

**INSTITUTO TECNOLÓGICO DE BUENOS AIRES – ITBA
ESCUELA DE INGENIERÍA INFORMÁTICA**

Análisis de sentimiento

**Comparación de algoritmos predictivos y métodos utilizando un
lexicon español**

AUTOR: Pauli, Pablo Agustin (Leg. Nº 51185)

TUTORA: Soliani, Valeria Ines

**TRABAJO FINAL PRESENTADO PARA LA OBTENCIÓN DEL TÍTULO DE
INGENIERO EN INFORMÁTICA**

Lugar: Buenos Aires, Argentina

Fecha: 29/07/2019

Tabla de contenido

1. Introducción	4
2. Objetivos	5
3. Estado del arte	6
4. Procesamiento del lenguaje natural	8
4.1 Niveles de análisis para el PLN	10
4.2 Aplicaciones del PLN	11
5. Análisis de sentimientos	13
5.1 Aplicaciones del análisis de sentimientos	15
5.2 Niveles de análisis de sentimientos	16
6. Análisis de documentos	17
6.1 Tipos de clasificación	17
6.1.1 Clasificación mediante aprendizaje supervisado	17
6.1.1.a Algoritmos de clasificación	18
6.1.1.b Opciones de muestreo para los algoritmos	20
6.1.2 Clasificación mediante aprendizaje no supervisado	Error! Bookmark not defined.
6.1.3 Métodos basados en diccionarios	21
6.1.4 Métodos basados en relaciones lingüísticas	22
6.2 Métricas y métodos de evaluación de resultados	22
6.3 Proceso de entrenamiento de los algoritmos	25
6.4 Preprocesamiento	25
7. Análisis del Data set	27
7.1 Corpus de entrenamiento	27
7.2 Análisis de datos obtenidos	28
7.2.a Negativos	28
7.2.b) Neutros	30
7.2.c) Positivos	31
8. Resultados	33
8.1 Pruebas de algoritmos de machine learning	33
NAIVE BAYES	33
REGRESIÓN LOGÍSTICA	34

SVM	34
Conclusión de resultados	35
8.2 Pruebas con lexicon	36
Primera versión del código	36
Segunda versión del código	37
Conclusión de resultados	39
9. Conclusiones	40
9.1 Conclusiones generales sobre el análisis de sentimiento y su futuro	40
9.2 Problemas hallados, posibles mejoras y resultados obtenidos	40
9.3 Trabajo futuro	42
10. Bibliografía	43
11. Enlaces consultados	44
12. Anexo	46

1. Introducción

El Análisis de Sentimientos es un área de investigación enmarcada dentro del campo del Procesamiento del Lenguaje Natural y cuyo objetivo fundamental es el tratamiento computacional de opiniones, sentimientos y subjetividad en textos. En este contexto, una opinión es una valoración positiva o negativa acerca de un producto, servicio, organización, persona o cualquier otro tipo de ente sobre la que se expresa un texto determinado. La llegada de la Web 2.0 y la popularización de redes sociales de microblogging como Twitter han catapultado este campo de investigación de la Inteligencia Artificial hacia las más altas cotas de interés y notoriedad debido a la indiscutible importancia que supone el poder obtener el grado de valoración de miles de personas en cada instante para empresas, organizaciones, gobiernos y consumidores. Esta gran cantidad de información junto al aumento de la potencia de computación de las computadoras han hecho posible la aplicación de técnicas de aprendizaje automático para la clasificación de los textos en base a su polaridad sentimental y han abierto una puerta a la que sin duda será una de las áreas de investigación y desarrollo más importantes de los próximos años.

El propósito de este proyecto es por un lado realizar una comparación entre distintos algoritmos de aprendizaje supervisado y ver que tan confiables son a la hora de clasificar la información luego de pasarlos por un proceso de entrenamiento.

