UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Estudio comparativo de herramientas para tareas de Procesamiento de Lenguaje Natural

TRABAJO FIN DE GRADO

Grado en Ingeniería Informática

Autor: Nicolás Díaz Roussel

Tutor: María José Castro Bleda, Tutor externo: José Ángel González Barba

Curso 2019-2020

Resum

En este treball s'analitzaran i compararan diferents eines de processament de llenguatge natural. En concret s'avaluaran els resultats obtinguts en les tasques d'anàlisi de sentiment i resum automàtic. Per a això s'utilitzaran APIs d'accés gratuït i una llibreria. Les eines que s'empraran són MeaningCloud, Google Cloud Natural Language, Microsoft Azure Text Analytics i la llibreria sumy. Per a l'anàlisi de sentiment s'usaran els corpora de la competició TASS 2019, mentres que per al resum automàtic emprarem les respostes a unes enquestes sobre aspectes a favor i a millorar de diverses assignatures.

Paraules clau: Processament de Llenguatge Natural, anàlisi de sentiment, resum automàtic

Resumen

En este trabajo se analizarán y compararán diferentes herramientas de Procesamiento de Lenguaje Natural. En concreto se evaluarán los resultados obtenidos en las tareas de análisis de sentimiento y resumen automático. Para ello se utilizarán APIs de acceso gratuito y una librería. Las herramientas que se emplearán son MeaningCloud, Google Cloud Natural Language, Microsoft Azure Text Analytics y la librería sumy. Para el análisis de sentimiento se usarán los corpora de la competición TASS 2019, mientras que para el resumen automático emplearemos las respuestas a unas encuestas sobre aspectos a favor y a mejorar de varias asignaturas.

Palabras clave: Procesamiento de Lenguaje Natural, análisis de sentimiento, resumen automático

Abstract

In this paper, different natural language processing tools will be analyzed and compared. In particular, the results obtained in the tasks of sentiment analysis and automatic summarization will be evaluated. For this purpose, free access APIs and a library will be used. The tools that will be used are MeaningCloud, Google Cloud Natural Language, Microsoft Azure Text Analytics and the sumy library. For the sentiment analysis we will use the corpora of the TASS 2019 competition, while for the automatic summarization we will use the answers to a survey on the pro and to improve aspects of several subjects.

Key words: Natural Language Processing, sentiment analysis, automatic summarization

Índice general

Índice de figuras

Índice de tablas

1. Introducción

El contenido no estructurado en forma de texto en formato libre, imágenes, audio y vídeo es la materia prima "natural" de las comunicaciones entre personas. Está comúnmente aceptado que el 80 % de la información relevante para las empresas se origina en forma no estructurada, principalmente texto, y que ese contenido no estructurado crece a una velocidad mucho mayor que los datos estructurados [1]. El Procesamiento de Lenguaje Natural se encarga de obtener información útil a partir de dichos datos. Esta rama de la inteligencia artificial consiste en el análisis de datos lingüísticos utilizando métodos computacionales. Ayuda a las computadoras a entender, interpretar y manipular el lenguaje humano. El objetivo del Procesamiento de Lenguaje Natural es generalmente construir una representación del texto que añada estructura al lenguaje natural no estructurado, aprovechando los conocimientos de la lingüística. Esta estructura puede ser de naturaleza sintáctica, capturando las relaciones gramaticales entre los elementos del texto, o más semántica, capturando el significado transmitido por el texto [2, 4].

Hoy en día hay muchas herramientas disponibles para realizar tareas de Procesamiento de Lenguaje Natural. El desafío es seleccionar cuál utilizar, de la gama de herramientas disponibles. Esta elección puede depender de varios aspectos como pueden ser el idioma, el lenguaje (formal o informal) o el tipo de texto (periodístico, científico, mensaje, publicación, etc.). En este trabajo se evalúan cuatro herramientas de PLN: MeaningCloud, Google Natural Language, Microsoft Azure Text Analytics y la librería sumy. Las tres primeras disponen de APIs de pago que realizan distintas tareas de PLN, aunque hasta un cierto número de peticiones mensuales son gratuitas, de forma que el usuario no se debe preocupar por implementar una solución para procesar sus datos. Lo único que debe hacer es enviárselos a la herramienta y analizar posteriormente los resultados. En el caso de la librería sumy ofrece únicamente funciones para el resumen automático y aunque emplearlo puede requerir cierta complejidad más, lo cierto es que en la práctica su uso no difiere bastante al de las otras herramientas en su versión gratuita.

1.1 Objetivos

El objetivo principal de este trabajo es analizar y comparar herramientas para tareas de Procesamiento de Lenguaje Natural. En particular se realizan dos análisis. Por un lado se estudia la aplicación de tres herramientas (MeaningCloud, Google Natural Language y Azure Text Analytics) a la tarea de análisis de sen-

timientos sobre un corpus de tweets propuesto en la competición TASS 2019. Se realiza un análisis comparativo en base a los resultados de cada herramienta para cada una de las variantes del español de distintos países (España, México, Uruguay, Costa Rica y Perú) en las que están redactados los tweets. Por otro lado se estudia el rendimiento de MeaningCloud y de la librería sumy con la tarea de resumen automático. Se evalúan los resultados de resúmenes generados a partir de respuestas a encuestas al respecto de los aspectos a favor y mejorar de varias asignaturas.

1.2 Estructura de la memoria

El presente documento describe la comparativa de herramientas de Procesamiento de Lenguaje Natural en las tareas de análisis de sentimiento y resumen automático. En primer lugar se introducen algunos conceptos sobre Procesamiento de Lenguaje Natural y se exponen los objetivos. En el siguiente capítulo se expone el estado del arte de este campo. A continuación se describen algunas de sus aplicaciones más comunes y tecnologías que realizan tareas de PLN. Posteriormente se detallan las tareas de análisis de sentimiento y de resumen automático, junto con su correspondiente experimentación y análisis de los resultados. Por último se presentan las conclusiones a las que se ha llegado.

2. Estado del arte y aplicaciones

En este capítulo presentaremos el estado del arte en el Procesamiento de Lenguaje Natural desde sus inicios [5, 14] hasta la actualidad [15, 6]. Seguidamente, mostraremos algunas de las aplicaciones más frecuentes, tales como la extracción de entidades, la clasificación de temas o la traducción automática.

2.1 Estado del arte

La investigación en el Procesamiento de Lenguaje Natural ha estado en marcha durante varias décadas que se remontan a finales de la década de 1940. La traducción automática (TA) fue la primera aplicación informática relacionada con el lenguaje natural. Aunque Weaver y Booth iniciaron uno de los primeros proyectos de TA en 1946 sobre la traducción informática basada en la experiencia en la ruptura de códigos enemigos durante la Segunda Guerra Mundial, se acordó en general que fue el memorando de Weaver de 1949 el que dio a conocer la idea de la TA e inspiró muchos proyectos. Sugirió usar ideas de la criptografía y la teoría de la información para la traducción del lenguaje. La investigación en el campo comenzó en varias instituciones de Estados Unidos en pocos años.

Los primeros trabajos en la TA adoptaron el punto de vista simplista de que las únicas diferencias entre los idiomas residían en sus vocabularios y en los órdenes de palabras permitidos, sin tener en cuenta la ambigüedad léxica inherente al lenguaje natural. Esto produjo resultados pobres. No fue hasta 1957, gracias a la publicación de *Estructuras Sintácticas* por Noam Chomsky [39], cuando el campo obtuvo una mejor comprensión de cómo la lingüística convencional podía ayudar a la TA.

Durante este período, otras áreas de aplicación del PLN comenzaron a surgir, como el reconocimiento de voz. La comunidad de procesamiento del lenguaje y la comunidad del habla se dividieron entonces en dos campos, con la primera dominada por la perspectiva teórica de la gramática generativa y reticente a los métodos estadísticos, y la segunda dominada por la teoría de la información estadística y hostil a la lingüística teórica.

Las insuficiencias de los sistemas existentes entonces condujeron al informe ALPAC (Comité Asesor de Procesamiento Automático del Lenguaje de la Academia Nacional de Ciencias - Consejo Nacional de Investigación) de 1966. El informe concluyó que la TA no era realizable a corto plazo y recomendó que no se financiara. Esto tuvo el efecto de detener la mayoría de los trabajos de PLN, al menos dentro de los Estados Unidos.

A pesar de ello, hubo algunos avances significativos. El trabajo teórico de finales de los años 60 y principios de los 70 se centró en la cuestión de cómo representar el significado y el desarrollo de soluciones computacionales trazables que las entonces existentes teorías de la gramática no eran capaces de producir. Junto con el desarrollo teórico, se desarrollaron muchos sistemas prototipo para demostrar la eficacia de determinados principios.

A principio de los 70 hubo demostraciones de que el PLN podía desarrollar sistemas operacionales con niveles reales de procesamiento lingüístico, en sistemas verdaderamente integrales aunque con aplicaciones simples. Fueron capaces de lograr sus limitados objetivos porque incluían todos los niveles de procesamiento del lenguaje en sus interacciones con los humanos. Estos sistemas de demostración inspiraron el nuevo campo, pero pasaron muchos años antes de que otros sistemas incluyeran los niveles más complejos de procesamiento en los sistemas del mundo real. A finales de la década de 1970, la atención se centró en las cuestiones semánticas, la generación de lenguaje natural, los fenómenos del discurso y los objetivos y planes de comunicación.

Hasta los años 80 se emplearon reglas escritas manualmente, que daban problemas para extraer significado del texto, así como a la hora de trabajar con textos específicos de una materia y textos de origen oral. Dichas reglas eran construidas por expertos lingüísticos o gramáticos para tareas particulares.

La década de 1980 dio lugar a una reorientación fundamental. La evaluación se hizo más rigurosa, los métodos de aprendizaje automático que usaban probabilidades se volvieron prominentes y grandes cuerpos de texto anotados (corpus) fueron empleados para entrenar los algoritmos de aprendizaje automático y proporcionaron estándares para la evaluación. El final de la década fue uno de los periodos de mayor crecimiento de la comunidad. Se dispuso de recursos prácticos, gramáticas y herramientas y analizadores. Algunas investigaciones en PLN marcaron temas importantes para el futuro como la desambiguación del sentido de la palabra, las redes probabilísticas, el trabajo sobre el léxico, el reconocimiento del habla y la comprensión del mensaje.

A principios de la década de 1990 el enfoque en el PLN cambió de lo que podría ser posible hacer en un idioma de forma gramaticalmente correcta, a lo que realmente se observa que ocurre en el texto natural. A medida que se disponía de más y mayores corpus, los métodos estadísticos se convirtieron en la norma para aprender las transformaciones que en los enfoques anteriores se realizaban mediante reglas construidas a mano. En el centro de este movimiento se encuentra la comprensión de que gran parte o la mayor parte del trabajo a ser realizado por los algoritmos de procesamiento de lenguaje es demasiado complejo para ser capturado por las reglas construidas por la generalización humana y requieren por tanto métodos de aprendizaje automático. La consiguiente disponibilidad de los recursos textuales de amplio alcance de la web, permitió además esta ampliación de los dominios de aplicación.

Paralelamente a estos avances en las capacidades estadísticas, pero avanzando a un ritmo más lento, se demostró que los niveles más altos de análisis del lenguaje humano son susceptibles de PLN. Los niveles del lenguaje son el método más explicativo para presentar lo que realmente ocurre dentro de un sistema de Procesamiento de Lenguaje Natural. Los niveles más bajos (morfológico, léxi-

co y sintáctico) se ocupan de unidades de análisis más pequeñas y se consideran más orientados a las reglas y, por lo tanto, más susceptibles de ser analizados estadísticamente, mientras que los niveles más altos (con la semántica como nivel medio y el discurso y la pragmática como niveles superiores) admiten una mayor libertad de elección y variabilidad en el uso. El punto clave es que el significado es transmitido por todos y cada uno de los niveles de lenguaje y que como se ha demostrado que los humanos usan todos los niveles de lenguaje para ganar comprensión, cuanto más apto sea un sistema de PLN, más niveles de lenguaje utilizará. Cada uno de estos niveles puede producir ambigüedades que pueden ser resueltas por el conocimiento de la frase completa.

En los últimos diez años del milenio, el campo creció rápidamente. Esto puede atribuirse a la mayor disponibilidad de grandes cantidades de texto electrónico, la disponibilidad de computadoras con mayor velocidad y memoria, y el advenimiento de Internet. Los enfoques estadísticos lograron resolver muchos problemas genéricos de la lingüística computacional y se convirtieron en norma en todo el PLN.

Las redes sociales como Twitter, Facebook e Instagram son hoy en día omnipresentes y juegan un papel imperativo en nuestra vida social. Los usuarios de estos sistemas pertenecen a todos los ámbitos de la vida y a menudo no siguen la regla sintáctica y gramatical de los lenguajes naturales a la hora de expresar sus opiniones. La extracción de información de estos sistemas con enfoques convencionales es más difícil debido a la gran ambigüedad del texto.

