UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

Universitat Politècnica de València

Departamento de Sistemas Informáticos y Computación

El impacto de las emociones en el análisis de la polaridad en textos con lenguaje figurado en Twitter

Proyecto Final de Máster

[Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital]

Autora:
Mª Amparo Escortell Pérez

Director:
Paolo Rosso

Valencia, Julio 2017

UNIVERSITAT POLITÈCNICA DE VALENCIA

Resumen

Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

En los últimos años hemos visto como el auge de la Web 2.0 y los medios sociales han provocado que los usuarios tomen cada vez más protagonismo en Internet, siendo éstos una fuente de generación de información que aumenta día tras día. Centrándonos en la red social Twitter, uno de los retos más complejos a los que se enfrenta el Procesamiento de Lenguaje Natural es el de determinar la polaridad de un tweet (positiva, negativa o neutra) cuando en él aparece lenguaje figurado.

Este trabajo presenta un estudio exhaustivo sobre la capacidad de distintos recursos léxicos de emociones para analizar la polaridad de un conjunto de datos extraídos de Twitter, detallando el impacto de cada uno de los recursos sobre distintas formas de lenguaje figurado como pueden ser la ironía y el sarcasmo que encontramos profusamente en este corpus.

Partimos de la hipótesis de que no todos los recursos disponibles favorecen la detección de la polaridad en igual medida, por lo tanto llevamos a cabo una serie de experimentos para evaluar cómo afectan diferentes recursos léxicos de emociones tanto en el lenguaje figurado como en el lenguaje literal. Nuestra metodología se llevará a cabo en dos fases: en la primera de ellas estudiaremos el impacto de los recursos léxicos sobre el entrenamiento de clasificadores que predigan la polaridad del conjunto completo de tweets de la tarea 11 de SemEval2015. En la segunda, evaluaremos en detalle el impacto de cada uno de los recursos para las diferentes tipologías del lenguaje figurado presentes en este corpus.

Los resultados obtenidos muestran indicios que apuntan a que la inclusión de información relativa a las emociones ayuda a clasificar correctamente la polaridad tanto a nivel global como a nivel del lenguaje figurado o literal. Por ello, puede ser de gran importancia desarrollar técnicas capaces de representar la información de manera que sea posible clasificar el sentimiento que el usuario intenta transmitir en un texto.

UNIVERSITAT POLITÈCNICA DE VALENCIA

Abstract

Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

During the recent years we have seen how the rise of Web 2.0 and social media have caused users to take more and more prominence on the Internet, being a source of information generation that increases day by day. Focusing on the social network Twitter, one of the most complex challenges facing Natural Language Processing is to determine the polarity of a tweet (positive, negative or neutral) when in it appears figurative language.

This work presents an exhaustive study about the capacity of different lexical resources of emotions to analyze the polarity of a set of data extracted from Twitter, detailing the impact of each one of the resources about different forms of figurative language such as irony and sarcasm we found profusely in this corpus.

We start from the hypothesis that not all the available resources favor the detection of polarity with an equal measure, therefore we carry out a series of experiments to evaluate how they affect different lexical resources of emotions in the figurative language as in the literal one. Our methodology will be carried out in two phases: in the first one, we will study the impact of lexical resources on the training of classifiers that predict the polarity of the complete set of tweets of task 11 of SemEval2015. In the second, we will evaluate in detail the impact of each of the resources for the different typologies of figurative language present in this corpus.

The obtained results show indications that the inclusion of information related to emotions helps to correctly classify polarity both globally and at the level of figurative or literal language. Therefore, it may be of big importance to develop techniques capable of representing the information in such a way that it is possible to classify the feeling that the user tries to transmit in a text.

Agradecimientos

Me gustaría agradecer a las personas que me han ayudado durante el tiempo que me ha llevado realizar los estudios del máster, ya que sin ellos el final no estaba, ni mucho menos, cercano.

Comenzaré por mi tutor Paolo Rosso, por confiar en mí por brindarme la oportunidad de hacer un proyecto bajo su tulelación. A Maite Giménez por la ayuda que me brindó en todo momento.

Al resto de profesores del máster por la formación y ayuda recibida en todo momento así como su implicación con los alumnos y la motivación que nos inculcaron. A los compañeros, por hacer más llevadero el día a día.

También me gustaría dar las a gracias a mi familia por su gran apoyo durante estos años animándome día a día. Mi madre, a la cual no sabría dar las gracias por algo en concreto, pero si por todo. A Silvia, mi hermana, porque siempre me ha animado con los estudios y con todas las facetas de mi vida. Al pequeño de la casa, Francisco, por acompañarme en el ordenador mientras trabajaba. A Francisca y Pedro, porque estáis también conmigo.

A mis amigos que me aguantan todos los días con toda la paciencia del mundo, Francisco y Alejandro, porque a pesar de la distancia para mi estáis muy cerca. A mi amigo Luis, por estar ahí desde incontables años.

A Vicente, por compartir su vida conmigo. Por apoyarme todos los dias y confiar en mí. Por regalarme su tiempo aunque yo no he podido darle todo el mío. Por su cariño y comprensión.

Sin duda son muchas las personas que han compartido su día a día conmigo, sus conocimientos, sus inquietudes. A todas ellas les doy las gracias porque han puesto su granito de arena.

Mª Amparo Escortell Pérez

Este trabajo se ha desarrollado en el marco del proyecto de investigación SomEMBED (TIN2015-71147-C2-1-P) del Ministerio de Economía y Sostenibilidad (MINECO).

Índice general

Resumen II

Abstract IV

Agradecimientos VI

Lista de Figuras XI

Lista de Tablas XII

Abreviaturas XIII

1. Introducción **1**
1.1. Descripción del problema, motivación y objetivos 1
1.2. Estructura de la tesis 4

2. Estado de la cuestión **7**
2.1. Análisis de sentimientos 7
2.2. Detección del lenguaje figurado 10
2.3. Competición SemEval 2015 11

3. Marco teórico **13**
3.1. Representación del texto 13
3.1.1. Bolsa de palabras 14
3.1.2. N-gramas 15
3.1.3. Term frequency - Inverse document frequency (TF-IDF) . . 17
3.2. Algoritmos de aprendizaje automático 19
3.2.1. Máquinas de Soporte Vectorial 20
3.2.1.1. Máquinas de Soporte Vectorial de Regresión 21
3.2.2. Árboles de decisión 21
3.2.3. Modelos lineales de regresión 22
3.2.3.1. Ridge 22

4. Metodología **25**
4.1. Entorno experimental 25
4.1.1. Corpus 25

4.1.2. Recursos 26
4.1.2.1. NRC Word-Emotion Association Lexicon (EmoLex) 27
4.1.2.2. Linguistic Inquiry and Word Count (LIWC) 27
4.1.2.3. Smilies 29
4.1.3. Scikit-learn 29
4.2. Entrenamiento 30
4.2.1. Paso 1: Estudio del vocabulario 31
4.2.2. Paso 2: Recursos 31
4.2.3. Paso 3: Experimentación 32
4.3. Medidas de evaluación 40
4.3.1. Error cuadrático medio 40
4.3.2. Distancia coseno 41
4.4. Resultados 42
4.5. Análisis 47

5. Conclusiones y trabajo futuro **51**

Bibliografía **54**

A. Publicación **61**

Índice de figuras

4.1. Fases de la experimentación. 32
4.2. Resultados obtenidos empleando la métrica ECM evaluando el subconjunto de hashtag #irony. Señalamos con la abreviatura "p/n" aquellas representaciones en las que únicamente se emplean las características "Positiva" o "Negativa" del recurso en cuestión, mientras que aquellos recursos en los que empleamos todas las categorías se señalan como "Todas". 43
4.3. Resultados obtenidos empleando la métrica ECM evaluando el subconjunto de hashtag #sarcasm. 44
4.4. Resultados obtenidos empleando la métrica ECM evaluando el subconjunto de hashtag #not. 45
4.5. Resultados obtenidos empleando la métrica ECM evaluando el subconjunto de hashtag #other. 46

Índice de tablas

4.1. Ejemplos extraídos del corpus de la tarea 11 de SemEval 2015. En la columna "Polaridad" se especifica la polaridad con la que se puntuó de media el tweet mostrado como ejemplo. 27
4.2. EmoLex: representación de la palabra *dark*. 28
4.3. LIWC: Ejemplo categorías *posemo* y *negemo* y las palabras *accepted* y *danger*. 28
4.4. Smilies: Clasificación de *smilies*. 29
4.5. Resultados del Error Cuadrático Medio de la baseline con y sin stopwords variando el tokenizador. 33
4.6. Parámetros de BoW y TF-IDF con el clasificador SVR. 36
4.7. Número de tweets del conjunto de datos de test que se tienen para cada hashtags. 38
4.8. Resultados de la distancia coseno de la evaluación de la tarea 11 en comparación a nuestro sistema con el mejor y peor sistema presentado en cada categoría. 47
4.9. Resultados del ECM de la evaluación de la tarea 11 en comparación a nuestro sistema con el mejor y peor sistema presentado en cada categoría. 47

Abreviaturas

PLN	**P**rocesamiento de **L**enguaje **N**atural
SA	**S**entiment **A**nalysis
SemEval	**Se**mantic **Eva**luation
SVM	**S**upport **V**ector **M**achine
BoW	**B**ag of **W**ords
TF-IDF	**T**erm **F**requency-**I**nverse **D**ocument **F**recuency
IR	**I**nformation **R**etrieval
SVR	**S**upport **V**ector **R**egression
DT	**D**ecision **T**ree
ECM	**E**rror **C**uadrático **M**edio

Capítulo 1

Introducción

Antes de adentrarnos de lleno en el trabajo que se ha realizado como proyecto final del "Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital", vamos a hacer una breve introducción para contextualizar el problema al que nos hemos enfrentado, así como las motivaciones y objetivos propuestos.

1.1. Descripción del problema, motivación y objetivos

En los últimos años ha habido un aumento considerable en el uso de las redes sociales, haciendo que los usuarios hayan aumentado su participación tanto en ellas como en otras páginas Web de interacción entre usuarios. Este auge de la Web 2.0 y redes sociales, ha provocado que cada vez se lleven a cabo más estudios sobre los textos que los usuarios aportan en las redes sociales, entre otros, para utilizarlos como fuentes de información sobre un gran abanico de temas. Es por esto, que durante los últimos años, se han desarrollado numerosas técnicas para extraer y analizar esta información por medio del análisis de sentimientos.

El análisis de sentimientos, *Sentiment Analysis (SA)*, es una tarea propia del Procesamiento de Lenguaje Natural (PLN), de la Lingüística Computacional y del Análisis de Textos, que trata de determinar la polaridad que un texto pretende transmitir. Generalmente, esta tarea se ha enfocado desde el PLN como una tarea de clasificación automática de un texto en tres clases de polaridad: positiva, negativa o neutra, a partir de la extracción de una serie de características.

Twitter[1] es una red social de microblogging creada en el 2006, en la que los usuarios pueden publicar contenido textual de hasta un máximo de 140 caracteres. Desde sus inicios, Twitter se convirtió en una de las plataformas más utilizadas para compartir información como experiencias, ideas, opiniones o noticias y eventos en tiempo real sobre cualquier tema.

Según las últimas estadísticas, esta red cuenta con 313 millones de usuarios activos mensuales [38] de los cuales el 79 % son de fuera de los EE.UU, y donde se utilizan más de 40 idiomas diferentes.

Debido a la gran cantidad de usuarios y la información que hay en esta red, se ha demostrado que es una herramienta capaz de movilizar opiniones dentro y fuera de Internet. Esto hace que, a partir de la información que se publica en Twitter se pueda estudiar el impacto de campañas publicitarias, la popularidad de las personas, reputación de las marcas u opiniones de productos. Esto provoca que sea de gran interés poder descubrir la opinión general sobre un determinado tema o producto. Sin embargo, la cantidad abismal de volumen de información que se tiene cada día, hace que no se pueda realizar este trabajo de forma manual y sea necesario automatizar esta tarea.

La competición SemEval (*Semantic Evaluation*)[2] se celebra desde el año 1998 organizada por la ACL (*Association for Computacional Linguisticts)*, y consiste en realizar una serie continua de evaluaciones a sistemas computacionales sobre el análisis semántico. Las evaluaciones pretenden explorar la naturaleza del significado en el lenguaje ya que, aunque el significado puede ser intuitivo para los seres

[1] https://twitter.com/
[2] https://en.wikipedia.org/wiki/SemEval Acceso: 27-06-2017

humanos, la transferencia de esas intuiciones al análisis computacional es difícil de alcanzar. Cada año se van incorporando nuevos problemas a esta competición y en este trabajo nos centraremos en concreto en la tarea 11 del año 2015 [9] [3]: *Sentiment Analysis of Figurative Language in Twitter*, cuyo objetivo es la detección de la polaridad en presencia de lenguaje figurado.

En el caso del lenguaje literal, las técnicas existentes logran resultados aceptables [16]. Sin embargo, esta tarea es especialmente compleja cuando en el texto encontramos lenguaje figurado, puesto que nos enfrentamos con distintos significados debido al uso de la ironía, la metáfora o el sarcasmo, por lo tanto la polaridad del significado literal puede contrastar fuertemente con el sentimiento que pretende transmitir el sentido figurado.

Incluso los seres humanos tenemos dificultad en decidir si una texto es irónico o metafórico. La ironía puede ser muy sutil, mientras que la metáfora se puede representar de muchas formas. Sin embargo, el sarcasmo es más fácil de detectar para los seres humanos, siendo en general más explícito.

El lenguaje figurado es especialmente común en los textos que podemos encontrar en la Web y en las redes sociales, especialmente en Twitter o Facebook. La limitación en la longitud en los textos de la red social Twitter, así como el uso de expresiones plagadas de argot y errores gramaticales, dificulta la comprensión del mensaje.

En definitiva, el lenguaje figurado presenta un desafío para el rendimiento de los sistemas de análisis de sentimientos convencionales basados en la semántica léxica de las palabras ya que a menudo resultan insuficientes para detectar los significados indirectos. En el trabajo de Hernández y Rosso, (2016) [12] puede encontrarse un estudio detallado del impacto de la ironía y el sarcasmo en el análisis de sentimientos.

En este trabajo se presenta un estudio del impacto de las emociones en la detección de la polaridad de un tweet. Partimos de la hipótesis de que no todos los recursos

[3] http://alt.qcri.org/semeval2015/task11/

disponibles favorecen la detección de la polaridad en igual medida, por lo tanto llevamos a cabo una serie de experimentos para evaluar cómo afectan diferentes recursos de emociones tanto en el lenguaje figurado como en el lenguaje literal.

Nuestra metodología está compuesta por dos fases: en la primera de ellas estudiaremos el impacto de los recursos léxicos sobre el entrenamiento de clasificadores que predigan la polaridad del conjunto completo de tweets de la tarea 11 de SemEval2015 y a continuación evaluaremos en detalle el impacto de cada uno de los recursos para las diferentes tipologías del lenguaje figurado presentes en este corpus.

1.2. Estructura de la tesis

Además del actual capítulo, el resto de la memoria consta de 4 capítulos y un apéndice distribuidos de la siguiente manera:

- Capítulo 2: Estado de la cuestión
  Este capítulo describe el estado de la cuestión en el estudio de las emociones en lenguaje figurado.

- Capítulo 3: Marco teórico
  En este capítulo se detallan y definen los conceptos en los que se basa el trabajo. Se verán las diferentes representaciones de texto que se han utilizado a lo largo de la experimentación acompañado de los diferentes algoritmos de clasificación que se han empleado durante el estudio.

- Capítulo 4: Metodología
  En este capítulo se detalla el proceso de experimentación. Primero se detallarán las diferentes partes que componen el entorno experimental como son el corpus, los recursos y las herramientas utilizadas. A continuación, se explicaran las diferentes fases que ha llevado la tarea de entrenamiento. Detallaremos las medidas de evaluación que se han utilizado para evaluar los modelos propuestos y obtener los resultados y un análisis sobre ellos.

- Capítulo 5: Conclusiones y trabajo futuro

  En el último capítulo se plantean las conclusiones a las que hemos llegado a partir de los resultados de la experimentación y se proponen diferentes líneas de investigación para el futuro realizadas con el trabajo presentado.

- Apéndice A: Publicación

  En este apéndice se muestra la publicación a la que ha dado lugar el trabajo presentado que se añade a la memoria.

Capítulo 2

Estado de la cuestión

En este capítulo se va a hacer una revisión del estado de la cuestión actual sobre el análisis de sentimientos. Se verán los resultados de otras investigaciones sobre este tema y situaremos este trabajo en el contexto global del análisis de sentimientos.

2.1. Análisis de sentimientos

Como ya hemos introducido, la definición más extendida de la tarea de análisis de sentimientos se centra en determinar la polaridad general de un texto tratando de determinar la actitud de un escritor con respecto a algún tema, es decir, trata de determinar el efecto emocional que el autor intenta causar en el lector. Para ello, el análisis de sentimientos clasifica los textos en tres categorías: textos positivos, negativos y neutros.

