (a) Escala continua. (b) Escala discreta.

7.3.4.3 Análisis del error

El análisis del error nos permite comprender cuales son las variedades más difíciles de discriminar y con cuales suele haber mayor confusión en caso de error. Tal y como se aprecia en la Tabla 7.11, la variedad de España es la más sencilla de discriminar. Por otro lado, la mayor confusión se produce entre Argentina, Chile y España. En la mayoría de los casos, la menor confusión se produce con Perú.

	Clasificado como				
Variedad	AR	CL	ES	MX	PE
AR	143	16	22	8	11
CL	17	151	11	11	10
ES	20	13	154	7	6
MX	20	18	18	131	13
PE	16	28	12	12	132

TABLA 7.11: Matriz de confusión en la clasificación a 5 clases.

En la gráfica (a) de la Figura 7.5 se muestran confrontados los valores de precisión y *recall* para cada una de las variedades. Además, se ha experimentado con clasificadores binarios, donde se clasifica cada una de las variedades del lenguaje frente a todas las demás, y se han obtenido los pares de valores de *accuracy* para cada experimento respecto a la correcta identificación de la variedad en cuestión, o la correcta identificación de los textos cuando no pertenecen a dicha variedad. Por ejemplo, si se clasifica la variedad de España frente al resto, se obtienen dos *accuracies*, cuando los ejemplos en variedad de España se clasifican como pertenecientes a la variedad de España, y cuando los ejemplos en otras variedades se clasifican como no pertenecientes a la variedad de España. Para cada una de las variedades

se representan estas dos *accuracies* en el gráfico (b) de la Figura 7.5, el primer caso en el eje X (clasificado como la variedad correspondiente) y el segundo en el eje Y (clasificado como la otra variedad).

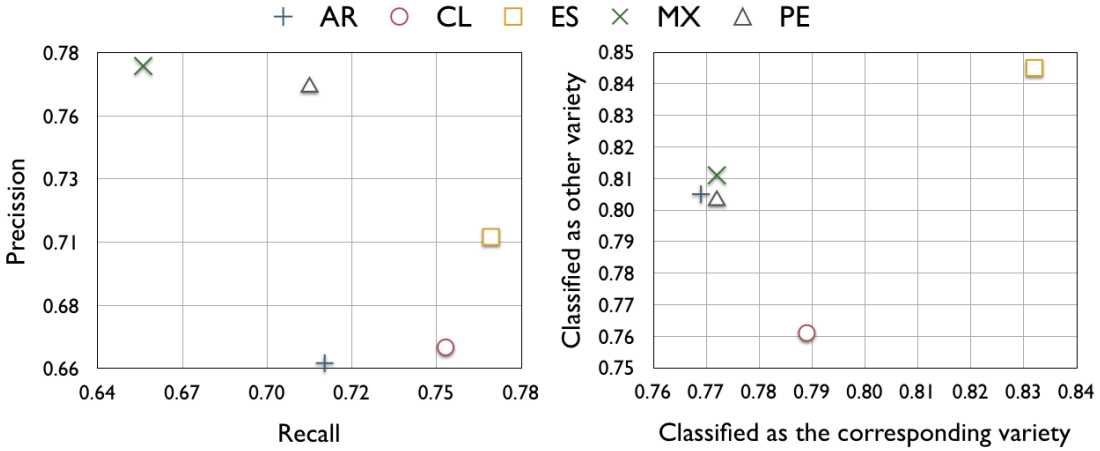

FIGURA 7.5: (a) Valores de *precision* y *recall* para la identificación de cada variedad; (b) Valores F1 para la identificación de cada variedad vs. las demás.

En la gráfica (*a*) se aprecia que las variedades de España y Chile son las que mayor *recall* obtienen, de manera que textos escritos en estas variedades tendrán menor probabilidad de ser confundidos con otras variedades. Sin embargo, las precisiones más elevadas se obtienen para las variedades de México y Perú, lo que implica que textos escritos en estas variedades serán más fácilmente reconocibles que en otras. Por otro lado, la gráfica (*b*) reafirma que la variedad de España es la más sencilla de discriminar, tanto para indicar la pertenencia como la no pertenencia de un texto a dicha variedad. Chile por el contrario es más difícil de discriminar en negativo, es decir, cuando un texto no pertenece a dicha variedad pero se confunde con ella. Quizás a ello se deba que la mayoría de los errores de clasificación apunten a esta variedad. Las de Argentina, México y Perú son variedades más fáciles de discriminar en negativo que en positivo, por lo que estas variedades tienen menor número de aciertos a favor de una menor confusión hacia otras variedades.

7.3.4.4 Características más discriminantes

La identificación de las características más discriminantes nos permite una mejor comprensión de la aportación de las diferentes medidas obtenidas para construir la representación LDR. Para ello, hemos obtenido la ganancia de información y ordenado las características en función la misma. En la Tabla 7.12 se puede apreciar que las características que mayor ganancia de información ofrecen son las de media, máximo y mínimo, así como desviación estándar, siendo las relativas a probabilidad y proporción las que menos ganancia de información tienen.

Con respecto a idiomas es curioso ver que las relativas a Chile aparecen a partir de la segunda columna, con ganancias de información relativamente inferiores a las del resto de variedades. Quizás sea causa de la menor precisión obtenida para esta variedad, aunque no justificaría el valor similar para la variedad de Argentina, por lo que es una simple conjetura.

Teniendo en cuenta lo anterior, hemos realizado una serie de experimentos con combinaciones de subconjuntos de características para mostrar su poder discriminativo real en la tarea. En la Figura 7.6 se puede apreciar, en línea con lo obtenido por la ganancia de información, que las características basadas en medias obtienen por sí solas valores de *accuracy* elevados (67,00%). En cambio, a pesar de que

Atributo	Ganancia	Atributo	Ganancia	Atributo	Ganancia
PE-avg	0,680 ± 0,006	ES-std	0,497 ± 0,008	PE-prob	0,152 ± 0,005
AR-avg	0,675 ± 0,005	CL-max	0,496 ± 0,005	MX-prob	0,151 ± 0,005
MX-max	0,601 ± 0,005	CL-std	0,495 ± 0,007	ES-prob	0,130 ± 0,011
PE-max	0,600 ± 0,009	MX-std	0,493 ± 0,007	AR-prob	0,127 ± 0,006
ES-min	0,595 ± 0,033	CL-min	0,486 ± 0,013	AR-prop	0,116 ± 0,005
ES-avg	0,584 ± 0,004	AR-std	0,485 ± 0,005	MX-prop	0,113 ± 0,006
MX-avg	0,577 ± 0,008	PE-std	0,483 ± 0,012	PE-prop	0,112 ± 0,005
ES-max	0,564 ± 0,007	AR-min	0,463 ± 0,012	ES-prop	0,110 ± 0,007
AR-max	0,550 ± 0,007	CL-avg	0,455 ± 0,008	CL-prop	0,101 ± 0,005
MX-min	0,513 ± 0,027	PE-min	0,369 ± 0,019	CL-prob	0,087 ± 0,010

TABLA 7.12: Ganancia de información de las características propuestas para la identificación de variedad del lenguaje.

no son de las características que mayor ganancia de información aportan, las basadas en desviaciones estándar obtienen los resultados más elevados obtenidos por características individuales (69,2%), así como cuya combinación con medias obtiene uno de los mejores resultados (70,8%). Vemos también que las características basadas en mínimos y máximos no obtienen por sí solas buenos resultados (48,3% y 54,7%), aunque su combinación produce un incremento significativo (61,1%). Es la combinación de las anteriores (71,00%) y la de todas juntas (71,10%) las que mejores resultados obtienen, con lo que queda demostrada la utilidad y aportación de cada una de ellas, y especialmente de su combinación, para obtener los mejores resultados.

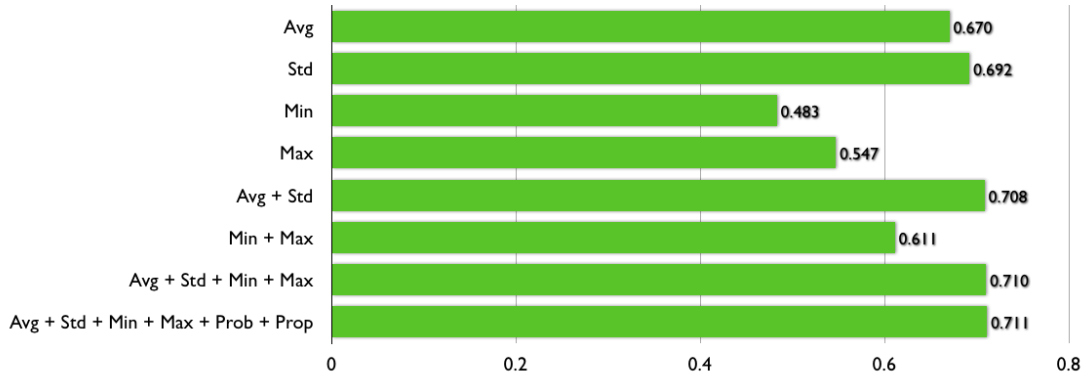

FIGURA 7.6: *Accuracy* con diferente combinación de características de LDR.

7.3.4.5 Análisis de costes

En este apartado se presenta un análisis de los costes computacionales de aplicar la representación de baja dimensionalidad propuesta. Hemos realizado el análisis desde dos perspectivas: *i*) el orden de complejidad del algoritmo para obtener los valores de la representación de un nuevo documento, esto es, cuando los pesos asociados a cada término del vocabulario de entrenamiento están calculados y se desean utilizar para representar un nuevo documento del conjunto de pruebas; y *ii*) el número de características necesarias para la representación de un documento, de manera comparativa con las representaciones vistas a lo largo del capítulo.

```
Sean c cada variedad del lenguaje del conjunto de variedades C,
     t cada termino del vocabulario T,
     s cada termino del documento D,
     W(t,c) el peso del termino t en la variedad c.
Sean las variables m, n, sum, val[], avg, std, min, max, prob, prop.

Inicio
    n <- numero de terminos en D
    para cada variedad c del conjunto de variedades C
        para cada termino s en el documento D
                si s contenido en T entonces
                sum <- sum + W(s, c)
                si min < W(s,c) entonces min <- W(s,c) fsi
                si max > W(s,c) entonces max <- W(s,c) fsi
                m <- m + 1
                val[m] <- W(s,c)
            fsi
        fpara
        avg <- sum / m; prob <- sum / n; prop <- m / n
        para cada valor v en val[m]
            std <- std + cuadrado(v - avg)
        fpara
        std <- raiz(std / (m - 1))
    fpara
Fin
```

LISTADO 7.1: Algoritmo para el cálculo de las características de LDR.

Para el cálculo del orden de complejidad, el Listado 7.1 muestra el algoritmo de cálculo de las características, del que se puede inferir que su coste va a depender de manera lineal del número de variedades del lenguaje que se desee discriminar ($c \in C$), y del número de términos que contenga el documento a representar ($s \in D$). Concretamente:

- El cálculo de los valores de media, mínimo, máximo, probabilidad y proporción es de orden $O(l \cdot n)$, donde l es el número de variedades del lenguaje $c \in C$ y n es el número de términos s del documento D.
- El cálculo de la desviación estándar es de orden $O(l \cdot m)$, donde l es el número de variedades del lenguaje $c \in C$ y m es el número de términos s del documento D que concuerdan con algún término t del vocabulario T.

Con lo anterior se tiene un orden global $O(l \cdot n) + O(l \cdot m)$, que por la regla de la suma $O(l \cdot n) + O(l \cdot m)$=$O(max(l \cdot n, l \cdot m)$ equivale a una complejidad del orden $O(l \cdot n)$, ya que el número de términos en el vocabulario será siempre mayor o igual que el número de términos del documento que coinciden con términos del vocabulario ($n >= m$). Además, puesto que el número de variedades será significativamente inferior al número de términos del documento ($l << n$), la complejidad global se puede determinar como lineal respecto al número de términos del documento a representar $O(n)$.

Con respecto al número de características necesarias para la representación de los documentos, en la Tabla 7.13 se aprecia la reducción significativa que proporciona el método propuesto. Además, frente a representaciones vectoriales basadas en los elementos más frecuentes, como los basados en n-gramas, la representación propuesta tiene en consideración todas las palabras del documento de entrada, evitando

así perder información proporcionada por términos poco frecuentes pero que pueden ser significativos para una variedad de lenguaje determinada (e.g. la significación del uso de la palabra *carro* frente a *coche* en un dominio diferente al automovilístico).

Representación	# Características
LDR	30
Skip-gramas	300
SenVec	300
EmoGraph	1.011
Char 4-gramas	10.000
BOW	10.000
tf-idf	10.000

TABLA 7.13: Número de características por representación.

7.3.4.6 Aplicabilidad a identificación de sexo y edad

El objetivo del apartado era investigar la viabilidad de EmoGraph para la tarea de identificación de variedad del lenguaje. Los resultados experimentales mostraron que las variaciones del lenguaje no se basan tanto en la expresión de emociones dentro del estilo discursivo sino más en la variedad léxica utilizada, para lo que representaciones como Skip-gramas, SenVec o la propuesta LDR obtienen resultados mucho más competitivos. En este punto hemos querido verificar la aplicabilidad de LDR a la tarea de edad y sexo, donde EmoGraph demostraba su superioridad. En la Tabla 7.14 se muestran los resultados. Se marca con * aquellos que son significativamente mejores al 95% de confianza en un test t-Student.

			Edad		**Sexo**	
Dataset	Género	Idioma	EmoGraph	LDR	EmoGraph	LDR
PAN-AP-2013	Social Media	Español	$66,24^*$	62,70	$63,65^*$	60,75
PAN-AP-2014	Social Media	Español	$45,9^*$	38,16	$68,6^*$	56,89
PAN-AP-2014	Social Media	Inglés	$34,2^*$	31,63	53,4	51,42
PAN-AP-2014	Blogs	Español	46,4	46,43	64,3	50,00
PAN-AP-2014	Blogs	Inglés	46,2	38,46	71,3	67,95
PAN-AP-2014	Twitter	Español	58,9	56,67	73,3	63,33
PAN-AP-2014	Twitter	Inglés	45,5	52,60	72,1	67,53
PAN-AP-2014	Revisiones	Inglés	30,8	32,28	66,1	67,11

TABLA 7.14: Resultados en social media en términos de precisión (*accuracy*).

Como se puede apreciar, en la mayoría de los casos EmoGraph obtiene resultados superiores a LDR, aunque destacan dos aspectos:

- LDR obtiene mejores resultados tanto para identificación de edad como de sexo en el dataset PAN-AP-2014, en el subconjunto de revisiones de hoteles. En el apartado 5.3.2 se argumenta que la limitación de este medio en cuanto a temáticas y/o emociones expresadas (e.g. quejas o alabanzas) con respecto a los servicios y producto ofrecidos por los hoteles, reduce el poder discriminativo de EmoGraph.

- LDR obtiene resultados más cercanos a EmoGraph en la tarea de identificación de edad, lo que sugiere que la variedad léxica utilizada es una características más ligada a la edad del autor que a su sexo.

Es importante mencionar que aunque LDR obtiene resultados inferiores a EmoGraph, sigue obteniendo resultados competitivos con respecto a las representaciones del estado del arte, tal y como se puede apreciar en la Tabla 7.15 donde se muestra la posición en el ranking en la que hubiese quedado LDR en las diversas tareas del PAN (se marca con * resultados obtenidos que coinciden con otros participantes).

Dataset	Género	Idioma	Edad	Sexo	# Participantes
PAN-AP-2013	Social Media	Español	3	6	21
PAN-AP-2014	Social Media	Español	5	9	9
PAN-AP-2014	Social Media	Inglés	6	9	10
PAN-AP-2014	Blogs	Español	3*	5*	9
PAN-AP-2014	Blogs	Inglés	3*	1*	10
PAN-AP-2014	Twitter	Español	2	2*	8
PAN-AP-2014	Twitter	Inglés	1	3	9
PAN-AP-2014	Revisiones	Inglés	5	5	10

Tabla 7.15: Posición en el ranking del PAN de los resultados obtenidos por LDR.

Se puede apreciar que el método obtiene resultados especialmente competitivos en Twitter y Blogs, obteniendo primeras posiciones en dos de los casos, y posiciones por debajo de la mitad del ranking en la mayoría de ellos.

7.4 Identificación de idiomas similares

La participación en la tarea *Discriminating between Similar Languages* (DSL) 2015[12] se motiva con un doble objetivo: *i*) analizar la robustez de las propuestas presentadas cuando se aplican no sólo a la clasificación de variedades de un mismo idioma sino de idiomas similares; y *ii*) determinar su validez cuando se trabaja con textos considerablemente más pequeños, como en este caso frases, a diferencia de los textos utilizados en las investigaciones previas. En los siguientes apartados se describe la tarea y el corpus proporcionado, y posteriormente se presentan y discuten los resultados obtenidos con las tres representaciones descritas: *i*) representaciones distribuidas [232]: Skip-gramas y SenVec; y *ii*) representación de baja dimensionalidad [233]: LDR.

7.4.1 Descripción de la tarea

Los organizadores de la tarea proporcionan el corpus DSLCC v.2.0 [234] consistente en frases extraídas de noticias en diferentes idiomas y dialectos. En la Tabla 7.16 se resumen los diferentes idiomas y variedades presentes en el corpus. El grupo codificado como *xx* se construye como compendio de diferentes idiomas no incluidos en los anteriores (por ejemplo el catalán).