En la actualidad el aprendizaje automático es la técnica más empleada en el campo. Este método implica el desarrollo de algoritmos que permiten a un programa inferir patrones a partir de datos de entrenamiento. Se utiliza un proceso iterativo para generar un modelo en base a unos parámetros. Este modelo se utiliza para hacer predicciones o clasificaciones generalizadas acerca de nuevos datos. La clave de la precisión de un modelo es la selección de las características o parámetros apropiados.

En general, el aprendizaje puede ser supervisado, cuando cada elemento de los datos de entrenamiento se etiqueta con la respuesta correcta, o no supervisado, donde no se etiqueta y el proceso de aprendizaje trata de reconocer patrones automáticamente. Un escollo de cualquier enfoque de aprendizaje es la posibilidad de un ajuste excesivo o sobreajuste: el modelo puede ajustarse casi perfectamente a los datos del ejemplo, pero hace malas predicciones para casos nuevos, no vistos anteriormente. Esto se debe a que puede aprender el ruido aleatorio en los datos de entrenamiento en lugar de sólo sus características esenciales. El riesgo de sobreajuste se reduce al mínimo mediante técnicas como la validación cruzada, que divide los datos de ejemplo al azar en conjuntos de entrenamiento y pruebas para validar internamente las predicciones del modelo. Este proceso de partición, entrenamiento y validación de datos se repite a lo largo de varias rondas, y los resultados de la validación se promedian luego a lo largo de las iteraciones.

La actual técnica dominante para abordar los problemas en PLN es el aprendizaje supervisado. Los modelos de aprendizaje automático más frecuentes que se utilizan comúnmente para la resolución de ambigüedades en el conocimiento lingüístico con las principales tareas del PLN son: el Modelo Oculto de Markov

(HMM), los Campos Aleatorios Condicionales (CRF), la Entropía Máxima (MaxEnt), las Máquinas de Vectores Soporte (SVM), los Árboles de Decisión (DT), el Clasificador Bayesiano Ingenuo (naïve Bayes) y el Aprendizaje Profundo (Deep Learning). Las técnicas no supervisadas se proponen cuando no es posible tener un conjunto inicial de documentos etiquetados para clasificar el resto de los artículos. La más común es la Agrupación (Clustering).

Aunque el uso del aprendizaje automático es mayoritario, dependiendo de la aplicación el problema se puede abordar con un método u otro. Por ejemplo para determinados casos puede ser suficiente con emplear una técnica del enfoque simbólico o basado en el léxico. Este utiliza recursos léxicos (términos, frases y expresiones) o sintácticos almacenados en bases de conocimiento (WordNet [1] , ConceptNet [2] , PropBank [3], DBPedia [4], etc.), que para el análisis de sentimiento se etiquetan con una polaridad asociada. En otros casos, las tareas pueden llegar a ser tan complejas que tal vez no sea posible elegir un solo método óptimo. No existe un método puramente estadístico. Cada uso de las estadísticas se basa en un modelo simbólico y las estadísticas por sí solas no son adecuadas para el PLN. Es por ello que los enfoques estadísticos no son incompatibles con los enfoques simbólicos. De hecho, son más bien complementarios. Como resultado, los investigadores desarrollan técnicas híbridas que utilizan las fortalezas de cada enfoque en un intento de abordar los problemas del PLN de manera más eficaz y flexible.

Los recientes avances en materia de inteligencia artificial han demostrado que los métodos eficaces utilizan los puntos fuertes de los circuitos electrónicos de alta velocidad y gran capacidad de memoria/disco, las técnicas de compresión de datos específicas para cada problema, las funciones de evaluación y la búsqueda altamente eficiente.

El Procesamiento de Lenguaje Natural proporciona tanto la teoría como las implementaciones para una gama de aplicaciones. De hecho, cualquier aplicación que utilice texto es un candidato para el PLN. En la siguiente sección se muestran algunas de las aplicaciones más frecuentes que utilizan PLN.

2.2 Aplicaciones

El Procesamiento de Lenguaje Natural engloba un conjunto de tareas que tienen multitud de aplicaciones prácticas, no sólo a nivel tecnológico o del lenguaje, si no también en ámbitos como la medicina, la política o los negocios. A continuación se enumeran algunas de esas tareas y ejemplos de aplicaciones en el mundo real.

2.2.1. Extracción de entidades o de temas

La extracción de temas extrae información relevante como entidades nombradas (personas, lugares, organizaciones, etc.), conceptos así como hechos (fechas,

[1] https://wordnet.princeton.edu
[2] https://conceptnet.io
[3] https://propbank.github.io
[4] https://wiki.dbpedia.org

expresiones de tiempo, cantidades, etc.) de los cuerpos de texto. Estos conceptos pueden ser más o menos específicos según el área de aplicación (por ejemplo nombres de proteínas en un texto científico). Algunos ejemplos en los que se emplea esta técnica son [46, 42, 41]:

- Comprender los recibos y las facturas: La extracción de entidades permite identificar las entradas más comunes en los recibos y las facturas, como las fechas, los números de teléfono, las empresas o los precios. Esto puede ayudar a comprender las relaciones entre las solicitudes y los comprobantes de pago.
- Extraer las entidades más importantes de los documentos: Usa la extracción de entidades personalizada para identificar las entidades específicas de cada dominio en los documentos (muchas de ellas no aparecen en los modelos de lenguaje estándar) sin necesidad de invertir tiempo ni dinero en análisis manuales.
- Información valiosa de los clientes: Identificar y etiquetar campos en documentos; por ejemplo, en correos electrónicos, chats o textos de redes sociales. De esta forma se puede tener una idea global del contenido de las críticas de los clientes.
- Analizar currículos: Los equipos de selección de personal pueden extraer instantáneamente la información más pertinente sobre los candidatos, desde información personal (como el nombre, la dirección, el número de teléfono, la fecha de nacimiento y el correo electrónico), hasta datos relacionados con su formación y experiencia (como certificaciones, títulos, nombres de empresas, aptitudes, etc.).

2.2.2. Clasificación del contenido

La clasificación textos o de contenido clasifica los textos en base a una categorización o taxonomía jerárquica. El dominio del texto no está acotado a priori, por lo que el campo de aplicación puede ser muy variado [6, 17]:

- Procesar y clasificar los incidentes de soporte técnico.
- Filtros de correo electrónico no deseado (spam).
- Mejora de resultados en motores de búsqueda.
- Clasificación de películas.

2.2.3. Análisis de sentimiento

El análisis de sentimiento consiste en el uso de tecnologías de Procesamiento de Lenguaje Natural, analítica de textos y lingüística computacional para identificar y extraer información subjetiva de contenido de diversos tipos. Es una de las tareas más empleadas por la información que aporta y la multitud de áreas en las que puede ser utilizada [16, 18, 42, 41]:

- Sistemas de recomendación: Los sistemas de recomendación pueden beneficiarse extrayendo la valoración del usuario del texto. El análisis de los sentimientos puede utilizarse como una tecnología de subcomponentes para los sistemas recomendadores al no recomendar objetos que reciben una retroalimentación negativa. Películas, series, canciones, anuncios, etc.

- Filtrado de mensajes/spam: En la comunicación online, nos encontramos con lenguaje abusivo y otros elementos negativos. Estos pueden ser detectados simplemente identificando un sentimiento altamente negativo, por ejemplo insultos, y tomando las medidas correspondientes contra él.

- Aplicaciones en negocios y comerciales: Hoy en día la gente tiende a mirar las reseñas de los productos que están disponibles en línea antes de comprarlos. Y para muchos negocios, la opinión online decide el éxito o el fracaso de su producto. Por lo tanto, el análisis de sentimiento juega un papel importante en los negocios. Las empresas también desean extraer el sentimiento de las reseñas en línea para mejorar sus productos y, a su vez, su reputación y ayudar a la satisfacción del cliente. El análisis de sentimiento también puede ser usado en la predicción de tendencias. Mediante el seguimiento de las opiniones del público, se pueden extraer datos importantes sobre las tendencias de ventas y la satisfacción del cliente.

- Aplicaciones en hogares inteligentes: Se supone que los hogares inteligentes son la tecnología del futuro. En el futuro, hogares enteros estarían conectados en red y la gente sería capaz de controlar cualquier parte de la casa usando su móvil o tablet. Recientemente ha habido mucha investigación en el Internet de las Cosas (IoT). El análisis de sentimiento también encontraría su camino en la IoT. Como por ejemplo, basado en el sentimiento o emoción actual del usuario, el hogar podría alterar su ambiente para crear un ambiente relajante y pacífico.

- Política: El análisis de los sentimientos permite el seguimiento de la opinión sobre los temas políticos. Por ejemplo puede ayudar a una organización política a entender qué temas son los más importantes para el votante.

- Inteligencia del Gobierno: Para una eficiente elaboración de normas, puede utilizarse para ayudar a analizar automáticamente las opiniones de la gente sobre las políticas pendientes o las propuestas de regulación del gobierno. Otra posible aplicación es la medición del estado de ánimo del público ante un escándalo o controversia.

2.2.4. Resumen automático

El resumen automático permite resumir el significado de un documento, extrayendo de él sus frases más relevantes. Algunas de sus aplicaciones más comunes son [45, 22, 48]:

- Obtener resúmenes de cualquier clase de texto: artículos de periódicos, libros, revistas, historias, eventos, artículos científicos, pronóstico del tiempo,

mercado de valores, noticias, currículum, libros, música, obras de teatro, películas y discursos.

- Publicación de contenidos: Para los medios y otros publicadores el poder dotar automáticamente de resúmenes a todos sus contenidos permite que sus lectores y visitantes se puedan centrar en la información que más les interesa, aumentando la calidad de servicio.
- Gestión del conocimiento: Allí donde el conocimiento de la organización se materializa en miles de documentos, poder extraer resúmenes automáticos de ellos permite gestionarlos y explotarlos mejor.
- Monitorización de medios: Cuando se necesita monitorizar cientos de fuentes de información, tanto tradicionales como sociales, la generación automática de resúmenes permite enfocarse en los documentos relevantes y ser más eficiente.
- Mejora de resultados en motores de búsqueda

2.2.5. Traducción automática

La traducción automática se refiere a la traducción por máquina de un texto de un idioma humano a otro. La aproximación más usada es la aproximación estadística. El desafío con las tecnologías de traducción automática no es traducir directamente las palabras como haría un diccionario, sino mantener el significado de las frases intacto junto con la gramática y los tiempos verbales. Aunque los conceptos detrás de la tecnología de traducción automática y las interfaces para usarlas son relativamente simples, la ciencia y las tecnologías detrás de ella son extremadamente complejas y reúnen varias tecnologías de vanguardia, en particular el aprendizaje profundo, la lingüística, computación en la nube y APIs Web [49, 6]. En muchos casos esta tarea va ligada a la detección automática del idioma del texto original. Algunas aplicaciones directas y bastante empleadas son:

- Traducción de documentos.
- Traducción de páginas web.
- Generación automática de subtítulos.

2.2.6. Respuesta a preguntas

La respuesta a preguntas tiene por objeto dar una respuesta específica a una pregunta formulada. Las necesidades de información deben estar bien definidas: fechas, lugares, etc. Aquí el Procesamiento de Lenguaje Natural intenta identificar el tipo de respuesta a dar, desambiguando la pregunta, analizando las restricciones establecidas, y con el uso de técnicas de extracción de información. Se considera que estos sistemas son los posibles sucesores de los actuales sistemas de recuperación de información [7].

Es una técnica utilizada principalmente en dos grandes campos:

- Motores de búsqueda.
- Asistentes inteligentes.

3. Herramientas PLN

En este capítulo describimos algunas de las herramientas existentes para realizar tareas de PNL. Por un lado varias que funcionan como servicios online a través de una API: MeaningCloud, Google Cloud Natural Language, Azure Text Analytics y Watson Natural Language Understanding. Por otro lado presentamos una librería abierta de resúmenes automáticos.

3.1 MeaningCloud

MeaningCloud [43] es un producto de APIs de pago por uso de análisis semántico, totalmente personalizable y que se proporciona en modo SaaS (Software-as-a-Service) y on-premises.

Dispone de SDKs para integrar MeaningCloud en una aplicación propia de manera rápida, publicados en tres lenguajes: Java, Python y PHP.

Las tareas que se pueden realizar a través de sus APIs son:

1. Análisis de sentimiento
2. Clasificación de texto
3. Clustering de texto
4. Categorización profunda
5. Extracción de temas
6. Identificación de idioma
7. Lematización, análisis morfológico y sintáctico
8. Análisis de reputación corporativa
9. Resumen automático
10. Análisis de la estructura de documentos

Para poder utilizar dichas APIs puede ser necesario contratar un plan de pago, en función del número de peticiones que realicemos. Una petición a MeaningCloud equivale al análisis de un texto de hasta 500 palabras para las API públicas y 125 palabras para las API Premium. Los planes van desde el gratuito, con el que

podemos hacer hasta 20,000 peticiones al mes y 2 peticiones por segundo, hasta el plan "Business" por $999, que nos permite hacer 4,200,000 peticiones al mes y 15 peticiones por segundo. Si se necesitan más peticiones existe también un plan a medida que se define junto con la empresa.