A partir del año 2000 [17] la investigación en el área de análisis de sentimientos comenzó a aumentar de forma considerable debido a la gran variedad de aplicaciones existentes como eran y son las opiniones y comparativas de productos, la gestión de reputaciones, experiencias personales, etc. provocado por el crecimiento de las redes sociales que hacen posible disponer infinidad de información.

Este problema de clasificación del SA, puede ser abordado con diferentes enfoques. En el caso de utilizar algoritmos supervisados, éste parte de un conjunto de datos etiquetados con la polaridad del texto. Sin embargo, como se ha comentado anteriormente, disponer de recursos bien etiquetados a los que se les vaya aportando más información es una tarea compleja y costosa. Dicha tarea también se puede abordar mediante el uso de algoritmos no supervisados.

Los trabajos pioneros (Pang, Lee, y Vaithyanathan, 2002) [27] abordaron esta tarea como un problema de clasificación supervisada en el que clasificaron opiniones de películas en dos clases: positiva y negativa. En este estudio utilizaron variables para la clasificación como son los vectores de palabras en los que se representaba el texto y la frecuencia de aparición de las palabras. Para clasificar hicieron uso de máquinas de soporte vectorial, *Support Vector Machines* (SVM) y Naïve Bayes. Distintos trabajos han tratado la tarea de SA utilizando algoritmos de aprendizaje automático como son las SVM, árboles de decisión, máxima entropía, Naïve Bates, etc. [1, 24, 45].

Sin embargo, en la literatura también podemos encontrar aproximaciones no supervisadas (Turney, 2002) [37] en la que se desarrolló una taxonomía para recomendar o no dependiendo de la polaridad de los textos en base a los adverbios y adjetivos que aparecían. La idea consistía en comparar las palabras consecutivas (bigramas) con patrones sintácticos prefijados por ser los más utilizados a la hora de expresar opiniones. A estos bigramas se les asociaba una polaridad en base a su distancia con la palabra positiva “excelente” y a la negativa “pobre”. La polaridad final del documento se obtenía calculando la media de todas las polaridades de los bigramas. Tanto los algoritmos supervisados como no supervisados disponen de unas características que hacen que ambas líneas sigan abiertas hoy en día.

Otros estudios se basan en el uso de diccionarios de polaridad o lexicones que han demostrado ser de gran utilidad en el SA. Los lexicones consisten en una colección de términos y frases etiquetadas con su polaridad. La clave para realizar la clasificación en este caso, consiste en identificar dichos términos en el texto y obtener la polaridad general del documento en función de las polaridades de los

términos encontrados. El problema que nos encontramos es que la gran mayoría de los lexicones se encuentran disponibles en inglés [15, 22, 44] y minoritariamente en otros idiomas [30, 41].

En el trabajo de Pang y Lee (2008) [26] se recoge un amplio estudio de las distintas técnicas que se han empleado para tratar de resolver la tarea del análisis de sentimientos sobre textos extraídos de Internet como por ejemplo la extracción de características, la subjetividad y los diferentes puntos de vista o la clasificación basada en la relación entre las sentencias del documento.

Según se recoge en Liu et al. (2012) [17], en el análisis de sentimientos se pueden considerar tres niveles de profundidad como son: a nivel de documento, de sentencia y de entidad-aspecto. El primero de ellos aborda el problema de clasificar la opinión general del texto asumiendo que cada texto expresa opiniones sobre un único tema y tan sólo está escrito por un único autor [27, 37]. En la aproximación de sentencias se considera cada frase como una unidad independiente y se asume que cada una de ellas contiene una única opinión. Esta tarea está relacionada con la clasificación de la subjetividad [43] cuyo objetivo es determinar si una frase es subjetiva u objetiva. Por último, la profundidad entidad-aspecto consigue extraer más información de las opiniones ya que, en lugar de atender a las construcciones propias del lenguaje como son los párrafos, frases, etc. se centra en la opinión de algún aspecto directamente, centrándose en que está compuesta por un sentimiento positivo o negativo y un objetivo. Las entidades se corresponde con la forma de representar los objetivos y sus atributos se definen como los aspectos a analizar. Esta aproximación de análisis se suele subdividir en varias fases: identificar las entidades y atributos, clasificar sus polaridades y clasificar la polaridad global del texto.

En el trabajo de Liu and Zhang [18] se recopilan los trabajos más relevantes en el área de SA del estado de la cuestión en el que respecto a los mejores sistemas, aún existe un gran margen de mejora.

Los estudios sobre el SA en Twitter comienzan años posteriores al inicio del SA ya que esta red social es más reciente. Los primeros estudios fueron publicados en el

2009 [40] cuando Twitter comenzó a ser más popular. En SemEval lanzaron varias tareas sobre el SA en Twitter como la tarea 9 del 2014 [1], la 10 del 2015 [2] y la 4 del 2016 [3] y 2017 [4]: *Sentiment Analysis in Twitter*.

2.2. Detección del lenguaje figurado

La detección del lenguaje figurado es una tarea en si misma, y distintas aproximaciones han intentado abordarla. El lenguaje permite expresar ideas o pensamientos de una manera más creativa, complicando la tarea del SA puesto que el sentido de las palabras puede no coincidir con lo que el autor ha querido transmitir. Esto implica que el lenguaje figurado puede invertir la polaridad de un texto, tal y como demostraron en su estudio Maynard and Greenwood [19].

Cuando se trata de analizar los textos, la información disponible en la Web se puede utilizar como una fuente de conocimiento para generar características auxiliares. En el trabajo de Veale and Hao (2007) [39] se describe una forma semiautomática de recopilar el conocimiento y la semántica de los estereotipos de la Web atacando directamente a las construcciones del lenguaje. Los autores demostraron que alrededor del 20 % de los símiles de la Web eran irónicos. Sin embargo, su trabajo no se puede utilizar para detectar la ironía de forma general ya que utilizaba las propias estructuras del lenguaje.

En el trabajo de Carvalho et al. (2009) [3] se investigan una serie de patrones para identificar frases irónicas de los comentarios enviados por los usuarios a un periódico en línea. El enfoque era identificar la ironía en oraciones con predicados positivos ya que son más susceptibles para contener ironía y enmascarar su verdadera polaridad. Los resultados que obtuvieron fueron que explorando ciertas pistas orales o gestuales del usuario como emoticonos, expresiones onomatopéyicas

[1] http://alt.qcri.org/semeval2014/task9/
[2] http://alt.qcri.org/semeval2015/task10/
[3] http://alt.qcri.org/semeval2016/task4/
[4] http://alt.qcri.org/semeval2017/task4/

o signos de puntuación se puede reconocer la ironía en su sistema con una precisión relativamente alta.

Tradicionalmente, el lenguaje figurado se ha intentado detectar explorando las características superficiales de los textos. Por una parte, existen estudios que intentan detectar el lenguaje figurado teniendo en cuenta el orden sintáctico, las propiedades léxicas o los elementos afectivos que componen el texto [32, 33]. Por otra parte, otros trabajos se centran en investigar cómo los hashtags de Twitter se emplean para remarcar una intención figurativa en el mensaje transmitido, en especial para la expresión de la ironía o sarcasmo [35].

Sin embargo, el afecto se puede estudiar de muchas maneras diferentes. En [11] se obtienen muy buenos resultados para clasificar entre textos irónicos y no irónicos. En él se propone un modelo que explora el uso de características afectivas basadas en una gama de recursos léxicos disponibles para el inglés, que reflejan diferentes facetas de las emociones.

2.3. Competición SemEval 2015

El interés que despierta la tarea de la detección de la polaridad así como el impacto que tiene sobre ésta el lenguaje figurado motivó en 2015 una tarea en la competición internacional para el la evaluación semántica (*SemEval*).

Quince equipos participaron en la tarea 11 de SemEval 2015 que fue abordada siguiendo múltiples perspectivas. La mayor parte de los participantes plantearon soluciones supervisadas para intentar resolver la tarea, predominando dos modelos de aprendizaje automático: las SVM y los modelos de regresión. Dichos modelos se entrenaron utilizando un conjuntos de características cuidadosamente seleccionadas para esta tarea como pueden ser: n-gramas de caracteres, n-gramas de palabras, valores extraídos de distintos léxicos, etc.[5].

[5]Para más información ver el artículo de Gosh et al. (2011).

Nuestro trabajo pretende extender la aproximación presentada por Hernández et al., (2015) [10], en la cual se abordó la tarea incorporando recursos externos adicionales. Los autores proponen representar un tweet mediante un conjunto de valores de características extraídas de recursos léxicos externos que modelan tanto las emociones como la información psicolingüística contenida en un tweet.

Asimismo, el trabajo de Sulis et al. (2016) [35] presenta un análisis de la distribución y correlación de un conjunto de características psicolingüísticas y emocionales extraídas de recursos léxicos para realizar la clasificación de tweets irónicos y sarcásticos.

Sin embargo, a diferencia de los citados trabajos sobre el estudio de las emociones en el lenguaje figurado, en este trabajo presentamos un estudio exhaustivo sobre la capacidad de diferentes recursos léxicos de emociones para predecir la polaridad del conjunto de datos de Twitter de la tarea 11 de SemEval 2015 detallando cómo afectan estos recursos a los tweets que contienen lenguaje figurado y lenguaje literal.

Capítulo 3

Marco teórico

En este capítulo se describen los distintos clasificadores que se han utilizado durante los experimentos para obtener la capacidad que tienen distintos recursos léxicos de emociones para analizar la polaridad de los tweets. Primero se detallará la representación del texto en base a n-gramas y los algoritmos de aprendizaje automático que se fundamentan en las Máquinas de Soporte Vectorial. Y, a continuación, las bases teóricas para los modelos presentados en este trabajo. Todas estas herramientas vienen implementadas en el toolkit *Scikit-learn* para Python.

3.1. Representación del texto

Los algoritmos de aprendizaje utilizados tan sólo trabajan con entradas de datos numéricas, sin embargo, los datos con los que se ha realizado la experimentación han sido extraídos de la red social Twitter. Por lo tanto, se necesitará desarrollar un método capaz de transformar los tweets en datos de entrada válidos para los algoritmos.

Existen multitud de técnicas para representar los datos al formato esperado por los algoritmos de aprendizaje. En este trabajo vamos a ver varias de estas técnicas, como son la representación de tipo bolsa de palabras (Bag of Words, BoW)

utilizando n-gramas o la ponderación mediante el método de frecuencia (*Term Frequency-Inverse Document Frecuency*, TF-IDF).

En este apartado vamos a describir en profundidad estas técnicas de representación del texto que se utilizarán posteriormente en la experimentación realizada en este trabajo final de máster.

3.1.1. Bolsa de palabras

Uno de los modelos de lenguaje más simples utilizados en el procesamiento de lenguaje natural es el modelo de bolsa de palabras. Esta aproximación de representación del texto, lo podemos encontrar en tareas de clasificación o en recuperación de información *Information Retrieval (IR)*.

En este modelo, la información del texto se representa teniendo únicamente en cuenta la aparición de palabras individuales, sin tener en cuenta la gramática ni el orden de las palabras. Es por ello que este modelo se utiliza comúnmente en los métodos de clasificación de documentos donde la frecuencia de ocurrencia de cada palabra se utiliza como una característica para la formación de un clasificador.

En el problema del procesamiento del lenguaje natural con el que estamos trabajando, en primer lugar se extraerá el vocabulario de la tarea, a continuación, para cada frase del corpus se formará un vector $v \, \epsilon \, \mathbb{N}^{|V|}$ donde $|V|$ es la talla del vocabulario, y se contará el número de veces que aparece una palabra del vocabulario en la frase. De esta manera para cada sentencia del corpus, se tendrá un vector indicando, entre otros, si aparece o no la unidad lingüística o la frecuencia de aparición.

Suponiendo que nuestro corpus se compone de estas dos frases:

- w_1 = "El gato saltó hasta el tejado."
- w_2 = "La mujer cerró la ventana para que no entrara el gato."

De estas dos sentencias, el vocabulario estará compuesto por:
{el, gato, saltó, hasta, tejado, la, mujer, cerró, ventana, para, que, no, entrara}.
Para obtener el BOW, contaremos el número de veces que aparece cada palabra en cada oración. Los vectores resultantes serán:

- $v_1 = \{2, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0\}$
- $v_2 = \{1, 1, 0, 0, 0, 2, 1, 1, 1, 1, 1, 1, 1\}$

Una vez se tienen los vectores para cada sentencia de nuestro corpus, entrenaremos los modelos con una matriz compuesta por estos vectores. Por tanto, la matriz tendrá la forma $V \epsilon \mathbb{N}^{|V|x|M|}$ siendo M la talla del vocabulario.

Al no tener en cuenta el orden de las palabras ni la gramática en la oración, es posible que frases que tengan un sentido totalmente opuesto, obtengan representaciones similares. A pesar de esto, al tratarse de un modelo simple y además eficiente computacionalmente, es un modelo muy utilizado y en este proyecto se han obtenido buenos resultados utilizándolo.

3.1.2. N-gramas

Una vez vista la representación del texto basada en bolsas de palabras, una mejora es hacer uso de n-gramas. Un modelo de n-gramas intenta predecir la próxima palabra de una oración a partir de las $N-1$ anteriores, de esta manera el orden de las palabras en la frase adquiere más importancia. Esta técnica ha sido aplicada con resultados convenientes en una gran diversidad de problemas diferentes, como son la traducción automática, la bioinformática o el reconocimiento automático del habla y escritura.

El proceso para aplicar esta técnica consiste en *tokenizar* el texto. *Tokenizar* consiste en dividir el texto en unidades más básicas como palabras o caracteres, que se llamarán *tokens*. Los n-gramas más comunes son los unigramas (talla 1), bigramas (talla 2) y trigamas (talla 3).

En cuanto a separar el texto en palabras, se puede obtener una mejora significa utilizando bigramas en lugar de dividir una oración después de palabras tales como "el" o "que". A continuación, se muestra un ejemplo de bigramas de palabras utilizando la representación de bolsas de palabras.

Suponiendo que tenemos el corpus del ejemplo anterior, el vocabulario está formado por las siguientes palabras:
{el gato, gato saltó, saltó hasta, hasta el, el tejado, tejado <EOF>, la mujer, mujer cerró, cerró la, la ventana, ventana para, para que, que no, no entrara, entrara el}.
Para obtener la bolsa de palabras, contaremos el número de veces que aparece cada palabra en cada oración. Los vectores resultantes serán:

- $v_1 = \{1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$
- $v_2 = \{1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$

Si en lugar de tokenizar en palabras, se tokeniza en elementos más pequeños como caracteres, nos encontramos que los caracteres que forman los n-gramas pueden pertenecer a más de una palabra. En este caso, el proceso consiste en recorrer el texto con una ventana de tamaño n. En cada iteración se almacenan las secuencias de n caracteres contenidas en la ventana.

Como de esta manera se obtiene un número elevado de términos, una aproximación consiste en mantener tan sólo aquellos que tengan una mayor frecuencia. Esto puede resultar beneficioso para problemas de clasificación de texto en las que la frecuencia de aparición de b-gramas es suficiente para realizar la clasificación [4]. Un ejemplo de esta técnica se muestra a continuación.

Partiendo de la palabra "ventana", se obtienen los siguientes unigramas, bigramas y trigamas:

- $n = 1 = \{$v, e, n, t, a$\}$
- $n = 2 = \{$ve, en, nt, ta, an, na$\}$
- $n = 3 = \{$ven, ent, nta, tan, ana$\}$

3.1.3. Term frequency - Inverse document frequency (TF-IDF)

Hasta ahora, las representaciones que se han explicado estaban basadas en la frecuencia en que los términos aparecen en los textos, pudiendo encontrar en el lenguaje qu palabras muy comunes, como son los artículos o preposiciones, aparecieran con una frecuencia mayor al resto. Para paliar este problema, se utiliza la técnica de TF-IDF (Term frequency – Inverse document frequency).

Esta técnica consiste en ponderar la frecuencia de aparición de un término en un función de su relevancia en el documento, por tanto, dado un término t y un documento d, se define el $tf(t, d)$ como el número de veces que el término t aparece en el documento d tal y como podemos ver a continuación:

$$tf(t, d) = \frac{f(t, d)}{\text{máx}\, f(w, d)\ \forall\ w\ \epsilon\ d} \tag{3.1}$$

Donde $maxf(w, d)$ representa la frecuencia de la palabra que aparece más veces en el documento. Se puede apreciar, que cuanto mayor es el número de documentos en que aparece t, el denominador será mayor por lo que el cociente será más cercano a 1.

Para determinar la relevancia de los términos y que los que sean poco comunes pero con gran relevancia se ponderen muy favorablemente, se introduce el factor de frecuencia inversa del documento, *idf*, que medirá la importancia de un término

con respecto al documento en cuestión. Mientras se calcula el tf, todos los términos se consideran igual de importantes.