Aunque este campo de investigación es muy reciente, su gran popularidad ha propiciado que se publiquen muchos estudios de diferentes universidades de todo el mundo, y a pesar de que no existe un método cuyos resultados destaquen con claridad por encima de los otros, una de las más importantes corrientes de investigación basa su trabajo en el uso de algoritmos de aprendizaje automático y en la habilidad que tienen este tipo de sistemas para la clasificación de textos a partir de las palabras y de las relaciones que se establecen entre ellas. Así, aquí nos centraremos en este tipo de soluciones y se ofrecerá un estudio comparativo en el que se tendrán en cuenta diversos algoritmos de aprendizaje automático supervisado y diferentes métodos de extracción de las características necesarias para su entrenamiento.

Es importante mencionar que este trabajo incluirá una parte teórica en la que se expondrán los conceptos más importantes y fundamentales sobre el análisis de sentimientos, sus problemas, limitaciones, así como un análisis del estado del arte.

Además, se va a realizar una comparación entre esos algoritmos y un lexicón español que contenga una relación palabra-positividad y observar los resultados de uno y el otro.

2. Objetivos

El objetivo principal de este trabajo es realizar una comparación entre distintos algoritmos y métodos para análisis de sentimientos utilizados para clasificar la información obtenida del corpus y evaluar con cuál de ellos se obtienen mejores resultados en cuanto a la precisión y fiabilidad de estos.

Las evaluaciones se realizarán comparando algoritmos de machine learning entre si y también comparándolos con un algoritmo que utilice un lexicon para clasificar la información.

Por otro lado, se generará un dataset clasificado que sirva como base para futuras investigaciones, ya que de momento no se cuenta con datasets ya armados y listos para utilizar relacionados con este dominio.

Por último, se desarrollará un lexicon en español adaptado al dominio de restaurantes y afines similares, que es en donde se centra este trabajo.

3. Estado del arte

En todo proceso de toma de decisiones las personas hacemos uso de la opinión de otros individuos a la hora de decantarnos por una opción u otro. El pedir ayuda, consejo u opinión no es más que un recurso que permite al ser humano ampliar su conocimiento sobre un determinado tema con el objetivo de minimizar el riesgo que supone el tomar una mala decisión.

Hace no mucho tiempo, antes de la llegada de Internet y de su actual omnipresencia en todo aspecto de nuestras vidas, las fuentes de opinión principales eran el conocimiento y la experiencia de las personas más cercanas que formaban parte de nuestro círculo de relaciones y amistades. A la hora de comprar un televisor o de alquilar una película en un videoclub, nuestra decisión por una u otra opción se basaba en la opinión de nuestros amigos, familiares o compañeros de trabajo/universidad. El boca a boca era el sistema que usábamos para transmitir opiniones, juicios y expresar las virtudes o defectos de productos y servicios. En esa época, publicaciones en papel especializadas en temas concretos nos asesoraban a la hora de adquirir una computadora, un auto o irnos de vacaciones a algún país remoto. La sección de cultura y espectáculos de los diarios servía como altavoz para publicar obras de teatro, exposiciones o cine.

A pesar de que estas fuentes de opinión tradicional no desaparecieron, es indudable que la implantación de Internet y especialmente la llegada de la Web 2.0 (concepto atribuido a Tim O'Reilly y nombrado en la conferencia sobre la Web 2.0 de O'Reilly Media, 2004) han supuesto un profundo cambio en la manera en que las personas buscamos opiniones que nos ayuden durante el proceso de toma de decisiones. Internet se ha convertido en un océano inmenso en donde millones de personas expresan su opinión sobre cualquier tema y en cualquier momento. Es precisamente la bidireccionalidad a la que hace referencia el concepto de Web 2.0 la que ha hecho posible que cualquier persona pueda conocer al instante la opinión de miles de usuarios sobre cualquier cuestión y además contribuir al debate con su propia opinión. Es esta manera de interactuar la que permite crear redes virtuales en donde varias personas relacionadas en base a un tema concreto y en un instante preciso pueden intercambiar su visión particular acerca de dicho tema. Las personas hemos pasado de pedir opinión a nuestro círculo de relaciones más cercano a buscar la opinión de absolutos desconocidos antes de tomar la decisión de comprar un televisor, alquilar una noche de hotel, ir al cine o incluso votar a un determinado partido político. El cambio ha sido tan contundente y efectivo que ya no existe sitio web que no cuente con una sección de opinión asociada a sus propias publicaciones y en donde los usuarios puedan expresar sus juicios acerca del tema a tratar, desde simples noticias de actualidad hasta videos amateur publicados por personas anónimas. La inmediatez de la información asociada al gran número de mensajes hace de Internet y de la Web 2.0 una absoluta revolución en cuestión de opiniones. Tanto es así que no solo los usuarios se han dado cuenta del gran valor que aporta esta información. También las organizaciones, empresas e incluso gobiernos saben de la importancia de estas opiniones y de su valor como herramienta para mejorar sus productos, servicios y su reputación general. Lo que hace años se tenía que obtener en base a largos y costosos procesos de análisis de encuestas ahora es posible conseguirlo sin tener que hacer inversiones en este tipo de