[12] http://ttg.uni-saarland.de/lt4vardial2015/dsl.html

Grupo	Idioma	Código
Eslavo sudoriental	Búlgaro	bg
	Macedonio	mk
Eslavo suroccidental	Bosnio	bs
	Croata	hr
	Serbio	sr
Eslavo occidental	Checo	cz
	Eslovaco	sk
Austranesio	Indonesio	id
	Malayo	my
Español	Argentino	es-AR
	Peninsular	es-ES
Portugués	Brasileño	pt-BR
	Europeo	pt-PT
Desconocido		xx

Tabla 7.16: Idiomas en el corpus DSLCC v.2.0.

La longitud de cada frase se encuentra entre 20 y 100 tokens. Para cada idioma o dialecto, el corpus contiene 18.000 instancias para entrenamiento, 2.000 para desarrollo y 1.000 para pruebas. Un resumen del número total de instancias se muestra en la Tabla 7.17. La partición de test se compone de dos partes, A y B. Ambas contienen el mismo número de instancias, pero la partición B ha sido procesada con un reconocedor de entidades nombradas para reemplazar cada entidad nombrada (NE, por sus siglas en inglés) por un marcador fijo #NE#, con el objeto de verificar la robustez de los métodos ante la ausencia de nombres propios que puedan aportar información de ciudades, personas, topónimos, etcétera. A esta partición se le llama *NE blinded*.

Entrenamiento	Desarrollo	Pruebas
252.000	28.000	14.000

Tabla 7.17: Número de instancias por partición.

A los participantes de la tarea se les proporcionan las particiones de entrenamiento y desarrollo con la información de clase para que puedan construir y validar sus modelos. La partición de pruebas se entrega sin la información de clase y los participantes deben enviar el resultado de la ejecución con la predicción para cada texto. La organización posteriormente evalúa esas ejecuciones y proporciona el ranking.

En la tarea se puede participar en dos modalidades:

- Abierta, donde se permite el entrenamiento de los modelos utilizando corpus externos además del proporcionado por la organización.
- Cerrada, donde sólo se permite el entrenamiento de los modelos utilizando los corpus proporcionados.

Por lo tanto, hasta cuatro son las ejecuciones que se pueden enviar, combinando la modalidad abierta y cerrada con la partición A y la B.

7.4.2 Representaciones distribuidas

Hemos aproximado la tarea en dos fases, en una primera fase se discrimina entre grupos de idiomas similares (clasificación intergrupo) para determinar en una segunda fase la variedad o idioma similar (clasificación intragrupo). En la primera fase se emplea una aproximación basada en distancias con prototipos del lenguaje para determinar el grupo, identificando el idioma con el modelo de Skip-gramas continuo. Con él se genera una representación distribuida de las palabras, un vector n-dimensional, y se ha utilizado la variación del vector de frases para generar la representación final de los documentos. Previo a estas fases se realiza un preprocesado de los textos, tokenizando y eliminando aquellos de longitud unitaria o correspondientes a números y signos de puntuación, pasando a minúsculas todo el texto y detectando de manera individual las frases que los componen. Destacar la participación con este modelo en la tarea de modalidad cerrada, pues no se hace uso de ningún corpus externo al facilitado por la organización de la tarea.

Clasificación intergrupo Para clasificar frases entre grupos de idiomas similares, se utiliza una versión simplificada de [235]. Dado un conjunto de entrenamiento Tr que contiene frases pertenecientes a uno de los grupos de lenguajes Lg, primero generamos el conjunto de prototipis $proto_{Lg}$ de cada grupo del lenguaje utilizando una representación de bolsa de palabras. A continuación, para cada nueva frase $t = (w_1, w_2, ..., w_n)$ del conjunto de pruebas Te, se computa el grupo de lenguaje g como sigue:

$$g = \text{argmax}_{pr_g \in proto_{Lg}} \sum_{i}^{n} |w_i \cap pr_g|, \tag{7.8}$$

donde básicamente se determina el grupo de idiomas de una frase como el grupo con el mayor número de palabras comunes. Nótese que la frase se representa como una lista y, consecuentemente, se permite la repetición de palabras, contrariamente a los prototipos.

Usando el método propuesto se obtiene un 99,99% de precisión determinando el grupo de idiomas en la partición de desarrollo, lo que muestra la trivialidad de la tarea de identificación de idioma.

Clasificación intragrupo Para identificar el idioma de las frases pertenecientes a idiomas similares o variedades del mismo, se adapta lo descrito en el apartado 7.2, generando la representación vectorial de las frases mediante los modelos Skip-gramas y SenVec descritos.

Hemos realizado una primera ejecución (run1) mediante la combinación de la representación Skip-gramas más un clasificador logístico (skip-gram + LG). Con el mismo modelo hemos realizado una segunda ejecución (run2) con máquinas de vectores soporte (SVM) con núcleo de base radial y coste 10 (skip-gram + SVM). Finalmente, se ha realizado una ejecución (run3) utilizando la representación SenVec mediante clasificadores logísticos (SenVec + LG).

En este punto se debe remarcar que las frases de pruebas contienen palabras que no estaban presentes en el conjunto de entrenamiento. Obviamente, para tales palabras no se ha aprendido un vector distribuido, por lo que se usa el vector aleatorio inicial. Aunque podríamos ignorar y eliminar dichas palabras, los experimentos con la partición de desarrollo muestran que no hay pérdida de rendimiento cuando se incorporan estos vectores, e incluso en algunas configuraciones (e.g. skip-gram + LG) proporcionan una ligera mejora (0,6%). Podemos hipotizar que esta inclusión de ruido en los vectores ayuda a los clasificadores a determinar las fronteras entre idiomas, por lo que lo dejamos así en los experimentos.

Resultados Como se puede ver en la Tabla 7.18, los idiomas similares son más fáciles de distinguir, con *accuracies* cercanas al 100%. Una tendencia similar se aprecia al clasificar el grupo "otros idiomas" que contiene instancias de diferentes idiomas como francés o catalán. Las variedades de idioma dentro de un mismo grupo son más difíciles de identificar, obteniendo valores en el rango 80–90%, siendo el más difícil de identificar el grupo serbocroata, seguido del español y el portugués.

	Pruebas A			**Pruebas B**		
	(run1)	**(run2)**	**(run3)**	**(run1)**	**(run2)**	**(run3)**
Búlgaro	**100**	**100**	98,5	**100**	**100**	99,8
Macedonio	**100**	**100**	99,9	**100**	**100**	99,8
Elavo sudoriental	**100**	**100**	99,2	**100**	**100**	99,8
Bosnio	**80,3**	79,5	74,4	**75,1**	75,0	64,1
Croata	**85,9**	83,7	84,7	**85,8**	85,3	76,9
Serbio	75,1	80,2	**91,2**	74,7	77,2	**87,1**
Eslavo suroccidental	80,4	81,1	**83,4**	78,5	**79,1**	76,0
Checo	**99,9**	**99,9**	99,8	**100**	**100**	**100**
Eslovaco	**100**	**100**	99,3	**100**	**100**	95,1
Eslavo occidental	**99,9**	**99,9**	99,5	**100**	**100**	97,6
Español (Peninsular)	82,1	**87,8**	86,3	80,6	**85,3**	79,6
Espanol (Argentino)	**90,3**	87,0	87,6	**84,7**	77,0	81,6
Español	86,2	**87,4**	86,9	**82,6**	80,6	80,6
Portugués (Brasileño)	**94,5**	92,6	87,6	**90,4**	90,0	78,3
Portugués (Europeo)	83,2	87,9	**90,0**	78,0	83,2	**86,6**
Portugués	88,8	**90,2**	88,8	84,2	**86,6**	82,4
Malayo	99,2	99,4	**99,8**	98,7	**99,0**	91,7
Indonesio	99,3	**99,6**	99,4	98,9	99,4	**99,6**
Austranesio	99,2	99,5	**99,6**	98,8	**99,2**	95,6
Otros idiomas	**99,8**	**99,8**	**99,8**	**99,8**	**99,8**	**99,8**
Global	92,1	**92,7**	**92,7**	90,5	**90,8**	88,5

TABLA 7.18: Resultados en *accuracy* de la identificación entre idiomas similares en las particiones de pruebas A y B.

Con respecto a la sustitución de entidades nombradas de la partición *NE blinded* se aprecia una leve reducción en la precisión, mayor para el modelo SenVec. Las diferencias entre los modelos y los clasificadores empleados son reducidas, aunque SVM proporciona cierta mejora en la identificación de idiomas. El modelo Skip-gramas es menos sensible a la sustitución de entidades nombradas y ofrece el mejor rendimiento de media. El resultado obtenido con este modelo es ligeramente inferior al del mejor participante, que obtuvo 95,54% y 94,01% respectivamente. Se puede comprobar la significatividad de los resultados con y sin entidades nombradas en la Tabla 7.19.

	R1A	**R2A**	**R3A**	**R1B**	**R2B**	**R3B**
R1A		=	=	*	*	*
R2A			=	*	*	*
R3A				*	*	*
R1B					=	*
R2B						*
R3B						

TABLA 7.19: Test de significación de las ejecuciones en las particiones A y B de pruebas. (= no significativa $p > 0.05$; * significativa $0.05 \geq p > 0.01$)

Discusión La participación en la tarea DSL con los modelos de representación distribuida nos permite observar ciertos comportamientos a mencionar. En primer lugar la adecuación del modelo a los

objetivos propuestos, es decir, su aplicabilidad a textos cortos y no sólo a variedades de un mismo idioma sino de idiomas similares. Además, el método se erige competitivo con los mejores participantes quedando a escasas décimas del mejor de ellos. Se ha podido apreciar que el modelo Skip-gramas obtiene resultados en media superiores a SenVec, y especialmente en el caso de las pruebas sin entidades nombradas, donde SenVec ligeramente reduce la precisión obtenida. Con respecto a las variedades de lenguajes las más difíciles de identificar han sido las serborcroatas, seguidas del español y el portugués. La identificación de grupos de idiomas, con precisiones cercanas al 100%, muestran la trivialidad de la tarea.

7.4.3 Representación de baja dimensionalidad

Siguiendo lo descrito en el apartado 7.3, calculamos los pesos por término para cada idioma y generamos la representación de baja dimensionalidad correspondiente, aprendiendo un modelo con una técnica de aprendizaje automático. Concretamente hemos probado diferentes métodos proporcionados por Weka. Posteriormente hemos construido una aplicación en Java para predecir el idioma de un nuevo documento de la partición de pruebas a partir de los modelos previamente entrenados. A continuación se muestran los resultados oficiales obtenidos en la participación en la tarea, así como la aproximación utilizada para la modalidad abierta y la cerrada. Se presentan resultados comparativos entre las particiones de desarrollo y las de pruebas A y B. Hemos llevado a cabo un análisis estadístico de significación entre ambas particiones de pruebas. Hemos usado la siguiente notación para los niveles de confianza: * al 95% y ** al 99%.

Modalidad abierta Hemos aproximado la tarea en modalidad abierta mediante un proceso en dos fases. En la primera fase, utilizamos un detector de idioma [236] para obtener el grupo principal. La *accuracy* de esta fase en la partición de desarrollo se muestra en la Tabla 7.20.

Grupo idiomático	**Accuracy**
bg	99,80
mk	100,00
es	99,96
pt	99,72
hr	99,73
id	99,92
cz	99,63
sk	99,65
xx	99,90
global	99,81

TABLA 7.20: *Accuracy* del detector de idioma en la partición de desarrollo.

En esta fase, se ha podido detectar el búlgaro (*bg*), checo (*cz*), macedonio (*mk*) y eslovaco (*sk*). Con respecto al resto de variedades, se han detectado del siguiente modo: *i*) eslavos sudoccidentales (croata, bosnio y serbio) se detectan como croata (*hr*); *ii*) austranesio (indonesio y malayo) se detectan como indonesio (*id*); *iii*) español (peninsular y argentino) y portugués (europeo y brasileño) como sus respectivos grupos (*es* y *pt*). Clasificamos como *xx* al resto. Una vez que el grupo principal ha sido identificado, hemos aplicado el método de baja dimensionalidad para detectar la variedad correspondiente. Los resultados para la partición de desarrollo y las particiones de pruebas A y B se muestran en la Tabla 7.21.

Idioma	Desarrollo	Pruebas A	Pruebas B
bg*	99,80	99,90	99,80
mk*	100,00	99,90	100,00
es-ES	88,00	84,70	79,50
es-AR*	87,50	88,00	87,70
pt-PT	88,60	87,40	94,00
pt-BR	90,10	90,03	68,50
bs*	78,35	78,00	74,40
hr*	86,15	85,80	85,40
sr**	86,40	86,40	82,70
id	99,40	99,40	92,90
my*	99,45	99,20	99,50
cz*	99,70	99,80	99,40
sk*	99,60	99,30	99,60
xx*	99,90	99,90	99,70
global	93,07	92,71	90,22

TABLA 7.21: *Accuracy* en la identificación para la modalidad abierta y las particiones de desarrollo y pruebas A y B.

Los resultados para los grupos de un sólo idioma (*bg*, *mk*, *cz* y *sk*) muestran precisiones sobre el 99% para ambas particiones, desarrollo y pruebas. Las precisiones para los grupos con más de una variedad son ligeramente inferiores. No es el caso del austronesio (*id*) donde los resultados obtenidos son superiores al 99% excepto para la variedad *id* en la partición de pruebas B. El peor resultado se obtiene para eslavo sudoccidental (*hr*) donde el clasificador debía discriminar entre tres clases. El test de significación muestra que el método es bastante robusto ante el blindaje de las entidades nombradas (partición de pruebas B) en los casos de eslavo sudoccidental (*bs*, *hr* y *sr*), malayo (*my*) y español de Argentina (*es-AR*).

Modalidad cerrada En la modalidad cerrada se aprende un multiclasificador para el conjunto de 14 idiomas diferentes. Los resultados se muestran en la Tabla 7.22. Se aprecia que los resultados globales para la partición de pruebas B (72,11%) son muy inferiores a los de las particiones de pruebas A (85,57%) y desarrollo (86,06%). En esta línea, los resultados para muchos de los idiomas son significativamente diferentes entre sí, excepto para el español de Argentina (*es-AR*), macedonio (*mk*) y el grupo otros (*xx*). Esto puede deberse a la pérdida de los pesos de los términos relativos a las entidades nombradas y que puedan proporcionar información discriminante entre variedades.

Comparación entre métodos En la Tabla 7.23 se muestran los resultados comparados entre las modalidades abiertas y cerradas en el conjunto de desarrollo. Es destacable que en ambas modalidades se obtienen resultados inferiores con determinados grupos de lenguas (*es*, *pt* y *hr*). En relación a grupos con un único idioma (*bg*, *mk*, *cz* y *sk*), se obtienen precisiones superiores al 95% en ambas modalidades, mostrando la robustez del método propuesto. Por otro lado, se ha llevado a cabo un análisis de significación estadística que no ha permitido afirmar que no haya diferencias entre los resultados obtenidos para ambas modalidades, concluyendo de este modo que el método en dos fases utilizado en la modalidad cerrada es más preciso que la aproximación que trata con todas las variedades a la vez.

Discusión De los resultados anteriores, donde se muestra la participación del método de baja dimensionalidad en las cuatro posibles modalidades de la tarea (abierta/cerrada, con/sin entidades nombradas), se desprende que el método en dos fases utilizado en la modalidad abierta obtiene mejores resultados,

Idioma	Desarrollo	Pruebas A	Pruebas B
bg	98,15	97,50	95,10
mk*	98,95	98,20	98,20
es-ES	87,55	84,80	48,70
es-AR**	67,05	70,00	74,10
pt-PT	82,15	81,20	58,30
pt-BR	72,45	72,50	65,90
bs	55,70	54,30	86,20
hr	80,85	78,88	13,10
sr	74,40	74,70	7,80
id	97,75	97,60	92,00
my	94,25	93,60	97,60
cz	98,45	98,40	94,40
sk	98,80	97,60	79,30
xx*	98,55	98,50	98,80
overall	86,06	85,57	72,11

TABLA 7.22: *Accuracy* en la identificación para la modalidad cerrada y las particiones de desarrollo y pruebas A y B.

Idioma	Abierta	Cerrada
bg	99,80	98,15
mk	100,00	98,95
es	87,75	77,30
pt	89,35	77,30
hr	83,63	70,32
id	99,43	96,00
cz	99,70	98,45
sk	99,60	98,80
xx	99,90	98,55
global	93,07	86,08

TABLA 7.23: *Accuracy* para las modalidades abierta y cerrada en el conjunto de desarrollo.

dividiendo el problema en dos fases: primero identificar idioma y después identificar variedad, frente al método que trata con todas las variedades a la vez.

Con relación a las variedades se aprecia que las más difíciles de predecir son las eslavas sudoccidentales, seguidas por las variedades de español y portugués. El grupo de las austronesias se identifica casi inequivocamente con ambas aproximaciones. Las mayores precisiones se obtienen para los grupos que sólo contienen una variedad, tanto en la modalidad abierta como en la cerrada, de lo que podemos inferir que son más variables lexicamente con respecto al resto y que el método propuesto es capaz de capturar de manera efectiva dichas variaciones.

7.4.4 Comparativa entre representaciones

En la Tabla 7.24 se muestran los resultados obtenidos por los métodos correspondientes a las representaciones distribuidas, Skip-gramas y SenVec, y la de baja dimensionalidad, en la modalidad cerrada y sobre el conjunto de desarrollo. Como se puede apreciar, la mayoría de resultados son muy similares. Hemos marcado la significación estadística según t-Student en aquellos casos en los que alguno de los resultados es significativamente mejor que el siguiente (e.g. para el caso del portugués de Portugal, los resultados de SenVec son significativamente mejores que los de LDR, que a su vez son significativamente mejores a

los de Skip-gram). Se puede comprobar, independientemente de la significación de los mismos, que en los casos en los que una determinada representación mejora para una determinada lengua o variedad, empeora para la otra u otras, obteniendo de media resultados muy similares.