Disponen de herramientas de personalización para adaptar estas tareas a las necesidades del usuario:

- Diccionarios: Crea nuevas entidades y conceptos, conectados en una ontología, para poder identificar su aparición en un texto
- Modelos de Categorización Profunda: Crea nuevas taxonomías utilizando todas las características del análisis morfosintáctico y semántico
- Modelos de Clasificación: Crea nuevas taxonomías y entrena/configura motores de clasificación para poder categorizar textos según ellas.
- Modelos de Sentimiento: Define la polaridad de (grupos de) palabras cuando aparecen en diferentes contextos y realizan diferentes funciones, para adaptar el análisis de sentimiento a un dominio determinado

Por otra parte, tienen una serie de integraciones que dan acceso a los análisis de MeaningCloud desde diferentes programas y plataformas, haciéndolos más accesibles y eliminando la necesidad de programar. Según las necesidades o herramientas con las que se trabaje, se puede elegir entre hojas de cálculo (Excel y Google Spreadsheets), datos estructurados (RapidMiner y Dataiku), automatización (Zapier), investigación (GATE) y analítica de datos (Qlik).

Por último, ofrecen packs, que son una serie de recursos preelaborados (diccionarios, modelos de categorización profunda, modelos de sentimiento) enfocados en una serie de escenarios habituales listos para su uso inmediato y que aportan una mejora en la precisión, cobertura y relevancia del análisis para esas aplicaciones. Estos packs tienen precios fijos que van desde los $100/mes hasta los $500/mes. Algunos ejemplos de estos packs son:

- Voz del cliente: Analiza el feedback de clientes desde todo tipo de canales (voz, email, redes sociales, etc.) para detectar el sentimiento, los principales temas de la conversación, intenciones de compra, etc.
- Analítica de medios sociales: Extrae información relevante desde medios sociales en tiempo real. Monitoriza lo que se dice sobre tu empresa, filtra los comentarios irrelevantes, analiza y categoriza el contenido de forma semántica y extrae tendencias.
- Analítica para RR.HH.: People Analytics, la solución de analítica para recursos humanos de MeaningCloud, ayuda a las empresas a extraer valiosas conclusiones a partir de la realimentación que sus empleados proporcionan.

3.2 GoogleAPI

La API de Cloud Natural Language [41] proporciona a los desarrolladores tecnologías de comprensión del lenguaje natural. Esta API forma parte de la familia más amplia de la API de Cloud Machine Learning.

Dispone de bibliotecas cliente en varios lenguajes: C#, Go, Java, Node, PHP, Python y Ruby.

Las tareas que se pueden realizar a través de sus APIs son:

1. Análisis de opiniones
2. Análisis de entidades
3. Análisis de opiniones de entidades
4. Análisis sintáctico
5. Clasificación de contenido

Los precios por el uso de Natural Language se calculan cada mes según la función de la API que se utilice y el número de unidades que se evalúen mediante sus funciones. Cada documento que se envía a la API para que se analice se considera una unidad como mínimo. Si los documentos tienen más de 1000 caracteres, se contabilizan como varias unidades, una por cada 1000 caracteres. Los precios están definidos como $/1000 unidades y van desde $0.50 - $2 a partir de 5000 peticiones hasta $0.10 - $0.5 a partir de 5 millones. Si el cliente necesita procesar más de 20 millones de unidades al mes (5 millones de unidades mensuales para la clasificación de contenido), hay que ponerse en contacto con un representante de ventas para que de una solución personalizada.

3.3 AzureAPI

Text Analytics API [50] es un servicio en la nube que proporciona procesamiento avanzado de lenguaje natural sobre texto sin formato.

La API forma parte de Azure Cognitive Services, una colección de algoritmos de aprendizaje automático y de inteligencia artificial en la nube para los proyectos de desarrollo.

Dispone de bibliotecas cliente en varios lenguajes: C#, Go, Java, JavaScript, Python y Ruby.

Las tareas que se pueden realizar a través de sus APIs o biblioteca cliente son:

1. Análisis de sentimiento
2. Extracción de frases clave
3. Reconocimiento de entidades
4. Identificación de idioma

Ofrecen distintos planes en función del número de transacciones mensuales. Una transacción corresponde al número de unidades de 1000 caracteres en un documento que se proporcionan como entrada en una solicitud de Text Analytics API. Los precios oscilan desde el plan gratuito con 5000 transacciones al mes, hasta los 4216.492€/mes por 10,000,000 transacciones. Además, con los planes de pago, si se supera el límite se pueden realizar más transacciones pagando una cuota por cada 1000 nuevas transacciones. Se incluye soporte técnico gratis de facturación y administración de suscripciones.

3.4 Watson Natural Language Understanding

Watson Natural Language Understanding [52] es un conjunto de recursos sofisticados de Procesamiento de Lenguaje Natural que permite a los usuarios analizar texto y extraer información de él.

Dispone de bibliotecas cliente en varios lenguajes: Android#, Java, Node, NET y Swift.

Las tareas que se pueden realizar a través de sus API o biblioteca cliente son:

1. Clasificación de texto
2. Identificación de conceptos de alto nivel
3. Análisis de emociones
4. Extracción de entidades
5. Extracción de palabras clave
6. Extracción de metadatos
7. Extracción de relaciones
8. Análisis semántico
9. Análisis de sentimiento
10. Extracción de entidades y relaciones con modelos personalizados

Los planes que ofrece la herramienta dependen del número de elemento de NLU. Este se define como el producto del número de unidades de datos (1 Unidad de datos = 1 grupo de 10.000 caracteres o menos) y el número funciones de enriquecimiento. Al igual que los otros toolkits, dispone de un plan gratuito hasta los 30,000 elementos al mes y un modelo personalizado incluido. En el plan de pago existen tres niveles que van desde los $0.003/elemento por un máximo de 250,000 elementos hasta $0.0002/elemento a partir de los 5 millones de elementos. Los tres niveles incluyen la opción de incluir modelos personalizados por $800/modelo/mes. Para disponer de más seguridad y aislamiento es necesario ponerse con el departamento de ventas para conseguir un plan personalizado.

3.5 Sumy

Sumy [51] es simplemente una librería y una utilidad de línea de comandos para extraer un resumen de las páginas HTML o de textos simples. Por lo tanto no ofrece servicios como en el caso de las herramientas anteriores y es completamente gratuita.

En su documentación explica qué métodos se pueden emplear para realizar el resumen automático:

1. Random: Método de prueba. Sólo se utiliza durante la evaluación de los resúmenes para la comparación con los otros algoritmos. La idea detrás de esto es que si algún resumen tiene una puntuación peor que este, probablemente sea un algoritmo realmente malo o haya algún fallo o error grave en la implementación.

2. Luhn: Método heurístico. El algoritmo más simple. Es el primero que se conoce y se basa en la suposición de que las frases más importantes son las que tienen las palabras más significativas. Las palabras significativas son las que están más a menudo en el texto pero al mismo tiempo no pertenecen a las ”stop words”. Por eso, si quieres usar esta, necesitas una lista de palabras clave para tu idioma. Sin ella, probablemente produciría muy malos resultados.

3. Edmundson [29]: Método heurístico que consiste en la mejora del método Luhn mencionado anteriormente. Edmundson añadió 3 heurísticas más al método para medir la importancia de las frases. Encontró las llamadas palabras pragmáticas, las palabras que están en los encabezados y la posición de los términos extraídos. Por lo tanto, este método tiene 4 sub-métodos y la combinación adecuada de ellos resulta en el método Edmundson. Este método es el que más depende del idioma porque necesita la lista de palabras de bonificación (comparativos, superlativos, términos de valor, términos de causalidad) y palabras de estigmatización (expresiones repetitivas, con poco significado, evasivas), excepto ”stop words” (llamadas aquí ”null words”, palabras irrelevantes).

4. LSA (Latent Semantic Analysis) [30]: El análisis semántico latente [26, 27] es un método algebraico que extrae estructuras semánticas ocultas de frases y palabras que se utilizan popularmente en la tarea de resumir textos. Es un enfoque de aprendizaje no supervisado que no requiere ningún tipo de conocimiento externo o de entrenamiento. LSA captura el texto del documento de entrada y extrae información como las palabras que frecuentemente aparecen juntas y las palabras que se ven comúnmente en diferentes oraciones. Un gran número de palabras comunes entre las frases ilustran que las frases están relacionadas semánticamente. La técnica LSA se aplica para extraer del informe las palabras relacionadas con el tema y el contenido importante que transmite las frases. La ventaja de adoptar los vectores LSA para la síntesis en lugar de los vectores de palabras es que las relaciones conceptuales, tal como se representan en el cerebro humano, se captan naturalmente en el LSA. La elección de la frase representativa de cada escala

de la capacidad asegura la relevancia de la frase para el documento y asegura la no redundancia. [21]. Es el método más avanzado pero también el más complicado computacionalmente. Es también el que utilizamos en este estudio.

5. LexRank [32] y TextRank [33]: Enfoque no supervisado inspirado en los algoritmos PageRank [34] y HITS [31] - algoritmos inspirados en la world wide web. Tratan de encontrar conexiones entre las frases e identificar las que están conectadas con las palabras/temas más significativos.

6. KL-Sum [35]: Método que de forma voraz añade frases a un resumen siempre y cuando disminuya la Divergencia KL (Kullback-Leibler).

7. Reducción: Resumen basado en grafos, en el que la relevancia de una frase se calcula como la suma de los pesos de sus aristas con respecto a otras frases. El peso de una arista entre dos oraciones se calcula de la misma manera que el TextRank.

4. Análisis de sentimiento

En este capítulo exponemos en qué consiste el análisis de sentimiento [9, 16, 18] así como los distintos métodos empleados para realizarlo [38] . Posteriormente describimos cómo se ha llevado a cabo la experimentación de esta tarea, las métricas empleadas y los resultados obtenidos con dichas métricas, en primer lugar comparando las herramientas por país y seguidamente por clase.

4.1 Descripción

Una parte vital de la era de la información ha sido averiguar las opiniones de otras personas. En la era pre-web las organizaciones realizaban encuestas de opinión, sondeos para conocer el sentimiento y la opinión del público en general sobre sus productos o servicios. La era de Internet y los avances de la tecnología web han cambiado la forma en que la gente expresa sus puntos de vista y emociones. Ahora se hace principalmente a través de entradas de blog, foros, sitios web de reseñas, redes sociales, etc. Los usuarios generan un gran volumen de datos de críticas, comentarios, recomendaciones, clasificaciones y respuestas. Sobre temas como productos, personas, organizaciones o un servicio. Esto presenta muchos desafíos ya que las organizaciones y los individuos intentan analizar y comprender la opinión colectiva de los demás. Desafortunadamente, encontrar fuentes de opinión, monitorizarlas y luego analizarlas son tareas hercúleas. No es posible encontrar manualmente las fuentes de opinión en línea, extraer los sentimientos de ellas y luego expresarlas en un formato estándar. Por lo tanto, surge la necesidad de automatizar este proceso y el análisis de los sentimientos es la respuesta a esta necesidad.

Las tareas de análisis de los sentimientos incluyen [8, 9, 38, 16, 18]:

- La detección de la polaridad de los sentimientos expresados en el texto: puede ser una clasificación binaria (positiva o negativa), una clasificación multiclase (extremadamente negativa, negativa, neutra, positiva o extremadamente positiva), aunque algunos métodos producen clasificaciones más finas o calificaciones de intensidad continua. Pueden ser específicas del dominio de análisis. Por ejemplo, opiniones sobre política, finanzas, un producto o un fármaco.
- El reconocimiento de las emociones: se centra en la extracción de un conjunto de etiquetas de emociones. Está muy interrelacionada con la clasificación de la polaridad y son interdependientes en la medida en que algunos modelos de clasificación de sentimientos lo tratan como una tarea única al inferir

la polaridad asociada a una frase directamente de las emociones que ésta transmite. En muchos casos, de hecho, el reconocimiento de las emociones se considera una subtarea de la detección de la polaridad.

- La identificación del objetivo/tema de los sentimientos: consiste en identificar los objetivos de la opinión en un texto de opinión, es decir, en detectar los aspectos específicos de un producto o servicio que el titular de la opinión está criticando.
- La identificación del titular de la opinión: reconocer las fuentes de opinión directas o indirectas.
- La identificación de si una entidad textual dada es subjetiva u objetiva.