estudios y con una mayor rapidez, midiendo en cada momento el sentir de las personas acerca de un determinado tema relacionado con su ámbito de negocio o actuación.

El éxito de este nuevo uso de Internet es indiscutible. Ya no compramos ningún producto en *Amazon*(www.amazon.es) sin mirar antes las opiniones de otras personas, no alquilamos un hotel en *Booking*(www.booking.com) sin analizar las puntuaciones que han dado otros usuarios sobre la comida, la limpieza o la situación física del alojamiento, *Menéame*(www.meneame.com) nos permite modelar nuestra opinión después de leer las propias de decenas de usuarios sobre las noticias publicadas en el sitio web y nos decantamos por ver una u otra película en base a la puntuación publicada por miles de desconocidos en *FilmAffinity*(www.filmaffinity.com). Además, este uso no sólo es válido para la selección de productos y servicios. En el terreno laboral *Glassdoor*(www.glassdoor.com) es una fuente para conocer la opinión de otros profesionales sobre una empresa y ayudar en el proceso de decisión antes de trabajar en ella. Otro ejemplo válido es *LinkedIn*(www.linkedin.com), que permite opinar sobre nuestros colegas de profesión mediante recomendaciones de manera que contribuyan a que su reputación aumente y sean más atractivos de cara al mercado laboral. Prácticamente todo está sujeto a la posibilidad de opinar sobre él, mejorando o empeorando la percepción que tienen las personas acerca del elemento en cuestión y ayudando en el proceso de selección.

Una fuente de opiniones fundamental dentro de este contexto son las redes sociales. Es innegable la ventaja que tiene para el proceso de toma de decisiones el poder contar en todo momento con los cientos de opiniones que sus usuarios expresan sobre cualquier tema en cada instante. Solo hay que detenerse por un momento a pensar el valor que tienen los tweets de todas estas personas para empresas, partidos políticos, gobiernos u organizaciones de cualquier tipo. Poder conocer en todo momento qué es lo que piensan los usuarios de alguno de tus productos ofrece una ventaja competitiva impensable de obtener hace tan solo diez años. La posibilidad de poder mejorar los aspectos menos populares de tus servicios o que un personaje público de cualquier índole pueda conocer si sus palabras o acciones son bien recibidas por sus seguidores es un instrumento de incuestionable utilidad.

Una vez conocida la importancia que tiene toda esta información de opinión y sentimiento vertida en cada instante en Internet, cabe preguntarse cómo poder trabajar con ella sin perderse en el vasto y extenso mar de opiniones y desfallecer en el intento. Procesar tal cantidad de datos requiere de nuevas tecnologías capaces de analizar y clasificar de manera automática el sentimiento y la polaridad descrita por los usuarios en sus opiniones. Es necesario encontrar un método que simplifique el proceso de entendimiento de todos estos textos escritos y haga viable el poder obtener una medida del sentir general que las personas expresan en la red para así poder explotar convenientemente esta valiosa información.