Idioma	**Skip-gram**	**SenVec**	**LDR**
bg	100	100	99,9
mk	100	100	99,9
es-ES	82,1	86,3*	84,7*
es-AR	90,3*	87,6	88,0
pt-PT	83,2	90,0*	87,4*
pt-BR	94,5*	87,6	90,0*
bs	80,3	74,4	78,0*
hr	85,9	84,7	85,8
sr	75,1	91,2	86,4*
id	99,3	99,4	99,4
my	99,2	99,8*	99,2
cz	99,9	99,8	99,8
sk	100*	99,3	99,3
xx	99,8	99,8	99,9

TABLA 7.24: Resultados comparados para las representaciones en términos de *accuracy* para la modalidad cerrada en el conjunto de desarrollo.

7.5 Conclusiones

El principal objetivo de este capítulo ha sido verificar si la expresión de las emociones en el marco del discurso puede llegar a ser un rasgo diferenciador entre usuarios de diferentes variedades de una misma lengua. Para ello hemos construido Hispablogs, un corpus con blogs en cinco variedades de español, y sobre él hemos experimentado con EmoGraph. Los resultados se muestran significativamente inferiores a modelos que capturan principalmente variedades lexicográficas como los basados en n-gramas, lo que nos permite argumentar a favor de una mayor diferenciación a nivel léxico que de estructura del discurso o expresión de emociones.

Partiendo de lo anterior, hemos modelado los documentos con dos representaciones distribuidas, y hemos verificado su superioridad en la tarea frente a las representaciones utilizadas en el estado del arte basadas en n-gramas. Esto pone de manifiesto que los modelos basados en variaciones léxicas del contenido proporcionan información más discriminativa a la hora de determinar variedades de una misma lengua, y que, además, modelos del lenguaje más elaborados, que tengan en consideración no sólo la secuencia de palabras y/o caracteres, sino también aspectos semánticos, permiten obtener resultados significativamente superiores, argumento a favor de que cambios no sólo a nivel de vocabulario, sino del uso y significado que se hace del mismo, aportan información valiosa para la discriminación entre variedades de una misma lengua.

Con lo anterior en mente, hemos propuesto una representación que tiene en cuenta todo el vocabulario completo, a diferencia de los modelos de n-gramas que acaban limitándose a un subconjunto del mismo, pero con el objetivo de reducir considerablemente la dimensionalidad de la representación y el coste computacional de obtenerla, de modo que pueda ser aplicada de manera óptima en entornos *big data*

como los medios sociales. Los resultados obtenidos muestran su superioridad respecto al estado del arte, con resultados significativamente superiores a los basados en n-gramas y comparables a los basados en representaciones distribuidas, pero con respecto a estos últimos, reduciendo al 10% su dimensionalidad. Esta representación ha puesto de manifiesto que únicamente las variaciones lexicográficas de los textos, representadas de modo que se tomen en consideración todas ellas en su conjunto, permiten discriminar de manera competitiva entre variedades de una misma lengua.

Finalmente, y con el objetivo de comprobar su adaptabilidad y extensibilidad, hemos participado en la tarea DSL para discriminar entre lenguas similares, además de variedades de una misma lengua, en otras lenguas diferentes al español. Hemos participado con las tres representaciones propuestas en el capítulo (Skip-gramas, SenVec y LDR), obteniendo resultados competitivos. La comparación entre las propuestas nos muestra resultados similares, verificando que cuando en ocasiones una representación obtiene mejores resultados en la discriminación de una lengua, la otra representación lo hace en la otra lengua.

Un último punto que llama la atención es la gran diferencia en las *accuracies* obtenidas en la tarea DSL (∼93%) frente a las obtenidas con el corpus HispaBlogs (70∼72%). Hemos identificado cuatro factores que podrían haber influido en tales diferencias: *i*) el mayor desafío de procesar el estilo de escritura de social media en comparación con el estilo más limpio empleado en los textos periodísticos; *ii*) la longitud de los textos, 10 artículos por blog frente a una única frase por instancia en la competición, en lugar de contribuir positivamente a la discriminación podría haber introducido ambigüedad y ruido; *iii*) el mayor número de clases en HispaBlogs (5) frente a la competición (3 en el peor de los casos), habiéndose mostrado ambas representaciones sensibles al incremento en dicho número; y *iv*) el posible sobreajuste al estilo de un mismo autor, ya que aunque en HispaBlogs los autores se separan completamente entre particiones, en el corpus DSLCC no parece haberse tenido en cuenta dicha restricción y se podrían haber encontrado textos de un mismo autor en ambas particiones, entrenamiento y pruebas.

Capítulo 8

Conclusiones y trabajo futuro

La revolución de los medios sociales está potenciando nuevos modelos de relación y comunicación entre las personas. La idiosincrasia inherente a estos medios sociales hace de ellos un entorno de comunicación especial, donde la libertad de expresión, la informalidad y la generación espontánea de temáticas y tendencias propician el acercamiento a la realidad diaria de las personas en su uso de la lengua. La posibilidad de analizar lo que en estos medios escriben los usuarios, en muchos casos de manera anónima o haciéndose pasar por otros, y poder inferir características como su edad, su sexo o la variedad de lenguaje que utilizan, permitiría afrontar con nuevas herramientas problemas forenses, de seguridad y de marketing.

Nuestro principal objetivo ha sido verificar la hipótesis de que el modo en que los usuarios articulan su discurso y expresan sus emociones tiene una fuerte dependencia con su edad y sexo. Para comprobarlo hemos propuesto EmoGraph, una representación que aprovecha la potencia de los grafos para modelar estructuras complejas, como por ejemplo el lenguaje natural, y obtener así información sobre la importancia relativa de sus elementos en función no sólo de su frecuencia de uso, sino de su posición con y en relación con el resto de elementos. La comparación de los resultados obtenidos por EmoGraph frente a las representaciones propuestas basadas en frecuencias de características de estilos y de n-gramas de partes del discurso, permite corroborar tal hipótesis. Los resultados obtenidos se muestra competitivos con los sistemas de mejores prestaciones del estado del arte, además de robustos a diferentes idiomas y medios sociales, lo que confirma nuestra hipótesis.

Del análisis de las características más discriminantes se desprenden conclusiones en línea con los estudios previos, pero a un nivel de detalle superior. Por ejemplo, en la identificación de sexo predominan como más significativas características relativas a la importancia de la organización del discurso en torno a determinadas categorías gramaticales como los nombres, verbos o adjetivos, junto con la carga emotiva que éstas pueden expresar. Por otro lado, en la identificación de edad predominan como más discriminativas las características relativas a la importancia de determinados elementos como los conectores del discurso, que ayudan en la construcción de las frases e imprimen un estilo particular, resultado fuertemente respaldado por el estado del arte donde se apunta a la mayor complejidad en la creación de estructuras lingüísticas cuanto más avanzado en edad se encuentre su autor. Además, en ambos casos las características relativas a las emociones están presentes entre las más de mayor poder discriminativo, lo que refuerza la hipótesis de su relación con la edad y el sexo de quien las expresa.

Un análisis del error nos ha mostrado que el método EmoGraph requiere de un mínimo de palabras para comenzar a ser eficaz; se pasa de errores superiores al 50% para un número menor a 50 palabras, a menos de la mitad cuando el número de palabras supera esta cifra, siendo mínimo a partir de 150-200 palabras. Los resultados obtenidos muestran que la capacidad discriminativa de EmoGraph es superior a las representaciones comúnmente usadas cuando se dispone de suficientes palabras, siendo esta dependencia mayor en el caso de la identificación de edad, aunque el análisis del error muestre a la identificación de sexo como más compleja. En el caso de que se disponga de pocas palabras, tal y como se ha mostrado en los ejemplos, con cualquier representación el problema de identificación se convierte en puro azar.

Es interesante apuntar el menor rendimiento alcanzado por EmoGraph en revisiones de hoteles, medio limitado en cuanto a temáticas y/o emociones expresadas (concretamente quejas o alabanzas sobre servicios relacionados con hoteles), frente a los mejores resultados obtenidos en Twitter y blogs, medios donde se dispone de mayor cantidad de textos por autor, lo que favorece el modelado de su estructura gramatical, y donde además la libre expresión de temáticas y emociones, enriquece la representación. Los resultados obtenidos en los diferentes medios, con la particularidad descrita anteriormente, argumenta en pro de la hipótesis de independencia del medio social siempre y cuando este medio social propicie la libre expresión de las emociones, como en los medios descritos.

Los resultados obtenidos para los diferentes idiomas se mantienen competitivos, lo que soporta la hipótesis de la independencia con respecto al idioma de la relación entre expresión de emociones y edad y sexo. La confirmación de esta hipótesis nos hacía conjeturar que la expresión de las emociones no pudiera ser una característica discriminativa a la hora de diferenciar usuarios de diferentes variedades de una misma lengua, postulando por contra que las diferencias se encontrarían principalmente a nivel léxico.

En esta línea hemos comprobado que EmoGraph no es capaz de capturar diferencias significativas como para discriminar de manera adecuada entre hablantes de diferentes variedades de una misma lengua. Por contra, tres de las cuatro representaciones basadas en contenido que hemos analizado (Skip-gramas, SenVec y LDR) han obtenido resultados significativamente superiores a los modelos más comunes del estado del arte. Estos resultados corroboran la hipótesis de que las variedades del lenguaje se diferencian en mayor medida en el vocabulario utilizado por los usuarios, que en la estructura gramatical y el modo en que dichos usuarios expresan sus emociones. Además, las dos representaciones distribuidas (Skip-gramas y SenVec) nos muestran que incorporar información sobre el uso y significado del vocabulario, propicia la correcta discriminación entre lenguas sobre el uso de secuencias de palabras y/o caracteres de los modelos de n-gramas.

Por otro lado y debido al gran volumen de información que se origina en los medios sociales en tiempo real, hemos propuesto la representación de baja dimensionalidad LDR que reduce drásticamente la dimensionalidad y el coste computacional de obtener las características para representar los textos, y hemos mostrado que no sólo obtiene resultados significativamente superiores a las representaciones más comúnmente utilizadas en el estado del arte, sino que es competitiva con las representaciones distribuidas tan en voga en la actualidad, pero que nunca antes ha sido utilizada para esta tarea. Esto además pone de manifiesto que únicamente las variaciones léxicas de los textos, cuando se toman en consideración todas ellas en su conjunto, permiten obtener resultados comparables a los basados en las representaciones distribuidas, pero reduciendo drásticamente su dimensionalidad al 10%.

8.1 Contribuciones

1. Hemos propuesto la representación EmoGraph para modelar el estilo discursivo y la expresión de las emociones en textos, y la hemos aplicado a la identificación de edad y sexo. Además, hemos verificado su aplicabilidad y robustez a diferentes medios sociales e idiomas. De este modo hemos verificado nuestra primera hipótesis (Capítulo 5).
2. Hemos investigado la aplicabilidad de la representación EmoGraph en la tarea de identificación de variedades de una misma lengua, comprobando que la expresión de emociones y el estilo discursivo no varía de modo discriminativo. Para verificar nuestra hipótesis de que las variaciones se producen principalmente a nivel léxico, hemos analizado varias representaciones: una basada en patrones (IG-WP), dos basadas en representaciones distribuidas sobre el conocido modelo de Skip-gramas continuos, y la representación de baja dimensionalidad (LDR) que hemos propuesto y que permite trabajar de manera eficiente en entornos *big data*. De esta manera hemos verificado nuestra segunda hipótesis (Capítulo 7).
3. Hemos creado los recursos necesarios para llevar a cabo la investigación, concretamente:
   (a) Con respecto a la identificación de edad y sexo, hemos colaborado en la organización y creación de un marco de evaluación en la tarea de identificación de edad y sexo del PAN en el CLEF (Capítulo 3), lo que ha permitido crear un conjunto de corpus recopilados de diferentes medios sociales (Twitter, blogs, revisiones online, redes sociales) y en diferentes idiomas (inglés, español, holandés e italiano), etiquetados con edad y sexo (Capítulo 3, apartados 3.1.1, 3.2.1 y 3.3.1).
   (b) Con respecto a la identificación de emociones en medios sociales, y concretamente en Twitter, hemos compilado el corpus Barcenas con tuits tratando un caso de corrupción política ocurrido en España entre el 9 de julio y el 2 de octubre de 2013, con un total de 4.397.023 tuits en español (Capítulo 4, apartado 4.2.1).
   (c) Con respecto a la identificación de emociones en medios sociales y su relación con el sexo, así como con el uso de la ironía, hemos generado el corpus EmIroGeFB con comentarios de Facebook anotados con las seis emociones básicas de Ekman, la presencia/ausencia de ironía, y el sexo de los autores de los comentarios. El corpus se enmarca dentro de tres temáticas (política, fútbol, famosos) y consta de 1.200 comentarios en español (Capítulo 4, apartado 4.3).
   (d) Por último, con respecto a la identificación de la variedad de lenguaje, hemos construido el corpus HispaBlogs con posts escritos en blogs personales en cinco variedades del español: Argentina, Chile, España, México y Panamá. El corpus consta de dos particiones de 2.400 y 1.000 autores por partición (Capítulo 7, apartado 7.1).

8.2 Trabajo futuro

Nuestro principal objetivo en líneas futuras es el de investigar aplicaciones de EmoGraph a otras tareas de *author profiling*, como por ejemplo la detección de perfiles de tipo *bot*. Los *bots* se programan para realizar publicaciones de manera (semi)automática sobre determinados temas cuando suceden determinadss eventos. Por ejemplo, publicar el precio de la gasolina, responder a un comentario sobre un producto o atacar masivamente a un político. La detección de la edad y el sexo del usuario no es suficiente para determinar si es un perfil de tipo *bot*, pues el lenguaje que utiliza puede estar condicionado por el usuario que lo ha programado. Así mismo, modelos basados en n-gramas o frecuencias de diferentes

elementos del texto, pueden sobreajustarse al dominio de entrenamiento, reduciendo considerablemente su rendimiento cuando se aplican a dominios diferentes[1]. Nuestra intuición es que los bots adolecen de homogeneidad emocional, y que su expresión y discurso es particularmente similar a través de sus publicaciones, a diferencia de un usuario humano que tenderá a hablar de los temas que en cada momento le produzcan emoción (como veíamos en el análisis de las tendencias). En este sentido proponemos utilizar EmoGraph para modelar su discurso, pero no a nivel de todos sus textos como hemos hecho en este trabajo, sino a nivel de cada uno de los textos que ha publicado, permitiendo así comparar el grado de similitud estructural y emotiva a lo largo de los mismos.

En la línea de la expresión de las emociones, planificamos utilizar EmoGraph para la detección de perfiles políticos[2]. Nuestra intuición es que los tópicos de los que hablan los usuarios y las emociones que suelen expresar cuando lo hacen, dependen fuertemente de su ideario político y del partido que gobierne en ese momento. Así pues, consideramos que ajustando los tópicos de EmoGraph a ontologías políticas del momento, nos permitirá capturar diferencias entre los diferentes perfiles como para permitir discriminarlos, en primer lugar a nivel de idearios (e.g. liberalismo, socialismo, etc.), y en segundo lugar a un nivel de detalle mayor, de partido político con el que simpatizan[3].

Una de las características que captura EmoGraph, y que ha ayudado a determinar la emotividad de los mensajes, es la tipología de los verbos utilizados por el usuario. Para su clasificación hemos seguido las teorías de Levin, donde los verbos se agrupan según respondan a una semántica comunicativa, emotiva, perceptiva, comprensiva o permisiva. Con clasificaciones alternativas podríamos aproximar la identificación de tipos de personas en función de su tendencia perceptiva y de aprendizaje (visuales, auditivas o kinestésicas) [237], lo que permitiría mejorar y personalizar la comunicación en ámbitos como la enseñanza, el marketing y las ventas, o los interrogatorios policiales [238]. De manera similar, la identificación de patrones en expresiones verbales (e.g. no recuerdo vs. no me acuerdo; me explico vs. me entiendes), podría ayudar en la identificación de testimonios [239] o revisiones de producto [240] falsos, así como a aplicar teorías como la de McClelland [241] para identificar el perfil de necesidades básicas (logro, reconocimiento y poder) a la hora de configurar equipos de trabajo de alto rendimiento.

Por otro lado, la identificación de lenguas similares y variedades de una misma lengua sigue siendo uno de nuestros principales intereses. Nuestro siguiente paso consiste en la construcción de un corpus de variedades del español en Twitter, donde el vocabulario utilizado es más restringido y espontáneo, por lo que pensamos que representa de manera más fiel la idiosincrasia cultural de los usuarios, y por lo tanto es más representativo de la realidad cotidiana de su lenguaje. Además, pretendemos bajar un nivel más en la escala entre lenguas para tratar con variedades dialectales; concretamente, estamos interesados en investigar las variaciones dialectales del catalán, valenciano y mallorquín, así como analizar el error que se comete cuando se contrasta cada una de ellas con el español, el francés, el italiano y el portugués, cuantificando así las posibles influencias culturales y geográficas.

[1] En esta línea hemos realizado un trabajo preliminar en la dirección de un trabajo de la asignatura de Aplicaciones de la Lingüística Computacional del Master en Inteligencia Artificial, Reconocimiento de Patrones e Imagen Digital: `https://www.upv.es/titulaciones/MUIARFID/`. En él hemos podido comprobar que los resultados decrecen considerablemente cuando se aplican en dominios cruzados. Se puede descargar la memoria del proyecto en la siguiente dirección: `http://ow.ly/XVREy`

[2] Hemos realizado, en el marco de la asignatura de Aplicaciones de la Lingüística Computacional mencionada anteriormente, un trabajo preliminar sobre identificación de militantes de los dos partidos políticos mayoritarios en el momento de estudio en España. Se puede descargar la memoria del trabajo preliminar en: `http://ow.ly/XVTnw`

[3] Nuestro interés se centra además en analizar la correlación o similitud entre el discurso emitido por los políticos, y el que repiten o hacen suyo los seguidores.