El análisis de los sentimientos se realiza en varias unidades de texto, desde palabras y conceptos, hasta frases y documentos completos [18]. Para ello existen diferentes técnicas según el caso de uso. Los primeros en surgir fueron los métodos basados en el léxico o en el conocimiento. Estos se basan principalmente en un léxico de sentimientos, es decir, una colección de términos, frases e incluso expresiones lingüísticas de sentimientos conocidos, a los cuales se les asigna un valor que describe su polaridad y su objetividad. Hay dos subclasificaciones para este enfoque en función de la procedencia del léxico. En primer lugar, el enfoque basado en el diccionario, que emplea términos que normalmente se recogen y anotan manualmente (p. ej. WordNet). Esta técnica tiene como mayor inconveniente que no puede tratar con orientaciones específicas de dominio y contexto. El segundo de los enfoques léxicos es el basado en el corpus, que tiene como objetivo proporcionar diccionarios relacionados con un dominio específico. Los diccionarios de este enfoque se generan a partir de un conjunto de términos de opinión iniciales que crece a través de la búsqueda de palabras relacionadas mediante el uso de técnicas estadísticas o semánticas. Estas técnicas basadas en el conocimiento son muy populares por su accesibilidad y economía. Sin embargo, necesitan una base de conocimiento integral que abarque lo suficiente para comprender la semántica asociada con el lenguaje natural o el comportamiento humano. Otra limitación de los enfoques basados en el conocimiento radica en la tipicidad de su representación del conocimiento, que suele estar estrictamente definida y no permite manejar los diferentes matices de los conceptos, ya que la inferencia de los rasgos semánticos y afectivos asociados a los conceptos está limitada por la representación fija y plana. Además, la validez de los enfoques basados en el conocimiento depende en gran medida de la profundidad y la amplitud de los recursos empleados.

Los métodos estadísticos, como las máquinas de vectores de apoyo (SVM) y el aprendizaje profundo (deep learning), han sido populares para la clasificación afectiva de los textos. Al alimentar un algoritmo de aprendizaje automático con un gran corpus de entrenamiento de textos anotados con su polaridad, es posible que el sistema no sólo aprenda el grado afectivo de las palabras (o grupos de palabras si se emplean bigramas o trigramas) clave, sino que también tenga en cuenta el grado de otras palabras clave arbitrarias y las frecuencias de presencia de términos. Sin embargo, los métodos estadísticos suelen ser semánticamente débiles, es decir, los elementos léxicos en un modelo estadístico tienen poco valor predictivo individualmente. En consecuencia, los clasificadores estadísticos de

texto sólo funcionan con una precisión aceptable cuando se les da una entrada de texto suficientemente grande. Así pues, si bien estos métodos pueden clasificar afectivamente el texto del usuario a nivel de página o de párrafo, no funcionan bien en unidades de texto más pequeñas como las oraciones o las frases.

Los enfoques híbridos de la computación afectiva y el análisis de los sentimientos, por último, explotan tanto las técnicas basadas en el conocimiento como los métodos estadísticos para realizar tareas como el reconocimiento de las emociones y la detección de la polaridad a partir de datos de texto o multimodales. Estos métodos tienen por objeto comprender mejor las reglas conceptuales que rigen el sentimiento y los indicios que pueden transmitir esos conceptos desde la realización hasta la verbalización en la mente humana. En los últimos años, estos enfoques están situando gradualmente la informática afectiva y el análisis de los sentimientos como campos interdisciplinarios entre la PLN y la comprensión del lenguaje natural, pasando gradualmente de las técnicas basadas en la sintaxis a marcos cada vez más conscientes de la semántica en los que se tienen en cuenta tanto el conocimiento conceptual como la estructura de las oraciones.

4.2 Corpora para experimentación

Para la experimentación utilizaremos los corpora del TASS 2019 [12], una competición organizada por la SEPLN (Sociedad Española para el Procesamiento de Lenguaje Natural) [1]. Estos consisten un conjuntos de tweets escritos en variantes del español: el hablado en España, México, Costa Rica, Perú y Uruguay. Cada tweet tiene su clase de sentimiento, que puede ser P (positivo), N (negativo), NEU (neutral) y NONE (sin sentimiento). Los principales desafíos que la herramienta de PLN debe afrontar son la falta de contexto (los tweets son cortos, de hasta 240 caracteres), la presencia de lenguaje informal como faltas de ortografía, onomatopeyas, emojis, hashtags, nombres de usuario, etc., las similitudes entre las variantes y el desequilibrio entre clases.

Tabla 4.1: Número de tweets por clase por país

	es	mx	cr	pe	uy
N	663	745	459	485	587
NEU	195	119	151	368	290
NONE	254	111	220	176	82
P	594	525	336	435	469
Total	1706	1500	1166	1464	1428

Como se puede ver en la tabla, los cuerpos de entrenamiento están desequilibrados y tienen un sesgo hacia las clases N y P. De hecho, en todas las variantes menos Perú la suma de neutral (NEU) y sin sentimiento (NONE) no alcanza el número de negativos ni de positivos por separado.

Los datos están almacenados en archivos CSV (Comma-separated values), uno por cada país, tal que en la primera columna tienen un número que sirve

[1] http://www.sepln.org/

de identificador, en la segunda está el contenido del tweet y el la tercera su polaridad. Para nuestro estudio únicamente nos interesan las dos últimas columnas.

Tabla 4.2: Tweets ejemplo

Tweet	Clase
me encantan los VAPES gracias por estar ahí todos los días y alegrarnos	P
yddeon la Universidad es fácil porque estudias lo que te gusta JAJAJAJAJAJAJAJAJAJAJA ES TODO MENTIRA	N
Acabo de descubrir a Filip van der Cruyssen y sus increibles #fotografia de #retratos ...yo quiero https://t.co/2T0S6DGxnm	P
para nada, mi cuerpo se programa solo y no duermo más	NEU
Tengo a la mejor persona a mi lado. Como lo sé?.. le encanta el rock, comprende los videojuegos y no se toma snaps con filtro de perro.	P
Con el tiempo he aprendido a que las ”Cosas a mi modo” salen mal y que en manos de Dios y a ”Su Voluntad” todo sale perfecto.	P
Me gusta ser la primera en llegar por que así la gente me busca a mi, y no al revés, por que estoy ciega y nunca encuentro a las personas	NEU
Si Mario no le regresa el Follow a Andrea, a Ixchel, a naye, a Julietta, mínimo a una de ellas, me mato muy netase lo merecen cañón	N
Pobre Wilmer. Si le va mal(muy posible con esa planilla) lo van a terminar crucificando	N
Anoche no dormí un solo segundo. Pero toca ir s la oficina	NONE
Lo cuál es cierto, triste, pero cierto Jonas LA fue un completo fracaso Ojalá la hubieran hecho mejor	N
Solo porque he estado diciéndole uno a uno a los usuarios de TW que no se embarquen con ellos	N
El sábado me dijeron ”yo te he visto antes, pero no te he hablado porque... tú eres renegona, ¿no? pero renegona FEO”	N
El Universo es infinito y como tal quiere que tu tambien lo seas, no importa lo que suceda, en tus manos esta el poder de cambiarlo	NEU
Mi amigo Adrián: Nadie en la uni quiere algo serio ,sobre todo los del salón. Yo :Me ofendes ,yo si quiero algo serio	N
Las remuneraciones económicas son lo que todos esperan, pero un ”me encanta lo que has hecho”, lo vale todo. El cliente es primero.	P
graaaaaaacias! Marido trabaja harta tarde así que creo q las pizzas o las compro o las hago yo	NEU
Por cierto, ahí no quiero escuchar lamentos compungidos de los arrepentidos. Aguanten al firme lo q hayan votado!	N
toda la tarde trabajando con buena música y seguiré un buen rato mas ...hasta que las velas ardan !	P
jajajaja, hay boludos q te siguen para dejarte enganchado siguiéndolos. Pues no, q se vayan ahí mismo!	N
lo he puntuado con 3.5 estrellas y me ha dolido bc tenía expectativas muy altas y me ha decepcionado un poco	N
te compras una base de pizza y le echas tomate natural y 0 grasas y listo	NONE
jejeje no pasa nada muchas gracias ‘	P
Es que el batido es mucho batido	NEU

Para obtener los resultados de cada herramienta leemos uno a uno cada tweet y realizamos una petición a la API. Una vez tenemos los valores devueltos por la API, los analizamos en base a una serie de métricas descritas en la siguiente sección.

4.3 Métricas para la experimentación

A continuación describimos las técnicas más comunes para la evaluación de resultados de modelos de clasificación. En nuestro estudio empleamos una librería de Python llamada scikitlearn [47], donde todas estas métricas, entre otras, están implementadas y simplemente tenemos que aportar los datos de entrada: valores reales, valores predichos y para algunas el parámetro promedio.

Matriz de confusión

La matriz de confusión es una tabla en la que se visualiza los resultados predichos para cada clase por un modelo de aprendizaje automático frente a los valores reales. Si tomamos como ejemplo la clase positivo (P) podemos observar dos valores posibles: verdadero positivo (TP), que será el número de veces que el modelo clasifica un valor positivo como positivo, y falso positivo (FP), que será el número de veces que clasifica un valor no positivo como positivo.

Tabla 4.3: Matriz de confusión

	Positivos predichos	Negativos predichos
Positivos reales	TP	FN
Negativos reales	FP	TN

Accuracy

El valor de la *accuracy* computa la exactitud de las predicciones correctas sobre el total de muestras. Si el conjunto completo de etiquetas pronosticadas para una muestra coincide estrictamente con el verdadero conjunto de etiquetas, entonces la precisión del subconjunto es 1; de lo contrario es 0. La *accuracy* es, intuitivamente, la capacidad del clasificador de clasificar correctamente una muestra cualquiera.

Precision

La *precision* es la relación entre valores de la clase clasificados correctamente y el total de valores etiquetados como de esa clase por el modelo. La *precision* es, intuitivamente, la capacidad del clasificador de no etiquetar como de una clase una muestra que no es de dicha clase. El mejor valor es 1 y el peor es 0.

$$\texttt{Precision} = \frac{TP}{TP + FP} \tag{4.1}$$

Recall

La *recall* es la relación entre valores de la clase clasificados correctamente y el total de valores reales de esa clase. La *recall* es intuitivamente la capacidad del

clasificador para encontrar todas las muestras positivas. El mejor valor es 1 y el peor valor es 0.

$$\texttt{Recall} = \frac{TP}{TP + FN} \tag{4.2}$$

F1 score

El valor de la F1 puede ser interpretado como un promedio ponderado de la *precision* y la *recall*, donde el valor de la F1 alcanza su mejor valor en 1 y el peor valor en 0. La contribución relativa de la *precision* y la *recall* a la puntuación de la F1 son iguales. La fórmula para la puntuación de la F1 es:

$$\texttt{F1} = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{4.3}$$

En nuestro caso de estudio hay varias clases en las que puede ser etiquetada cada muestra (P, N, NEU, NONE). Por ello realizamos una ponderación de las métricas calculadas para cada clase. En concreto empleamos como parámetro promedio la ponderación macro, aunque también hay otras opciones:

- micro: Calcula las métricas globalmente contando el total de verdaderos positivos, falsos negativos y falsos positivos.
- macro: Calcula las métricas de cada etiqueta, y encuentra su media no ponderada. Esto no tiene en cuenta el desequilibrio de las etiquetas.
- ponderada: Calcula las métricas de cada etiqueta, y encuentra su promedio ponderado por el soporte (el número de instancias verdaderas de cada etiqueta). Esto altera la *macro* para tener en cuenta el desequilibrio de la etiqueta; puede dar lugar a una puntuación F que no está entre la *precision* y la *recall*.

4.4 Experimentación con las distintas herramientas

Cada herramienta de PLN tiene sus particularidades que deben ser tenidas en cuenta a la hora de realizar el análisis. Lo que nos interesa en nuestro caso es básicamente obtener los resultados de cada herramienta y transformarlos para que sean homogéneos, adaptándolos a las clases de los datos de test del TASS 2019 (P, N, NEU, NONE).

Los posibles valores que devuelve la API de MeaningCloud como resultado del análisis de sentimiento son: P+ (muy positivo); P (positivo); NEU (neutral); N (negativo); N+ (muy negativo); NONE (sin sentimiento). La conversión de valores es sencilla, pues simplemente debemos convertir los P+ y los N+ a P y N respectivamente.

GoogleAPI no representa la polaridad de forma tan similar a nuestros datos de test. Emplea dos valores numéricos, *score* y *magnitude*. El *score* de las opiniones oscila entre -1.0 (negativo) y 1.0 (positivo), y corresponde a la tendencia emocional general del texto. La *magnitude* indica la intensidad general de la emoción (tanto

positiva como negativa) en un determinado texto, entre 0.0 y +inf. A diferencia del *score*, la *magnitude* no está normalizada; cada expresión de emoción en el texto (tanto positiva como negativa) contribuye a la magnitude de este (por ende, las magnitudes podrían ser mayores en bloques de texto más extensos). Una puntuación neutral (*score* alrededor de 0.0) podría indicar que un documento tiene un nivel de emoción bajo o emociones mixtas, con valores tanto positivos como negativos que se cancelan unos a otros. Por lo general, se pueden usar valores *magnitude* para eliminar la ambigüedad de estos casos, ya que los documentos verdaderamente neutros (NONE en nuestro caso) tendrán un valor de *magnitude* bajo, mientras que los documentos mixtos (NEU en nuestro caso) tendrán valores de magnitud mayores.