Los conceptos mencionados fueron desarrollados en mayor detalle en el libro *Sentiment analysis and opinion mining*, por Liu, B (2012).

4. Procesamiento del lenguaje natural

En este apartado se tratará el campo del procesamiento del lenguaje natural. Esta área de estudio es en la que se enmarca el análisis de sentimientos y, como se verá a continuación, aplicaciones muy populares como buscadores web, traductores multi idioma o asistentes virtuales dependen de sus métodos, técnicas y avances de investigación. En esta sección de la memoria se incluye un breve recorrido por su historia, algunos de los trabajos más destacados, los niveles comunes de análisis que existen dentro de todas las aplicaciones que procesan de alguna manera lenguajes naturales y los usos más importantes de esta área de investigación.

El procesamiento del lenguaje natural (PLN o NPL por sus siglas del inglés Natural Language Processing), es un campo enmarcado dentro del área de la inteligencia artificial, la computación y la lingüística. Su objetivo fundamental es facilitar y hacer eficaz la comunicación entre las personas y los computadores mediante el uso de protocolos como los lenguajes naturales. Estos lenguajes son los usados por las personas para comunicarse entre sí tanto de forma oral como escrita. La comunicación es un elemento esencial para establecer relaciones entre individuos o entidades, sean éstas del mismo tipo o no. Es sencillo deducir que la comunicación entre elementos de la misma naturaleza, como entre personas, máquinas o animales de la misma especie, es más simple, directa y efectiva que cuando se produce entre entidades de diferente origen. Por esta razón y debido a la relación existente entre las personas y los computadores, se hace necesaria la búsqueda y el estudio de protocolos que faciliten la comunicación e interacción entre ambos objetos para así mejorar sus relaciones. Es el área del PLN quien se encarga de esta tarea.

La historia del PLN tiene su origen a mediados del siglo XX con la aparición de una nueva disciplina dentro de las ciencias de la computación. Su objetivo era el desarrollo de sistemas lo suficientemente inteligentes para que la comunicación entre personas y máquinas se hiciese mediante el uso de lenguaje natural. Por aquel entonces, justo después de la Segunda Guerra Mundial, era conocida la importancia de poseer algún sistema que permitiese traducir textos entre diferentes idiomas y de manera automática. Uno de los sistemas que se crearon aquella época fue el conocido como Experimento de Georgetown-IBM, en 1954. Desarrollado conjuntamente por la Universidad de Georgetown e IBM, el experimento consistió en una demostración de traducción automática entre los idiomas inglés y ruso. Contaba con un conjunto de reglas gramaticales y un par de cientos de elementos de vocabulario para llevar a cabo las traducciones. A través de una interfaz rudimentaria, un operador sin conocimientos de lengua rusa introdujo una serie de frases sobre política, ciencia o matemáticas que fueron procesadas por un ordenador IBM 701, generando una impresión con las frases traducidas al inglés. Aunque es indiscutible el hito que supuso esta prueba, es necesario decir que las oraciones a traducir fueron especialmente escogidas para la prueba. El sistema no llevaba a cabo ningún tipo de análisis sintáctico que detectase la estructura de las frases y el enfoque usado estaba

basado en diccionarios en donde las palabras estaban asociadas a reglas muy específicas. Aun así, los resultados de la prueba generaron unas altas expectativas ya que sus autores afirmaban que el problema de la traducción automática estaría resuelto en pocos años, haciendo que la inversión en este tipo de sistemas se disparase. Pasados mas de diez años desde aquella prueba, los investigadores en esta materia reconocían que sus avances eran mucho mas lentos de lo esperado, de manera que los fondos invertidos para la investigación se redujeron radicalmente.