8.3 Publicaciones

A continuación mostramos las publicaciones realizadas en revistas y congresos, agrupándolas por cada uno de los objetivos propuestos:

Identificación de edad y sexo

- Francisco Rangel and Paolo Rosso. On the Impact of Emotions on Author Profiling. In: Information Processing & Management 52(1):73–92, 2016
- Francisco Rangel and Paolo Rosso. On the Multilingual and Genre Robustness of EmoGraphs for Author Profiling in Social Media. In: 6th International Conference of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction (CLEF'15), Springer-Verlag, LNCS(9283), pp. 274–280, 2015
- Francisco Rangel and Paolo Rosso. Use of Language and Author Profiling: Identification of Gender and Age. In: 10th International Workshop on Natural Language Processing and Cognitive Science (NLPCS'13) pp. 177–186, 2013
- Francisco Rangel. Author Profile in Social Media: Identifying Information about Gender, Age, Emotions and beyond*. In: Future Directions in Information Access Symposium (FDIA'13) at ESSIR 2013

Procesamiento afectivo

- Soren Volgmann, Francisco Rangel, Oliver Niggemann and Paolo Rosso. Emotional Trends in Social Media – A State Space Approach. In: 21st Conference on Artificial Intelligence (ECAI'14) pp. 1123-1124, 2014
- Francisco Rangel, Irazú Hernández, Paolo Rosso and Antonio Reyes. Emotions and Irony per Gender in Facebook. In: Workshop on Emotion, Social Signals, Sentiment & Linked Open Data (ES3LOD - LREC'14) pp. 66-73, 2014
- Francisco Rangel and Paolo Rosso. On the Identification of Emotions in Facebook Comments. In: First International Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM'13) A Workshop of the XIII International Conference of the Italian Association for Artificial Intelligence (AI*IA'13) pp. 34–46, 2013

Identificación de variedad del lenguaje

- Francisco Rangel and Paolo Rosso. A Low Dimensionality Representation to Language Variety Identification. In Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'16), Springer-Verlag, LNCS(), pp. , 2016
- Marc Franco-Salvador, Francisco Rangel, Paolo Rosso, Mariona Taulé and Antonia Martí. Language Variety Identification using Distributed Representations of Words and Documents. In: 6th International Conference of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction (CLEF'15), Springer-Verlag, LNCS(9283), pp. 28–40, 2015

- Marc Franco Salvador, Paolo Rosso and Francisco Rangel. Distributed Representations of Words and Documents for Discriminating Similar Languages. In: Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4CloseLang2 – VarDial 2). Recent Advances in Natural Language Processing (RANLP), 2015
- Raül Fabra, Francisco Rangel and Paolo Rosso. NLEL_UPV_Autoritas participation at Discrimination between Similar Languages (DSL) 2015 Shared Task. In: Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4CloseLang2 – VarDial 2). Recent Advances in Natural Language Processing (RANLP), 2015

Generación de recursos y marco de evaluación

- Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstatios Stamatatos and Giacomo Inches. Overview of the Author Profiling Task at PAN 2013. In: Forner P., Navigli R., Tufis D. (Eds.), CLEF 2013 Labs and Workshops, Notebook Papers. CEUR-WS.org, vol. 1179 pp. 352–365, 2013
- Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven and Walter Daelemans. Overview of the 2nd Author Profiling Task at PAN 2014. In: Cappellato L., Ferro N., Halvey M., Kraaij W. (Eds.) CLEF 2014 Labs and Workshops, Notebook Papers. CEUR-WS.org, vol. 1180 pp. 898-827, 2014
- Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein and Walter Daelemans. Overview of the 3rd Author Profiling Task at PAN 2015. In: Cappellato L., Ferro N., Jones G., San Juan E. (Eds.) CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1391 pp. , 2015
- Tim Gollub, Martin Potthast, Anna Beyer, Matthias Busse, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos and Benno Stein. Recent Trends in Digital Text Forensics and its Evaluation: Plagiarism Detection, Author Identification, and Author Profiling. In: Forner P., Navigli R., Tufis D.(Eds.), Notebook Papers of CLEF 2013 LABs and Workshops, CLEF-2013, Springer-Verlag, LNCS(8138), pp. 282–302, 2013
- Martin Potthast, Tim Gollub, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos and Benno Stein. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: 5th International Conference of CLEF on Information Access Evaluation meets Multilinguality, Multimodality, and Interaction (CLEF'14), Springer-Verlag, LNCS(8685), pp. 268-299, 2014
- Efstathios Stamatatos, Martin Potthast, Francisco Rangel, Paolo Rosso and Benno Stein. Overview of the PAN/CLEF 2015 Evaluation Lab. In: 6th International Conference of CLEF on Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF'15), Springer-Verlag, LNCS(9283), pp. 518–538, 2015

Otros

- Francisco Rangel and Paolo Rosso. El Uso del Lenguaje en los Diferentes Canales de Internet. In: IV Congreso de Redes Sociales Comunica2.0, Valencia: Editorial UPV, 2013

8.4 Impacto en medios

A continuación mostramos el impacto de la investigación en medios como la televisión, la prensa digital o la prensa impresa.

Televisión

- ¿Es Internet un cerebro?. Programa Informe Semanal de Televisión Española (TVE) (http://ow.ly/XBrFK)
- La UPV desarrolla, junto a Autoritas Consulting, una herramienta contra la pederastia que permite deducir sexo y edad de los usuarios de redes sociales. UPV TV. (http://ow.ly/XVVrw)

Prensa digital

- Uncovering Plagiarism - Author Profiling at PAN. ERCIM News. (http://ow.ly/XBUrI)
- La UPV desarrolla, junto a Autoritas Consulting, una herramienta contra la pederastia que permite deducir sexo y edad de los usuarios de redes sociales. UPV. (http://ow.ly/XVVrw)
- Los mensajes anónimos de pederastas, al descubierto. Las Provincias (http://ow.ly/XBrI3)
- Una herramienta detecta en internet perfiles falsos y amenazas de seguridad. Levante-EMV. (http://ow.ly/XBrPO)
- Investigadores de la UPV crean un programa que detecta perfiles falsos en redes sociales. Levante-EMV. (http://ow.ly/XBrWD)
- Crean en España app para detectar perfiles falsos en la Red. Computer Hoy. (http://ow.ly/XBMcy)
- Una herramienta detecta en internet perfiles falsos y amenazas de seguridad. 20minutos. (http://ow.ly/XBMjM)
- Una herramienta detecta en internet los perfiles falsos, amenazas de seguridad y posibles casos de pederastia. La Sexta. (http://ow.ly/XBMvV)
- Una nova eina detecta amenaces de seguretat i perfils falsos a Internet. Diari de Girona. (http://ow.ly/XBMBc)
- ¿Está sobrevalorado el poder de los 'influencers' en Twitter? Expansión. (http://ow.ly/XVW5S)

Prensa impresa

Se puede descargar PDF con recortes escaneados de la siguiente dirección: http://ow.ly/YlvpC

- 02/11/2015 Una nova eina detecta amenaces de seguretat i perfils falsos a Internet. Diari de Girona.
- 02/11/2015 Una aplicación detecta perfiles falsos en las redes sociales. Canarias 7.
- 01/11/2015 Un detective contra los perfiles falsos 3. El Correo de Andalucía.
- 01/11/2015 Lanzan una herramienta para detectar perfiles falsos on-line. El Correo Gallego.
- 01/11/2015 Investigadores de la UPV crean un programa que detecta perfiles falsos en redes sociales. Levante EMV.
- 01/11/2015 Una herramienta detecta periles falsos en Internet. La Opinión de Murcia.

- 01/11/2015 Los mensajes anónimos de pederastas, al descubierto. Las Provincias Alicante.
- 01/11/2015 Los mensajes anónimos de pederastas, al descubierto. Las Provincias.
- 31/10/2015 Los mensajes anónimos de pederastas, al descubierto. Las Provincias Valencia.
- 31/10/2015 Una herramienta detecta en internet los perfiles falsos, amenazas de seguridad y posibles casos de pederastia. La Sexta.
- 31/10/2015 Una herramienta detecta en internet perfiles falsos y amenazas de seguridad. Levante EMV.
- 31/10/2015 Una herramienta detecta en internet perfiles falsos y amenazas de seguridad. Levante TV.
- 31/10/2015 Una herramienta detecta en internet perfiles falsos y amenazas de seguridad. 20minutos.
- 31/10/2015 Una herramienta detecta en internet perfiles falsos y amenazas seguridad. La Vanguardia.
- 31/10/2015 Una herramienta detecta en internet perfiles falsos y amenazas seguridad. Las Provincias Valencia.
- 31/10/2015 Una herramienta detecta en internet perfiles falsos y amenazas seguridad. ABC.
- 19/08/2014 ¿Está sobrevalorado el poder de los 'influencers' en Twitter? Expansión. (http://ow.ly/YlvNM)

8.5 Difusión en medios sociales

En esta tesis nos hemos querido centrar en los medios sociales debido, como remarcábamos en la introducción, a su proliferación e importancia en y para la sociedad actual. En esta misma línea, hemos querido contribuir a través de estos mismos medios sociales a difundir cuestiones relativas al desarrollo de esta investigación, con el doble objetivo de llegar a personal no científico y mostrando su aplicación práctica e implicaciones para la sociedad. A continuación describimos los tres medios utilizados para la difusión, con una selección de cinco artículos representativos publicados en cada uno de ellos.

Socialancer (http://www.socialancer.com/) Blog líder en el ámbito de las redes sociales y el marketing digital con más de 40.000 seguidores de todo el mundo, donde hemos difundido información relativa al análisis y generación de inteligencia a partir de los medios sociales.

- Cómo utilizar los perfiles de usuario para hacer Escucha Activa en redes sociales. (http://ow.ly/XBPpv)
- Influencers en redes sociales: cómo encontrar a los de tu sector. 5 perfiles. (http://ow.ly/XBPPK)
- Las 3 grandes mentiras de las herramientas de Social Media. Lo que no te puedes creer. (http://ow.ly/XBPA4)
- Cómo buscar empleados con tu herramienta de monitorización de Social Media. (http://ow.ly/XBQbH)
- Reputación online: 10 claves para hacer un buen diagnóstico. (http://ow.ly/XBQmd)

Coolhunting (http://coolhunting.autoritas.net/) Blog de Autoritas generado de manera colaborativa por expertos que hablan de tendencias en inteligencia organizacional. En él abordamos problemáticas técnicas o tendencias relativas a avances en el mundo científico en las áreas de interés.

- Dime cómo escribes y te digo de dónde eres. (http://ow.ly/XBQAV)
- Author Profiling y La Vanguardia. (http://ow.ly/XBQEM)
- Big Data y Social Media en retrospectiva. (http://ow.ly/XBQPz)
- Big Data y el Sesgo de Confirmación. (http://ow.ly/XBQYd)

Hablemos de I+D (http://www.kicorangel.com/) Blog personal donde hablamos de cuestiones relativas a la investigación, tanto desde la perspectiva científica como de su aplicación práctica en la industria.

- Author Profiling en el mundo real. (http://ow.ly/XBSsD)
- Cifras del análisis de las emociones en Facebook. (http://ow.ly/XBSEl)
- Big Five y de cómo predecir la personalidad de los internautas. (http://ow.ly/XBSPC)
- Are you interested in knowing how people see you in Twitter? (http://ow.ly/XBS0q)

Bibliografía

[1] Howard Rheingold. *Smart mobs: the next social revolution*. Basic books, 2007.

[2] Soraya Mehdizadeh. Self-presentation 2.0: narcissism and self-esteem on facebook. *Cyberpsychology, behavior, and social networking*, 13(4):357–364, 2010.

[3] Daria J Kuss and Mark D Griffiths. Online social networking and addiction—a review of the psychological literature. *International journal of environmental research and public health*, 8(9): 3528–3552, 2011.

[4] Pierre Lévy. *Collective intelligence*. Plenum/Harper Collins, 1997.

[5] Andrea De Mauro, Marco Greco, and Michele Grimaldi. What is big data? a consensual definition and a review of key research topics. In *AIP conference proceedings*, volume 1644, pages 97–104, 2015.

[6] Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. Influence and passivity in social media. In *Machine learning and knowledge discovery in databases*, pages 18–33. Springer, 2011.

[7] Giacomo Inches and Fabio Crestani. Overview of the international sexual predator identification competition at pan-2012. In *CLEF (Online working notes/labs/workshop)*, volume 30, 2012.

[8] Dasha Bogdanova, Paolo Rosso, and Thamar Solorio. Exploring high-level features for detecting cyberpedophilia. *Computer speech & language*, 28(1):108–120, 2014.

[9] Donato Hernández Fusilier, Manuel Montes-y Gómez, Paolo Rosso, and Rafael Guzmán Cabrera. Detecting positive and negative deceptive opinions using pu-learning. *Information processing & management*, 51(4):433–443, 2015.

[10] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[11] Krzysztof Kredens and Malcolm Coulthard. Corpus linguistics in suthorship identification. *The oxford handbook of language and law*, pages 504–516, 2012.

[12] Lawrence M Solan and Peter M Tiersma. Author identification in american courts. *Applied linguistics*, 25(4):448–465, 2004.

[13] Jono Fischbach. With a little bit of heart and soul analyzing the role of humint in the post cold war era. In *Woodrow wilson school policy conference 401a*, 1997.

[14] James W Pennebaker, Roger J Booth, and Martha E Francis. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, 2007.

[15] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *Text-The hague then Amsterdam then Berlin-*, 23(3):321–346, 2003.

[16] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412, 2002.

[17] Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. Author profiling for english emails. In *Proceedings of the 10th conference of the pacific association for computational linguistics (PACLING'07)*, pages 263–272, 2007.

[18] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[19] Dominique Estival, Tanja Gaustad, Ben Hutchinson, Son Bao Pham, and Will Radford. Author profiling for english and arabic emails. 2008.

[20] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006.

[21] Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. Stylometric analysis of bloggers' age and gender. In *Third international AAAI conference on weblogs and social media*, 2009.

[22] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2009.

[23] Xiang Yan and Ling Yan. Gender classification of weblog authors. In *AAAI spring symposium: computational approaches to analyzing weblogs*, pages 228–230, 2006.

[24] Arjun Mukherjee and Bing Liu. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 207–217. Association for Computational Linguistics, 2010.

[25] Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of the fifteenth conference on computational natural language learning*, pages 78–86. Association for Computational Linguistics, 2011.

[26] Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 conference short papers*, pages 38–42. Association for Computational Linguistics, 2010.

[27] Juan Soler-Company and Leo Wanner. Multiple language gender identification for blog posts. In *The annual meeting of the cognitive science society COGSCI*, 2015.

[28] Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on search and mining user-generated contents*, pages 37–44. ACM, 2011.

[29] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[30] Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and H Andrew Schwartz. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP'14)*, pages 1146–1151, 2014.

[31] Arthur E Hoerl and Robert W Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[32] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: a library for large linear classification. *The journal of machine learning research*, 9:1871–1874, 2008.

[33] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, pages 240–242, 1895.

[34] Jahna Otterbacher. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proceedings of the 19th ACM international conference on information and knowledge management*, pages 369–378. ACM, 2010.

[35] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. "how old do you think i am?"; a study of language and age in twitter. In *Proceedings of the seventh international AAAI conference on weblogs and social media*. AAAI Press, 2013.

[36] Cathy Zhang and Pengyu Zhang. Predicting gender from blog posts. Technical report, Technical report. University of Massachusetts Amherst, USA, 2010.

[37] Dong Nguyen, Noah A Smith, and Carolyn P Rosé. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, pages 115–123. Association for Computational Linguistics, 2011.

[38] Francisco Rangel, Paolo Rosso, Moshe Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the author profiling task at pan 2013. In *Forner P., Navigli R., Tufis D. (Eds.), CLEF 2013 labs and workshops, notebook papers. CEUR-WS.org, vol. 1179*, 2013.

[39] Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. Overview of the 2nd author profiling task at pan 2014. In *Cappellato L., Ferro N., Halvey M., Kraaij W. (Eds.) CLEF 2014 labs and workshops, notebook papers. CEUR-WS.org, vol. 1180*, 2014.

[40] Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. Overview of the 3rd author profiling task at pan 2015. In *Cappellato L., Ferro N., Jones G., San Juan E. (Eds.) CLEF 2015 labs and workshops, notebook papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1391*, 2015.

[41] John Houvardas and Efstathios Stamatatos. N-gram feature selection for authorship identification. In *Artificial intelligence: methodology, systems, and applications*, pages 77–86. Springer, 2006.

[42] A. Pastor Lopez-Monroy, Manuel Montes-Y-Gomez, Hugo Jair Escalante, Luis Villasenor-Pineda, and Esau Villatoro-Tello. INAOE's participation at PAN'13: author profiling task—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[43] A. Pastor López-Monroy, Manuel Montes y Gómez, Hugo Jair-Escalante, and Luis Villase nor Pineda. Using intra-profile information for author profiling—Notebook for PAN at CLEF 2014. In Cappellato et al. [243].

[44] Miguel-Angel Álvarez-Carmona, A.-Pastor López-Monroy, Manuel Montes-Y-Gómez, Luis Villaseñor-Pineda, and Hugo Jair-Escalante. Inaoe's participation at pan'15: author profiling task—notebook for pan at clef 2015. In Cappellato et al. [244].

[45] A Pastor López-Monroy, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, and Efstathios Stamatatos. Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-based systems*, 89:134–147, 2015.