Las opiniones ”claramente positivas” y ”claramente negativas” varían según los clientes y los casos prácticos. Es posible obtener resultados diferentes para cada situación específica. Se recomienda definir un límite que funcione para el caso particular y luego ajustarlo después de probar y verificar los resultados. Por esto empleamos un umbral de 0.05, ya que es el valor que mejor resultados aporta. De esta forma lo valores cuya puntuación es menor que -0.05 son considerados negativos, los mayores que 0.05 son positivos y los intermedios neutros.

Es necesario precisar que con el criterio descrito anteriormente para los resultados de GoogleAPI ningún tweet es clasificado como neutro, por lo que todas las métricas para esa clase tendrán el valor 0 e influirán a la baja en la media de las métricas globales.

La API Text Analytics de Azure ofrece dos versiones de análisis de sentimiento, la versión 2 y la versión 3. La versión 3 de análisis de sentimiento aporta importantes mejoras a la precisión y al detalle de la puntuación y categorización del texto de la API. Es esta última la que empleamos en nuestra experimentación. Esta API clasifica el texto con etiquetas de opinión. Las puntuaciones devueltas representan la confianza del modelo de que el texto es positivo, negativo o neutro. Cuanto más alto sea el valor, que oscila entre 0 y 1, mayor será la confianza. En consecuencia, tomamos como positivos o negativos los valores cuya puntuación más alta sea la positiva o negativa respectivamente. A priori los valores neutrales serían todos aquellos cuya puntuación de la etiqueta neutra sea la más alta. Sin embargo, debemos determinar de alguna manera qué tweets consideramos como NONE. Para ello empleamos la misma técnica que anteriormente con la API de Google. En este caso el criterio para considerar un tweet como neutro es tomar los que tienen sentimiento positivo y negativo suficientemente elevado (mayor que un epsilon). El resto los analizamos como NONE. El valor que mejores resultados arroja es epsilon = 0.05. En consecuencia etiquetamos como neutros (NEU) los tweets con puntuación neutra mayor que la positiva y la negativa, y con estas dos últimas mayores que 0.05. Los demás los consideramos sin sentimiento (NONE).

4.5 Resultados y análisis

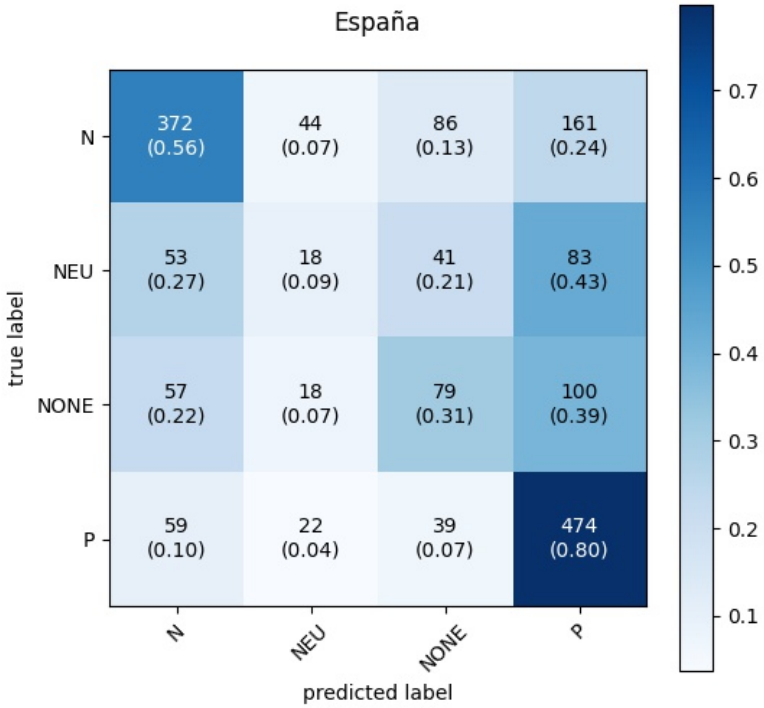

Figura 4.1: Matriz de confusión ponderada de los tweets de España. Generada con scikit-learn.

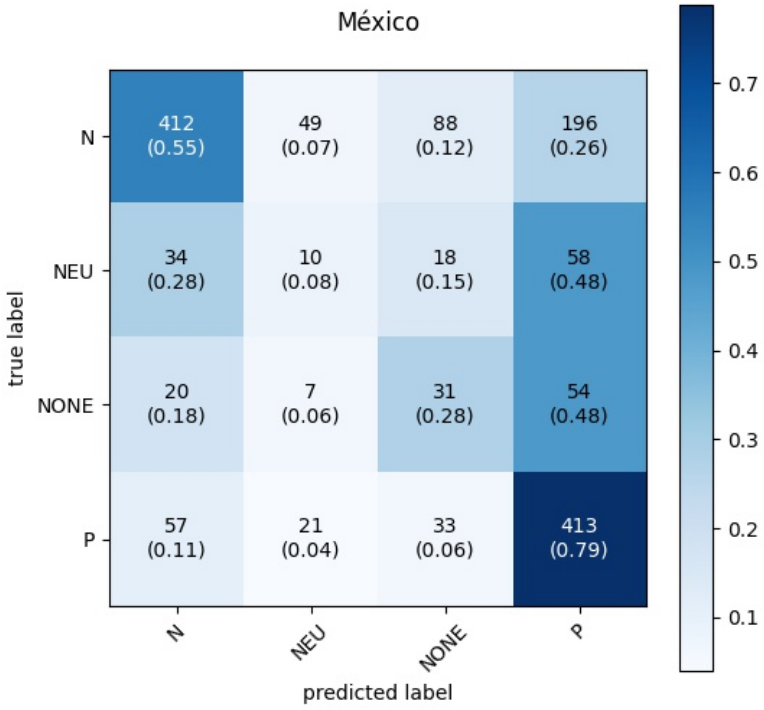

Figura 4.2: Matriz de confusión ponderada de los tweets de México. Generada con scikit-learn

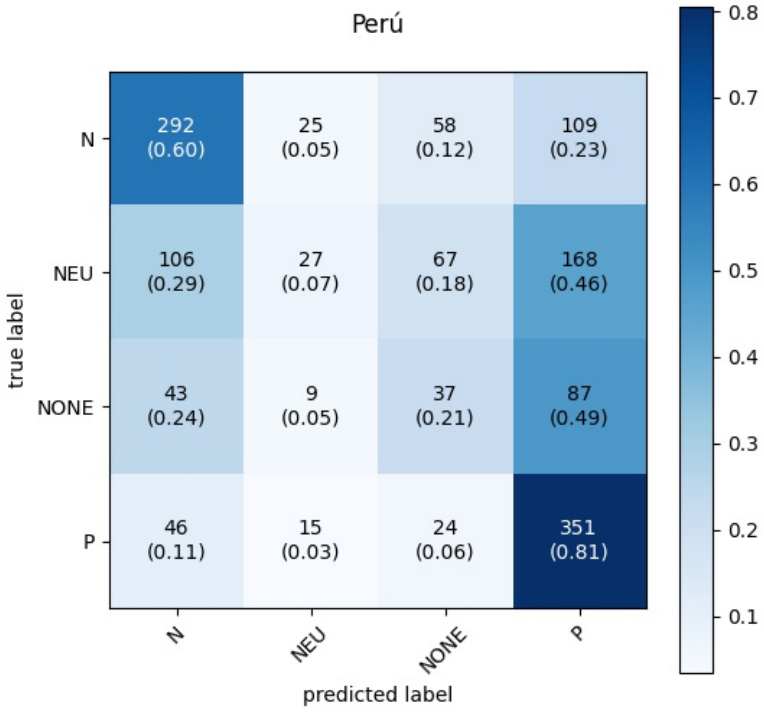

Figura 4.3: Matriz de confusión ponderada de los tweets de Perú. Generada con scikit-learn

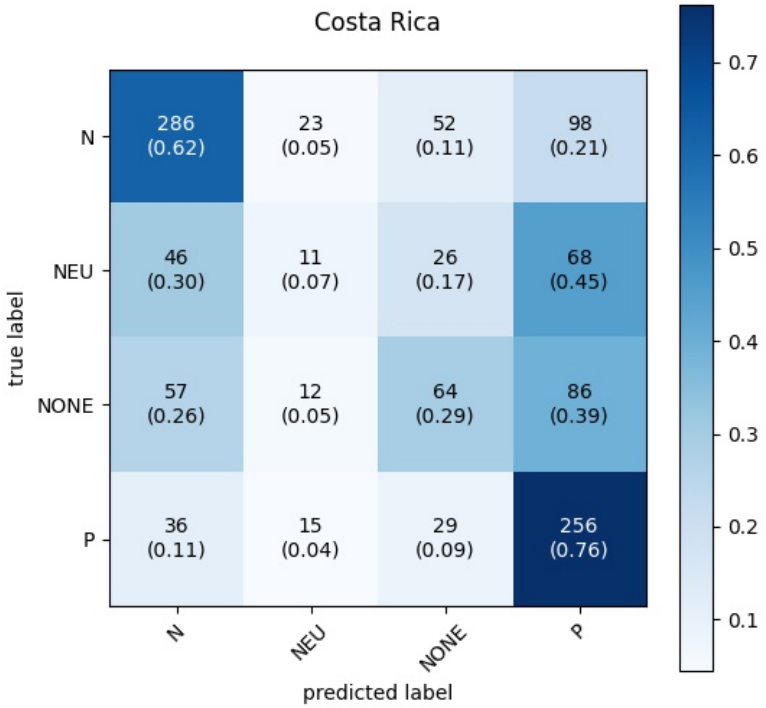

Figura 4.4: Matriz de confusión ponderada de los tweets de Costa Rica. Generada con scikit-learn

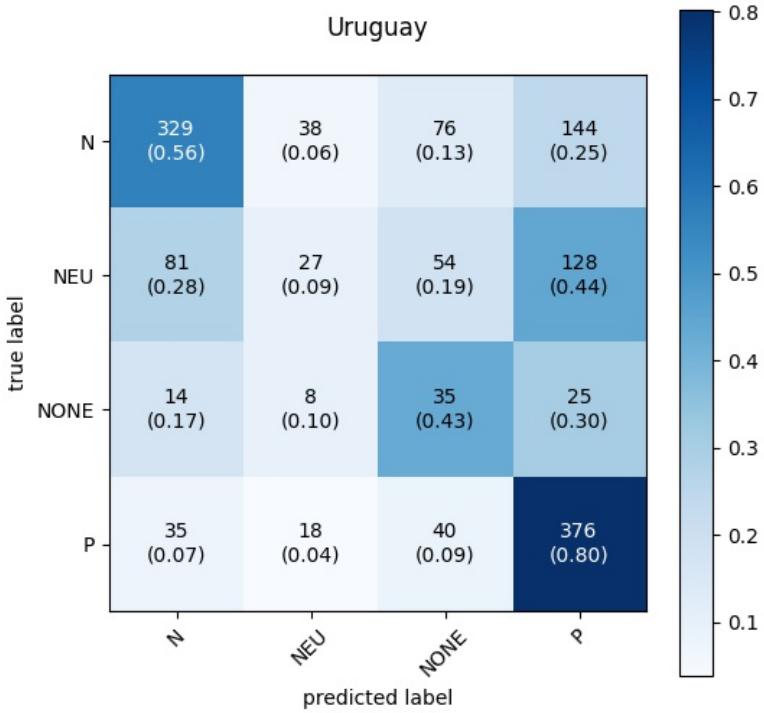

Figura 4.5: Matriz de confusión ponderada de los tweets de Uruguay. Generada con scikit-learn

Las anteriores son matrices de confusión ponderadas. Esto es, se indica el porcentaje de valores clasificados como de una clase sobre el conjunto de valores reales de cada clase. Para todos los países la mayoría de muestras de clase real positiva se clasifican por las herramientas como positivas, concretamente cerca del 80 %. El segundo mejor valor es para las muestras negativas, que son estimadas como tal en casi un 60 % de los casos. Por otra parte, las muestras de clase NONE y NEU no obtienen buenos resultados y tienen como clase predicha más común la P, salvo en el caso de Uruguay, donde las NONE son mayoritariamente (43 %) etiquetadas correctamente.

Los resultados son bastante parecidos entre países. Como diferencias significativas cabe destacar que México y Perú clasifican como P casi el 50 % de valores NONE, mientras que Uruguay es la única que clasifica correctamente como NONE más del 40 % de estos.

Tabla 4.4: Resultados para la tarea TASS 2019 de análisis de sentimiento en español (España) con los distintos toolkits.