Siendo D el número total de documentos en el corpus, se tiene:

$$tf(t, D) = \log \frac{|D|}{|d \; \epsilon \; D : \; t \; \epsilon \; d|} \tag{3.2}$$

Por último, para calcular la importancia de cada término se hará el cálculo de $tf - idf$ tal y como se puede apreciar a continuación:

$$tf - idf(t, d, D) = tf(t, d) \; \times \; idf(t, D) \tag{3.3}$$

Variaciones de esta técnica son empleadas frecuentemente por los motores de búsqueda como herramienta fundamental para medir la relevancia de un documento dada una consulta del usuario, estableciendo así una ordenación o ranking de los mismos, ya que mejora la estimación de los modelos basados en el conteo o frecuencia de aparición de términos.

De los vectores resultantes es interesante conocer la longitud que tienen. Por ello, se define el operador norma como la longitud o magnitud del vector a considerar.

Se pueden definir una infinidad de normas en $\mathbb{R}^n$, sin embargo, las más utilizadas son las llamadas normas de Hölder o normas l_p. En el caso de TF-IDF, los valores de norma que puede emplear son la l_1 o l_2.

Suponiendo que tenemos el vector x representado de la siguiente manera $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ perteneciente a $\mathbb{R}^n$, se tiene que:

Norma l_1:

$$||x||_1 = \sum_{i=1}^{n} |x_i| \tag{3.4}$$

Norma l_2, conocida como norma Euclídea:

$$||x||_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} \tag{3.5}$$

3.2. Algoritmos de aprendizaje automático

Tal y como definió Tom Mitchel en el prefacio de su libro sobre Machine Learning, "El campo del aprendizaje automatizado se refiere a la cuestión de cómo construir software que mejora automáticamente con la experiencia" [21].

Los algoritmos de aprendizaje automático se pueden dividir principalmente en algoritmos supervisados y no supervisados:

- Los algoritmos supervisados son aquellos en los que a partir de datos de entrenamiento previamente etiquetados, el sistema es capaz de aprender. El usuario con unos datos de entrenamiento en una máquina puede deducir, dado un conjunto de datos de entrada del sistema, a qué clase pertenecen los datos en su salida. Este problema puede clasificarse en dos categorías:
  - Regresión: los valores de salida consisten en una o más variables continuas.
  - Clasificación: las muestras pertenecen a dos o más clases y se pretende aprender de lo que ya se conoce cómo clasificar nuevas muestras.
- Los algoritmos no supervisados no disponen de un conjunto de entrenamiento que permita conocer las etiquetas de los datos, así pues, para intentar construir estas etiquetas se hace necesario el uso de técnicas de agrupamiento (cluster). Estas técnicas tienen como finalidad catalogar los objetos en conjuntos tales que los que estén en el mismo sean muy semejantes entre sí, mientras que el grado de semejanza entre grupos diferentes sea bajo.

En este trabajo, al disponer de un corpus etiquetado, se ha abordado la experimentación con el uso de algoritmos supervisados. Además, la tarea de análisis

de sentimientos se ha abordado con técnicas de regresión ya que se trata de una problemática en la que se deben predecir los valores numéricos dado un dato de entrada, por lo que se debe entrenar una función $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

En esta sección, se detallarán los distintos algoritmos de clasificación utilizados en la experimentación, todos ellos implementados en herramienta *Scikit-Learn* para Python. Se han utilizado máquinas de soporte vectorial, métodos en los que debe realizar una selección de características como árboles de decisión (*Decision Tree (DT)*, y modelos lineales de regresión (Ridge).

3.2.1. Máquinas de Soporte Vectorial

Las máquinas de soporte vectorial [5] son un conjunto de algoritmos que pueden aplicarse a problemas de clasificación o regresión. Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra.

Una SVM trata de determinar un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alto (o incluso infinito) que separe la muestra de dos clases diferentes, de forma que el margen entre la distancia de las muestras más próximas de cualquier clase y el hiperplano sea la máxima, esto es, aquel hiperplano que se encuentre más lejos de las muestras de entrenamiento.

La manera más simple de realizar la separación es mediante una línea recta, un plano recto o un hiperplano N-dimensional. Sin embargo, en los problemas nos encontramos con que un algoritmo de SVM debe tratar con más de dos variables predictoras, casos donde los conjuntos de datos no pueden ser completamente separados o clasificaciones en más de dos categorías.

Debido a las limitaciones computacionales de las máquinas de aprendizaje lineal, éstas no pueden ser utilizadas en la mayoría de las aplicaciones. La representación por medio de funciones Kernel ofrece una solución a este problema, proyectando la información a un espacio de características de mayor dimensión el cual aumenta

la capacidad computacional de la máquinas de aprendizaje lineal. Es decir, mapearemos el espacio de entradas a un nuevo espacio de características de mayor dimensionalidad.

3.2.1.1. Máquinas de Soporte Vectorial de Regresión

Las SVM fueron desarrolladas para resolver problemas de clasificación pero posteriormente se extendieron a problemas de regresión. La adaptación de las SVM para regresión se denominan máquinas de soporte vectorial de regresión, *Support Vector Regression machines (SVR)* [6].

Su funcionamiento es análogo a las SVM pero su principal diferencia radica en las variables utilizadas. SVM asigna una variable a cada punto del dato de entrenamiento, mientras que SVR utiliza dos variables de holgura para cada punto de datos de entrenamiento.

En el apartado 4.2.3 veremos en detalle el ajuste de parámetros a la hora de utilizar SVR en los modelos propuestos.

3.2.2. Árboles de decisión

Los árboles de decisión, DT, [34] son un método de aprendizaje supervisado sin parametrizar utilizado para la clasificación y regresión. El objetivo es crear un modelo que prediga el valor de una variable objetivo mediante el aprendizaje de reglas de decisión simples deducidas de las características de los datos de entrenamiento.

El funcionamiento de este método consiste en clasificar una muestra de entrada seleccionando una única característica. Dependiendo del valor de la característica se seleccionará uno de sus hijos y se evaluará otra característica. Esto se hará de forma recursiva hasta que que el algoritmo alcance alguna hoja. Estas hojas contendrán la etiqueta que se le asignará a la muestra de entrada.

3.2.3. Modelos lineales de regresión

La regresión lineal es uno de los métodos de aprendizaje supervisado más utilizado para predecir el comportamiento de los datos o, al menos, intentar hacerlo.

El método que se utiliza para analizar los datos es de regresión lineal cuando los datos de salida son números reales. La idea es ajustar los datos experimentales a una curva, sin limitarlo a una recta, de esta manera en el peor de los casos se pueden tener polinomios.

En este trabajo se ha utilizado la técnica de Ridge [14] que usualmente se utiliza para la creación de modelos en presencia de un gran número de características, entendiendo como grande el que pueda producirse overfitting o la existencia un reto computacional.

Esta técnica, usualmente denominada técnica de "regularización", trabaja penalizando la magnitud de coeficientes de características mientras que trata de minimizar el error entre las observaciones previstas y reales. La elección del parámetro de penalización (λ) es fundamental y es necesario un procedimiento que estime el valor del parámetro a partir de los datos.

3.2.3.1. Ridge

El método Ridge tiende a contraer los coeficientes de regresión al incluir el término de penalización en la función objetivo. Cuanto mayor sea λ mayor penalización y, por tanto, mayor contracción de los coeficientes. Se emplea una penalización L_2, de modo que el estimador de la regresión es:

$$\widehat{\beta}_{RR} = arg\min_{\beta} \left\{ L(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{3.6}$$

Donde β es el vector de coeficientes de regresión y L la función log-verosimilitud negativa. El parámetro de penalización $\lambda \geq 0$ determina la fuerza de la penalización ($\lambda = 0$ no hay contracción y $\lambda \to \infty$ resulta en todos los parámetros a 0).

Uno de los inconvenientes de este método es que contrae todos los coeficientes hacia cero, pero sin conseguir la nulidad de ninguno de ellos. Por tanto, no se produce selección de variables, permaneciendo en el modelo todas las variables. En aquellos estudios que se tiene un elevado número de variables resulta un inconveniente, pero en este trabajo se han obtenido buenos resultados con esta técnica como veremos en los posteriores capítulos.

Capítulo 4

Metodología

En este capítulo describiremos la metodología que empleamos para el estudio del impacto de ciertos recursos léxicos sobre la detección de la polaridad. Veremos cuáles son los recursos de los que disponemos y que información contienen. Detallaremos los modelos que hemos implementado y como hemos utilizado los algoritmos de clasificación vistos en el capítulo 3. Por último, veremos los resultados obtenidos a la hora de evaluarlos.

4.1. Entorno experimental

Para la parte experimental, se ha trabajado con diferentes recursos: LIWC, EmoLex y Smilies y el corpus proporcionado por la tarea 11 de SemEval. Todos los modelos fueron creados utilizando la librería Scikit-Learn [28].

A continuación veremos en detalle cada uno de estos recursos y herramientas.

4.1.1. Corpus

En la tarea 11 de SemEval 2015 se utilizó un corpus de lenguaje figurativo extraído del servicio de Twitter. Estos textos, llamados tweets, están formados por un

máximo 140 caracteres con un gran número de ironías, sarcasmos o metáforas, sin embargo, ningún tweet en particular puede garantizar que se manifieste cualquiera de estos fenómenos.

La ironía y el sarcasmo normalmente se utilizan para criticar o burlarse y, por lo tanto, sesgar la percepción del sentimiento hacia un valor negativo, por lo que no es suficiente para un sistema determinar simplemente si el sentimiento de un tweet dado es positivo o negativo.

Este conjunto de datos fue recogido durante 4 semanas, del 1 de junio al 30 de junio de 2014, a través de la API Twitter4J, que soporta la recolección de tweets en tiempo real mediante la búsqueda de determinadas consultas. Se eliminaron aquellos tweets que no cumplieran una serie de condiciones como por ejemplo, no contener al menos 30 caracteres sin incluir el hashtag. Los hashtag son etiquetas precedidas por una almohadilla (#) que se utilizan en un tweet para que, tanto el sistema como los usuarios identifiquen de forma rápida la temática sobre la que trata. Asimismo, se filtró también únicamente aquellos tweets que estuvieran escritos en inglés, por lo tanto se trata de una tarea monolingüe.

Cada tweet fue etiquetado por siete anotadores, tres de los cuales eran hablantes nativos de inglés y el resto de los anotadores eran competentes en el idioma. A todos ellos se les pidió asignar una puntuación que oscilaba desde -5 a 5, donde 0 es el valor neutro para aquellos tweets que tienen el mismo valor negativo que positivo. El sentimiento general de cada tweet se calculó como una media ponderada de las siete puntuaciones donde las puntuaciones de los nativos del inglés valían el doble.

El conjunto de tweets de entrenamiento y test está compuesto por 8000 y 4000 tweets respectivamente. Un ejemplo de varios tweets se muestra en la Tabla 4.1.

4.1.2. Recursos

Los recursos que hemos estudiado almacenan distintos niveles de información respecto a las palabras. El nivel más básico es la información sobre si una palabra es

Tweet	Polaridad
There is nothing better than Pitbull singing 'playoffs' as Timber plays in the background. **#sarcasm**	-2.5
Updated my router and it froze. Now I can't access the internet to google a solution. **#irony** #thankfulforsmartphones	-4.14
I've had a lot of wake up calls in my day, but I've always been good at hitting the snooze **#metaphor** #nailedit	0.22

TABLA 4.1: Ejemplos extraídos del corpus de la tarea 11 de SemEval 2015. En la columna "Polaridad" se especifica la polaridad con la que se puntuó de media el tweet mostrado como ejemplo.

"positiva" o "negativa" aunque también incluyen otras categorías que indican qué emociones están vinculadas a las palabras.

4.1.2.1. NRC Word-Emotion Association Lexicon (EmoLex)

El recurso NRC Emotion Lexicon [23], es una lista de palabras en inglés con sus correspondientes asociaciones con las ocho emociones básicas de Plutchik [31]: ira, miedo, anticipación, confianza, sorpresa, tristeza, alegría y el disgusto (*anger, fear, anticipation, trust, surprise, sadness, joy y disgust*) y dos sentimientos: positivo y negativo (*negative y positive*).

Todas las anotaciones de este recurso fueron hechas manualmente mediante anotación voluntaria manual (*crowdsourcing*). Si la palabra pertenece a la categoría se indica con un 1, en caso contrario con un 0. En la totalidad de este recurso podemos encontrar 14.182 palabras etiquetadas.

En la Tabla 4.2 se muestra un ejemplo de cómo se codifica la información en este recurso.

4.1.2.2. Linguistic Inquiry and Word Count (LIWC)

El recurso LIWC [29] le asocia a cada palabra una serie de categorías. En total hay un conjunto de 64 categorías diferentes y se muestra la asociación para un

Palabra	**Categoría**	**Asociación**
dark	anger	0
dark	anticipation	0
dark	disgust	0
dark	fear	0
dark	joy	0
dark	negative	0
dark	positive	0
dark	sadness	1
dark	surprise	0
dark	trust	0

TABLA 4.2: EmoLex: representación de la palabra *dark*.

total de 4.485 palabras.

La estructura que sigue es la siguiente: la primera sección, separada por la cadena "%%", define las categorías y les asigna a cada una un índice numérico, mientras que en la segunda sección se encuentra la palabra y los índices de las categorías a los que pertenece.

En la Tabla 4.3 se muestra un ejemplo para la categoría *posemo* (emoción positiva) y *negemo* (emoción negativa). Según este recurso sabemos que, por ejemplo, la palabra "accepted" connota una emoción positiva y "danger" negativa.

Categoría
...
126 posemo
127 negemo
...
Palabras pertenecientes
accepted 11 13 125 126 131 132
danger 125 127 129

TABLA 4.3: LIWC: Ejemplo categorías *posemo* y *negemo* y las palabras *accepted* y *danger*.

4.1.2.3. Smilies

El recurso *Smilies* [36] clasifica 176 *smilies* diferentes según la emoción asociada a los mismos. En lugar de asociar un smilie a una de las seis emociones básicas definidas en la teoría de Ekman (alegría, ira, miedo, asco, sorpresa, tristeza) [8], los autores utilizan los ocho tipos de emociones avanzadas definidas en la teoría de Plutchik. A partir de estos ocho tipos de emociones y utilizando una lista con los hashtags emocionales más frecuentes, los autores de este recurso seleccionaron quince categorías para etiquetar los diferentes *smilies*: feliz, risueño, amoroso, enfadado, triste, llanto, disgustado, sorpresa, beso, guiño, lengua, escéptico, indeciso, avergonzado y maligno (*happy, laugh, love, annoyed, sad, cry, disgust, surprise, kiss, wink, tongue, skeptical, indecision, embarassed y evil*).

Se puede apreciar un ejemplo de este recurso en la Tabla 4.4.

Emoticono	**Emoción**
:D	LAUGH
:@	SAD
;-)	WINK
3:-)	EVIL

Tabla 4.4: Smilies: Clasificación de *smilies*.

4.1.3. Scikit-learn

Scikit-learn [1] [28] es una librería de software libre para el aprendizaje automático utilizado con el lenguaje de programación Python. Cuenta con varios algoritmos de clasificación, regresión y clustering incluyendo algoritmos de máquinas de soporte vectorial, random forest, gradientes, k-medias y DBSCAN y está diseñado para operar con bibliotecas numéricas y científicas como NumPy y SciPy.

Inicialmente, fue desarrollado por David Cournapeau, como un proyecto de Google en 2007. Más tarde Matthieu Brucher se unió al proyecto y comenzó a utilizarlo

[1] Para más información ver `http://scikit-learn.org/stable/`

como parte de su trabajo de tesis. En 2010 INRIA se involucró y el primer lanzamiento público (v0.1 beta) fue publicado a finales de enero de 2010. El proyecto cuenta ahora con más de 30 contribuyentes activos y tiene patrocinios de INRIA, Google, Tinyclues y la Python Software Foundation.

Está disponible bajo una licencia BSD simplificada más permisiva y se proporciona en diversas distribuciones de Linux, fomentando el uso académico y comercial.

4.2. Entrenamiento

En este apartado detallaremos en qué ha consistido la fase de experimentación que realizamos para seleccionar y ajustar los modelos propuestos a nuestro problema. Nuestra experimentación se ha llevado a cabo en varias fases. En la primera de ellas se ha estudiado el impacto de los recursos sobre el conjunto completo de tweets y a continuación se ha evaluado el grado de impacto para cada uno de los diferentes conjuntos de tweets con lenguaje figurado en el corpus.

El procedimiento que hemos llevado a cabo para evaluar el impacto de cada recurso sobre la detección de la polaridad consistió en una vez tokenizados los datos de entrenamiento y test, desarrollar un estudio ablativo que evalúa cómo el uso de diferentes técnicas, como la bolsa de palabras o TF-IDF, así como los recursos para representar un tweet, afectan a la calidad de la clasificación. Este proceso se explicará en detalle en los siguientes apartados.