[46] Jean-Valère Cossu, Nicolas Dugué, and Vincent Labatut. Detecting real-world influence through twitter. In *Network intelligence conference (ENIC), 2015 second european*, pages 83–90. IEEE, 2015.

[47] Michal Meina, Karolina Brodzinska, Bartosz Celmer, Maja Czokow, Martyna Patera, Jakub Pezacki, and Mateusz Wilk. Ensemble-based classification for author profiling using various features—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[48] Edson RD Weren, Anderson U Kauer, Lucas Mizusaki, Viviane P Moreira, J Palazzo M de Oliveira, and Leandro K Wives. Examining multiple features for author profiling. *Journal of information and data management*, 5(3):266, 2014.

[49] Suraj Maharjan, Prasha Shrestha, Thamar Solorio, and Ragib Hasan. A straightforward author profiling approach in mapreduce. In *Advances in artificial intelligence (IBERAMIA'14)*, pages 95–107. Springer, 2014.

[50] James W. Pennebaker. *The secret life of pronouns: what our words say about us*. Bloomsbury USA, 2013. ISBN 9781608194964.

[51] W Nelson Francis and Henry Kucera. Frequency analysis of english usage: lexicon and grammar, 1982.

[52] Cindy Chung and James W Pennebaker. The psychological functions of function words. *Social communication*, pages 343–359, 2007.

[53] Francisco Rangel and Paolo Rosso. El uso del lenguaje en los diferentes canales de internet. In *IV congreso de redes sociales comunica2.0, valencia: editorial UPV*, pages 120–135.

[54] Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, 2004.

[55] Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.

[56] Francisco Rangel and Paolo Rosso. Use of language and author profiling: identification of gender and age. *Natural language processing and cognitive science (NLPCS'13)*, pages 177–186, 2013.

[57] Tim Gollub, Benno Stein, and Steven Burrows. Ousting ivory tower research: towards a web framework for providing experiments as a service. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval*, pages 1125–1126. ACM, 2012.

[58] Tim Gollub, Benno Stein, Steven Burrows, and Dennis Hoppe. Tira: configuring, executing, and disseminating information retrieval experiments. In *Database and expert systems applications (DEXA), 2012 23rd international workshop on*, pages 151–155. IEEE, 2012.

[59] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

[60] Janet Holmes and Miriam Meyerhoff. *The handbook of language and gender*, volume 25. John Wiley & sons, 2008.

[61] Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 783–792. ACM, 2010.

[62] Braja Gopal Patra, Somnath Banerjee, Dipankar Das, Tanik Saikh, and Sivaji Bandyopadhyay. Automatic author profiling based on linguistic and stylistic features—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[63] Erwan Moreau and Carl Vogel. Style-based distance features for author profiling—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[64] Edson Weren, Viviane P. Moreira, and Jose Oliveira. Using simple content features for the author profiling task—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[65] Aditya Pavan, Aditya Mogadala, and Vasudeva Varma. Author profiling using LDA and maximum entropy—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[66] Lucie Flekova and Iryna Gurevych. Can we hide in the web? Large scale simultaneous age and gender author profiling in social media—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[67] Wee Yong Lim, Jonathan Goh, and Vrizlynn L. L. Thing. Content-centric age and gender profiling—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[68] Andres Alfonso Caurcel Diaz and Jose Maria Gomez Hidalgo. Experiments with SMS translation and stochastic gradient descent in Spanish text author profiling—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[69] Delia Irazu Hernandez Farias, Rafael Guzman-Cabrera, Antonio Reyes, and Martha Alicia Rocha. Semantic-based features for author profiling identification: first insights—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[70] Upendra Sapkota, Thamar Solorio, Manuel Montes-Y-Gomez, and Gabriela Ramirez-De-La-Rosa. Author profiling for English and Spanish text—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[71] Fermin Cruz, Rafa Haro, and Javier Ortega. ITALICA at PAN 2013: an ensemble learning approach to author profiling—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[72] Maria De-Arteaga, Sergio Jimenez, George Duenas, Sergio Mancera, and Julia Baquero. Author profiling using corpus statistics, lexicons and stylistic features—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[73] Yuridiana Aleman, Nahun Loya, Darnes Vilarino Ayala, and David Pinto. Two methodologies applied to the author profiling task—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[74] K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma. Author profiling: predicting age and gender from blogs—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[75] Lee Gillam. Readability for author profiling?—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[76] Mechti Seifeddine, Jaoua Maher, and Hadrich Belghith Lamia. Author profiling using style-based features—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[77] Magdalena Jankowska, Vlado Keselj, and Evangelos Milios. CNG text classification for authorship profiling task—Notebook for PAN at CLEF 2013. In Forner et al. [242].

[78] Enrique Amigó, Jorge Carrillo De Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten De Rijke, and Damiano Spina. Overview of replab 2013: evaluating online reputation monitoring systems. In *Information access evaluation. Multilinguality, multimodality, and visualization*, pages 333–352. Springer, 2013.

[79] Enrique Amigó, Jorge Carrillo-de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Edgar Meij, Maarten de Rijke, and Damiano Spina. Overview of replab 2014: author profiling and reputation dimensions for online reputation management. In *Information access evaluation. Multilinguality, multimodality, and interaction*, pages 307–322. Springer, 2014.

[80] Suraj Maharjan, Prasha Shrestha, and Thamar Solorio. A simple approach to author profiling in mapreduce—Notebook for PAN at CLEF 2014. In Cappellato et al. [243].

[81] James Marquardt, Golnoosh Fanardi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, and Martine De Cock. Age and gender identification in social media—Notebook for PAN at CLEF 2014. In Cappellato et al. [243].

[82] Christopher Ian Baker. Proof of concept framework for prediction—Notebook for PAN at CLEF 2014. In Cappellato et al. [243].

[83] Gilad Gressel, Hrudya P, Surendran K, Thara S, Aravind A, and Prabaharan Poomachandran. Ensemble learning approach for author profiling—Notebook for PAN at CLEF 2014. In Cappellato et al. [243].

[84] Edson R.D. Weren, Viviane P. Moreira, and José P.M. de Oliveira. Exploring information retrieval features for author profiling—Notebook for PAN at CLEF 2014. In Cappellato et al. [243].

[85] Julio Villena-Román and José-Carlos González-Cristóbal. DAEDALUS at PAN 2014: guessing tweet author's gender and age—Notebook for PAN at CLEF 2014. In Cappellato et al. [243].

[86] Seifeddine Mechti, Maher Jaoua, and Lamia Hadrich Belguith. Machine learning for classifying authors of anonymous tweets, blogs and reviews—Notebook for PAN at CLEF 2014. In Cappellato et al. [243].

[87] Paul T Costa and Robert R MacCrae. *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): professional manual.* Psychological assessment resources, 1992.

[88] Paul T Costa and Robert R McCrae. *The NEO personality inventory: manual, form S and form R.* Psychological assessment resources, 1985.

[89] Paul T Costa and Robert R McCrae. Normal personality assessment in clinical practice: the neo personality inventory. *Psychological assessment*, 4(1):5, 1992.

[90] Mounica Arroju, Aftab Hassan, and Golnoosh Farnadi. Age, gender and personality recognition using tweets in a multilingual setting—notebook for pan at clef 2015. In Cappellato et al. [244].

[91] Andreas Grivas, Anastasia Krithara, and George Giannakopoulos. Author profiling using stylometric and structura feature groupings—notebook for pan at clef 2015. In Cappellato et al. [244].

[92] Fahad Najib, Arshad Cheema, Rao Muhammad, and Adeel Nawab. Author's traits prediction on twitter data using content based approach—notebook for pan at clef 2015. In Cappellato et al. [244].

[93] Hafiz Rizwan-Iqbal, Muhammad Adnan-Ashraf, and Rao-Muhammad Adeel-Nawab. Predicting an author's demographics from text using topic modeling approach—notebook for pan at clef 2015. In Cappellato et al. [244].

[94] Carlos-E. González-Gallardo, Azucena Montes, Gerardo Sierra, Antonio Núñez, Salinas Adolfo, and Juan Ek. Tweets classification using corpus dependent tags, character and pos n-grams—notebook for pan at clef 2015. In Cappellato et al. [244].

[95] Suraj Maharjan and Thamar Solorio. Using wide range of features for author profiling—notebook for pan at clef 2015. In Cappellato et al. [244].

[96] Scott Nowson, Julien Perez, Caroline Brun, Shachar Mirkin, and Claude Roux. Xrce personal language analytics engine for multilingual author profiling—notebook for pan at clef 2015. In Cappellato et al. [244].

[97] Alberto Bartoli, Andrea De-Lorenzo, Alessandra Laderchi, Eric Medvet, and Fabiano Tarlao. An author profiling approach based on language-dependent content and stylometric features—notebook for pan at clef 2015. In Cappellato et al. [244].

[98] Adam Poulston, Mark Stevenson, and Kalina Bontcheva. Topic models and n-gram language models for author profiling—notebook for pan at clef 2015. In Cappellato et al. [244].

[99] Edson Weren. Information retrieval features for personality traits—notebook for pan at clef 2015. In Cappellato et al. [244].

[100] Juan-Pablo Posadas-Durán, Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, Ildar Batyrshin, Alexander Gelbukh, and Obdulia Pichardo-Lagunas. Syntactic n-grams as features for the author profiling task—notebook for pan at clef 2015. In Cappellato et al. [244].

[101] Yasen Kiprov, Momchil Hardalov, Preslav Nakov, and Ivan Koychev. Su@pan'2015: experiments in author profiling—notebook for pan at clef 2015. In Cappellato et al. [244].

[102] Octavia-Maria Sulea and Daniel Dichiu. Automatic profiling of twitter users based on their tweets—notebook for pan at clef 2015. In Cappellato et al. [244].

[103] Maite Gimenez, Irazú-Hernández, and Ferran Plá. Segmenting target audiences: authomatic author profiling using tweets—notebook for pan at clef 2015. In Cappellato et al. [244].

[104] Alonso Palomino-Garibay, Adolfo T. Camacho-González, Ricardo A. Fierro-Villaneda, Irazú Hernández-Farias, Davide Buscaldi, and Ivan V. Meza-Ruiz. A random forest approach for authorship profiling. In Cappellato et al. [244].

[105] Lesly Miculicich. Statistical learning methods for profiling analysis—notebook for pan at clef 2015. In Cappellato et al. [244].

[106] Ifrah Pervaz, Iqra Ameer, Abdul Sittar, Rao Muhammad, and Adeel Nawab. Identification of author personality traits using stylistic features—notebook for pan at clef 2015. In Cappellato et al. [244].

[107] Piotr Przybyla and Pawel Teisseyre. What do your look-alikes say about you? explotting strong and weak similarities for author profiling—notebook for pan at clef 2015. In Cappellato et al. [244].

[108] Caitltin McCollister, Bo Luo, and Shu Huang. Building topic models to predict author attributes from twitter messages—notebook for pan at clef 2015. In Cappellato et al. [244].

[109] Mirco Kocher. Unine at clef 2015: author profiling—notebook for pan at clef 2015. In Cappellato et al. [244].

[110] Saif M Mohammad. # emotional tweets. In *Proceedings of the first joint conference on lexical and computational semantics-volume 1: proceedings of the main conference and the shared task, and Volume 2: proceedings of the sixth international workshop on semantic evaluation*, pages 246–255. Association for Computational Linguistics, 2012.

[111] Harold D Lasswell and J Zvi Namenwirth. The lasswell value dictionary. *New Haven*, 1969.

[112] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. The general inquirer: a computer approach to content analysis. 1966.

[113] Alison Huettner and Pero Subasic. Fuzzy typing for document management. *ACL 2000 companion volume: tutorial abstracts and demonstration notes*, pages 26–27, 2000.

[114] Cynthia M Whissel. The dictionary of affect in language, emotion: theory, research and experience: vol. 4, the measurement of emotions, r. *Plutchik and H. Kellerman, Eds., New York: Academic*, 1989.

[115] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology, university of Florida, 1999.

[116] Carlo Strapparava, Alessandro Valitutti, et al. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086, 2004.

[117] Saif M Mohammad and Peter D Turney. Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics, 2010.

[118] James W Pennebaker, Roger J Booth, and Martha E Francis. Linguistic inquiry and word count: Liwc2007 - operator's manual. In *Austin, TX: LIWC.net*, 2007.

[119] Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. The spanish adaptation of anew (affective norms for english words). *Behavior research methods*, 39(3):600–605, 2007.

[120] Grigori Sidorov, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, Noé Castro-Sánchez, Francisco Velásquez, I Díaz Rangel, Sergio Suárez Guerra, Alejandro Trevino, and Juan Gordon. Empirical study of opinion mining in spanish tweets. *Invited paper. MICAI*, 2012.

[121] Paul Eckman. Universal and cultural differences in facial expression of emotion. In *Nebraska symposium on motivation*, volume 19, pages 207–284. University of Nebraska Press Lincoln, 1972.

[122] Alexander Osherenko and Elisabeth André. Lexical affect sensing: are affect dictionaries necessary to analyze affect? In *Affective computing and intelligent interaction*, pages 230–241. Springer, 2007.

[123] François-Régis Chaumartin. Upar7: a knowledge-based system for headline sentiment tagging. In *Proceedings of the 4th international workshop on semantic evaluations*, pages 422–425. Association for Computational Linguistics, 2007.

[124] Zornitsa Kozareva, Borja Navarro, Sonia Vázquez, and Andrés Montoyo. Ua-zbsa: a headline emotion classification through web information. In *Proceedings of the 4th international workshop on semantic evaluations*, pages 334–337. Association for Computational Linguistics, 2007.

[125] Phil Katz, Matthew Singleton, and Richard Wicentowski. Swat-mp: the semeval-2007 systems for task 5 and task 14. In *Proceedings of the 4th international workshop on semantic evaluations*, pages 308–313. Association for Computational Linguistics, 2007.

[126] Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: affective text. In *Proceedings of the 4th international workshop on semantic evaluations*, pages 70–74. Association for Computational Linguistics, 2007.

[127] Clark Davidson Elliott. The affective reasoner: a process model of emotions in a multi-agent system. 1992.

[128] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on empirical methods in natural language processing*, volume 10, pages 79–86. Association for Computational Linguistics, 2002.

[129] Hugo Liu, Henry Lieberman, and Ted Selker. Automatic affective feedback in an email browser. *In MIT media lab software agents group', ACM*, 2002.

[130] K Dhaliwal, M Gillies, J O'connor, A Oldroyd, D Robertson, and L Zhang. Facilitating online role-play using emotionally expressive characters. artificial and ambient intelligence. In *Proceedings of the AISB annual convention*, pages 179–186, 2007.

[131] D Garcıa and F Alıas. Emotion identification from text using semantic disambiguation. *Procesamiento del lenguaje natural*, 50:75–82, 2008.

[132] Futoshi Sugimoto and Masahide Yoneyama. A method for classifying emotion of text based on emotional dictionaries for emotional reading. In *Artificial intelligence and applications*, pages 91–96, 2006.

[133] Ismael Díaz Rangel. *Detección de afectividad en texto en espanol basada en el contexto lingıstico para sıntesis de Voz*. PhD thesis, Tesis doctoral. Instituto politécnico nacional. México, 2013.

[134] Saif M Mohammad and Tony Wenda Yang. Tracking sentiment in mail: how genders differ on emotional axes. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis (ACL-HLT 2011*, pages 70–79, 2011.

[135] Nargis Pervin, Fang Fang, Anindya Datta, Kaushik Dutta, and Debra Vandermeer. Fast, scalable, and context-sensitive detection of trending topics in microblog post streams. *ACM transactions on management information systems (TMIS)*, 3(4):19, 2013.

[136] Krisztian Balog, Gilad Mishne, and Maarten De Rijke. Why are they excited?: identifying and explaining spikes in blog mood levels. In *Proceedings of the eleventh conference of the european chapter of the association for computational linguistics: posters & demonstrations*, pages 207–210. Association for Computational Linguistics, 2006.

[137] Giuseppe Amodeo, Roi Blanco, and Ulf Brefeld. Hybrid models for future event prediction. In *Proceedings of the 20th ACM international conference on information and knowledge management*, pages 1981–1984. ACM, 2011.

[138] Sören Volgmann, Francisco M Rangel Pardo, Oliver Niggemann, and Paolo Rosso. Emotional trends in social media - a state space approach. In *21st. european conference on artificial intelligence (ECAI'14)*, pages 1123–1124, 2014.

[139] William L Brogan. *Modern control theory*. Pearson education india, 1974.

[140] Peter J Brockwell and Richard A Davis. *Introduction to time series and forecasting*. Springer science & business media, 2006.

[141] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of fluids engineering*, 82(1):35–45, 1960.

[142] Francisco Rangel and Paolo Rosso. On the identification of emotions and authors' gender in facebook comments on the basis of their writing style. In *ESSEM workshop on emotion and sentiment in social and expressive media, AIxIA, CEUR-WS.org*, volume 1096, pages 34–46, 2013.

[143] Leslie Greenberg. Emociones: una guía interna. *Ed. Descleé de Brouwer*, 2000.

[144] Andrew Ortony and Terence J Turner. What's basic about basic emotions? *Psychological review*, 97(3):315, 1990.

[145] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

[146] Francisco Rangel, Irazú Hernández, Paolo Rosso, and Antonio Reyes. Emotions and irony per gender in facebook. In *Workshop on emotion, social signals, sentiment & linked open data (ES3LOD)*, pages 68–73, 2014.

[147] Cornelis Joost Van Rijsbergen. Foundation of evaluation. *Journal of documentation*, 30(4):365–373, 1974.

[148] Byron C Wallace. Computational irony: a survey and new perspectives. *Artificial intelligence review*, 43(4):467–483, 2013.