Tool	Accuracy	Precision	Recall	Macro F1
MeaningCloud	0.536	0.447	0.449	0.437
Google	0.539	0.355	0.397	0.364
Azure	0.585	0.485	0.479	0.480

Tabla 4.5: Resultados para la tarea TASS 2019 de análisis de sentimiento en español (México) con los distintos toolkits

Tool	Accuracy	Precision	Recall	Macro F1
MeaningCloud	0.536	0.409	0.423	0.398
Google	0.574	0.358	0.387	0.358
Azure	0.623	0.454	0.466	0.455

Tabla 4.6: Resultados para la tarea TASS 2019 de análisis de sentimiento en español (Uruguay) con los distintos toolkits

Tool	Accuracy	Precision	Recall	Macro F1
MeaningCloud	0.510	0.443	0.506	0.418
Google	0.517	0.312	0.383	0.333
Azure	0.585	0.480	0.526	0.486

Tabla 4.7: Resultados para la tarea TASS 2019 de análisis de sentimiento en español (Perú) con los distintos toolkits

Tool	Accuracy	Precision	Recall	Macro F1
MeaningCloud	0.455	0.381	0.405	0.368
Google	0.462	0.304	0.397	0.336
Azure	0.531	0.461	0.468	0.441

Tabla 4.8: Resultados para la tarea TASS 2019 de análisis de sentimiento en español (Costa Rica) con los distintos toolkits

Tool	Accuracy	Precision	Recall	Macro F1
MeaningCloud	0.509	0.429	0.435	0.422
Google	0.508	0.351	0.402	0.359
Azure	0.570	0.480	0.476	0.471

Azure es la herramienta que obtiene los mejores resultados en todas las métricas para todas las variantes. Google obtiene valores bajos de forma general salvo en la *accuracy*. Recordemos que GoogleAPI no predice ningún valor como neutro, de forma que las métricas para esa clase son todas 0 y hacen que la media de las métricas globales sea inferior. Los valores con MeaningCloud son intermedios, con la accuracy menor aunque no de forma significativa.

Tabla 4.9: Resultados para la tarea TASS 2019 de análisis de sentimiento de los tweets N de todos los países con los distintos toolkits

Tool	Precision	Recall	Macro F1
MeaningCloud	0.701	0.512	0.589
Google	0.678	0.566	0.615
Azure	0.703	0.661	0.679

Tabla 4.10: Resultados para la tarea TASS 2019 de análisis de sentimiento de los tweets P de todos los países con los distintos toolkits

Tool	Precision	Recall	Macro F1
MeaningCloud	0.534	0.756	0.625
Google	0.497	0.842	0.624
Azure	0.609	0.775	0.681

Tabla 4.11: Resultados para la tarea TASS 2019 de análisis de sentimiento de los tweets NEU de todos los países con los distintos toolkits

Tool	Precision	Recall	Macro F1
MeaningCloud	0.204	0.100	0.128
Google	0	0	0
Azure	0.240	0.150	0.178

Tabla 4.12: Resultados para la tarea TASS 2019 de análisis de sentimiento de los tweets NONE de todos los países con los distintos toolkits

Tool	Precision	Recall	Macro F1
MeaningCloud	0.248	0.405	0.293
Google	0.170	0.165	0.161
Azure	0.337	0.346	0.329

Analizando por clase de sentimiento se observan diferencias importantes entre clases. En primer lugar, lo valores de las métricas son mucho más altos para las clases P y N que para NEU y NONE. Los más bajos se dan para la clase neutra, es decir, las herramientas no son capaces de clasificar bien si una muestra tiene el mismo nivel de sentimiento positivo y negativo. También de forma genérica ocurre que las herramientas aciertan al clasificar muestras como negativas más que al clasificarlas como positivas (*precision* de N >*precision* de P), mientras que encuentran mejor los valores positivos que los negativos (*recall* P >*recall* N).

Si nos fijamos por herramienta cabe destacar que Google obtiene resultados muy malos con la clase neutra, lo cual ya sabíamos debido a que no predice ninguna muestra como neutra. Algo similar sucede con la clase NONE aunque con valores ligeramente superiores. Esto quiere decir que tampoco clasifica correctamente valores que no tienen sentimiento. No obstante, los resultados con P y N son bastante mejores. Particularmente da un resultado muy bueno con la *recall* de P, lo que quiere decir que predice correctamente la mayoría de valores positivos reales.

5. Resumen automático

En este capítulo exponemos en qué consiste el resumen automático [23, 28, 21, 22], cómo se clasifica [23] así como los distintos métodos empleados para realizarlo [21, 23, 28, 22]. Posteriormente describimos cómo se ha llevado a cabo la experimentación de esta tarea, las métricas empleadas y los resultados obtenidos con dichas métricas.

5.1 Descripción

Con el crecimiento exponencial de la información en línea la extracción de información o el resumen los textos se ha convertido en algo necesario para los usuarios. A medida que las tecnologías de la información y las comunicaciones crecen a gran velocidad, se dispone de un gran número de documentos electrónicos en línea y los usuarios tienen dificultades para encontrar la información pertinente. Además, Internet ha proporcionado grandes colecciones de texto sobre una variedad de temas. Esto explica la redundancia de los textos disponibles en línea. El resumen de texto es el proceso de extraer información relevante del texto original y presentar esa información al usuario en forma de resumen. El resumen, tal como lo hacen los humanos, implica leer y comprender un artículo, sitio web o documento para encontrar los puntos clave. Los puntos clave se utilizan luego para generar nuevas frases, que forman el resumen. Para los seres humanos, generar un resumen es un proceso sencillo, pero consume mucho tiempo. Por lo tanto, la necesidad de resúmenes automatizados es cada vez más evidente para generar automáticamente el resumen y obtener la idea general de los datos textuales extensos. Huang et al. [36] consideran cuatro objetivos principales en un resumen: cobertura de la información, importancia de la información, redundancia de la información y cohesión del texto.

El resumen automático puede clasificarse a grandes rasgos en dos clases principales: el resumen extractivo y el resumen abstractivo. El resumen extractivo se refiere al proceso de resumen en el que, en primer lugar, se identifican las frases clave y/o las oraciones importantes a partir del texto. La implicación de las frases se determina en función de múltiples características de la frase, como las léxicas, sintácticas, estadísticas y lingüísticas. Más tarde, estas frases importantes se concatenan, normalmente en el orden original, para generar un resumen. Es un método simple y robusto para el resumen de texto. El resumen extractivo también puede ser percibido como un problema de clasificación en el que tenemos que clasificar cada frase como parte parte del resumen o no. Los principales problemas del resumen extractivo son la información irrelevante en el resumen

generado debido a las largas frases, la resolución de la anáfora (la frase en resumen podría referirse a algún concepto en las frases precedentes) y las frases contradictorias.

El resumen abstractivo, por otra parte, utiliza el Procesamiento de Lenguaje Natural para producir un resumen abstracto que incluye palabras y frases diferentes a las que aparecen en el documento de origen. Este método requiere un análisis profundo de los documentos para identificar mejor los conceptos subyacentes que residen en el documento para un resumen eficaz. Por consiguiente, se requieren métodos lingüísticos para el análisis y la interpretación exhaustivos del texto. Después de identificar la información clave de un documento, el objetivo final del resumen abstractivo es producir un resumen gramaticalmente sólido y coherente. Esta tarea requiere técnicas avanzadas en el ámbito de la generación y modelización de lenguajes. Por lo tanto, es mucho más complejo que el resumen extractivo. Así, el resumen extractivo, debido a su mayor viabilidad, ha alcanzado un estándar en el resumen de documentos.

Sobre la base del número de documentos, los resúmenes de uno y varios documentos son las dos categorías importantes de resúmenes. El resumen se genera a partir de un documento único en el resumen de un solo documento, mientras que en el resumen de varios documentos se utilizan muchos documentos para generar un resumen. Se considera que el resumen de un solo documento se extiende para generar el resumen de múltiples documentos. Pero la tarea de resumir varios documentos es más difícil que la tarea de resumir documentos individuales. La redundancia es uno de los mayores problemas de resumir múltiples documentos. Hay algunos sistemas que tratan con la redundancia seleccionando inicialmente las frases al principio del párrafo y luego midiendo la similitud de la siguiente frase con las frases ya elegidas y si esta frase consiste en un nuevo contenido relevante, entonces sólo se selecciona la primera.

Los resúmenes también pueden ser de dos tipos: genéricos o centrados en la consulta. Los resúmenes centrados en temas o centrados en el usuario son los otros nombres de los resúmenes centrados en la consulta. Dicho resumen incluye el contenido relacionado con la consulta, mientras que en un resumen genérico se proporciona un sentido general de la información presente en el documento.

En función del estilo de salida, hay dos tipos de resúmenes: los resúmenes indicativos y los resúmenes informativos. Los resúmenes indicativos dicen de qué trata el documento. Dan información sobre el tema del documento. Los resúmenes informativos, aunque cubren los temas, dan toda la información en forma elaborada. Existe un resumen más similar a ellos, los resúmenes de evaluación crítica. Estos resúmenes consisten en opiniones de los autores sobre un tema en particular y contienen opiniones, revisiones, recomendaciones, retroalimentaciones, etc. Por ejemplo, los revisores examinan el trabajo de investigación para las revistas y conferencias y envían sus valiosos comentarios al candidato que incluyen la aceptación, el rechazo o la aceptación del trabajo con alguna modificación.

En base al idioma, hay tres tipos de resúmenes: resúmenes multilingües, monolingües y multilingües cruzados. Cuando el idioma del documento de origen y el de destino es el mismo, es un sistema de resumen monolingüe. Cuando el documento de origen está en varios idiomas como el inglés, el árabe y el francés y el resumen también se genera en estos idiomas, entonces se denomina sistema

de resumen multilingüe. Si el documento fuente está en inglés y el resumen generado está en árabe o en cualquier otro idioma que no sea el inglés, entonces se denomina sistema de resumen multilingüe cruzado.

Los métodos empleados en el resumen abstractivo se pueden clasificar en enfoques basados en la estructura y en enfoques basados en la semántica. Los enfoques basados en estructuras captan la información importante en forma de esquemas cognitivos como plantillas, reglas, grafos, árboles, ontologías, etc. Los enfoques basados en estructuras se clasifican además en métodos basados en árboles, en plantillas, en ontologías, en frases hechas y en reglas basadas en el esquema subyacente utilizado. Por otra parte, los enfoques basados en la semántica utilizan la información semántica para generar texto. En los enfoques basados en la semántica, las frases de los verbos y los sustantivos suelen identificarse mediante datos lingüísticos. Los enfoques basados en la semántica se clasifican además sobre la base de la construcción de modelos semánticos. Los principales enfoques para la construcción de modelos semánticos incluyen el modelo semántico multimodal, el método basado en los elementos de información y los métodos basados en grafos semánticos.

La calidad del resumen se mejora en el enfoque basado en la estructura, ya que produce un resumen coherente, menos redundante y con mayor alcance. El método basado en la estructura puede tener algunos problemas gramaticales, ya que no tiene en cuenta la representación semántica del documento. El modelo basado en la semántica proporciona una mejor calidad lingüística al resumen ya que implica la representación semántica del documento de texto capturando las relaciones semánticas. El método semántico supera los problemas de la estructura basada en la estructura, es decir, reduce la redundancia en el resumen, asegura una mejor cohesión y también proporciona un contenido rico en información con una mejor calidad lingüística.

Las principales técnicas para el resumen extractivo en un solo documento incluyen clasificadores bayesianos, clasificadores de árboles de decisión, modelos ocultos de Markov, modelos log-lineales y redes neuronales. Por otra parte, los enfoques para el resumen de documentos múltiples incluyen enfoques basados en plantillas, fusión de información, máxima relevancia marginal, técnicas basadas en grafos y técnicas de agrupación. Según tengamos o no datos de entrenamiento para nuestro modelo, podemos clasificar el enfoque como supervisado o no supervisado. Los enfoques supervisados se clasifican además en enfoques bayesianos, enfoques basados en redes neuronales y campos aleatorios condicionales. Los enfoques no supervisados se clasifican en enfoques basados en lógica difusa, agrupación (clustering), basados en conceptos, basados en grafos y basados en análisis semántico.

En los últimos años se han utilizado métodos algebraicos como el Análisis Semántico Latente (LSA), la Factorización Matricial No Negativa (NMF) y la Descomposición de Matriz Semidiscreta para resumir los documentos. Entre estos algoritmos el más conocido es el LSA, que se basa en la descomposición de Valor Singular (SVD). En este algoritmo se extraen las similitudes entre las frases y las similitudes entre las palabras. Además de la integración, el algoritmo LSA también se utiliza para la agrupación de documentos y el filtrado de información [26].

A la luz de las investigaciones existentes, los enfoques ampliamente utilizados incluyen la heurística y los enfoques supervisados y no supervisados para el resumen extractivo. Los estudios existentes apuntan a un desarrollo relativamente menor del resumen abstractivo en comparación con el resumen extractivo [20].

5.2 Corpora para experimentación

Los textos a resumir consisten en una serie de encuestas a alumnos sobre varias asignaturas. En concreto son las respuestas a dos preguntas por asignatura: aspectos a favor y aspectos a mejorar. Todas las respuestas han sido anonimizadas.