Este estudio se ha realizado tanto a nivel de todo el corpus, como a nivel de los distintos tipos de lenguaje figurado presentes en este corpus. Para dividir los tweets entre aquellos que contienen lenguaje figurado o lenguaje literal utilizamos los hashtags, asumiendo que el usuario etiqueta su propio tweet con el tipo de lenguaje empleado facilitando su comprensión, siguiendo la aproximación presentada por Sulis et al. (2016).

A continuación, vamos a ver en detalle el procedimiento llevado a cabo.

4.2.1. Paso 1: Estudio del vocabulario

Antes de proceder a proponer un modelo para la tarea, se estudió el vocabulario de los diferentes recursos. Para ello, se siguieron los siguientes pasos:

1. Se tokeniza el corpus utilizando la librería NLTK ([2]).
2. Se eliminan las palabras que no aportan información discursiva (*stopwords*).
3. Se convierte todo el texto que no sean *smilies* a minúsculas.

A continuación, se han obtenido las representaciones BoW y TF-IDF de los tweets utilizando la librería Scikit-learn.

4.2.2. Paso 2: Recursos

Una vez que se lleva a cabo la fase 1 vista en el apartado anterior, el siguiente paso ha consistido en recuperar la información de los recursos léxicos.

Para los recursos de EmoLex y LIWC se han creado diccionarios que representan eficientemente la información. Cada entrada del diccionario corresponde con una categoría emocional y en ella se almacenan las palabras que forman parte de dicha categoría.

Para utilizar el recurso de *smilies* se ha tenido que crear un tokenizador ad hoc con cada una de las expresiones regulares necesarias para identificar todos los *smilies*. Para ello, se han identificado todos los *smilies* del corpus y con ellos se formaron las expresiones. Un ejemplo para la categoría *evil* seria:

$$evil = \text{“ } >:)| >;\ )| >: -)|\} : -)|\} :)|3 : -)|3 :)''$$

Una vez se han obtenido los diccionarios de cada recurso, se han elaborado representaciones vectoriales de las muestras de entrenamiento y test. Cada uno de estos

vectores indican para cada tweet, el número de veces que aparece una palabra de las categorías que se tienen en el diccionario. De esta manera se tiene un vector diferente para cada recurso que posteriormente se combinará con los demás vectores para realizar la experimentación. La combinación de estos vectores consiste simplemente en agregar al final del vector del primer recurso el vector del segundo.

4.2.3. Paso 3: Experimentación

Utilizando las estructuras comentadas en el apartado anterior, se han realizado varios experimentos sobre todo el conjunto del corpus. Primero utilizando únicamente las categorías "positivo" y "negativo", y a continuación, utilizando todas las categorías disponibles de cada uno de los recursos.

En la Figura 4.1 se puede observar los pasos que se han seguido para la experimentación.

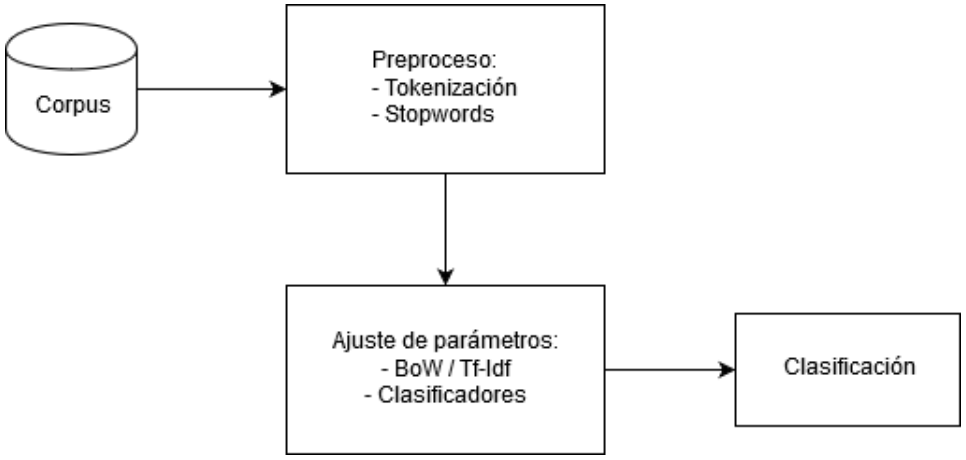

FIGURA 4.1: Fases de la experimentación.

Ajuste de parámetros

La primera fase antes de empezar de lleno con la experimentación de los recursos, ha sido realizar una serie de modelos para determinar qué tokenizador utilizar, el grado de mejora de eliminar o no las stopwords, los valores de la representación del texto (BoW y TF-IDF), ajustar los parámetros de cada uno de los clasificadores

que hemos visto en el apartado 3.2 y que son los que se van a utilizar para llevar a cabo la experimentación.

Para determinar qué tokenizador emplear, se utilizaron los tokenizadores *TweetTokenizer* y *word_tokenize* de la librería *nltk* de Scikit-Learn y se combinaron junto a la opción de eliminar o no las stopwords utilizando un clasificador SVM para realizar una baseline. Un ejemplo del resultado de aplicar ambos tokenizadores es el siguiente:

Suponiendo que nuestro corpus se compone de la frase:

- Esto es un ejemplo de un tweet con su #hashtag y algunos smilies :-) :-P <3

Al aplicar los dos tokenizadores diferentes se obtiene:

- word_tokenize: [Esto, es, un, ejemplo, de, un, tweet, con, su, #hashtag, y, algunos, smilies, :-), :-P, <3]
- TweetTokenize: [Esto, es, un, ejemplo, de, un, tweet, con, su, #, hashtag, y, algunos, smilies, :, -,), :, -, P, <, 3]

Los resultados obtenidos se muestran en la siguiente tabla:

	Con stopwords	**Sin stopwords**
TweetTokenizer	**4,6060**	4,6152
word_tokenize	4,6062	4,6368

Tabla 4.5: Resultados del Error Cuadrático Medio de la baseline con y sin stopwords variando el tokenizador.

Dado que con las stopwords se obtienen mejores resultados, se decide para las siguientes experimentaciones no eliminarlas. Además, existe muy poca diferencia entre ambos tokenizadores, pero puesto que en ambos casos con TweetTokenizer se obtienen mejores resultados, se utilizará éste para tokenizar el corpus.

En cuanto al ajuste de parámetros de los clasificadores y la representación de BoW y TF-IDF, se ha utilizado la función *pipeline* de Python. Dicha función consiste en, dada una serie de características y los valores que queremos que se prueben

de esas características, realiza todas las combinaciones posibles de valores para los parámetros e indica con cuáles de ellos se ha obtenido el mejor resultado. Este proceso es laborioso y requiere un tiempo de computo elevado, pero de esta forma nos aseguramos de obtener los resultados óptimos.

De esta manera, para obtener los parámetros al utilizar Bag of Words, se utilizará la función *CountVectorizer()* con el clasificador *SVR()* (como uso de clasificador de SVM) para realizar la baseline. Para ello, el *pipeline* tendrá dos características *vect* y *clf* que serán el BoW y el clasificador respectivamente.

Los parámetros que se probaran de BoW serán:

- *max_df*: puede tener el rango de valores [0.0, 1.0]. Por defecto su valor es el 1.0. Este parámetro indica que al construir el vocabulario, ignore aquellos términos que tienen una frecuencia de documento estrictamente superior al umbral dado.
- *max_features*: Por defecto *None* (Ninguno). Si no es *None*, construye un vocabulario que consiste en en las *max_features* ordenadas por frecuencia.
- *ngram_range*: tupla (min_n, max_n). Indican el límite de los n-valores para diferentes n-gramas a extraer. Todos los valores de n se encuentran entre $min\_n <= n <= n <= max\_n$.

En la variable *parameters* es la que contendrá el rango de valores a probar. Cada uno de los parámetros anteriores, tendrá un rango de valores que es lo que se irá combinando para ver cuál de ellas es la que da mejor resultados.

Un ejemplo del código resultante para determinar el valor de estos parámetros con el clasificador SVR es el siguiente:

```
1  pipeline = Pipeline([
2              ('vect', CountVectorizer()),
3              ('clf', SVR())
4              ])
```

```
5
6   parameters = {
7               'vect__max_df': [0.3, 0.4, 0.5, 0.6]
8               'vect__max_features': [None, 1000, 5000, 10000],
9               'vect__ngram_range': [(1,1), (1,2), (1,3), (1,4)
      , (1,5), (1,6), (1,7)]
10              }
```

En el caso de TF-IDF, los parámetros que se probaron son:

- *use_idf*: Este parámetro habilita o no la reponderación de frecuencia inversa del documento.
- *norm*: Puede obtener los valores *l1*, *l2* o *None*. Se utiliza para normalizar los vectores de términos.
- *ngram_range*: Se utiliza igual que para BoW visto anteriormente.

Por ello, se tendrán los siguientes valores:

```
1   pipeline = Pipeline([
2               ('vect', TfidfVectorizer()),
3               ('clf', SVR())
4               ])
5
6   parameters = {
7               'vect__use_idf': [0.3, 0.4, 0.5, 0.6]
8               'vect__norm': [None, 1000, 5000, 10000],
9               'vect__ngram_range': [(1,1), (1,2), (1,3), (1,4)
      , (1,5), (1,6), (1,7)]
10              }
```

El resultado que se obtiene con cada uno de estos sistemas se muestra en la tabla 4.6.

BoW	**TF-IDF**
$max\_df = 0{,}5$	$use\_idf = False$
$max\_features = None$	$norm =' l2'$
$ngram\_range = (1, 6)$	$ngram\_range = (1, 2)$

TABLA 4.6: Parámetros de BoW y TF-IDF con el clasificador SVR.

Una vez se tienen los parámetros del primer paso que es obtener los vectores de los datos de entrenamiento y test, el siguiente paso es ajustar los parámetros de los clasificadores.

Para el clasificador de máquinas de soporte vectorial se han utilizado los siguientes parámetros:

- *gamma*: Coeficiente para los kernels *rbf*, *poly* y *sigmoid*.
- *degree*: Coeficiente de la función polinomial del kernel *poly*.
- C: Indica la tolerancia del sistema. Un valor C grande puede conllevar que muestras de entrenamiento estén mal clasificadas. Sin embargo, un valor pequeño incluye muestras bien clasificadas, por lo que seleccionará menos vectores soporte.
- *kernel*: Especifica el tipo de kernel que se utilizará en el algoritmo.

El resultado que se obtuvo al utilizar la técnica de *pipeline* visto anteriormente, indicó que los mejores valores para cada parámetro son:

- $gamma = 0{,}1$
- $kernel = rbf$
- $degree = 1$ Este parámetro se puede obviar ya que el kernel a utilizar es *rbf*.
- $C = 8$

En cuanto a los clasificadores lineales de regresión como Ridge, se estimaron los valores de:

- *normalize*: Si es verdadero, los regresores se normalizan antes de la regresión. Esto significa que los hiperparametros aprendidos sean más robustos y casi independientes del número de muestras.
- *solver*: Se probaron los valores: *auto, svd, cholesky, sparse_cg, lsqr, sag.* Cada uno de ellos difiere en el modo de utilizar los datos.

El valor para cada uno de estos parámetros que se obtuvo fue:

- $normalize = False$
- $solver = sag$. Utiliza un descenso por gradiente estocástico. También utiliza un procedimiento iterativo y suele ser más rápido que otros *solvers* cuando el número de muestras y características son elevados.

Y por último, para DecsisionTreeRegressor, los valores fueron:

- *max_features*: Con los valores *auto, sqrt, log2* o *None* se indica el número máximo de características máximo del sistema.
- *max_deph*: Profundidad máxima del árbol. Si se indica *None*, los nodos se expanden hasta que todas la hojas tengo menos muestras que el valor de *min_samples_split.*
- *min_samples_split*: Indica el número mínimo de muestras necesario para dividir un nodo.

En este caso, los valores devueltos por el sistema fueron:

- $max\_features = auto$. El sistema utilizará todas las características.
- $max\_deph = 2$
- $min\_samples\_split = 2$

Modelos

Dada esta experimentación preliminar para obtener los valores óptimos, la segunda fase ha consistido en determinar el grado de impacto de cada uno de los recursos sobre las diferentes expresiones de lenguaje figurado más habituales en este corpus.

Tal y como se ha mencionado anteriormente, se tiene un especial interés por la ironía y el sarcasmo pero, ¿cómo afectan los tweets con el hashtag #not en el lenguaje figurado? ¿Se puede asociar a tweets irónicos o sarcásticos o es una categoría por sí misma? Es importante categorizar este hashtag porque puede resultar clave para determinar la polaridad del tweet.

Para ello se han diferenciado un total de cuatro grupos de tweets, en función de la aparición de los siguientes hashtags: *#irony*, *#sarcasm*, *#not* y *otros*. Si un tweet tiene varios hashtags pertenecerá a ambos conjuntos.

En la Tabla 4.7 se muestra el número de tweets por cada uno de los hashtags que aparecen en el conjunto de test de nuestro sistema.

Hashtag	**#Tweet**
#irony	765
#sarcasm	536
#not	981
other	1718

TABLA 4.7: Número de tweets del conjunto de datos de test que se tienen para cada hashtags.

En este último conjunto, *otros*, se agrupan aquellos los tweets que no forman parte de los tres primeros grupos. Esta separación se ha hecho sobre los datos del test del corpus, aceptando que el usuario ha empleado el hashtag para autoetiquetar el tipo de lenguaje que su tweet contenía.

La organización de SemEval reportó resultados sobre un conjunto de test con metáforas, pero no aparece el hashtag #metaphor en el test y no hemos podido llevar a cabo una separación automática de este conjunto de tweets. Por lo que en

el conjunto *otros* aparecerán metáforas, que según las actas de las tarea [9] es una de las formas de lenguaje figurado más difíciles de clasificar.

Una vez separados los tweets, se ha llevado a cabo la segunda fase de la experimentación en la que se han utilizado como datos de entrenamiento todos los tweets del conjunto de entrenamiento de la tarea, pero como datos de test se ha utilizado cada uno de los grupos acabamos de describir.

Las combinaciones que se han hecho para determinar como afectan los recursos son:

- BoW/TF-IDF + EmoLex p/n (positivo/negativo)
- BoW/TF-IDF + EmoLex Todas (Todas las categorias)
- BoW/TF-IDF + LIWC p /n
- BoW/TF-IDF + LIWC Todas
- BoW/TF-IDF + EmoLex + LIWC p/n
- BoW/TF-IDF + EmoLex + LIWC Todas
- BoW/TF-IDF + Smilies
- BoW/TF-IDF + EmoLex + LIWC + Smilies p/n
- BoW/TF-IDF + EmoLex + LIWC + Smilies Todas

Para ello, los clasificadores disponen de una serie de funciones que se ejecutan paso a paso. Siendo *nombre-modelo* la variable del clasificador. Se tiene que:

```
1 nombre-modelo.fit()
2 nombre-modelo.predict()
3 nombre-modelo.score()
```

Con la función *fit()* entrenamos el modelo para obtener los parámetros que utilizaremos sobre los datos de test con la función *predict()*. Finalmente, con *score()*

podremos obtener una estimación de la capacidad de acierto de nuestro modelo sobre los datos de trabajo.

En el siguiente apartado indicaremos los scores utilizados para obtener los resultados de los modelos.

4.3. Medidas de evaluación

Una vez se han descrito los algoritmos de clasificación empleados durante el proyecto, un punto fundamental es determinar el comportamiento de los modelos que se proponen. Por ello, en este apartado se definen una serie de métricas de evaluación que han sido utilizadas para obtener los resultados y realizar comparativas entre los distintos modelos creados teniendo en cuenta que nos enfrentamos a un problema de regresión.

4.3.1. Error cuadrático medio

El error cuadrático medio (ECM) define el error cometido por el vector de predicciones $\hat{Y} \in \mathbb{R}^n$ con respecto al vector con los valores correctos para esas n muestras $Y \in \mathbb{R}^n$, es decir, la diferencia entre el estimador y lo que se estima.

Se calcula de la siguiente manera:

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{4.1}$$

Donde nuestro objetivo será minimizar el ECM de las predicciones realizadas por los modelos.

4.3.2. Distancia coseno

Dados dos vectores, donde uno de ellos es un vector con las predicciones $\hat{A}$ y otro B con los valores reales, es posible evaluar el algoritmo de aprendizaje entre estos vectores utilizando la distancia coseno en un mismo espacio vectorial D. Esta evaluación se aplica en un espacio n-dimensional, siendo n el número de muestras que se han clasificado.