[149] Herbert P Grice. *Logic and conversation*. na, 1970.

[150] Salvatore Attardo. Irony as relevant inappropriateness. *Journal of pragmatics*, 32(6):793–826, 2000.

[151] Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. Overview of the evalita 2014 sentiment polarity classification task. *Proceedings of the 4th evaluation campaign of natural language processing and speech tools for Italian (EVALITA'14). Pisa, Italy*, 2014.

[152] Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, Antonio Reyes, and John Barnden. Semeval-2015 task 11: sentiment analysis of figurative language in twitter. In *International workshop on semantic evaluation (SemEval-2015)*, 2015.

[153] Antonio Reyes and Paolo Rosso. Making objective decisions from subjective data: detecting irony in customer reviews. *Decision support systems*, 53(4):754–760, 2012.

[154] Antonio Reyes, Paolo Rosso, and Tony Veale. A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268, 2013.

[155] Cristina Bosco, Viviana Patti, and Andrea Bolioli. Developing corpora for sentiment analysis: the case of irony and senti-tuit. *IEEE intelligent systems*, (2):55–63, 2013.

[156] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[157] William Sealy Gosset. Student's. *Collected papers. london: biometrika office, university college*, 1943.

[158] Francisco Rangel and Paolo Rosso. On the impact of emotions on author profiling. *Information processing & management*, 52(1):73–92, 2016.

[159] Yi Liu. Graph-based learning models for information retrieval: a survey, 2006.

[160] Marc Franco-Salvador, Paolo Rosso, and Roberto Navigli. A knowledge-based representation for cross-language document retrieval and categorization. In *Proceedings of EACL*, pages 414–423, 2014.

[161] David Pinto, Helena Gómez-Adorno, Darnes Vilariño, and Vivek Kumar Singh. A graph-based multi-level linguistic representation for document understanding. *Pattern recognition letters*, 41: 93–102, 2014.

[162] Rada Mihalcea and Dragomir Radev. *Graph-based natural language processing and information retrieval*. Cambridge University Press, 2011.

[163] Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics, 1997.

[164] Janyce Wiebe and Rada Mihalcea. Word sense and subjectivity. In *Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics*, pages 1065–1072. Association for Computational Linguistics, 2006.

[165] Diego R Amancio, Renato Fabbri, Osvaldo N Oliveira Jr, Maria GV Nunes, and Luciano da F Costa. Distinguishing between positive and negative opinions with complex network features. In *Proceedings of the 2010 workshop on graph-based methods for natural language processing*, pages 83–87. Association for Computational Linguistics, 2010.

[166] Swapna Somasundaran, Galileo Namata, Lise Getoor, and Janyce Wiebe. Opinion graphs for polarity and discourse classification. In *Proceedings of the 2009 workshop on graph-based methods for natural language processing*, pages 66–74. Association for Computational Linguistics, 2009.

[167] EAGLES (1996a). Recommendations for the morphosyntactic annotation of corpora. In *Eag-tcwg-mac/r, ILC-CNR, Pisa*, 1996.

[168] EAGLES (1996b). Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. a common proposal and applications to european languages. In *Eag-clwg-morphsyn/r, ILC-CNR, Pisa*, 1996.

[169] Beth Levin. *English verb classes and alternations: a preliminary investigation*. University of chicago press, 1993.

[170] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[171] John Aldrich. Eigenvalue, eigenfunction, eigenvector, and related terms. *Earliest known uses of some of the words of mathematics*, 2006.

[172] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[173] Jérémie Bouttier, Philippe Di Francesco, and Emmanuel Guitter. Geodesic distance in planar graphs. *Nuclear physics b*, 663(3):535–567, 2003.

[174] Oystein Ore and Yystein Ore. *Theory of graphs*, volume 38. American mathematical society Providence, 1962.

[175] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[176] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *nature*, 393(6684):440–442, 1998.

[177] Matthieu Latapy. Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theoretical computer science*, 407(1):458–473, 2008.

[178] Ulrik Brandes. A faster algorithm for betweenness centrality*. *Journal of mathematical sociology*, 25(2):163–177, 2001.

[179] Stephen P Borgatti. Centrality and network flow. *Social networks*, 27(1):55–71, 2005.

[180] Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of mathematical sociology*, 2(1):113–120, 1972.

[181] Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420, 1997.

[182] Andrew J Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information theory, IEEE transactions on*, 13(2):260–269, 1967.

[183] Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, pages 1554–1563, 1966.

[184] Witold Litwin. Linear hashing: a new tool for file and table addressing. In *VLDB*, volume 80, pages 1–3, 1980.

[185] Lester R Ford Jr. Network flow theory. Technical report, DTIC document, 1956.

[186] Richard Bellman. On a routing problem. Technical report, DTIC document, 1956.

[187] RV Mises and Hilda Pollaczek-Geiringer. Praktische verfahren der gleichungsauflösung. *Journal of applied mathematics and mechanics/zeitschrift für angewandte mathematik und mechanik (ZAMM)*, 9(1):58–77, 1929.

[188] Francisco Rangel and Paolo Rosso. On the multilingual and genre robustness of emographs for author profiling in social media. In *6th international conference of CLEF on experimental IR meets multilinguality, multimodality, and interaction*, pages 274–280. Springer-Verlag, LNCS(9283), 2015.

[189] E Mark Gold. Language identification in the limit. *Information and control*, 10(5):447–474, 1967.

[190] William B Cavnar, John M Trenkle, et al. N-gram-based text categorization. *Ann arbor mi*, 48113 (2):161–175, 1994.

[191] Ted Dunning. *Statistical identification of language*. Computing research laboratory, new mexico state university, 1994.

[192] Marco Lui and Timothy Baldwin. Cross-domain feature selection for language identification. In *In proceedings of 5th international joint conference on natural language processing*. Citeseer, 2011.

[193] Marco Lui and Timothy Baldwin. langid. py: an off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics, 2012.

[194] Canasai Kruengkrai, Prapass Srichaivattana, Virach Sornlertlamvanich, and Hitoshi Isahara. Language identification based on string kernels. In *Communications and information technology, 2005. ISCIT 2005. IEEE international symposium on*, volume 2, pages 926–929. IEEE, 2005.

[195] Baden Hughes, Timothy Baldwin, SG Bird, Jeremy Nicholson, and Andrew MacKinlay. Reconsidering language identification for written language resources. 2006.

[196] Sylviane Granger, Estelle Dagneaux, Fanny Meunier, et al. The international corpus of learner english. handbook and cd-rom. 2002.

[197] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, volume 1, pages 180–189. Association for Computational Linguistics, 2011.

[198] Shin'ichiro Ishikawa. A new horizon in learner corpus studies: the aim of the icnale project. *Korea*, 404:89–168, 2011.

[199] Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. Toefl11: a corpus of non-native english. *ETS research report series*, 2013(2):i–15, 2013.

[200] Yukio Tono. International corpus of crosslinguistic interlanguage: project overview and a case study on the acquisition of new verb co-occurrence patterns. *Developmental and crosslinguistic perspectives in learner corpus research*, pages 27–46, 2012.

[201] Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the 8th workshop on innovative use of NLP for building educational applications*, pages 22–31, 2013.

[202] Shin Ishikawa. Vocabulary in interlanguage: a study on corpus of english essays written by asian university students (ceeaus). *Phraseology, corpus linguistics and lexicography*, pages 87–100, 2009.

[203] Mick Randall and Nicholas Groom. The buid arab learner corpus: a resource for studying the acquisition of l2 english spelling. In *Proceedings of the corpus linguistics conference (CL)*, 2009.

[204] Moshe Koppel, Jonathan Schler, and Kfir Zigdon. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining*, pages 624–628. ACM, 2005.

[205] Parham Tofighı, Cemal Köse, and Leila Rouka. Author's native language identification from web-based texts. 2012.

[206] Charles Carpenter Fries. *The structure of english: an introduction to the construction of English sentences*. Longman, 1973.

[207] Joel Tetreault, Daniel Blanchard, and Aoife Cahill. A report on the first native language identification shared ttask. In *Proceedings of the 8th workshop on innovative use of NLP for building educational applications*, pages 48–57. Citeseer, 2013.

[208] Julian Brooke and Graeme Hirst. Using other learner corpora in the 2013 nli shared task. In *Proceedings of the 8th workshop on innovative use of NLP for building educational applications*, pages 188–196, 2013.

[209] Serhiy Bykh and Detmar Meurers. Exploring syntactic features for native language identification: a variationist perspective on feature encoding and ensemble optimization. COLING, 2014.

[210] Preslav Nakov, Petya Osenova, and Cristina Vertan. *Proceedings of the EMNLP'2014 workshop on language technology for closely related languages and language variants*. Association for computational linguistics, Doha, Qatar, October 2014. URL http://www.aclweb.org/anthology/W14/W14-42.

[211] Marcos Zampieri, Liling Tan, Nikola Ljubešic, and Jörg Tiedemann. A report on the dsl shared task 2014. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, pages 58–67, 2014.

[212] Marcos Zampieri, Liling Tan, Nikola Ljubešic, Jörg Tiedemann, and Preslav Nakov. Overview of the dsl shared task 2015. In *Joint workshop on language technology for closely related languages, varieties and dialects*, page 1, 2015.

[213] Cyril Goutte, Serge Léger, and Marine Carpuat. The nrc system for discriminating similar languages. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, pages 139–145, Dublin, Ireland, August 2014. Association for computational linguistics. URL http://www.aclweb.org/anthology/W14-5316.

[214] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[215] Shervin Malmasi and Mark Dras. Language identification using classifier ensembles. In *Proceedings of the joint workshop on language technology for closely related languages, varieties and dialects (LT4VarDial'15)*, page 35–43, 2015.

[216] Marcos Zampieri and Binyam Gebrekidan Gebre. Automatic identification of language varieties: the case of portuguese. In *The 11th conference on natural language processing (KONVENS)*, pages 233–237. Österreichischen Gesellschaft für Artificial Intelligende (ÖGAI), 2012.

[217] Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. Automatic identification of arabic language varieties and dialects in social media. *Proceedings of SocialNLP*, page 22, 2014.

[218] Wolfgang Maier and Carlos Gómez-Rodrıguez. Language variety identification in spanish tweets. *LT4CloseLang 2014*, page 25, 2014.

[219] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of workshop at international conference on learning representations*, 2013.

[220] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155, 2003.

[221] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *11th annual conference of the international speech communication association (INTERSPEECH'10)*, pages 1045–1048, 2010.

[222] Marc Franco-Salvador, Francisco Rangel, Paolo Rosso, Mariona Taulé, and M Antònia Martí. Language variety identification using distributed representations of words and documents. In *Experimental IR meets multilinguality, multimodality, and interaction*, pages 28–40. Springer, 2015.

[223] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems 26*, pages 3111–3119, 2013.

[224] Richard S Sutton and Andrew G Barto. *Reinforcement learning: an introduction*, volume 1. MIT press Cambridge, 1998.

[225] Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *Proceedings of the international workshop on artificial intelligence and statistics*, pages 246–252. Citeseer, 2005.

[226] Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The journal of machine learning research*, 13(1):307–361, 2012.

[227] Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. *Proceedings of the 29th international conference on machine learning (ICML'12)*, pages 1751–1758, 2012.

[228] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st international conference on machine learning*, 2014.

[229] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the international conference on empirical methods in natural language processing*, 2014.

[230] M. Antònia Martí, Manuel Bertran, Mariona Taulé, and Maria Salamó. Distributional approach based on syntactic dependencies for discovering constructions. Technical report, CLiC, Centre de Llenguatge i Computació, Departament de Lingüística General, Universitat de Barcelona, 2015.

[231] Francisco Rangel, Paolo Rosso, and Marc Franco-Salvador. A low dimensionality representation for language variety identification. In *17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing*. Springer-Verlag, LNCS(), 2016.

[232] Marc Franco-Salvador, Paolo Rosso, and Francisco Rangel. Distributed representations of words and documents for discriminating similar languages. In *Joint workshop on language technology for closely related languages, varieties and dialects*, page 11, 2015.

[233] Raül Fabra-Boluda, Francisco Rangel, and Paolo Rosso. Nlel upv autoritas participation at discrimination between similar languages (dsl) 2015 shared task. In *Joint workshop on language technology for closely related languages, varieties and dialects*, page 52, 2015.

[234] Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. Merging comparable data sources for the discrimination of similar languages: the dsl corpus collection. In *7th workshop on building and using comparable corpora building resources for machine translation research (BUCC'14)*, pages 6–10, Reykjavik, Iceland, 2014.

[235] Paul McNamee. Language identification: a solved problem suitable for undergraduate instruction. *Journal of computing sciences in colleges*, 20(3):94–101, 2005.

[236] Nakatani Shuyo. Language detection library for java, 2010. URL http://code.google.com/p/language-detection/.

[237] Alan D Baddeley. *The psychology of memory*. Basic books, 1976.

[238] Tommaso Fornaciari and Massimo Poesio. Automatic deception detection in italian court cases. *Artificial intelligence and law*, 21(3):303–340, 2013.

[239] Tommaso Fornaciari and Massimo Poesio. Decour: a corpus of deceptive statements in italian courts. In *LREC*, pages 1585–1590, 2012.

[240] Tommaso Fornaciari and Massimo Poesio. Identifying fake amazon reviews as learning from crowds. In *EACL*, pages 279–287, 2014.

[241] David C McClelland. Toward a theory of motive acquisition. *American psychologist*, 20(5):321, 1965.

[242] Pamela Forner, Roberto Navigli, and Dan Tufis, editors. *CLEF 2013 evaluation labs and workshop. CEUR-WS.org vol. 1179*, 2013.

[243] Linda Cappellato, Nicola Ferro, Martin Halvey, and Wessel Kraaij, editors. *CLEF 2014 evaluation labs and workshop. CEUR-WS.org vol. 1180*, 2014.

[244] Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San-Juan, editors. *CLEF 2015 labs and workshops, notebook papers. CEUR-WS.org vol. 1391*, 2015.

Apéndice I. Author Profiling en PAN 2014

A Significación estadística pareada de sistemas

A continuación se indica la codificación de los niveles de significación utilizados en las tablas:

Símbolo	Nivel de significación		
=	$p > 0.05$	$\sim$	no significativo
*	$0.05 \geq p > 0.01$	$\sim$	significativo
**	$0.01 \geq p > 0.001$	$\sim$	muy significativo
***	$p \leq 0.001$	$\sim$	altamente significativo

Tabla A1: Significación de la diferencia entre pares de sistemas para la identificación de edad en el corpus completo.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	***	***	***	***	***	***	***	***	=
baker			=	***	***	=	***	***	***	***	***
castillojuarez				***	***	*	***	***	***	***	***
liau					=	***	***	=	***	**	***
lopezmonroy						***	***	=	*	=	***
marquardt							***	***	***	***	***
mechti								***	***	***	***
shrestha									**	=	***
villenaroman										=	***
weren											***
baseline											

Tabla A2: Significación de la diferencia entre pares de sistemas para la identificación de edad en social media en inglés.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	***	***	***	**	*	***	***	***	***
baker			**	***	***	=	=	***	***	***	***
castillojuarez				***	***	=	***	***	***	***	***
liau					=	***	***	=	=	=	***
lopezmonroy						***	***	=	=	=	***
marquardt							***	***	***	***	***
mechti								***	***	***	***
shrestha									=	=	***
villenaroman										=	***
weren											***
baseline											

Tabla A3: Significación de la diferencia entre pares de sistemas para la identificación de edad en social media en español.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	***	***	***	***	***	***	***	***	***
baker			**	***	***	=	***	**	=	***	*
castillojuarez				***	***	***	=	***	***	***	=
liau					=	***	***	**	***	*	***
lopezmonroy						***	***	=	***	=	***
marquardt							***	***	=	***	*
mechti								***	***	***	**
shrestha									**	=	***
villenaroman										**	**
weren											***
baseline											

Tabla A4: Significación de la diferencia entre pares de sistemas para la identificación de edad en blogs en inglés.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	=	=	=	=	=	=	**	=
baker			=	=	=	=	=	=	=	=	**
castillojuarez				=	=	=	=	=	=	=	**
liau					=	=	=	=	=	=	**
lopezmonroy						=	*	=	=	=	***
marquardt							=	=	=	**	=
mechti								*	*	**	=
shrestha									=	=	**
villenaroman										=	**
weren											***
baseline											

Tabla A5: Significación de la diferencia entre pares de sistemas para la identificación de edad en blogs en español.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	***	***	***	***	***	***	***	***	**
baker			=	=	=	=	=	=	=	*	**
castillojuarez				=	*	*	=	*	*	=	=
liau					=	=	=	=	=	*	**
lopezmonroy						=	=	=	=	*	**
marquardt							=	=	=	***	***
mechti								=	=	=	=
shrestha									=	*	***
villenaroman										*	**
weren											=
baseline											

Tabla A6: Significación de la diferencia entre pares de sistemas para la identificación de edad en Twitter en inglés.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	***	*	=	=	***	=	=	=	*
baker			***	**	**	=	***	=	=	=	=
castillojuarez				***	***	***	***	***	***	***	***
liau					=	*	***	=	*	**	***
lopezmonroy						**	***	=	=	**	***
marquardt							***	=	=	=	=
mechti								***	***	***	***
shrestha									=	*	***
villenaroman										=	**
weren											=
baseline											

Tabla A7: Significación de la diferencia entre pares de sistemas para la identificación de edad en Twitter en español.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	=	***	***	***	***	***	***	***	***
baker			***	=	=	=	**	*	=	=	=
castillojuarez				***	***	***	***	***	***	***	***
liau					=	=	**	**	=	=	=
lopezmonroy						=	***	=	=	=	=
marquardt							**	=	=	=	=
mechti								***	**	***	**
shrestha									**	=	*
villenaroman										=	=
weren											=
baseline											