Los datos están organizados en páginas de un archivo Excel, de forma que en cada página están las respuestas a una pregunta de una asignatura.

Tabla 5.1: Ejemplo de respuestas del corpus de encuestas. En la primera columna, aparece el nombre de la asignatura. En la segunda columna, si es una respuesta sobre la pregunta "Aspectos a favor" (F) o "Aspectos a mejorar" (M). En la tercera columna se indica el texto correspondiente.

Asignatura	F/M	Respuesta
A	F	Ninguno.
A	F	Me gusta mucho como se imparten las clases de teoría.
A	M	Si no vas a la rama de hardware vas a ver cosas que no te importan mucho.
A	M	Mejorar los boletines de prácticas.
B	F	Buen aprendizaje y profundidad de SQL.
B	F	Asignatura entretenida y con un profesorado motivado.
B	M	Carga excesiva de trabajos.
B	M	Algunos profesores y la corrección de los exámenes es demasiado estricta.
C	F	Las prácticas son muy entretenidas y aplicadas.
C	F	Buenos profesores, contenidos interesantes y muy útiles.
C	M	Planificación de las prácticas y mínimos en los exámenes.
C	M	Recuperaciones Complejas.
C	M	La asistencia a las prácticas obligatoria sirve de poco si luego de todas formas se tiene que dedicar tiempo extra para su preparación (informes). Hay profesores que no tienen interés en que los alumnos aprendan.
D	F	Buena introducción a la rama de Ingeniería del software.
D	F	Se transmite su utilidad de forma clara. Interesante para los alumnos.
D	M	Es necesario más práctica y cambiar el entorno de desarrollo.
D	M	La enorme carga de trabajo que supone el proyecto de laboratorio es difícil de compaginar con otras asignaturas.
E	F	Tiene vídeos explicativos sobre cada tema.
E	F	Muy bien organizada.
E	M	No hay examen de recuperación. Las prácticas no son muy claras y no ayudan a aprender.
E	M	Las clases podrían ser más entretenidas con ejemplos más entretenidos.
F	F	Javascript resulta interesante.
F	F	Buen profesor, ejercicios propuestos muy buenos que ayudaban a comprender los aspectos vistos en teoría.
F	F	Cubre temas actuales e interesantes que probablemente el alumno usará en el futuro.
F	M	Necesidad de memorizar la api de librerías específicas cuando para nada se corresponde a un ejemplo real.
F	M	Demasiado temario para un cuatrimestre, se exige un conocimiento que no hay tiempo para adquirir.
F	M	Hay demasiadas pruebas distintas, la prueba de laboratorio a veces no tienen nada que ver.

Por otra parte tenemos los resúmenes de referencia correspondientes a las respuestas. Estos se organizan de forma análoga, con las frases que componen el resumen en lugar de las respuestas.

Tabla 5.2: Ejemplo de resúmenes del corpus de resúmenes de referencia. En la primera columna, aparece el nombre de la asignatura. En la segunda columna, si es un resumen a respuestas sobre la pregunta "Aspectos a favor" (F) o "Aspectos a mejorar" (M). En la tercera columna se indica el texto correspondiente.

Asignatura	F/M	Resumen
A	F	Buen Profesorado.
A	F	Difícil pero aceptada por los alumnos.
A	M	No hay evaluación continua. No se evalúa el trabajo diario.
A	M	Mejorar los boletines de prácticas.
B	F	Aprender SQL.
B	F	Buen Profesorado.
B	M	Muy estrictos en la corrección.
B	M	Poco peso de las prácticas (en la nota final).
C	F	Útil y Temario Interesante.
C	F	Buen Profesorado.
C	M	Tests de Prácticas. Difíciles.
C	M	Recuperaciones Complejas.
C	M	Mejorar las explicaciones de las prácticas (facilitar su comprensión).
D	F	Desarrollar un Proyecto de Software "grande". Les gusta el enfoque de programación.
D	F	Asequible-Fácil para los alumnos.
D	M	Los créditos de prácticas no se corresponden con las horas reales de trabajo que hay que realizar.
D	M	Entorno de Desarrollo. Visual Studio.
E	F	Vídeos Flip accesibles a todos los grupos.
E	F	Múltiples actos de evaluación y actividades permiten mejorar la nota.
E	M	Evaluación: El último examen tiene mucho peso. ¿Te lo juegas todo? ¿No recuperable? Mucha carga de trabajo en las evaluaciones. ¿Se evalúan 14 temas?
E	M	Temario Excesivo. Segunda parte del temario difícil.
F	F	Javascript, Node.js, Docker, ZMQ.
F	F	Temario Interesante.
F	M	Mucho Temario (teoría) para tan poco tiempo. Muchos conceptos prácticos y teóricos. Innecesariamente larga y complicada.
F	M	Lo explicado en clases o prácticas no se corresponde con las preguntas de los exámenes. Actos de evaluación desacordes con la teoría, seminarios desproporcionados, prácticas en las que se piden mucho y se explica poco.
F	M	En general, los alumnos no tienen una buena percepción de los profesores.

El número de respuestas y el tamaño de sus correspondientes resúmenes es bastante similar para todas las asignaturas y cuestiones, salvo para la pregunta de aspectos a mejorar de la asignatura F, que es algo superior.

Tabla 5.3: Número de respuestas por encuesta y número de frases del resumen de referencia

	Número de respuestas	Número de frases resumen
A-Afavor	65	3
A-Amejor	63	5
B-Afavor	61	4
B-Amejor	64	5
C-Afavor	61	4
C-Amejor	60	5
D-Afavor	72	4
D-Amejor	76	6
E-Afavor	64	4
E-Amejor	68	5
F-Afavor	74	2
F-Amejor	90	13

Para llevar a cabo la experimentación, juntamos las respuestas de cada asignatura y cuestión en un mismo texto. De esta forma podemos hacer la petición a la API de MeaningCloud, pasándole como parámetro el tamaño del resumen en número de frases, idéntico al de los resúmenes de referencia. Hacemos lo propio con la librería sumy, aunque este caso no es necesario hacer ninguna petición ya que la utilizamos como librería local. Cada resumen que nos devuelven procedemos a evaluarlo en base a las métricas descritas en la sección siguiente.

5.3 Métricas para la experimentación

Tradicionalmente, la evaluación del resumen implica juicios humanos de diferentes parámetros de calidad, por ejemplo, coherencia, concisión, gramática, legibilidad y contenido [40]. Esto es muy costoso y difícil de llevar a cabo con frecuencia. Por eso existen técnicas que permiten medir dichos parámetros de forma automática. Una de las más comunes es ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [19] que tiende a medir la similitud entre el resumen de referencia, que suele ser generado por los seres humanos, y el resumen generado por el sistema. Hay tres variantes de ROUGE ampliamente utilizadas ROUGE-1, ROUGE-2 y ROUGE-L donde los resultados se computan con respecto a los unigramas, los bigramas y la subsecuencia común más larga respectivamente.

ROUGE-N

Formalmente, el ROUGE-N es una *recall* de n-gramas entre un resumen candidato y un conjunto de resúmenes de referencia. ROUGE-N se calcula de la siguiente manera:

`ROUGE-N (recall)=`

$$\frac{\sum_{S\in\{Resmenesdereferencia\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{Resmenesdereferencia\}}\sum_{gram_n\in S} Count(gram_n)}$$

Donde *n* significa la longitud del n-grama, *gramn*, y *Countmatch (gramn)* es el número máximo de n-gramas co-ocurrentes en un resumen candidato y un conjunto de resúmenes de referencia. Está claro que ROUGE-N es una medida relacionada con la *recall* porque el denominador de la ecuación es la suma total del número de n-gramas que ocurren en el lado del resumen de referencia.

Obsérvese que el número de n-gramas en el denominador de la fórmula ROUGE-N aumenta a medida que añadimos más referencias. Esto es intuitivo y razonable porque pueden existir múltiples buenos resúmenes. También nótese que el numerador suma sobre todos los resúmenes de referencia. Esto efectivamente da más peso a la coincidencia de n-gramas que ocurren en múltiples referencias. Por lo tanto, un resumen candidato que contiene palabras compartidas por más referencias es favorecido por la medida ROUGE-N. Esto es de nuevo muy intuitivo y razonable porque normalmente preferimos un resumen candidato que es más similar al consenso entre los resúmenes de referencia.

Podemos calcular el ROUGE-N también como una *precision* observando cuantos n-gramas del resumen candidato están presentes en los resúmenes de referencia. De esta forma sabemos si los resúmenes generados contienen información en exceso o no relevante.

`ROUGE-N (precision)=`

$$\frac{\sum_{S\in\{Resumenesdereferencia\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{Resumenescandidatos\}}\sum_{gram_n\in S} Count(gram_n)}$$

A la hora de comparar el rendimiento de distintos modelos se suele emplear la f-measure:

`ROUGE-N (f-measure)=`

$$\frac{(1+\beta^2)ROUGE-N_{recall}ROUGE-N_{precision}}{ROUGE-N_{recall}+\beta^2 ROUGE-N_{precision}}$$

En nuestro estudio empleamos la F1, es decir, f-measure con β igual a 1.

ROUGE-L

La L hace referencia a la primera letra en inglés de la subsecuencia común más larga (Longest Common Subsequence). Dadas dos secuencias X e Y, una subsecuencia común más larga de X e Y es una subsecuencia común con una longitud máxima. Para aplicar el LCS en la evaluación de resumen, vemos una oración de resumen como una secuencia de palabras. La intuición es que cuanto más larga es la LCS de dos frases de resumen, más similares son los dos resúmenes. Asumiendo que X es un resumen de referencia de longitud m y que Y es un resumen candidato de longitud n, se calcula de la siguiente manera:

`R(lcs)=`

$$\frac{LCS(X,Y)}{m}$$

P(lcs)=

$$\frac{LCS(X,Y)}{n}$$

F(lcs)=

$$\frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}$$

Nótese que ROUGE-L es 1 cuando X = Y; mientras que ROUGE-L es cero cuando LCS(X,Y) = 0, es decir, no hay nada en común entre X e Y.

Una ventaja de usar el LCS es que no requiere coincidencias consecutivas sino coincidencias en secuencia que reflejan el orden de las palabras del nivel de la oración como n-gramas. La otra ventaja es que incluye automáticamente los n-gramas comunes más largos de la secuencia, por lo tanto no es necesario predefinir la longitud de los n-gramas.

5.4 Experimentación con las distintas herramientas

Tabla 5.4: Resultados de la métrica ROUGE con unigramas para todos los resúmenes de las respuestas a las dos preguntas y a todas las asignaturas con los distintos toolkits

Encuesta	Herramienta	ROUGE-1-R	ROUGE-1-P	ROUGE-1-F
A-Afavor	MeaningCloud	0.400	0.162	0.231
	sumy	0.067	0.024	0.035
A-Amejor	MeaningCloud	0.185	0.111	0.139
	sumy	0.148	0.093	0.114
B-Afavor	MeaningCloud	0.538	0.159	0.246
	sumy	0.385	0.109	0.170
B-Amejor	MeaningCloud	0.286	0.178	0.219
	sumy	0.250	0.159	0.194
C-Afavor	MeaningCloud	0.294	0.122	0.172
	sumy	0.235	0.087	0.127
C-Amejor	MeaningCloud	0.517	0.306	0.385
	sumy	0.207	0.140	0.167
D-Afavor	MeaningCloud	0.313	0.233	0.267
	sumy	0.188	0.143	0.162
D-Amejor	MeaningCloud	0.455	0.455	0.455
	sumy	0.250	0.229	0.239
E-Afavor	MeaningCloud	0.346	0.184	0.240
	sumy	0.269	0.163	0.203
E-Amejor	MeaningCloud	0.174	0.211	0.190
	sumy	0.217	0.238	0.227
F-Afavor	MeaningCloud	0.143	0.023	0.040
	sumy	0.000	0.000	0.000
F-Amejor	MeaningCloud	0.451	0.500	0.474
	sumy	0.235	0.273	0.253

De forma general los valores de ROUGE-1 son superiores con MeaningCloud. Se obtienen resultados significativamente buenos con MeaningCloud en *B-Afavor* y *C-Amejor*. Concretamente una ROUGE-1-R superior a 0.5, lo que nos indica que más de la mitad de palabras presentes en el resumen de referencia están también en el resumen generado por la herramienta.