También denominada similaridad coseno, la distancia coseno mide el coseno del ángulo que forman estos dos vectores. Dados dos vectores de atributos A y B, la distancia coseno $\cos(\theta)$, se representa como sigue:

$$
\begin{aligned}
similarity &= \cos(\theta) && (4.2)\\
&= \frac{A \cdot B}{\parallel A \parallel \parallel B \parallel} && (4.3)\\
&= \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} && (4.4)
\end{aligned}
$$

Donde A_i y B_i son componentes del vector A y B respectivamente.

Si el ángulo comprendido entre los dos vectores es 0, ambos vectores presentan la misma dirección y sentido, por lo que el $\cos 0 = 1$, es decir, la similitud de ambos vectores es máxima. En caso de que los vectores difieran y el algoritmo no prediga las etiquetas correctamente, los vectores serán perpendiculares y la distancia coseno será $\cos 90 = 0$.

Si los vectores fueran ortogonales el coseno se anularía, y si apuntaran en sentido contrario su valor sería -1. De esta forma, el valor de esta métrica se encuentra entre -1 y 1, es decir en el intervalo cerrado [-1,1].

La distancia coseno no es muy eficiente en cuanto a computación, pero permite tener un sentido de la dirección de ambos vectores. Por ello, esta distancia es muy popular, a pesar de que no satisface la condición de desigualdad triangular y, por tanto, no es una métrica completa.

4.4. Resultados

En este apartado vamos a ver los resultados obtenidos de los modelos propuestos para los distintos clasificadores y en el siguiente realizaremos un análisis más exhaustivo de estos resultados.

Se van a mostrar de forma gráfica los resultados para evaluar de una manera más clara y concisa los modelos propuestos utilizando los diferentes recursos tal y como se detalló en el apartado 4.2.3 para cada uno de los 4 hashtag vistos (#irony, #sarcasm, #not y other).

Comenzando por #irony, en la figura 4.2 podemos observar gráficamente el ECM que se ha obtenido para cada una las combinaciones vistas y en colores diferentes se detallan los clasificadores utilizados (SVR, Ridge y DecisionTree).

Se puede observar como el clasificador SVR destaca por ser el que peores resultados obtiene en la mayoría de combinaciones. Por otra parte, los clasificadores Ridge y DecisionTree se encuentran más igualados, siendo en este caso Ridge el que menos error obtiene en casi todas las combinaciones.

En cuanto a la comparación de la representación BoW y TF-IDF, BoW funciona mejor con el clasificador SVR con una gran diferencia respecto a TF-IDF. Con los otros clasificadores el resultados es no adquiere tanta diferencia, pero en general se sigue obteniendo un mejor resultado con la aproximación de BoW.

Dependiendo del recurso utilizado, la diferencia entre la utilización de las dos categorías “posemo” y “negemo” y el uso de todas las disponibles, resulta en que en algunos casos como EmoLex, utilizar todas proporciona mejores resultados, pero en el caso de LIWC es totalmente lo contrario.

En el caso del hashtag #sarcasm los resultados se pueden ver en la figura 4.3. En este caso se obtienen mejores resultados que con #irony. Se observa como con la representación TF-IDF los errores son claramente más elevados para todos los clasificadores excepto con Decicion Tree, con el cual se observan resultados similares entre BoW y TF-IDF.

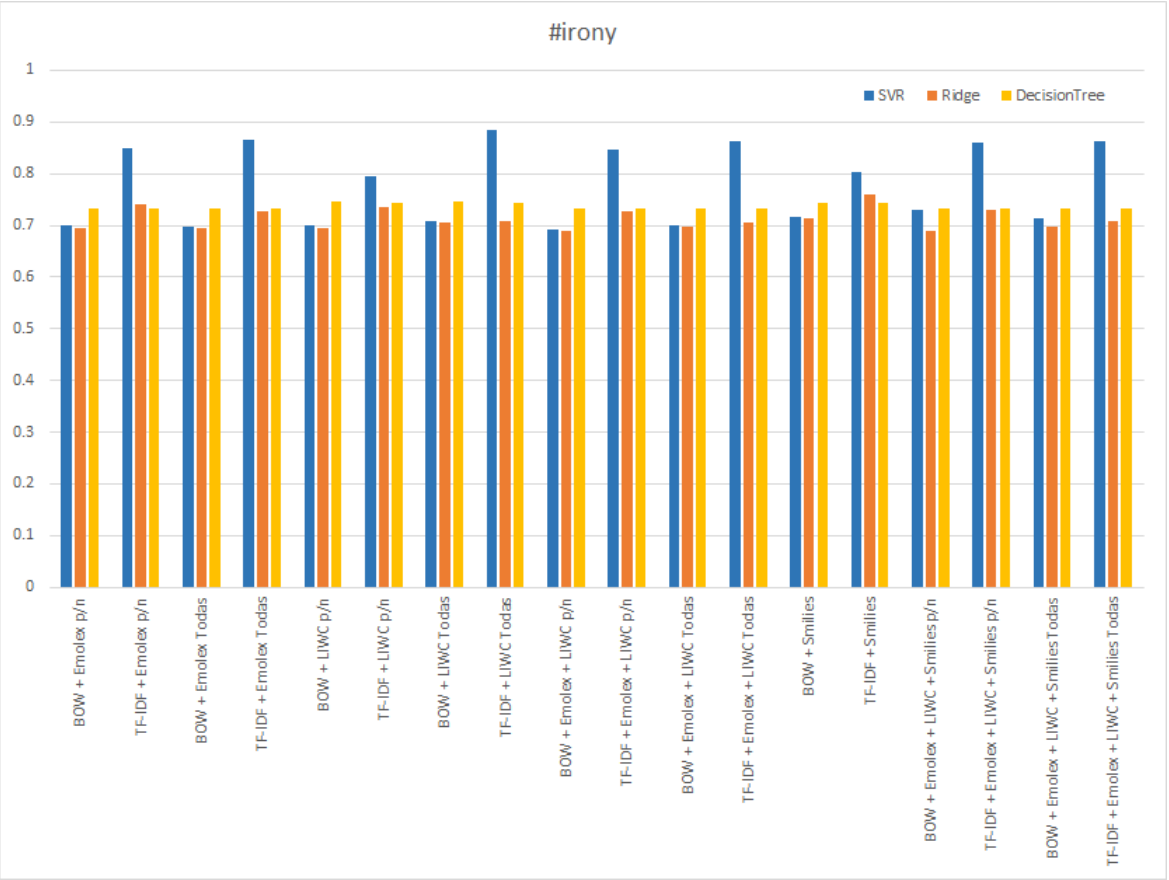

FIGURA 4.2: Resultados obtenidos empleando la métrica ECM evaluando el subconjunto de hashtag #irony. Señalamos con la abreviatura "p/n" aquellas representaciones en las que únicamente se emplean las características "Positiva" o "Negativa" del recurso en cuestión, mientras que aquellos recursos en los que empleamos todas las categorías se señalan como "Todas".

Utilizando los clasificadores SVR y Ridge se puede ver como el error se dispara, en gran medida con el uso de TF-IDF pero, en ambos casos la diferencia entre estos dos clasificadores y DecisionTree es significativa.

Al igual que en el caso anterior, dependiendo del recurso utilizado y de las categorías que se han empleado para el modelo, se obtienen mejores o peores resultados de manera diferente, es decir, en unos casos se obtienen mejores resultados utilizando todas las categorías y en otros no.

Para el hashtag #not, se muestran los resultados en la figura 4.4. En este caso, el valor del error aumenta considerablemente respecto a los dos hashtag anteriores. Se ha pasado de un error comprendido entre [0,5-0,7] a tener unas cifras alrededor del 3,8 de ECM.

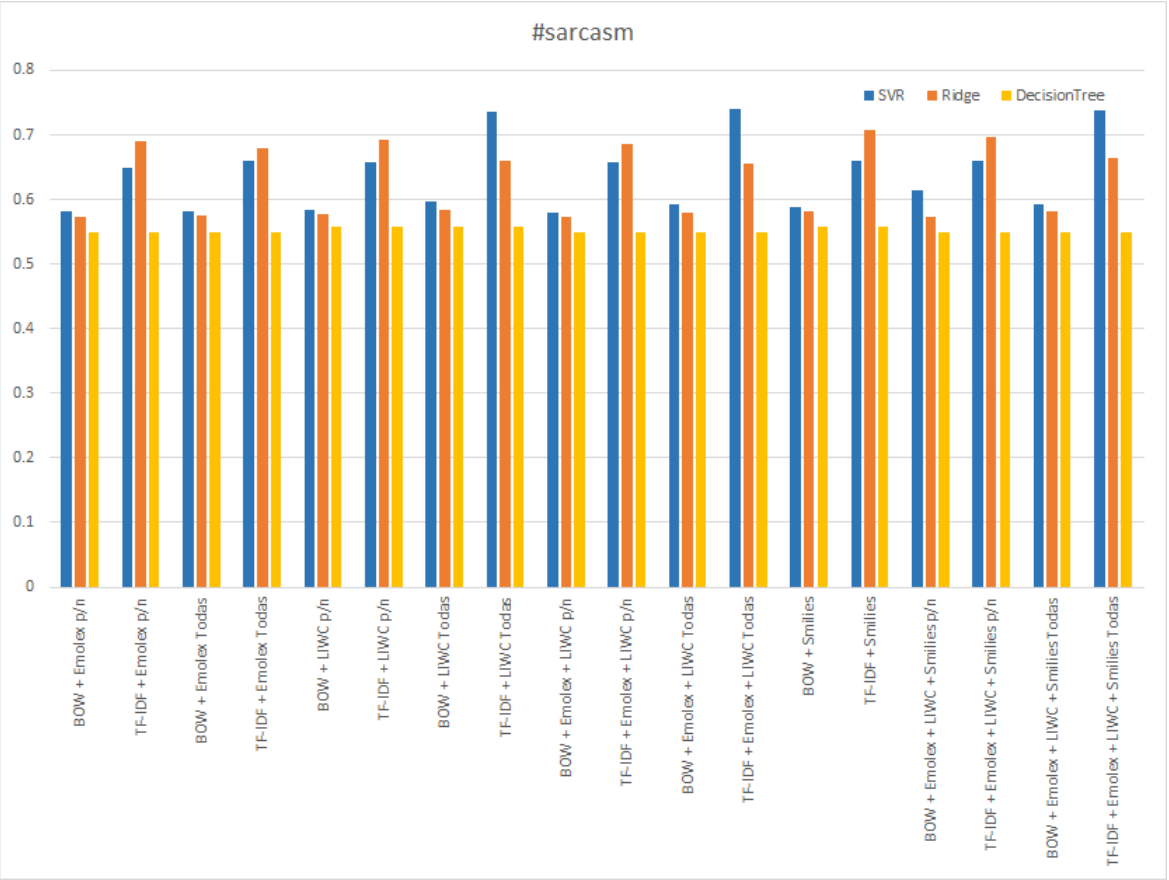

FIGURA 4.3: Resultados obtenidos empleando la métrica ECM evaluando el subconjunto de hashtag #sarcasm.

Destaca que no hay demasiada diferencia respecto a BoW y TF-IDF en los primeros modelos, pero después con SVR se dispara el error, llegando incluso hasta un valor mayor a 5.0. Esto no sucede con los otros dos clasificadores, en los que el error va oscilando dependiendo de la combinación de los recursos utilizada, siendo con DecisionTree con los que se obtienen los resultados más parejos entre BoW y TF-IDF y, en general, los mejores resultados para todas las combinaciones de recursos y categorías utilizadas. Se puede apreciar como, al contrario que en los casos anteriores, el uso de todas las categorías de los recursos nos hace obtener mejores resultados.

Por último, en la categoría other se muestran los resultados obtenidos en la figura 4.5. Como vimos en el apartado anterior (4.2.3), en este conjunto se agrupan los tweets que no forman parte de los otras categorías.

Como se puede observar, a simple vista se puede ver que el clasificador DecisionTree

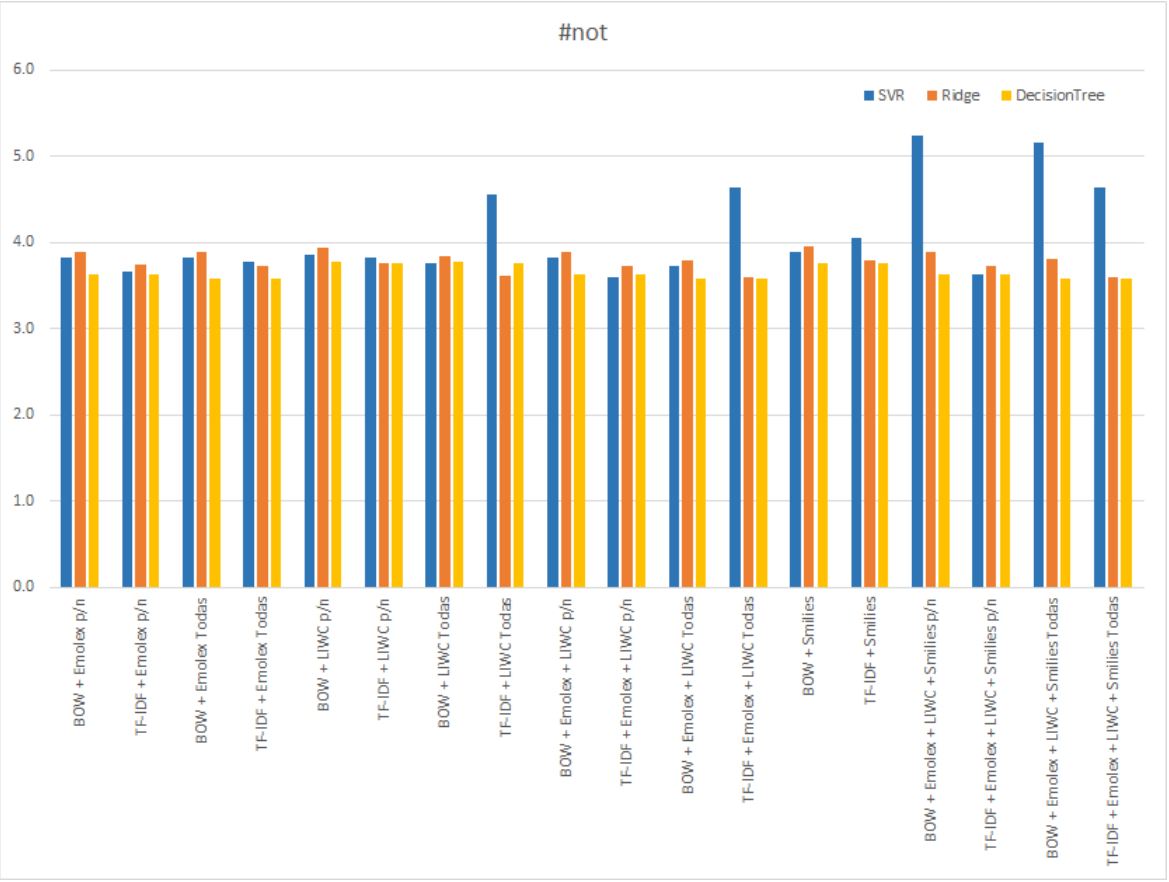

FIGURA 4.4: Resultados obtenidos empleando la métrica ECM evaluando el subconjunto de hashtag #not.

es el que peor resultados obtiene. Así mismo, dependiendo de la combinación de recursos, nos encontramos que SVR nos da tanto los mejores resultados, como los peores, dependiendo de los recursos y categorías utilizadas.

El uso de BoW y TF-IDF es prácticamente similar en la mayoría de modelos, sin embargo, existen diferencias entre utilizar todas las categorías o tan sólo emociones positivas "posemo" y negativas "negemo", siendo con estas últimas cuando mejor resultados se obtienen.

Los organizadores de la tarea 11 de SemEval 2015 desarrollaron tres sistemas [9] para medir el baseline de la tarea 11 y así comparar los resultados con los enviados por los participantes. Estos sistemas utilizaban la representación de bolsa de palabras y se basaron en:

- SVM con la que obtuvieron un 0,390 de distancia coseno.