Tabla A8: Significación de la diferencia entre pares de sistemas para la identificación de edad en revisiones de hotel en inglés.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	***	***	=	***	***	***	***	*
baker			=	***	***	=	***	***	***	***	=
castillojuarez				***	***	=	***	***	***	***	*
liau					=	***	***	=	*	=	***
lopezmonroy						***	***	=	=	=	***
marquardt							***	***	***	***	*
mechti								***	***	***	***
shrestha									=	=	***
villenaroman										=	*
weren											***
baseline											

Tabla A9: Significación de la diferencia entre pares de sistemas para la identificación de sexo en el corpus completo.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	**	***	***	***	***	***	***	***	***
baker			*	***	***	***	=	***	***	***	***
castillojuarez				***	***	***	***	***	***	***	***
liau					***	***	***	**	***	**	***
lopezmonroy						***	***	=	=	=	*
marquardt							***	***	***	***	*
mechti								***	***	***	***
shrestha									=	=	*
villenaroman										=	*
weren											*
baseline											

Tabla A10: Significación de la diferencia entre pares de sistemas para la identificación de sexo en social media en inglés.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	=	=	=	=	=	=	=	=
baker			=	**	=	=	=	**	**	**	=
castillojuarez				**	=	=	=	**	***	*	=
liau					=	=	=	=	=	=	*
lopezmonroy						=	=	=	*	=	=
marquardt							=	=	=	=	=
mechti								=	*	=	=
shrestha									=	=	*
villenaroman										=	**
weren											*
baseline											

Tabla A11: Significación de la diferencia entre pares de sistemas para la identificación de sexo en social media en español.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	***	***	***	***	***	***	***	***	***
baker			=	***	***	***	**	***	*	***	***
castillojuarez				***	***	***	**	***	=	***	***
liau					=	=	***	*	***	**	=
lopezmonroy						=	*	=	**	=	=
marquardt							*	=	**	=	=
mechti								*	=	=	*
shrestha									**	=	=
villenaroman										**	***
weren											=
baseline											

Tabla A12: Significación de la diferencia entre pares de sistemas para la identificación de sexo en blogs en inglés.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	**	**	=	=	=	*	**	=
baker			=	=	*	=	=	=	*	=	=
castillojuarez				=	*	=	=	=	=	=	=
liau					=	*	=	=	=	=	=
lopezmonroy						**	=	=	=	=	=
marquardt							=	=	*	=	=
mechti								=	=	=	=
shrestha									=	=	=
villenaroman										=	=
weren											=
baseline											

Tabla A13: Significación de la diferencia entre pares de sistemas para la identificación de sexo en blogs en español.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	***	***	***	***	***	***	***	***	***
baker			=	=	=	=	=	=	=	=	=
castillojuarez				=	=	=	=	=	=	=	=
liau					=	=	=	=	=	=	=
lopezmonroy						=	=	=	=	=	=
marquardt							=	=	=	=	=
mechti								=	=	=	=
shrestha									=	=	=
villenaroman										=	=
weren											=
baseline											

Tabla A14: Significación de la diferencia entre pares de sistemas para la identificación de sexo en Twitter en inglés.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	***	***	**	=	=	**	=	=	**
baker			***	***	**	=	=	**	=	=	*
castillojuarez				***	***	***	***	***	***	***	***
liau					=	***	***	=	***	**	**
lopezmonroy						***	***	=	***	**	*
marquardt							=	*	=	=	=
mechti								*	=	=	=
shrestha									**	=	=
villenaroman										=	*
weren											=
baseline											

Tabla A15: Significación de la diferencia entre pares de sistemas para la identificación de sexo en Twitter en español.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	=	***	***	***	***	***	***	***	***
baker			***	=	=	=	=	*	=	=	=
castillojuarez				***	***	***	***	***	***	***	***
liau					=	=	=	=	=	=	=
lopezmonroy						=	=	=	=	=	=
marquardt							=	=	=	=	=
mechti								*	=	=	=
shrestha									=	=	**
villenaroman										=	=
weren											=
baseline											

Tabla A16: Significación de la diferencia entre pares de sistemas para la identificación de sexo en revisiones de hotel en inglés.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	***	***	***	=	***	***	***	***
baker			=	***	***	***	***	***	***	***	***
castillojuarez				***	***	*	=	***	***	***	***
liau					***	***	***	***	***	***	***
lopezmonroy						***	***	=	=	=	=
marquardt							***	***	***	***	***
mechti								***	***	***	***
shrestha									=	=	=
villenaroman										=	=
weren											=
baseline											

Tabla A17: Significación de la diferencia entre pares de sistemas para la identificación conjunta en el corpus completo.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	**	***	***	***	***	***	***	***	***
baker			*	***	***	***	=	***	***	***	***
castillojuarez				***	***	***	***	***	***	***	***
liau					***	***	***	**	***	***	***
lopezmonroy						***	***	=	=	=	**
marquardt							***	***	***	***	*
mechti								***	***	***	***
shrestha									=	=	**
villenaroman										=	*
weren											*
baseline											

Tabla A18: Significación de la diferencia entre pares de sistemas para la identificación conjunta en social media en inglés.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	=	=	=	=	=	=	=	=
baker			=	**	=	=	=	**	**	**	=
castillojuarez				**	=	=	=	**	**	*	=
liau					=	=	=	=	=	=	**
lopezmonroy						=	=	=	=	=	=
marquardt							=	=	=	=	=
mechti								=	*	=	=
shrestha									=	=	*
villenaroman										=	**
weren											*
baseline											

Tabla A19: Significación de la diferencia entre pares de sistemas para la identificación conjunta en social media en español.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	***	***	***	***	***	***	***	***	***
baker			=	***	***	***	**	***	=	***	***
castillojuarez				***	***	***	**	***	=	***	***
liau					=	=	***	=	***	**	=
lopezmonroy						=	*	=	**	=	=
marquardt							*	=	**	=	=
mechti								*	=	=	*
shrestha									**	=	=
villenaroman										**	***
weren											=
baseline											

Tabla A20: Significación de la diferencia entre pares de sistemas para la identificación conjunta en blogs en inglés.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	**	**	=	=	=	*	**	=
baker			=	=	*	=	=	=	*	=	=
castillojuarez				=	*	=	=	=	=	=	=
liau					=	*	=	=	=	=	=
lopezmonroy						**	=	=	=	=	=
marquardt							=	=	*	=	=
mechti								=	=	=	=
shrestha									=	=	=
villenaroman										=	=
weren											=
baseline											

Tabla A21: Significación de la diferencia entre pares de sistemas para la identificación conjunta en blogs en español.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	***	***	***	***	***	***	***	***	***
baker			=	=	=	=	=	=	=	=	=
castillojuarez				=	=	=	=	=	=	=	=
liau					=	=	=	=	=	=	=
lopezmonroy						=	=	=	=	=	=
marquardt							=	=	=	=	=
mechti								=	=	=	=
shrestha									=	=	=
villenaroman										=	=
weren											=
baseline											

Tabla A22: Significación de la diferencia entre pares de sistemas para la identificación conjunta en Twitter en inglés.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	***	***	***	=	=	**	=	=	**
baker			***	***	***	=	=	**	=	=	**
castillojuarez				***	***	***	***	***	***	***	***
liau					=	**	**	=	***	**	*
lopezmonroy						***	**	=	**	**	*
marquardt							=	*	=	=	=
mechti								*	=	=	=
shrestha									**	=	=
villenaroman										=	**
weren											=
baseline											

Tabla A23: Significación de la diferencia entre pares de sistemas para la identificación conjunta en Twitter en español.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		***	=	***	***	***	***	***	***	***	***
baker			***	=	=	=	=	*	=	=	=
castillojuarez				***	***	***	***	***	***	***	***
liau					=	=	=	=	=	=	=
lopezmonroy						=	=	=	=	=	=
marquardt							=	=	=	=	=
mechti								*	=	=	=
shrestha									=	=	*
villenaroman										=	=
weren											=
baseline											

Tabla A24: Significación de la diferencia entre pares de sistemas para la identificación conjunta de revisiones de hotel en inglés.

	ashok	baker	castillojuarez	liau	lopezmonroy	marquardt	mechti	shrestha	villenaroman	weren	baseline
ashok		=	=	***	***	***	=	***	***	***	***
baker			=	***	***	***	***	***	***	***	***
castillojuarez				***	***	**	=	***	***	***	***
liau					***	***	***	***	**	***	***
lopezmonroy						***	***	=	=	=	=
marquardt							***	***	***	***	***
mechti								***	***	***	***
shrestha									=	=	=
villenaroman										=	=
weren											=
baseline											

B Distancias en la identificación de edad

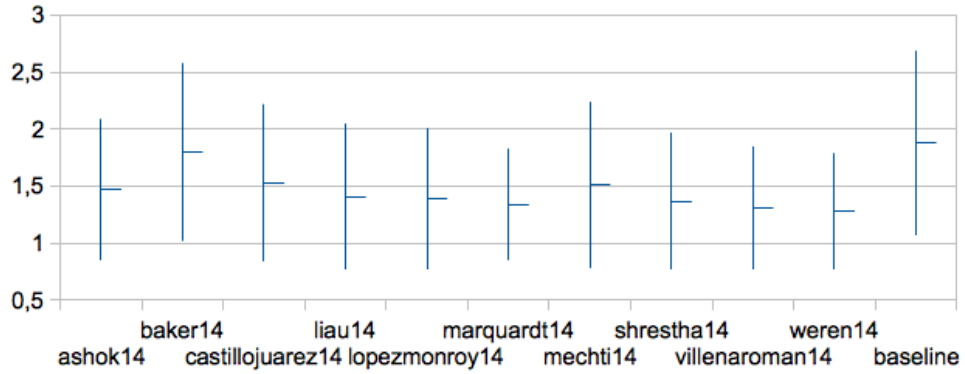

FIGURA B1: Distancias entre la predicción y la clase real en social media en inglés.

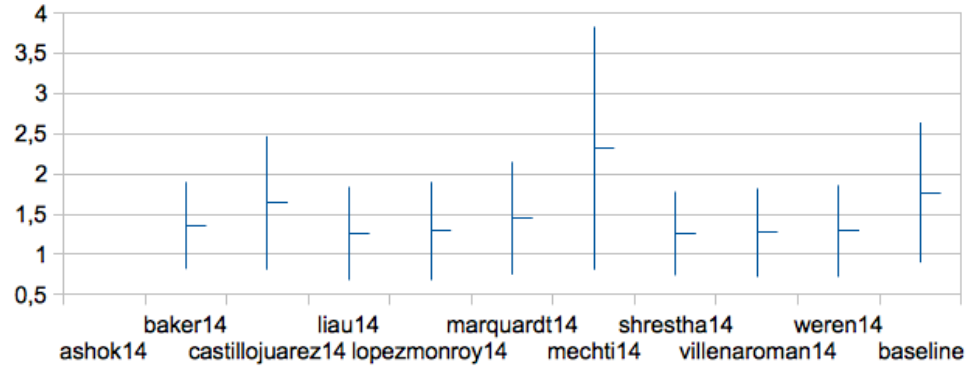

FIGURA B2: Distancias entre la predicción y la clase real en social media en español.

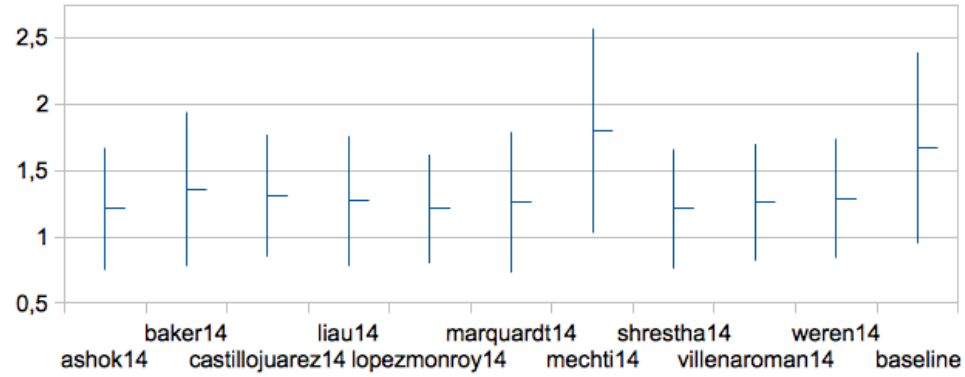

FIGURA B3: Distancias entre la predicción y la clase real en blogs en inglés.

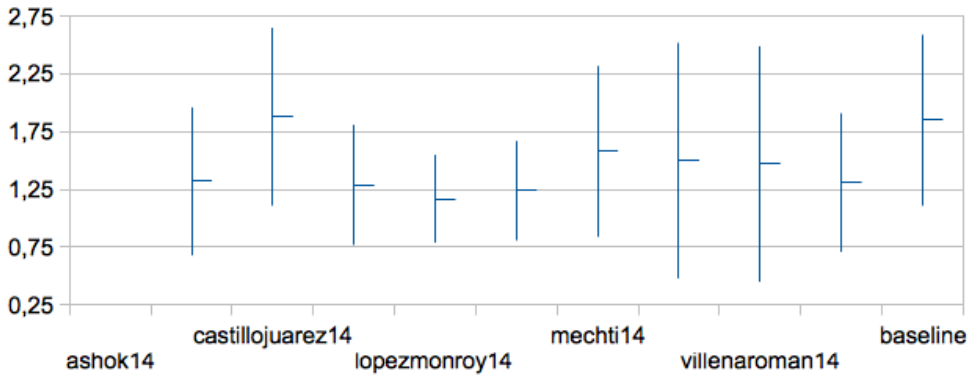

FIGURA B4: Distancias entre la predicción y la clase real en blogs en español.

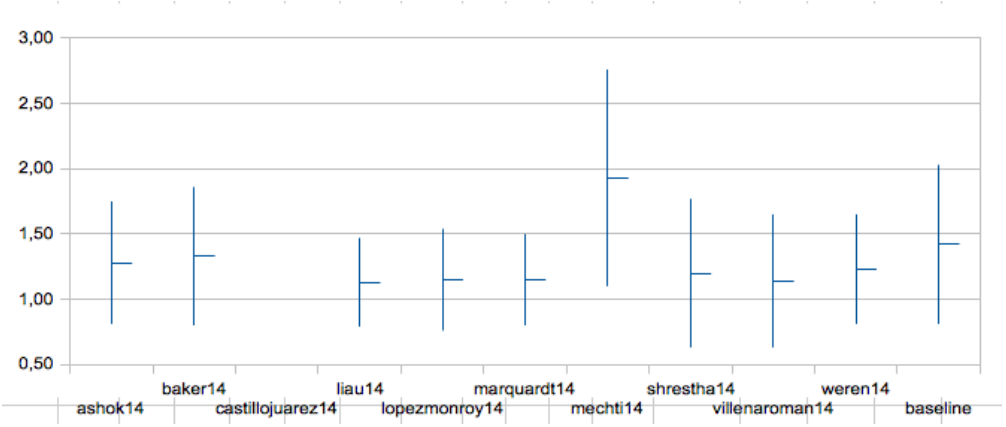

FIGURA B5: Distancias entre la predicción y la clase real en Twitter en inglés.

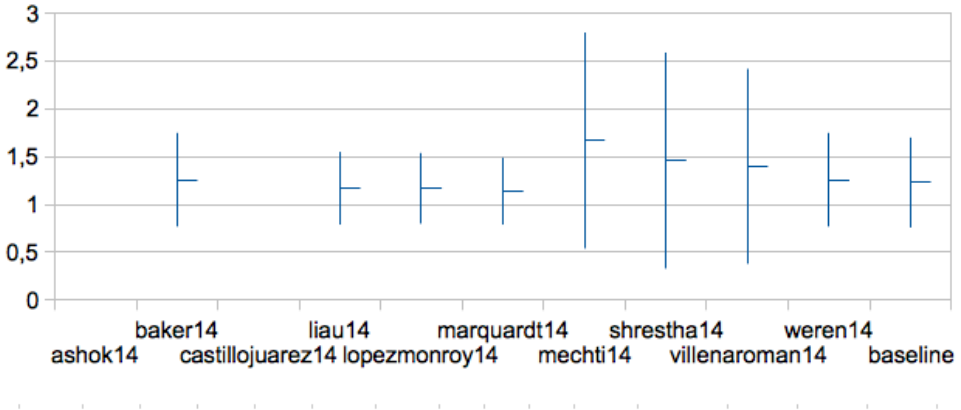

FIGURA B6: Distancias entre la predicción y la clase real en Twitter en español.

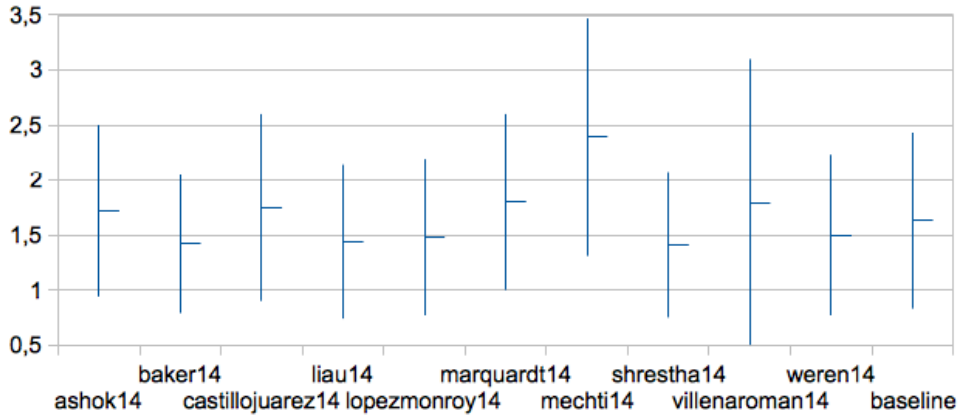

FIGURA B7: Distancias entre la predicción y la clase real en revisiones de hotel en inglés.