Tabla 5.5: Resultados de la métrica ROUGE con bigramas para todos los resúmenes de las respuestas a las dos preguntas y a todas las asignaturas con los distintos toolkits

Encuesta	Herramienta	ROUGE-2-R	ROUGE-2-P	ROUGE-2-F
A-Afavor	MeaningCloud	0.143	0.056	0.080
	sumy	0.000	0.000	0.000
A-Amejor	MeaningCloud	0.077	0.045	0.057
	sumy	0.000	0.000	0.000
B-Afavor	MeaningCloud	0.083	0.023	0.036
	sumy	0.083	0.022	0.035
B-Amejor	MeaningCloud	0.037	0.023	0.028
	sumy	0.000	0.000	0.000
C-Afavor	MeaningCloud	0.063	0.025	0.036
	sumy	0.063	0.022	0.033
C-Amejor	MeaningCloud	0.286	0.167	0.211
	sumy	0.036	0.024	0.029
D-Afavor	MeaningCloud	0.097	0.071	0.082
	sumy	0.000	0.000	0.000
D-Amejor	MeaningCloud	0.209	0.209	0.209
	sumy	0.070	0.064	0.067
E-Afavor	MeaningCloud	0.040	0.021	0.027
	sumy	0.120	0.071	0.090
E-Amejor	MeaningCloud	0.044	0.054	0.049
	sumy	0.044	0.049	0.047
F-Afavor	MeaningCloud	0.000	0.000	0.000
	sumy	0.000	0.000	0.000
F-Amejor	MeaningCloud	0.160	0.178	0.168
	sumy	0.060	0.070	0.065

Análogamente a lo que sucede con ROUGE-1, para ROUGE-2 los resultados de las métricas con MeaningCloud son superiores a los obtenidos con la librería *sumy*. No obstante, los valores son bastantes más bajos y cercanos a 0, especialmente para *sumy*. Que ROUGE-2 sea 0 quiere decir que el resumen generado por la herramienta y el resumen de referencia no contienen ningún bigrama en común.

Tabla 5.6: Resultados de la métrica ROUGE con la subsecuencia común más larga para todos los resúmenes de las respuestas a las dos preguntas y a todas las asignaturas con los distintos toolkits

Encuesta	Herramienta	ROUGE-L-R	ROUGE-L-P	ROUGE-L-F
A-Afavor	MeaningCloud	0.333	0.135	0.192
	sumy	0.067	0.024	0.035
A-Amejor	MeaningCloud	0.148	0.089	0.111
	sumy	0.111	0.070	0.086
B-Afavor	MeaningCloud	0.462	0.136	0.211
	sumy	0.308	0.087	0.136
B-Amejor	MeaningCloud	0.286	0.178	0.219
	sumy	0.179	0.114	0.139
C-Afavor	MeaningCloud	0.235	0.098	0.138
	sumy	0.176	0.065	0.095
C-Amejor	MeaningCloud	0.517	0.306	0.385
	sumy	0.207	0.140	0.167
D-Afavor	MeaningCloud	0.281	0.209	0.240
	sumy	0.188	0.143	0.162
D-Amejor	MeaningCloud	0.386	0.386	0.386
	sumy	0.205	0.188	0.196
E-Afavor	MeaningCloud	0.308	0.163	0.213
	sumy	0.231	0.140	0.174
E-Amejor	MeaningCloud	0.109	0.132	0.119
	sumy	0.130	0.143	0.136
F-Afavor	MeaningCloud	0.143	0.023	0.040
	sumy	0.000	0.000	0.000
F-Amejor	MeaningCloud	0.431	0.478	0.454
	sumy	0.196	0.227	0.211

Los resultados con la subsecuencia común más larga son similares a los de ROUGE-1, aunque ligeramente inferiores. Esto es, de nuevo la métrica nos indica que MeaningCloud genera mejores resúmenes que sumy.

6. Conclusiones

En este trabajo nos hemos planteado investigar distintas herramientas para tareas de PLN. En particular hemos estudiado las herramientas MeaningCloud, Google Natural Language API, Azure Text Analytics API y la librería sumy. Nos hemos centrado en las tareas de análisis de sentimiento, en la que hemos empleado las tres primeras herramientas, y resumen automático, en la que hemos empleado MeaningCloud y sumy. Para el estudio del análisis de sentimiento hemos utilizado corpora del Taller de Análisis Semántico en la SEPLN del año 2019, también llamado TASS-2019. Para el estudio del resumen automático hemos utilizado encuestas y resúmenes de las respuestas a estas recopilados por nosotros.

Encontramos que los mejores resultados para la primera de las tareas se obtienen con Azure Text Analytics, siendo estos similares a los logrados por los mejores sistemas en el TASS-2019. Además observamos que existe una diferencia significativa entre clases de sentimiento, de forma que las clases positiva y negativa ofrecen resultados mucho mejores que la neutra y la clase sin sentimiento. Por la parte de resumen automático, nuestros resultados indican que MeaningCloud genera resúmenes de mayor calidad que la librería sumy en base la métrica ROUGE.

Bibliografía

[1] Analítica de texto. Consultado en https://www.meaningcloud.com/es/soluciones/analitica-de-texto.

[2] Qué es el Procesamiento de Lenguaje Natural. Consultado en https://www.sas.com/es_es/insights/analytics/what-is-natural-language-processing-nlp.html.

[3] Verspoor K., Cohen K.B.. *Natural Language Processing. Dubitzky W., Wolkenhauer O., Cho KH., Yokota H. (eds) Encyclopedia of Systems Biology. Springer, New York, NY*.

[4] Verspoor K., Cohen K.B.. Natural Language Processing *Encyclopedia of Systems Biology*. Springer, New York, NY, 2013.

[5] Elizabeth D. Liddy Natural Language Processing Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc, 2001

[6] Diksha Khurana, Aditya Koli, Kiran Khatter, Sukhdev Singh. Natural Language Processing: State of The Art, Current Trends and Challenges 2017

[7] Mari Vallez, Rafael Pedraza-Jimenez. Natural Language Processing in Textual Information Retrieval and Related Topics Hipertext.net, num. 5, 2007

[8] David Zimbra, Ahmed Abbasi, Daniel Zeng, Hsinchun Chen. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation *ACM Transactions on Management Information Systems*, 9:2:1–29, 2018.

[9] Harsh Thakkar, Dhiren Patel. Approaches for Sentiment Analysis on Twitter: A State-of-Art study Department of Computer Engineering, National Institute of Technology. 2015

[10] Liz Liddy, Eduard Hovy, Jimmy Lin, John Prager, Dragomir Radev, Lucy Vanderwende, Ralph Weischedel. Natural Language Processing *Encyclopedia of Library and Information Science*. 2003

[11] Saif M. Mohammad. Socio-Affective Computing *A Practical Guide to Sentiment Analysis*, 5:4:61–83, 2017.

[12] Manuel Carlos Díaz-Galiano, Manuel García-Vega, Edgar Casasola, Luis Chiruzzo, Miguel García-Cumbreras, Eugenio Martínez Cámara, Daniela Moctezuma, Arturo Montejo Ráez, Marco Antonio Sobrevilla Cabezudo, Eric Tellez, Mario Graff, Sabino Miranda. Overview of TASS 2019: One More

Further for the Global Spanish Sentiment Analysis Corpus *CEUR Workshop Proceedings*, 2421:550–560, 2019.

[13] Alexandre Pinto, Hugo Gonçalo Oliveira, Ana Oliveira Alves. Comparing the performance of different PLN toolkits in formal and social media text *penAccess Series in Informatics*, 51:3:31–316, 2016.

[14] Prakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman. Natural language processing: an introduction *Journal of the American Medical Informatics Association*, 18:5:544–551, 2011.

[15] Wahab Khan, Ali Daud, Jamal A. Nasir, Tehmina Amjad. A survey on the state-of-the-art machine learning models in the context of PLN *Kuwait J. Sci.*, 43:4:95–113, 2016.

[16] Vishal A. Kharde, S.S. Sonawane. Sentiment Analysis of Twitter Data: A Survey of Techniques *International Journal of Computer Applications*, 139:11:5–15, 2016.

[17] Mita K. Dalal, Mukesh A. Zaveri. Automatic Text Classification: A Technical Review *International Journal of Computer Applications*, 28:2:37–40, 2011.

[18] Akshi Kumar, Teeja Mary Sebastian. Sentiment Analysis: A Perspective on its Past, Present and Future *International Journal of Intelligent Systems and Applications*, 4:10:1–14, 2012.

[19] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries Chin-Yew *Text Summarization Branches Out*, 74–81, 2004.

[20] Zara Nasar, Syed Waqar Jaffry, Muhammad Kamran Malik. Textual keyword extraction and summarization: State-of-the-art *Information Processing and Management*, 56:6, 2019.

[21] N. Moratanch, S. Chitrakala. A survey on extractive text summarization *International Conference on Computer, Communication, and Signal Processing: Special Focus on IoT, ICCCSP 2017*, 1–6, 2017.

[22] N. Moratanch, S. Chitrakala. A Survey on Abstractive Text Summarization *International Conference on Circuit, Power and Computing Technologies [ICCPCT] 2016*, 2016.

[23] Mahak Gambhir, Vishal Gupta. Recent automatic text summarization techniques: a survey *Artificial Intelligence Review*, 47:1:1–66, 2017.

[24] Horacio Saggion, Dragomir Radev, Simone Teufel, Wai Lam. Meta-evaluation of Summaries in a Cross-lingual Environment using Content-based Metrics *Proceedings of the 19th international conference on Computational linguistics*, 2002.

[25] Dipanjan Das, André F.T. Martins. A survey on Automatic Text Summarization *Journal of AI and Data Mining*, 2007.

[26] Makbule Gulcin Ozsoy, Ferda Nur Alpaslan, Ilyas Cicekli. Text summarization using latent semantic analysis *Journal of Information Science*, 37:4:405–417, 2011.

[27] I. V. Mashechkin, M. I. Petrovskiy, D. S. Popov, D. V. Tsarev. Automatic text summarization using latent semantic analysis *Programming and Computer Software*, 37:6:299–305, 2011.

[28] A. Khan, N. Salim. A Review on Abstractive Summarization Methods *Journal of Theoretical and Applied Information Technology*, 59:1:64–72, 2014.

[29] H. P. Edmundson. New Methods in Automatic Extracting *Journal of the ACM (JACM)*, 16:2:264–285, 1969.

[30] Josef Steinberger, Karel Ježek. Using latent semantic analysis in text summarization and summary evaluation *Proceedings of the 7th International Conference ISIM*, 93–100, 2004.

[31] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment *Journal of the ACM (JACM)*, 46:5:604–632, 1999.

[32] Günes Erkan, Dragomir R. Radev. LexRank: Graph-based lexical centrality as salience in text summarization *Journal of Artificial Intelligence Research*, 46:5:604–632, 1999.

[33] Rada Mihalcea, Paul Tarau. TextRank: Bringing Order into Texts *Proceedings of EMNLP*, 85:404–411, 2004.

[34] Sergey Brin, Lawrence Page. The anatomy of a large-scale hypertextual Web search engine *Computer Networks and ISDN Systems*, 30:1-7:107–117, 1998.

[35] Aria Haghighi, Lucy Vanderwende. Exploring content models for multi-document summarization *ANAACL HLT 2009 - Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 362–370, 2009.

[36] Huang L, He Y, Wei F, Li W. Modeling document summarization as multi-objective optimization *Proceedings of the third international symposium on intelligent information technology and security informatics*, 382–386, 2010.

[37] Martin Hausl, Johannes Forster, Maximilian Auch, Marcel Karrasch, and Peter Mandl. *An evaluation concept for named entity recognition and keyword apis in social media analysis. In Proceedings of the Second International Workshop on Entrepreneurship in Electronic and Mobile Business, IWEMB 2018, pages 79–96*. PubliQation Academic Publishing, New York, NY, 2019.

[38] Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. *Affective Computing and Sentiment Analysis. In Socio-Affective Computing, pages 1–10*. A Practical Guide to Sentiment Analysis , 2017.

[39] Noam Chomsky. Syntactic structures , 1957.

[40] Inderjeet Mani. Automatic Summarization , 2001.

[41] Google Natural Language API Consultado en https://cloud.google.com/natural-language.

[42] Ejemplos de escenarios de usuario de Text Analytics API Consultado en https://docs.microsoft.com/es-es/azure/cognitive-services/text-analytics/text-analytics-user-scenarios.

[43] MeaningCloud Consultado en https://www.meaningcloud.com/es.

[44] Productos de MeaningCloud Consultado en https://www.meaningcloud.com/es/productos.

[45] Resumen automático en MeaningCloud Consultado en https://www.meaningcloud.com/es/productos/resumen-automatico.

[46] Aplicaciones de la extracción de entidades Consultado en https://monkeylearn.com/blog/named-entity-recognition/.

[47] Documentación de la librería scikit-learn Consultado en https://scikit-learn.org/stable/modules/model_evaluation.html.

[48] Aplicaciones del resumen automático Consultado en https://blog.frase.io/20-applications-of-automatic-summarization-in-the-enterprise/.

[49] Traducción automática Microsoft Consultado en https://www.microsoft.com/es-es/translator/business/machine-translation/.

[50] Azure Text Analytics Consultado en https://docs.microsoft.com/es-es/azure/cognitive-services/text-analytics/.

[51] Sumy library Consultado en https://github.com/miso-belica/sumy.

[52] Watson Natural Language Understanding Consultado en https://www.ibm.com/es-es/cloud/watson-natural-language-understanding.