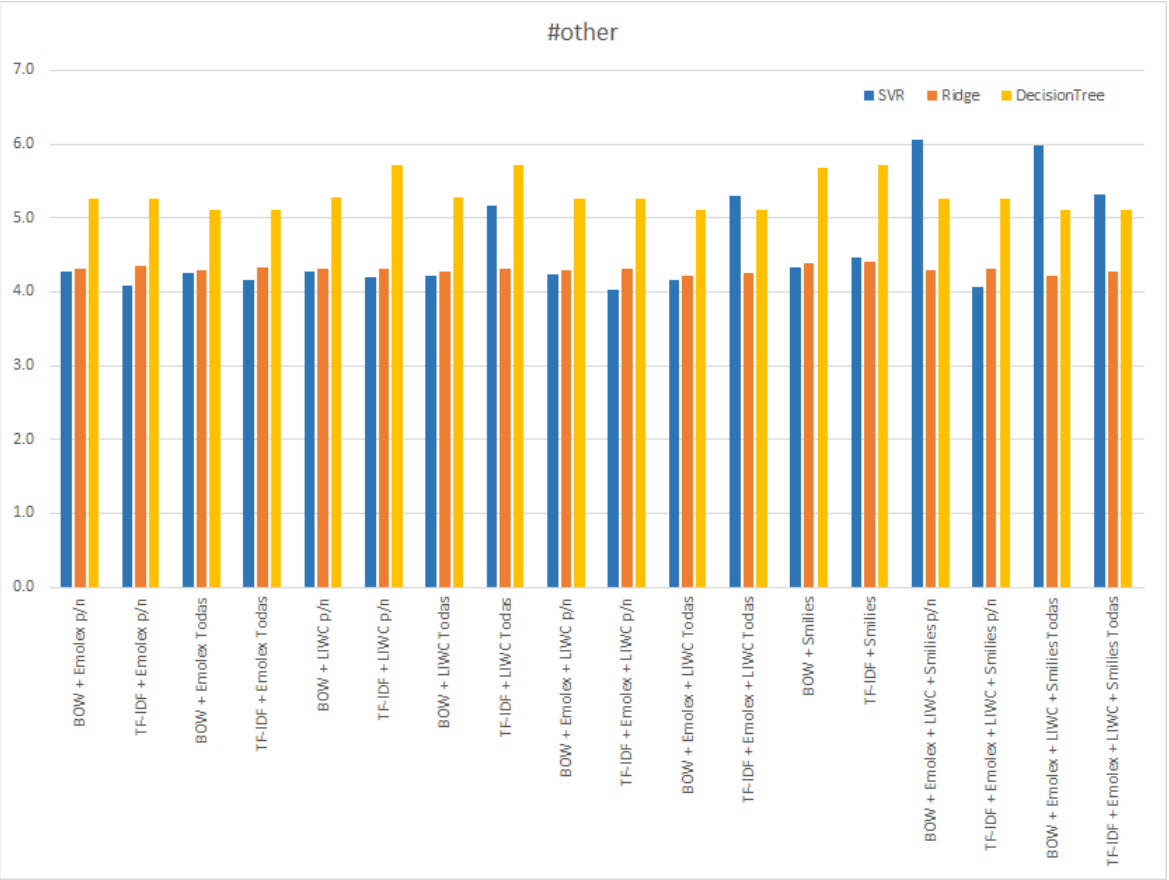

FIGURA 4.5: Resultados obtenidos empleando la métrica ECM evaluando el subconjunto de hashtag #other.

- Máxima entropía con la que se logró un 0,426 de distancia coseno.
- Árboles de decisión con la que alcanzaron un 0,547 de distancia coseno.

En la tabla 4.3.2 podemos ver una comparativa con los resultados obtenidos por nuestro modelo "BoW + EmoLex + LIWC + Smilies Todas" basado en árboles de decisión, que utiliza todas las categorías de los recursos estudiados, para cada las categorías de ironía, sarcasmo y otros, en comparación a los mejores y peores sistemas que se enviaron a la organización de la tarea 11 de SemEval 2015.

La comparación con la categoría #not no se ha podido realizar ya que no se contempló en la tarea 11. Sin embargo, para el resto de categorías se adquieren buenas posiciones tanto para la ironía como para el sarcasmo. Para la categoría otros el sistema tiene cae hasta la séptima posición.

Subconjunto	Distancia coseno	Posición	Mejor	Peor
Ironía	0,916	2	**0,918**	-0,209
Sarcasmo	**0,944**	**1**	0,904	0,412
Otros	0,311	7	0,584	-0,025

TABLA 4.8: Resultados de la distancia coseno de la evaluación de la tarea 11 en comparación a nuestro sistema con el mejor y peor sistema presentado en cada categoría.

A pesar de que la medida de evaluación oficial de la tarea 11 era la distancia coseno vista en el apartado 4.3.2, también evaluaron el comportamiento de los sistemas utilizando el ECM definido en el apartado 4.3.1. En este caso la comparativa se muestra en la tabla 4.3.1.

Subconjunto	ECM	Posición	Mejor	Peor
Ironía	0,732	2	0,671	7,609
Sarcasmo	**0,549**	**1**	0,934	4,375
Otros	5,113	6	3,411	12,16

TABLA 4.9: Resultados del ECM de la evaluación de la tarea 11 en comparación a nuestro sistema con el mejor y peor sistema presentado en cada categoría.

En este caso se obtienen unas posiciones similares al caso de la comparación con la distancia coseno. Tanto con ironía como con sarcasmo se obtienen buenos resultados y con el conjunto de otros se bajan posiciones.

4.5. Análisis

En este apartado se detallará más exhaustivamente los resultados vistos en el apartado anterior. Debido a la naturaleza de la tarea 11 de SemEval 2015, se realizó el estudio para las categorías de emociones "posemo" y "negemo" y posteriormente se utilizaron todas las emociones que aparecían en los recursos. Con los resultados obtenidos se han afianzado los resultados obtenidos en [7], en los que se llegó a un análisis similar a los experimentos de este proyecto.

Como se ha visto en el apartado anterior, a pesar de que el ECM varía considerablemente en función del subconjunto del lenguaje que se está considerando, la inclusión de nuevos recursos conlleva mejorar significativamente el comportamiento de los modelos que se han entrenado.

Todos los modelos que hemos presentados consiguen mejorar el modelo de control o *baseline*, lo cual nos indica que, efectivamente, los recursos léxicos que aportan información acerca de las emociones ayudan a mejorar la predicción del sentimiento comunicado en un tweet.

Cabe destacar que, cuando incluimos la información respecto a todas las emociones disponibles en un recurso léxico, y no únicamente las categorías "positivas" y "negativas", conseguimos mejorar el comportamiento del modelo entrenado. Además, a pesar de que estamos ante un corpus con una baja frecuencia de *smilies* y por lo tanto la cobertura del léxico smilies es escasa, este recurso también consigue mejorar el sistema.

Las aproximaciones que incluyen la bolsa de palabras han obtenido, generalmente, mejores resultados que los basados en TF-IDF. Debido a, como se ha comentado anteriormente, la cobertura del léxico de smilies, los mejores sistemas se encuentran entre "BoW + EmoLex + LIWC + Smilies Todas" y "BoW + EmoLex + LIWC Todas".

El mejor sistema participante en la tarea, CLaC [25], obtuvo un EMC de 2,117 para lo cual se desarrolló un complejo proceso para la extracción de la polaridad de las palabras en función del contexto en el que aparezcan. En nuestro caso, el sistema desarrollado comparte características con el sistema denominado ValenTo [13] como considerar los emoticonos para determinar cómo se expresa una determinada emoción o las frecuencias de los signos de puntuación.

El estudio de cómo se distribuye el error entre los distintos subgrupos de tweets nos indica que la mayor parte del error se concentra en el subgrupo que no contenía ningún tipo de hashtags del conjunto de hastags estudiados (#irony, #sarcarm,

#not). Esto puede explicarse debido a que no se puede analizar de modo independiente el impacto de la metáfora sobre el conjunto *otros*. No obstante, se requiere un estudio pormenorizado de los tweets y la polaridad para explicar este fenómeno.

Capítulo 5

Conclusiones y trabajo futuro

Como trabajo final del máster, se ha presentado una tarea en el área de investigación, del análisis de sentimientos a partir de textos de redes sociales. Debido al crecimiento de las redes sociales, páginas web de opiniones sobre items, comparativas, etc. donde el usuario interacciona y escribe sus opiniones, la tarea del análisis de sentimientos cada vez está adquiriendo una mayor importancia no sólo en el ámbito académico, por lo que desarrollar sistemas automáticos que sean capaces de analizar los textos de las redes en tiempo real para adquirir conocimiento puede resultar interesante.

En este trabajo, se ha presentado un estudio sobre la capacidad de distintos recursos léxicos de emociones para predecir la polaridad de un conjunto de datos extraídos de Twitter. Se ha visto el impacto de cada uno de ellos sobre las distintas formas de lenguaje figurado como la ironía y el sarcasmo y la importancia de desarrollar técnicas capaces de representar esta información para clasificar el sentimiento que el autor escribió en un texto. Se han obtenido unos resultados que apuntan a que la inclusión de información relativa a las emociones ayuda a clasificar correctamente la polaridad tanto a nivel global como a nivel del lenguaje figurado.

Como trabajo futuro, se puede estudiar cómo se puede aumentar la cobertura de los recursos léxicos, es decir el número de palabras que encontramos en el diccionario,

utilizando técnicas como por ejemplo la corrección automática del texto, para eliminar, en la medida de lo posible los errores gramaticales presentes en Twitter.

Además, seria interesante expandir y adaptar nuevos recursos a otros idiomas para poder resolver tareas de PLN para todos ellos, o al menos, sobre los idiomas más utilizados en Internet, ya que la mayor parte de los recursos de los que disponemos como pueden ser diccionarios de polaridad o de emociones, las bases de conocimiento, etc. se encuentran sobre todo para textos en inglés.

No sólo eso, si no que la mayor parte de los textos no recogen el uso de lenguaje que hacen los usuarios en las redes sociales, por lo que, los recursos de los que se disponen tienen una baja cobertura sobre el lenguaje. Recoger la polaridad del argot utilizado en las redes, haciendo por tanto que se tengan recursos con un lenguaje más común y no tan normativo, podría ayudar a desarrollar modelos capaces de mejorar sustancialmente los resultados obtenidos.

A pesar de que nuestros sistemas se basan en aproximaciones supervisadas, en las redes sociales nos encontramos con una gran cantidad de datos no etiquetados y, aunque dispongamos de recursos con datos etiquetados, el esfuerzo económico y temporal que se requiere para ello, hace que la diferencia de cantidad entre etiquetados y no etiquetados sea abismal.

Por ello, en futuros trabajos se podrían incorporar técnicas de aprendizaje automático que permitieran incorporar más conocimiento como puede ser mediante el uso de técnicas de online learning. También se podría hacer uso de algoritmos de clustering que, mediante el agrupamiento de texto no etiquetado y viendo su similitud a las diferentes clases obtenidas utilizando datos etiquetados, permitieran clasificar datos no etiquetados. De esta manera se podrían obtener sistemas semisupervisados, que a partir de datos etiquetados, pudieran mejorar los sistemas al utilizar datos no etiquetados.

En los sistemas que se han presentado se han utilizado máquinas de soporte vectorial y algoritmos de regresión, puesto que se debían obtener resultados entre

un rango de valores continuos. Se han utilizado aproximaciones basadas en bolsas de palabras de n-gramas y coeficientes TF-IDF. En ambos casos se obtienen resultados competitivos, pero aún queda mucho margen de mejora. Por ello, se podría experimentar con métodos de representación de texto continuos (word embeddings) con los que resolver problemas PLN.

Con los resultados del estudio realizado, tanto para la ironía como para el sarcasmo se han obtenido muy buenos resultados al añadir todos los recursos que se han utilizado para la experimentación ya que cada uno de ellos aporta más información y mejora los resultados. Se ha visto que la mayor parte del error se concentra en el grupo otros, grupo en el cual se encuentra la presencia de metáforas. Se asume con esto, que los modelos que se han utilizado en la experimentación se adaptan a los datos de entrenamiento de la tarea, lo que los limita para obtener resultados ante nuevos tweets no vistos. Dado esto, una mejora de estos sistemas seria desarrollar modelos capaces de estudiar el impacto de variantes de lenguaje figurado sobre el sentimiento global de un tweet.

Bibliografía

[1] Barbosa, L. and Feng, J., *Robust Sentiment Detection on Twitter from Biased and Noisy Data*, In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, páginas 36–44, 2010.

[2] Bird, S., Klein, E. and Loper, E., *Natural Language Processing with Python*, O,Reilly Media, Inc., 2009.

[3] Carvalho, P., Sarmento, L., Silva, M. J. and DeOliveira, E., *Clues for Detecting Irony in User-Generated Contents: Oh...!! It's so easy;-)*, In Proceedings of the 1st international CIKM workshop on Topic-sentiment analys for mass opinion, ACM, páginas 53–56, 2009.

[4] Cavnar, W. B. and Trenkle, J. M., *N-Gram-Based Text Categorization*, Ann Arbor MI, 48113(2), páginas 161–175, 1994.

[5] Cortes, C. and Vapnik, V., *Support-Vector Networks*, Machine learning, 20(3), páginas 273-297, 1995.

[6] Drucker, H., Burges, C., Kaufman, L., Smola, A. J., and Vapnik, V. N. *Support Vector Regression Machines*, Advances in Neural Information Processing Systems 9, páginas 155–161, MIT Press, 1997.

[7] Escortell M., Giménez M. and Rosso P., *El impacto de las emociones en el análisis de la polaridad en textos con lenguaje figurado en Twitter*, Procesamiento del Lenguaje Natural, 58, páginas 85-92, 2017.

[8] Ekman, P., *Universals and Cultural Differences in Facial Expressions of Emotions*, Nebraska Symposium on Motivation, 19, páginas 207-283, 1972.

[9] Ghosh, A., Li G., Veale, T., Rosso, P., Shutova, E., Barnden, J. and Reyes, A., *SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter*, Proc. 9th Int. Workshop on Semantic Evaluation (SemEval 2015) páginas 470-478, 2015.

[10] Hernández, I., Benedí, J. M., and Rosso, P., *Pattern Recognition and Image Analysis: 7th Iberian Conference, IbPRIA 2015, Santiago de Compostela, Spain, June 17-19, 2015*, Proceedings, chapter Applying Basic Features from Sentiment Analysis for Automatic Irony, Springer International Publishing, Cham, Detection, páginas 337–344. 2015.

[11] Hernández, F., Patti, V. and Rosso, P., *Irony Detection in Twitter: The Role of Affective Content*, ACM Transactions on Internet Technology. 16 (3), páginas 1-24, 2016.

[12] Hernández, I. and Rosso, P., Capítulo 7 en: *Irony, Sarcasm, and Sentiment Analysis*, Sentiment Analysis in Social Networks, F.A., Pozzi, E. Fersini, E. Messina, and B. Liu (Eds.), Elsevier Science and Technology, páginas 113-128, 2016.

[13] Hernández, D., Sulis, E., Patti, V., Ruffo, G. and Bosco, C., *ValenTo: Sentiment Analysis of Figurative Language Tweets with Irony and Sarcasm*, SemEval-2015, páginas 694-698, 2015.

[14] Hoerl, A. and Kennard., R., *Ridge Regression: Biased Estimation of Nonorthogonal Problems*, Technometrics, 12, páginas 55-67, 1970.

[15] Hu, M. and Liu, B., *Mining and Summarizing Customer Reviews*, In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, páginas 168–177, 2004.

[16] Kiritchenko, S., Mohammad, S. and Salameh, M., *SemEval-2016 Task 7: Determining sentiment intensity of english and arabic phrases*, Proceedings of

the International Workshop on Semantic Evaluation (SemEval), San Diego, California, Junio, páginas 42-51, 2016.

[17] Liu, B., *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012.

[18] Liu, B. and Zhang, L., *A Survey of Opinion Mining and Sentiment Analysis*, In Mining text data, Springer, páginas 415–463, 2012.

[19] Maynard, D. and Greenwood, M., *Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis*, In Language Resources Evaluation Conference, páginas 4238–4243, 2014.

[20] Metsis, V., Androutsopoulos, I. and Paliouras, G., *Spam Filtering with Naive Bayes-wWhich Naive Bayes?*, Proceedings of the 3rd Conference on Email and Anti-Spam, CEAS 2006, páginas 27–28, 2006.

[21] Mitchell, T. M., *Machine Learning*, McGraw-Hill Boston, 1997.

[22] Mohammad, S., Dunne, C. and Dorr, B., *Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus*, In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2, Association for Computational Linguistics, páginas 599–608, 2009.

[23] Mohammad, S. and Turney, P., *Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon*, Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text, páginas 26-34, 2010.

[24] O'Connor, B., Balasubramanyan, R., Routledge, B. and Smith, N., *From Tweets To Polls: Linking Text sentiment To Public Opinion Time Series*, International Conference on Web and Social Media (ICWSM), 11, páginas 122-129, 2010.

[25] Ozdemir, C. and Bergler, S., *CLaC-SentiPipe: SemEval2015 Subtasks 10 B, E, and Task 11*, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), páginas 479-485, 2015.

[26] Pang, B. and Lee, L., *Opinion Mining and Sentiment Analysis*, Foundations and trends in information retrieval, (2), páginas 1-135, 2008.

[27] Pang, B., Lee, L. and Vaithyanathan, S., *Thumbs up?: Sentiment Classification using Machine Learning Techniques*, Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 10, páginas 79-86, 2002.

[28] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and otros, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, 12, páginas 2825-2830, 2011.

[29] Pennebaker, J., Francis, M. and Booth, R., *Linguistic Inquiry and Word Count: LIWC 2001*, Mahway: Lawrence Erlbaum Associates, 71, 2001.