Apéndice II. Author Profiling en PAN 2015

C Significación estadística pareada entre sistemas

A continuación se indica la codificación de los niveles de significación utilizados en las tablas:

Símbolo	Nivel de significación		
=	$p > 0.05$	$\sim$	no significativo
*	$0.05 \geq p > 0.01$	$\sim$	significativo
**	$0.01 \geq p > 0.001$	$\sim$	muy significativo
***	$p \leq 0.001$	$\sim$	altamente significativo

	alvarezcarmona15	ameer15	arroju15	bartoli15	bayot15	cheema15	gimenez15	gonzalesgallardo15	grivas15	kocher15	maharjan15	mccollister15	mezaruiz15	miculicich15	nowson15	poulston15	sulea15	teisseyre15	weren15
alvarezcarmona15		***	***	***	***	***	***	=	=	***	***	***	***	**	***	***	***	***	***
ameer15			=	=	**	=	***	***	***	*	**	=	=	***	**	=	***	=	*
arroju15				=	***	=	***	***	***	=	=	=	**	**	*	=	*	=	=
bartoli15					***	=	***	***	***	=	=	=	*	***	*	=	**	=	=
bayot15						*	***	***	***	***	***	***	=	***	***	***	***	**	***
cheema15							***	***	***	***	**	=	=	***	***	*	***	=	**
gimenez15								***	***	***	***	***	***	***	***	***	***	***	***
gonzalesgallardo15									=	***	***	***	***	**	***	***	**	***	***
grivas15										***	***	***	***	***	***	***	***	***	***
kocher15											=	=	***	=	=	=	=	*	=
maharjan15												=	***	=	=	=	=	*	=
mccollister15													**	***	*	=	**	=	=
mezaruiz15														***	***	**	***	=	***
miculicich15															=	**	=	***	=
nowson15																=	=	**	=
poulston15																	*	=	=
sulea15																		**	=
teisseyre15																			=
weren15																			

Tabla A1: Significación de la diferencia entre pares de sistemas para la identificación de sexo en el corpus completo.

	alvarezcarmona15	ameer15	arroju15	bartoli15	bayot15	cheema15	gimenez15	gonzalesgallardo15	grivas15	kocher15	maharjan15	mccollister15	mezaruiz15	miculicich15	nowson15	poulston15	sulea15	teisseyre15	weren15
alvarezcarmona15		***	***	***	***	***	***	*	***	***	***	***	***	***	***	***	*	***	***
ameer15			=	=	=	=	=	**	*	=	=	=	***	=	*	=	***	=	=
arroju15				=	**	=	***	*	=	=	=	*	***	=	**	=	**	*	=
bartoli15					=	=	=	***	**	=	=	=	***	=	=	=	***	=	=
bayot15						=	=	***	***	**	=	=	**	*	=	**	***	=	**
cheema15							=	***	**	=	=	=	***	=	=	=	***	=	=
gimenez15								***	***	*	=	=	**	*	=	*	***	=	**
gonzalesgallardo15									=	**	***	***	***	**	***	*	=	***	=
grivas15										=	**	***	***	=	***	=	=	***	=
kocher15											=	=	***	=	**	=	**	=	=
maharjan15												=	***	=	=	=	***	=	=
mccollister15													***	=	=	=	***	=	=
mezaruiz15														***	*	***	***	***	***
miculicich15															**	=	**	=	=
nowson15																**	***	=	**
poulston15																	**	*	=
sulea15																		***	*
teisseyre15																			*
weren15																			

Tabla A2: Significación de la diferencia entre pares de sistemas para la identificación de edad en el corpus completo.

	alvarezcarmona15	ameer15	arroju15	bartoli15	bayot15	cheema15	gimenez15	gonzalesgallardo15	grivas15	kocher15	maharjan15	mccollister15	mezaruiz15	miculicich15	nowson15	poulston15	sulea15	teisseyre15	weren15
alvarezcarmona15		***	***	***	***	***	***	=	*	***	***	***	***	***	***	***	***	***	***
ameer15			=	=	***	=	***	***	***	=	=	=	***	*	=	=	***	=	*
arroju15				**	***	*	***	***	***	=	=	*	***	=	**	=	**	=	=
bartoli15					**	=	***	***	***	**	**	=	**	***	=	*	***	=	**
bayot15						***	***	***	***	***	***	***	=	***	**	***	***	***	***
cheema15							***	***	***	**	*	=	***	***	=	*	***	=	***
gimenez15								***	***	***	***	***	***	***	***	***	***	***	***
gonzalesgallardo15									=	***	***	***	***	**	***	***	*	***	**
grivas15										***	***	***	***	*	***	***	=	***	*
kocher15											=	**	***	=	***	=	*	=	=
maharjan15												*	***	=	***	=	*	=	=
mccollister15													***	***	=	*	***	=	**
mezaruiz15														***	**	***	***	***	***
miculicich15															***	=	=	**	=
nowson15																**	***	=	***
poulston15																	**	=	=
sulea15																		***	=
teisseyre15																			*
weren15																			

Tabla A3: Significación de la diferencia entre pares de sistemas para la identificación conjunta en el corpus completo.

D Significación estadística pareada de sistemas por idioma

A continuación se indica la codificación de los niveles de significación utilizados en las tablas:

Símbolo	Nivel de significación		
=	$p > 0.05$	$\sim$	no significativo
*	$0.05 \geq p > 0.01$	$\sim$	significativo
**	$0.01 \geq p > 0.001$	$\sim$	muy significativo
***	$p \leq 0.001$	$\sim$	altamente significativo

	alvarezcarmona15	ameer15	arroju15	bartoli15	bayot15	cheema15	gimenez15	gonzalesgallardo15	grivas15	kocher15	maharjan15	mccollister15	mezaruiz15	miculicich15	nowson15	poulston15	sulea15	teisseyre15	weren15
alvarezcarmona15		***	*	***	***	***	***	=	=	**	**	**	***	=	*	***	*	=	*
ameer15			=	=	**	=	***	***	***	=	=	=	**	*	=	=	=	**	=
arroju15				*	***	***	***	=	*	=	=	=	***	=	=	=	=	=	=
bartoli15					*	=	***	***	***	=	=	=	*	*	*	=	*	***	=
bayot15						=	***	***	***	***	***	***	=	***	***	**	***	***	***
cheema15							***	***	***	*	*	**	=	***	***	=	**	***	**
gimenez15								***	***	***	***	***	***	***	***	***	***	***	***
gonzalesgallardo15									=	**	*	**	***	=	=	***	=	=	*
grivas15										**	**	**	***	=	*	***	*	=	*
kocher15											=	=	***	=	=	=	=	**	=
maharjan15												=	***	=	=	=	=	=	=
mccollister15													***	=	=	=	=	*	=
mezaruiz15														***	***	**	***	***	***
miculicich15															=	=	=	=	=
nowson15																=	=	=	=
poulston15																	=	**	=
sulea15																		=	=
teisseyre15																			=
weren15																			

Tabla B1: Significación de la diferencia entre pares de sistemas para la identificación de sexo en inglés.

	alvarezcarmona15	ameer15	arroju15	bartoli15	bayot15	cheema15	gimenez15	gonzalesgallardo15	grivas15	kocher15	maharjan15	mccollister15	mezaruiz15	miculicich15	nowson15	poulston15	sulea15	teisseyre15	weren15
alvarezcarmona15		**	***	*	***	***	***	=	*	***	***	***	***	***	***	**	=	*	**
ameer15			=	=	**	=	**	=	=	=	=	=	***	=	***	=	=	=	=
arroju15				=	*	=	*	*	=	=	=	=	***	=	***	=	**	=	=
bartoli15					***	=	**	=	=	=	*	=	***	*	***	=	=	=	=
bayot15						=	=	***	***	=	=	**	**	=	=	**	***	***	**
cheema15							=	*	=	=	=	=	***	=	**	=	**	=	=
gimenez15								***	***	=	=	**	**	=	=	**	***	***	*
gonzalesgallardo15									=	**	**	=	***	***	***	=	=	=	=
grivas15										*	*	=	***	**	***	=	=	=	=
kocher15											=	=	***	=	**	=	**	*	=
maharjan15												=	***	=	**	=	**	*	=
mccollister15													***	=	***	=	=	=	=
mezaruiz15														***	=	***	***	***	***
miculicich15															**	*	***	**	=
nowson15																***	***	***	***
poulston15																	=	=	=
sulea15																		=	*
teisseyre15																			=
weren15																			

Tabla B2: Significación de la diferencia entre pares de sistemas para la identificación de edad en inglés.

	alvarezcarmona15	ameer15	arroju15	bartoli15	bayot15	cheema15	gimenez15	gonzalesgallardo15	grivas15	kocher15	maharjan15	mccollister15	mezaruiz15	miculicich15	nowson15	poulston15	sulea15	teisseyre15	weren15
alvarezcarmona15		***	***	***	***	***	***	=	=	***	***	***	***	***	***	***	*	=	**
ameer15			=	=	***	=	***	***	**	=	=	=	***	=	*	=	*	**	=
arroju15				=	***	*	***	***	**	=	=	=	***	=	***	=	=	*	=
bartoli15					***	=	***	***	***	=	=	=	***	=	=	=	**	***	=
bayot15						***	***	***	***	***	***	***	=	***	*	***	***	***	***
cheema15							***	***	***	=	=	=	***	*	=	=	***	***	*
gimenez15								***	***	***	***	***	***	***	***	***	***	***	***
gonzalesgallardo15									=	***	***	***	***	***	***	***	=	=	**
grivas15										***	**	***	***	**	***	**	=	=	*
kocher15											=	=	***	=	**	=	=	**	=
maharjan15												=	***	=	**	=	=	*	=
mccollister15													***	=	**	=	*	**	=
mezaruiz15														***	**	***	***	***	***
miculicich15															**	=	=	*	=
nowson15																**	***	***	**
poulston15																	=	*	=
sulea15																		=	=
teisseyre15																			=
weren15																			

Tabla B3: Significación de la diferencia entre pares de sistemas para la identificación conjunta en inglés.

	alvarezcarmona15	ameer15	arroju15	bartoli15	bayot15	cheema15	gimenez15	gonzalesgallardo15	grivas15	kocher15	maharjan15	mccollister15	mezaruiz15	miculicich15	nowson15	poulston15	sulea15	teisseyre15	weren15
alvarezcarmona15		***	***	*	***	**	***	=	=	**	***	***	**	=	***	**	*	***	**
ameer15			=	*	=	*	***	**	***	=	=	=	*	***	=	*	**	=	*
arroju15				=	=	=	***	*	**	=	=	=	=	**	=	=	*	*	=
bartoli15					***	=	***	=	=	=	=	***	=	=	=	=	=	***	=
bayot15						**	***	***	***	**	*	=	**	***	=	***	***	=	***
cheema15							***	=	*	=	=	*	=	=	=	=	=	***	=
gimenez15								***	***	***	***	***	***	***	***	***	***	***	***
gonzalesgallardo15									=	=	=	***	=	=	*	=	=	***	=
grivas15										**	**	***	*	=	**	=	=	***	*
kocher15											=	=	=	*	=	=	=	***	=
maharjan15												=	=	*	=	=	=	**	=
mccollister15													*	***	=	*	**	=	*
mezaruiz15														=	=	=	=	***	=
miculicich15															**	=	=	***	=
nowson15																=	=	**	=
poulston15																	=	***	=
sulea15																		***	=
teisseyre15																			***
weren15																			

Tabla B4: Significación de la diferencia entre pares de sistemas para la identificación de sexo en español.

	alvarezcarmona15	ameer15	arroju15	bartoli15	bayot15	cheema15	gimenez15	gonzalesgallardo15	grivas15	kocher15	maharjan15	mccollister15	mezaruiz15	miculicich15	nowson15	poulston15	sulea15	teisseyre15	weren15
alvarezcarmona15		***	*	***	***	***	***	=	**	*	***	***	***	=	*	***	=	***	=
ameer15			=	=	=	=	=	**	*	*	=	=	=	*	=	=	**	*	*
arroju15				***	=	=	*	=	=	=	=	***	**	=	=	=	=	***	=
bartoli15					*	*	=	***	***	***	*	=	=	***	**	*	***	=	***
bayot15						=	=	*	*	=	=	=	=	*	=	=	*	**	=
cheema15							=	*	*	=	=	=	=	*	=	=	**	**	*
gimenez15								**	*	=	=	=	=	*	=	=	**	**	*
gonzalesgallardo15									=	=	*	***	**	=	=	**	=	***	=
grivas15										=	=	***	**	=	=	=	=	***	=
kocher15											=	*	*	=	=	=	=	***	=
maharjan15												=	=	=	=	=	*	**	=
mccollister15													=	***	*	=	***	*	**
mezaruiz15														***	*	=	***	*	**
miculicich15															=	=	=	***	=
nowson15																=	=	***	=
poulston15																	**	**	=
sulea15																		***	=
teisseyre15																			***
weren15																			

Tabla B5: Significación de la diferencia entre pares de sistemas para la identificación de edad en español.

	alvarezcarmona15	ameer15	arroju15	bartoli15	bayot15	cheema15	gimenez15	gonzalesgallardo15	grivas15	kocher15	maharjan15	mccollister15	mezaruiz15	miculicich15	nowson15	poulston15	sulea15	teisseyre15	weren15
alvarezcarmona15		***	***	***	***	***	***	=	=	**	***	***	***	=	***	***	*	***	*
ameer15			=	=	=	=	***	***	***	*	=	=	=	**	=	=	***	**	**
arroju15				=	=	=	***	***	***	*	=	*	=	**	=	=	**	***	*
bartoli15					=	=	***	***	***	***	**	=	=	***	=	**	***	=	***
bayot15						=	***	***	***	***	*	=	=	***	=	**	***	*	***
cheema15							***	***	***	*	=	*	=	**	=	=	**	***	*
gimenez15								***	***	***	***	***	***	***	***	***	***	***	***
gonzalesgallardo15									=	=	*	***	***	=	***	**	=	***	=
grivas15										=	*	***	***	=	**	*	=	***	=
kocher15											=	***	**	=	=	=	=	***	=
maharjan15												**	=	*	=	=	=	***	=
mccollister15													=	***	*	**	***	=	***
mezaruiz15														***	=	=	***	**	**
miculicich15															**	=	=	***	=
nowson15																=	**	***	*
poulston15																	=	***	=
sulea15																		***	=
teisseyre15																			***
weren15																			

Tabla B6: Significación de la diferencia entre pares de sistemas para la identificación conjunta en español.

	alvarezcarmona15	ameer15	arroju15	bartoli15	bayot15	cheema15	gimenez15	gonzalesgallardo15	grivas15	kocher15	maharjan15	mccollister15	mezaruiz15	miculicich15	nowson15	poulston15	sulea15	teisseyre15	weren15
alvarezcarmona15		=	=	*	=	=	***	=	=	=	=	=	=	=	=	=	=	*	=
ameer15			=	=	=	=	***	**	*	=	=	=	=	=	=	=	=	=	=
arroju15				=	=	=	***	**	*	*	=	=	=	=	=	=	=	=	=
bartoli15					=	=	***	***	**	**	=	=	=	=	*	*	=	=	=
bayot15						=	***	**	*	=	=	=	=	=	*	=	=	=	=
cheema15							***	*	*	=	=	=	=	=	*	=	=	=	=
gimenez15								***	***	***	***	***	***	***	***	***	***	***	***
gonzalesgallardo15									=	=	*	*	**	*	=	=	*	***	*
grivas15										=	=	*	*	=	=	=	=	***	*
kocher15											=	*	=	=	=	=	=	*	=
maharjan15												=	=	=	=	=	=	=	=
mccollister15													=	=	*	=	=	=	=
mezaruiz15														=	**	=	=	=	=
miculicich15															=	=	=	=	=
nowson15																=	=	**	*
poulston15																	=	*	=
sulea15																		=	=
teisseyre15																			=
weren15																			

Tabla B7: Significación de la diferencia entre pares de sistemas para la identificación de sexo en italiano.

	alvarezcarmona15	ameer15	arroju15	bartoli15	bayot15	cheema15	gimenez15	gonzalesgallardo15	grivas15	kocher15	maharjan15	mccollister15	mezaruiz15	miculicich15	nowson15	poulston15	sulea15	teisseyre15	weren15
alvarezcarmona15		**	***	*	***	**	***	=	=	=	=	=	**	=	=	***	=	**	*
ameer15			=	=	=	=	***	**	***	=	=	=	=	=	=	=	*	=	=
arroju15				=	=	=	***	***	***	**	=	*	=	*	*	=	*	=	=
bartoli15					=	=	***	*	**	=	=	=	=	=	=	=	=	=	=
bayot15						=	***	**	***	*	=	*	=	*	=	=	*	=	=
cheema15							***	***	***	*	*	**	=	*	*	=	**	=	=
gimenez15								***	***	***	***	***	***	***	***	***	***	***	***
gonzalesgallardo15									=	=	=	=	***	=	=	**	=	**	**
grivas15										=	=	=	***	=	*	***	=	**	**
kocher15											=	=	*	=	=	*	=	=	=
maharjan15												=	=	=	=	=	=	=	=
mccollister15													**	=	=	**	=	=	=
mezaruiz15														*	*	=	*	=	=
miculicich15															=	*	=	=	=
nowson15																*	=	=	=
poulston15																	*	=	=
sulea15																		*	=
teisseyre15																			=
weren15																			

Tabla B8: Significación de la diferencia entre pares de sistemas para la identificación de sexo en holandés.