[30] Perez-Rosas, V., Banea, C. and Mihalcea, R., *Learning sentiment lexicons in spanish*, Conference: Eighth International Conference on Language Resources and Evaluation (LREC-2012), At Istanbul, Turkey, Volumen: Proceedings of the Eighth International Conference on Language Resources and Evaluation, 2012.

[31] Plutchik, R., *Emotion: Theory, Research, and Experience*, Theories of Emotion, 1, 1980.

[32] Reyes, A., Rosso, P. and Buscaldi, D., *From Humor Recognition to Irony Detection: The Figurative Language of Social Media*, Data & Knowledge Engineering, (74), páginas 1-12, 2012.

[33] Reyes, A., Rosso, P. and Veale, T., *A multidimensional Approach for Detecting Irony in Twitter*, Language resources and evaluation, 47 (1), páginas 239-268, 2013.

[34] Rokach, L. and Maimon, O., *Data Mining with Decision Trees: Theory and Applications*, World Scientific Pub Co Inc, 2008.

[35] Sulis, E., Hernández, I., and Rosso, P., Patti, V. and Ruffo, G., *Figurative Messages and Affect in Twitter: Differences Between #irony, #sarcasm and #not*, Knowledge-Based Systems, 108 páginas 132-143, 2016.

[36] Suttles, J. and Ide, N., *Distant Supervision for Emotion Classification with Discrete Binary Value*, Computational Linguistics and Intelligent Text Processin, páginas 121-136, 2013.

[37] Turney, P., *Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews*, Proceedings of the 40th annual meeting on association for computational linguistics, páginas 417-424, 2002.

[38] Twitter, Inc., *Twitter: About the company*, 2016, Acceso: 27-06-2017. `https://about.twitter.com/es/company`

[39] Veale, T. and Hao, Y., *Learning to Understand Figurative Language: from Similes to Metaphors to Irony*, Proceedings of CogSci, 2007.

[40] Vinodhini, g. and Chandrasekaran, R., *Sentiment Analysis and Opinion Mining: A Survey*, International Journal, 2(6), páginas 282-292, 2012.

[41] Waltinger, U., *GermanPolarityClues: A Lexical Resource for German Sentiment Analysis*, Conferencia: Proceedings of the International Conference, Language Resources and Evaluation, 2010.

[42] Wang, Y., Hodges, J. and Tang, B., *Classification of Web documents using a naive Bayes method*, Proceedings on 15th IEEE International Conference on Tools with Artificial Intelligence, páginas 560–564, IEEE, 2003.

[43] Wiebe, J., Bruce, R. and O'Hara, T., *Development and Use of a Gold-Standard Data Set for Subjectivity Classifications*, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, (ACL-99), páginas 246–253, 1999.

[44] Wiebe, J., Wilson T. and Cardie, C., *Annotating Expressions of Opinions and Emotions in Language*, Language resources and evaluation, 39(2-3), páginas 165–210, 2005.

[45] Zhu, X., Kiritchenko, S. and Mohammad, S., *NRC-Canada-2014: Recent Improvements in the Sentiment Analysis of Tweets*, In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), páginas 443–447, 2014.

Apéndice A

Publicación

Las investigaciones descritas en este proyecto han permitido la publicación de:

Escortell M., Giménez M. & Rosso P. (2017) El impacto de las emociones en el análisis de la polaridad en textos con lenguaje figurado en Twitter. Procesamiento del Lenguaje Natural, 58, 85-92. Consultado en `http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5416`

En la publicación se detalla una primera experimentación en la que se obtuvieron resultados similares a los que se han expuesto en este proyecto.

A continuación se añade dicho artículo.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 85-92 recibido 30-11-2016 revisado 12-01-2017 aceptado 09-02-2017

El impacto de las emociones en el análisis de la polaridad en textos con lenguaje figurado en Twitter

The Impact of Emotions on Polarity Analysis using Figurative Language in Twitter

Mª Amparo Escortell Pérez
Universitat Politècnica de València
Camino de Vera, s/n, València, Spain
mapees3@fiv.upv.es

Maite Giménez Fayos, Paolo Rosso
PRHLT Research Center
Universitat Politècnica de València
Camino de Vera, s/n, València, Spain
{mgimenez, prosso}@dsic.upv.es

Resumen: Uno de los retos más complejos a los que se enfrenta el Procesamiento de Lenguaje Natural es el de determinar la polaridad de un tweet (positiva, negativa o neutra) cuando en éste aparece lenguaje figurado, es particularmente complejo en los textos cortos y agramaticales que podemos encontrar en las redes sociales. Este trabajo presenta un estudio exhaustivo sobre la capacidad de distintos recursos léxicos de emociones para analizar la polaridad de un conjunto de datos extraidos de Twitter, detallando el impacto de cada uno de los recursos sobre distintas formas de lenguaje figurado como pueden ser la ironía y el sarcasmo que encontramos profusamente en este corpus. Los resultados obtenidos muestran indicios que apuntan a que la inclusión de información relativa a las emociones ayuda a clasificar correctamente la polaridad tanto a nivel global como a nivel del lenguaje figurado o literal.
Palabras clave: Análisis de sentimientos, emociones, lenguaje figurado, twitter, ironía, sarcasmo, semeval, polaridad.

Abstract: One of the most challenging tasks in Natural Language Processing is to determine the polarity of a tweet (positive, negative or neutral) when figurative language is present, especially in the short and ungrammatical texts that can be found in social media. In this paper we present a comprehensive study of the capacity of several emotional lexicons for Sentiment Analysis of Figurative Language in Twitter, detailing how each resource impacts on different figurative language devices such as sarcasm and irony. There are indications in our results that suggest that using emotional information improves the performance of a Sentiment Analysis model regardless of the presence or not of figurative language in the texts analyzed.
Keywords: Sentiment analysis, emotions, figurative language, twitter, irony, sarcasm, semeval, polarity.

1 Introducción

El análisis de sentimientos es una tarea propia del Procesamiento de Lenguaje Natural (PLN) que trata de determinar la polaridad que un texto pretende transmitir. Generalmente, esta tarea se ha enfocado desde el PLN como una tarea de clasificación automática de un texto en tres clases de polaridad: positiva, negativa o neutra. En el caso del lenguaje literal, las técnicas existentes logran resultados aceptables (Kiritchenko, Mohammad, y Salameh, 2016). Sin embargo, esta tarea es especialmente compleja cuando en el texto encontramos lenguaje figurado, puesto que nos enfrentamos con distintos significados debido al uso de la ironía, la metáfora o el sarcasmo, por lo tanto la polaridad del significado literal puede contrastar fuertemente con el sentimiento que pretende transmitir el sentido figurado. Incluso los seres humanos tenemos dificultad en decidir si una texto es irónico o metafórico. La ironía puede ser muy sutil, mientras que la metáfora se puede representar de muchas formas. Sin embargo, el sarcasmo es más fácil de detectar

ISSN 1135-5948

para los seres humanos. El lenguaje figurado es especialmente común en los textos que podemos encontrar en la web y en las redes sociales, especialmente en Twitter o Facebook. La limitación en la longitud en los textos de la red social Twitter, así como el uso de expresiones plagadas de argot y errores gramaticales, dificulta la comprensión del mensaje. En definitiva, el lenguaje figurado presenta un desafío para el rendimiento de los sistemas de análisis de sentimientos convencionales basados en la semántica léxica de las palabras ya que a menudo resultan insuficientes para detectar los significados indirectos. En el trabajo de Hernández et al., (2016) puede encontrarse un estudio detallado del impacto de la ironía y el sarcasmo en el análisis de sentimientos.

En este trabajo se presenta un estudio del impacto de las emociones en la detección de la polaridad de un tweet. Partimos de la hipótesis de que no todos los recursos disponibles favorecen la detección de la polaridad en igual medida, por lo tanto llevamos a cabo una serie de experimentos para evaluar cómo afectan diferentes recursos de emociones tanto en el lenguaje figurado como en el lenguaje literal. Nuestra metodología está compuesta por dos fases: en la primera de ellas estudiaremos el impacto de los recursos léxicos sobre el entrenamiento de clasificadores que predigan la polaridad del conjunto completo de tweets de la tarea 11 de SemEval2015[1] y a continuación evaluaremos en detalle el impacto de cada uno de los recursos para las diferentes tipologías del lenguaje figurado presentes en este corpus.

El resto del artículo está organizado de la siguiente manera: en primer lugar describiremos el estado de la cuestión en el apartado 2; la metodología se describe en detalle en el apartado 3; a continuación en el apartado 4 se presenta el conjunto de datos sobre el cual validaremos nuestra metodología, que como ya hemos introducido se trata de la tarea 11 de SemEval2015; en el apartado 5 se presentan los experimentos que se han llevado a cabo; a continuación en el apartado 6 evaluaremos los resultados obtenidos y, por último, extraeremos las debidas conclusiones en el apartado 7.

[1] http://alt.qcri.org/semeval2015/task11/

2 Estado de la cuestión

Como ya hemos introducido, la definición más extendida de la tarea de análisis de sentimientos se centra en clasificar los textos en tres categorías: textos positivos, negativos y neutros. Los trabajos pioneros (Pang, Lee, y Vaithyanathan, 2002) abordaron esta tarea como un problema de clasificación supervisada aunque en la literatura también podemos encontrar aproximaciones no supervisadas (Turney, 2002). En el trabajo de Pang y Lee (2008) se recoge un amplio estudio de las distintas técnicas que se han empleado para tratar de resolver la tarea del análisis de sentimientos sobre textos extraídos de Internet.

La detección del lenguaje figurado es una tarea en si misma, y distintas aproximaciones han intentado abordarla. Cuando se trata de analizar los textos, la información disponible en la web se puede utilizar como una fuente de conocimiento para generar características auxiliares. En el trabajo de Veale y Hao (2007) se describe una forma semiautomática de recopilar el conocimiento y la semántica de los estereotipos de la web atacando directamente a las construcciones del lenguaje. Los autores demostraron que alrededor del 20 % de los símiles de la web eran irónicos. Sin embargo, su trabajo no se puede utilizar para detectar la ironía de forma general ya que utilizaba las propias estructuras del lenguaje. Tradicionalmente, el lenguaje figurado se ha intentado detectar explorado las características superficiales de los textos. Por una parte, existen estudios que intentan detectar el lenguaje figurado teniendo en cuenta el orden sintáctico, las propiedades léxicas o los elementos afectivos que componen el texto (Reyes, Rosso, y Buscaldi, 2012; Reyes, Rosso, y Veale, 2013). Por otra parte, otros trabajos se centran en investigar como los hashtags de Twitter se emplean para remarcar una intención figurativa en el mensaje transmitido, en especial para la expresión de la ironía o sarcasmo (Sulis et al., 2016). El interés que despierta la tarea de la detección de la polaridad así como el impacto que tiene sobre ésta el lenguaje figurado motivó en 2015 una tarea en la competición internacional para el la evaluación semántica (*Semantic Evaluation - (SemEval)*) (Ghosh et al., 2015).

Quince equipos participaron en la tarea 11 de SemEval 2015 que fue abordada siguiendo múltiples perspectivas. La mayor parte de los participantes plantearon soluciones super-

visadas para intentar resolver la tarea, predominando dos modelos de aprendizaje automático: las Máquinas de Soporte Vectorial (MSV) y los Modelos Regresión. Dichos modelos se entrenaron utilizando un conjuntos de características cuidadosamente seleccionadas para esta tarea como pueden ser: n-gramas de caracteres, n-gramas de palabras, valores extraídos de distintos léxicos, etc.[2].

Nuestro trabajo pretende extender la aproximación presentada por Hernández et al., (2015), en la cual se abordó la tarea incorporando recursos externos adicionales. Los autores proponen representar un tweet mediante un conjunto de valores de características extraídas de recursos léxicos externos que modelan tanto las emociones como la información psicolingüística contenida en un tweet. Asimismo, el trabajo de Sulis et al., (2016) presenta un análisis de la distribución y correlación de un conjunto de características psicolingüísticas y emocionales extraídas de recursos léxicos para realizar la clasificación de tweets irónicos y sarcásticos.

Sin embargo, a diferencia de los citados trabajos sobre el estudio de las emociones en el lenguaje figurado, en este artículo presentamos un estudio exhaustivo sobre la capacidad de diferentes recursos léxicos de emociones para predecir la polaridad del conjunto de datos de Twitter de la tarea 11 de SemEval2015 detallando cómo afectan estos recursos a los tweets que contienen lenguaje figurado y lenguaje literal.

3 Metodología

En este apartado describiremos la metodología que empleamos para el estudio del impacto de ciertos recursos léxicos sobre la detección de la polaridad.

Se ha trabajado con diferentes recursos: LIWC, EmoLex y Smilies, que se detallarán en el siguiente apartado. Los recursos que hemos estudiado almacenan distintos niveles de información respecto a las palabras. El nivel más básico es la información sobre si una palabra es "positiva" o "negativa" aunque también incluyen otras categorías que indican que emociones están vinculadas a las palabras. Primeramente, se ha realizado una serie de experimentos para determinar la polaridad de los tweets utilizando únicamente las categorías de "positivo" y "negativo" de dichos recursos por separado y a continuación para todas las categorías de los recursos.

El procedimiento que hemos llevado a cabo para evaluar el impacto de cada recurso sobre la detección de la polaridad consistió en una vez tokenizados los datos de entrenamiento y test, desarrollar un estudio ablativo que evalúa cómo el uso de diferentes técnicas, como la bolsa de palabras (*Bag of words (BOW)*) o *Term frequency – Inverse document frequency (Tf-Idf)*, así como los recursos para representar un tweet, afectan a la calidad de la clasificación. Este proceso se explicará en detalle en el apartado 5.

Independientemente de la representación elegida, se ha entrenado un sistema de clasificación automático para inferir la polaridad utilizando la librería scikit-learn (Pedregosa et al., 2011). Teniendo en cuenta que el sistema debía predecir un valor de polaridad continuo, se ha empleado una Máquina de Soporte Vectorial adaptada para regresión (MSVR).

Este estudio se ha realizado tanto a nivel de todo el corpus, como a nivel de los distintos tipos de lenguaje figurado presentes en este corpus. Para dividir los tweets entre aquellos que contienen lenguaje figurado o lenguaje literal utilizamos los hashtags, asumiendo que el usuario etiqueta su propio tweet con el tipo de lenguaje empleado facilitando su comprensión, siguiendo la aproximación presentada por Sulis et al., (2016).

3.1 Recursos

Se han utilizado varios recursos para obtener los sentimientos y emociones de los tweets. Estos recursos son los siguientes:

NRC Word-Emotion Association Lexicon (EmoLex)(Mohammad y Turney, 2010): El recurso NRC Emotion Lexicon es una lista de palabras en inglés con sus correspondientes asociaciones con las ocho emociones básicas de Plutchik (Plutchik, 1980): ira, miedo, anticipación, confianza, sorpresa, tristeza, alegría y el disgusto (*anger, fear, anticipation, trust, surprise, sadness, joy y disgust*) y dos sentimientos: positivo y negativo (*negative y positive*). Este recurso fue manualmente anotado. Si la palabra pertenece a la categoría se indica con un 1, en caso contrario con un 0. En este recurso podemos encontrar 14,182 palabras etiquetadas. En la Tabla 1

[2]Para más información ver el artículo de Gosh et al., (2011).

se muestra un ejemplo de como se codifica la información en este recurso.

Palabra	Categoría	Asociación
dark	anger	0
dark	anticipation	0
dark	disgust	0
dark	fear	0
dark	joy	0
dark	negative	0
dark	positive	0
dark	sadness	1
dark	surprise	0
dark	trust	0

Tabla 1: EmoLex: representación de la palabra *dark*

Linguistic inquiry and word count (LIWC) (Pennebaker, Francis, y Booth, 2001): Este recurso le asocia a cada palabra una serie de categorías. En total hay un conjunto de 64 categorías diferentes y se muestra la asociación para un total de 4485 palabras.

Smilies (Suttles y Ide, 2013): También se utilizó este recurso de *smilies* que clasifica 176 *smilies* diferentes según la emoción asociada a los mismos. En este trabajo, en lugar de asociar un smilie a una de las seis emociones básicas definidas en la teoría de Ekman (alegría, ira, miedo, asco, sorpresa, tristeza) (Ekman, 1972), los autores utilizan los ocho tipos de emociones avanzadas definidas en la teoría de Plutchik. A partir de estos ocho tipos de emociones y utilizando una lista con los hashtags emocionales más frecuentes, los autores de este recurso seleccionaron quince categorías para etiquetar los diferentes *smilies*: feliz, risueño, amoroso, enfadado, triste, llanto, disgustado, sorpresa, beso, guiño, lengua, escéptico, indeciso, avergonzado y maligno (*happy, laugh, love, annoyed, sad, cry, disgust, surprise, kiss, wink, tongue, skeptical, indecision, embarassed y evil*). Se puede apreciar un ejemplo en la Tabla 2.

Emoticono	Emoción
:D	LAUGH
:@	SAD
;-)	WINK
3:-)	EVIL

Tabla 2: Smilies: Clasificación de *smilies*

4 Descripción de la tarea

En la tarea 11 de SemEval 2015 se utilizó un corpus de lenguaje figurado extraído de la red social Twitter el cual presenta un gran número de ironías, sarcasmos o metáforas, sin embargo, no se puede garantizar que se manifieste cualquiera de estos fenómenos en cada uno de los tweets.

La ironía y el sarcasmo normalmente se utilizan para criticar o burlarse y, por lo tanto, sesgar la percepción del sentimiento hacia un valor negativo, por lo que no es suficiente para un sistema determinar simplemente si el sentimiento de un tweet dado es positivo o negativo atendiendo únicamente al lenguaje literal presente en el mismo.

Los organizadores de la tarea proporcionaron el corpus etiquetado siguiendo una escala de 11 puntos que oscilaban desde -5 (muy negativo, para tweets con significados muy críticos) a 5 (muy positivo, tweets con significados muy optimistas). El punto cero de esta escala se utiliza para determinar los tweets neutros.

Los sistemas se evaluaron utilizando dos métricas la distancia coseno y el Error Cuadrático Medio (ECM), ambas métricas apropiadas para problemas de regresión. Por simplicidad computacional se ha decidido evaluar los sistemas aquí presentados únicamente el Error Cuadrático Medio, que define el error cometido por el vector de predicciones $\hat{Y} \in \mathbb{R}^n$ con respecto al vector con los valores correctos para esas n muestras $Y \in \mathbb{R}^n$ siguiendo la siguiente fórmula:

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \qquad (1)$$

Para más información acerca de los detalles de la tarea se pueden consultar las actas de la tarea (Ghosh et al., 2015).

El reto de nuestro sistema es determinar cómo un recurso o la combinación de estos, pueden influir en la capacidad del sistema de clasificación que desarrollemos para predecir el sentimiento presente en un tweet. Validaremos esta metodología propuesta sobre el conjunto de datos que describiremos en detalle en el siguiente apartado.

4.1 Corpus

El conjunto de datos empleado en la tarea 11 de SemEval fueron recolectados a través de la API Twitter4J, que soporta la recolección de tweets en tiempo real mediante la búsqueda consultas. Se utilizaron consultas de hashtags como #sarcasm, #sarcastic y #irony para obtenerlo.

Este conjunto de datos fue recogido durante 4 semanas, del 1 de junio al 30 de junio de 2014. Se eliminaron aquellos tweets que no cumplieran una serie de condiciones como por ejemplo, no contener al menos 30 caracteres sin incluir el hashtag. Asimismo, se filtró también únicamente aquellos tweets que estuvieran escritos en inglés, por lo tanto se trata de una tarea monolingüe.

Cada tweet fue etiquetado por siete anotadores, tres de los cuales eran hablantes nativos de inglés y el resto de los anotadores eran competentes en el idioma. A todos ellos se les pidió asignar una puntuación que oscilaba desde -5 a 5, donde 0 es el valor neutro para aquellos tweets que tienen el mismo valor negativo que positivo.

El sentimiento general de cada tweet se calculó como una media ponderada de las siete puntuaciones donde las puntuaciones de los nativos del inglés valían el doble.

El conjunto de tweets de entrenamiento y test está compuesto por 8000 y 4000 tweets respectivamente. Un ejemplo de varios tweets se muestra en la siguiente tabla:

Tweet	Polaridad
"@erikaekengren: From 50 to 100 degrees in less than a week #kansas" #cantwait **#sarcasm**	-3
Updated my router and it froze. Now I can't access the internet to google a solution. **#irony** #thankfulforsmartphones	-3.48
I've had a lot of wake up calls in my day, but I've always been good at hitting the snooze **#metaphor** #nailedit	0.22

Tabla 3: Ejemplos extraídos del corpus de la tarea 11 de SemEval 2015. En la columna "Polaridad" se especifica la polaridad con la que se puntuó de media el tweet mostrado como ejemplo

5 Experimentación

Nuestra experimentación se ha llevado a cabo en varias fases. En la primera de ellas se ha estudiado el impacto de los recursos sobre el conjunto completo de tweets y a continuación se ha evaluado el grado de impacto para cada uno de los diferentes conjuntos de tweets con lenguaje figurado en el corpus.

Se ha tokenizado el corpus utilizando la librería NLTK (Bird, Klein, y Loper, 2009), se han eliminado las palabras que no aportan información discursiva (*stopwords*) y se ha convertido todo el texto que no fueran *smilies* a minúsculas. A continuación, se han obtenido las representaciones BOW y Tf-Idf de los tweets utilizando la librería scikit-learn (Pedregosa et al., 2011).

Para los recursos de EmoLex y LIWC se han creado diccionarios que representan eficientemente la información. Cada entrada del diccionario corresponde con una categoría emocional y en ella se almacenan las palabras que forman parte de dicha categoría. Para utilizar el recurso de *smilies* se ha tenido que crear un tokenizador ad hoc con cada una de las expresiones regulares necesarias para identificar todos los *smilies*.

Una vez se han obtenido los diccionarios de cada recurso, se han elaborado representaciones vectoriales de las muestras de entrenamiento y test. Cada uno de estos vectores indican para cada tweet, el número de veces que aparece una palabra de las categorías que se tienen en el diccionario. De esta manera se tiene un vector diferente para cada recurso que posteriormente se combinaran para realizar la experimentación. La combinación de estos vectores consiste simplemente en agregar al final del vector del primer recurso el vector del segundo.

Utilizando estas estructuras, se han realizado varios experimentos sobre todo el conjunto del corpus. Primero utilizando únicamente las categorías "positivo" y "negativo", y, a continuación, utilizando todas las categorías disponibles de cada uno de los recursos.

En la última columna de la Tabla 4 se muestran los resultados alcanzados tras combinar diferentes recursos para entrenar una SVR y calculando el resultado utilizando el ECM. Además, para poder comparar los re-

Recurso	**#irony**	**#sarcasm**	**#not**	**otros**	**total**
BOW + EmoLex p/n	0,8466	0,5790	5,8184	6,8359	4,6025
TF-IDF + EmoLex p/n	0,8781	0,5794	5,9199	6,9498	4,6825
BOW + EmoLex Todas	0,8459	0,5787	5,8136	6,8269	4,5972
TF-IDF + EmoLex Todas	0,8770	0,5800	5,9101	6,9374	4,6744
BOW + LIWC p/n	0,8471	0,5817	5,8257	6,8421	4,6074
TF-IDF + LIWC p/n	0,8754	0,5821	5,9148	6,9451	4,6789
BOW + LIWC Todas	**0,8331**	0,5780	5,8076	6,8187	4,5897
TF-IDF + LIWC Todas	0,8583	0,5782	5,8972	6,9265	4,6628
BOW + EmoLex + LIWC p/n	0,8451	0,5788	5,8203	6,8421	4,6022
TF-IDF + EmoLex + LIWC p/n	0,8756	0,5796	5,9173	6,9460	4,6796
BOW + EmoLex + LIWC Todas	0,8338	0,5759	**5,7883**	**6,7972**	**4,5756**
TF-IDF + EmoLex + LIWC Todas	0,8586	**0,5755**	5,8778	6,9054	4,6487
BOW + *Smilies*	0,8484	0,5817	5,8227	6,8407	4,6360
TF-IDF + *Smilies*	0,8748	0,5813	5,9041	6,9355	4,6719
Baseline: Naïve Bayes	-	-	-	-	5,6720

Tabla 4: Resultados obtenidos empleando la métrica ECM evaluando cada uno de subconjuntos de tipos de lenguaje figurado que podemos encontrar en este corpus. Señalamos con la abreviatura "p/n" aquellas representaciones en las que únicamente se emplean las características "positiva" y "negativa" del recurso en cuestión, mientras que aquellos recursos en los que empleamos todas las categorías se señalan como "Todas"

sultados se incluye el ECM de un sistema de control (*Baseline*) Naïve Bayes entrenado con una aproximación de bolsa de palabras facilitado por la organización de la tarea.

Dada esta experimentación preliminar, la segunda fase ha consistido en determinar el grado de impacto de cada uno de los recursos sobre las diferentes expresiones de lenguaje figurado más habituales en este corpus. Para ello se han diferenciado un total de cuatro grupos de tweets, en función de la aparición de los siguientes hashtags: *#irony* (765 tweets), *#sarcasm* (536 tweets), *#not* (981 tweets) y *otros* (1718 tweets). Si un tweet tiene varios hashtags pertenecerá a ambos conjuntos.

En este último conjunto,*otros* se agrupan aquellos los tweets que no forman parte de los tres primeros grupos y que por lo tanto asumimos que se trata de lenguaje literal. Esta separación se ha hecho sobre los datos del test del corpus, aceptando que el usuario ha empleado el hashtag para auto-etiquetar el tipo de lenguaje que su tweet contenía. La organización de SemEval reportó resultados sobre un conjunto de test con metáforas, pero no aparece el hashtag #metaphor en el test y no hemos podido llevar a cabo una separación automática de este conjunto de tweets. Por lo que en el conjunto *otros* aparecerán metáforas, que según las actas de las tarea (Ghosh et al., 2015) es una de las formas de lenguaje figurado más difíciles de clasificar.

Una vez separados los tweets, se ha llevado a cabo la segunda fase de la experimentación en la que se han utilizado como datos de entrenamiento todos los tweets del conjunto de entrenamiento de la tarea, pero como datos de test se ha utilizado cada uno de los grupos acabamos de describir. Al igual que en la primera fase de experimentos, se ha probado cada recurso por separado así como la combinación de ellos. En la Tabla 4 se muestran para cada uno de los grupos los resultados obtenidos con los diferentes recursos.

6 Análisis de los Resultados

Como ha podido comprobarse en los resultados de la experimentación que hemos presentado en el apartado anterior, la inclusión de nuevos recursos nos conduce a mejorar significativamente el comportamiento de los mo-

delos que entrenemos. Sin embargo, el ECM varía considerablemente en función del subconjunto de lenguaje que estemos considerando. Todos los modelos que hemos presentados consiguen mejorar el modelo de control o *baseline*, lo cual nos indica que, efectivamente, los recursos léxicos que aportan información acerca de las emociones ayudan a mejorar la predicción del sentimiento comunicado en un tweet. Cabe destacar que, cuando incluimos la información respecto a todas las emociones disponibles en un recurso léxico, y no únicamente las categorías "positivas" y "negativas", conseguimos mejorar el comportamiento del modelo entrenado independientemente de si estamos ante lenguaje figurado o literal. Además, a pesar de que estamos ante un corpus con una baja frecuencia de *smilies* y por lo tanto la cobertura del léxico smilies es escasa, este recurso también consigue mejorar el sistema. La aproximación que incluye la bolsa de palabras y los recursos afectivos EmoLex y LIWC con todas las emociones obtiene resultados satisfactorios aunque en el caso de los tweets en los que aparece el *hashtag #irony* la aproximación no utiliza el recurso EmoLex es la que presenta un mejor comportamiento mientras que en el caso de los tweets con el *hashtag #sarcasm* el mejor sistema utiliza una representación de palabras basada en Tf-Idf. Sin embargo, la diferencia entre estos sistemas no es significativa y podemos concluir que el mejor sistema es el que emplea las características "BOW + EmoLex + LIWC Todas". El mejor sistema participante en la tarea, ClaC (Ozdemir y Bergler, 2015), obtuvo un EMC 2.117 para lo cual se desarrolló un complejo proceso para la extracción de la polaridad de las palabras en función del contexto en el que aparezcan. Nuestro sistema comparte características con el sistema denominado ValenTo (Farıas et al., 2015), aunque en este trabajo se incluyen más recursos que pretendemos explorar en trabajos futuros.

El estudio de cómo se distribuye el error entre los distintos subgrupos de tweets ha arrojado resultados sorprendentes: la mayor parte del error se concentra en el subgrupo que no contenía ningún tipo de hashtags del conjunto de hashtags estudiando (#irony, #sarcarm, #not), lo cual puede explicarse porque no se puede analizar de modo independiente el impacto de la metáfora sobre el conjunto *otros*. No obstante, se requiere un estudio pormenorizado de los tweets y la polaridad para explicar este fenómeno.

7 Conclusiones y Trabajo Futuro

En este trabajo, se ha presentado un estudio sobre la capacidad de distintos recursos léxicos de emociones para predecir la polaridad de un conjunto de datos extraídos de Twitter. Se ha visto el impacto de cada uno de ellos sobre las distintas formas de lenguaje figurado como la ironía y el sarcasmo y la importancia de desarrollar técnicas capaces de representar esa información para clasificar el sentimiento que el autor emitió en un texto. Se han obtenido unos resultados que apuntan a que la inclusión de información relativa a las emociones ayuda a clasificar correctamente la polaridad tanto a nivel global como a nivel del lenguaje figurado o literal .

Como trabajo futuro, se pretende extender el estudio a distintos algoritmos de aprendizaje automático para comprobar cómo afecta a su comportamiento la inclusión de información emocional recurrente o ruidosa, puesto que las MSV son capaces de descartar aquellas muestras no significativas para la clasificación y son más robustas respecto al ruido. Asimismo, se evaluará de forma sistemática la contribución de la representación de palabras y de nuevos recursos léxicos en la detección de la polaridad de un tweet. Igualmente, se estudiará cómo se puede aumentar la cobertura de los recursos léxicos, es decir el número de palabras que encontramos en el diccionario, utilizando técnicas como por ejemplo la corrección automática del texto, para eliminar, en la medida de lo posible los errores gramaticales presentes en Twitter.

Agradecimientos

Este trabajo se ha desarrollado en el marco del proyecto de investigación SomEMBED (TIN2015-71147-C2-1-P) del Ministerio de Economía y Sostenibilidad (MINECO). Asimismo, el trabajo de la segunda autora ha sido financiado a través del Programa de Ayudas de Investigación y Desarrollo de la Universitat Politècnica de València (PAID 2015).

Bibliografía

Bird, S., E. Klein, y E. Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

Ekman, P. 1972. Universals and cultural differences in facial expressions of emo-

tions. *Nebraska Symposium on Motivation*, 19:207–283.

Farıas, D. I. H., E. Sulis, V. Patti, G. Ruffo, y C. Bosco. 2015. Valento: Sentiment analysis of figurative language tweets with irony and sarcasm. *SemEval-2015*, página 694.

Ghosh, A., L. G., T. Veale, P. Rosso, E. Shutova, J. Barnden, y A. Reyes. 2015. SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. *Proc. 9th Int. Workshop on Semantic Evaluation (SemEval 2015)*, páginas 470–478.

Hernández, I. y P. Rosso. 2016. Irony, sarcasm, and sentiment analysis. En F. Pozzi E. Fersini E. Messina, y B. Liu, editores, *Sentiment Analysis in Social Networks*. Morgan Kaufmann, capítulo 7, páginas 113–128.

Kiritchenko, S., S. Mohammad, y M. Salameh. 2016. Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases. En *Proceedings of the International Workshop on Semantic Evaluation (SemEval), San Diego, California, June*.

Mohammad, S. M. y P. D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. En *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, páginas 26–34. Association for Computational Linguistics.

Ozdemir, C. y S. Bergler. 2015. Clacsentipipe: Semeval2015 subtasks 10 b, e, and task 11. En *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, páginas 479–485.

Pang, B. y L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Pang, B., L. Lee, y S. Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. En *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, páginas 79–86. Association for Computational Linguistics.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, y others. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Pennebaker, J. W., M. E. Francis, y R. J. Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.

Plutchik, R. 1980. Emotion: Theory, research, and experience. *Theories of Emotion*, 1.

Reyes, A., P. Rosso, y D. Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Reyes, A., P. Rosso, y T. Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268.

Rosenthal, S., P. Nakov, S. Kiritchenko, S. Mohammad, A. Ritter, y V. Stoyanov. 2016. SemEval-2015 Task 10: Sentiment Analysis in Twitter. *Proc. 9th Int. Workshop on Semantic Evaluation (SemEval 2015)*.

Sulis, E., I. Hernández, P. Rosso, V. Patti, y G. Ruffo. 2016. Figurative Messages and Affect in Twitter: Differences Between #irony, #sarcasm and #not. *Knowledge-Based Systems*, 108:132–143.

Suttles, J. y N. Ide. 2013. Distant Supervision for Emotion Classification with Discrete Binary Values. *Computational Linguistics and Intelligent Text Processing*, páginas 121–136.

Turney, P. D. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. En *Proceedings of the 40th annual meeting on association for computational linguistics*, páginas 417–424. Association for Computational Linguistics.

Veale, T. y Y. Hao. 2007. Learning to understand figurative language: from similes to metaphors to irony. *Proceedings of CogSci 2007*.