Pocos años después del experimento de Georgetown-IBM, en 1957, Noam Chomsky publicó su libro *Syntactic Structures* el cual supuso un gran acontecimiento y ejerció una enorme influencia en el campo de la lingüística. Chomsky creía que el cerebro humano contaba con una facilidad innata para usar y entender el lenguaje y esto se debía a que existían un conjunto de reglas y normas universales, comunes a todas las lenguas, que permitían operar con el lenguaje y con las que las personas ya contábamos a la hora de nacer. Estas reglas conforman la llamada “gramática universal” y fueron la base para que Chomsky introdujera la conocida como “gramática generativa transformacional”. La influencia de estas ideas creó una corriente de investigación del PLN en donde sus integrantes afirmaban que el éxito de este tipo de sistemas radicaba en el uso de reglas y patrones estructurales gramaticales formados entre las palabras de los textos. Estas reglas se debían combinar con diccionarios para resolver tareas como traducir textos, buscar información o interactuar con computadoras usando un lenguaje similar al usado por humanos. Un ejemplo paradigmático de este enfoque es el programa ELIZA, desarrollado a mediados de la década de 1960 por Joseph Weizenbaum en el MIT. ELIZA era un programa que procesaba lenguaje natural y recreaba la sensación de una conversación coherente con un interlocutor humano. El funcionamiento consistía en extraer palabras clave de la frase introducida por el usuario y contestar con otra relacionada que mantenía en una base de datos interna. La conversación llegaba a ser tan convincente que producía la sensación de estar hablando con otro interlocutor humano. No obstante, tenía problemas cuando las palabras del usuario no figuraban en su base de datos. En estos casos se limitaba a reformular la frase del humano en forma de pregunta. Además, si la conversación se alargaba demasiado, esta empezaba a ser incoherente. ELIZA puede ser considerado como un precursor primitivo de los asistentes virtuales más actuales como Siri o Alexa.

Las teorías de Chomsky permanecieron vigentes en el PLN durante las tres siguientes décadas, hasta bien entrada la década de 1980. En esta etapa, existía otro grupo de investigadores que confiaban en los modelos probabilísticos basados en datos para hallar soluciones a los problemas planteados dentro del PLN. Este enfoque trataba de buscar relaciones matemáticas entre los componentes de los textos, como letras, palabras u oraciones y calculaba la probabilidad de que éstas apareciesen en determinados contextos. En base a estas probabilidades se puede llegar a deducir cuál será el siguiente componente lingüístico dentro de una secuencia sin necesidad de recurrir a reglas gramaticales. Uno de los proyectos de mayor éxito que forma parte de esta corriente de investigación es el programa CANDIDE, desarrollado en el año 1991 por investigadores del Thomas J. Watson Center de IBM en Nueva York. Este programa pretendía generar traducciones automáticas sin usar herramientas más allá de sistemas puramente estadísticos. Para ello, se hizo uso del conjunto de actas del Parlamento de Canadá y que constaba por aquel entonces de tres millones de oraciones escritas tanto en inglés como en francés. El proceso consistió en alinear palabras y oraciones de ambos idiomas y calcular la probabilidad de que una

palabra de una oración en un idioma se correspondiese con otras palabras en el otro idioma. Los resultados fueron sorprendentemente positivos debido a que la mitad de las frases traducidas se correspondían de manera exacta o tenían un significado similar a las del texto original. Esta investigación es considerada un hito en el campo de la traducción automática y el uso de sistemas estadísticos dentro del PLN y puede ser considerada como la precursora de herramientas más modernas como el servicio de traducción multi idioma Google Translate.

Con el paso de los años, el PLN estadístico se ha ido imponiendo con gran éxito dejando atrás las ideas chomskianas y el uso de reglas de transformación gramáticas escritas por humanos. Esto es debido principalmente al aumento de la potencia de cálculo de los ordenadores y a la gran cantidad de información que existe a disposición de los mismos y que es necesaria para construir los sistemas probabilísticos. Así, mediante el uso de los sistemas de aprendizaje automático, son los propios computadores los que aprenden el lenguaje natural, infiriendo las reglas y normas que gobiernan los lenguajes naturales.

4.1 Niveles de análisis para el PLN

Todo sistema de PLN debe llevar a cabo un conjunto de tareas de análisis del lenguaje que faciliten el entendimiento entre el usuario y el propio sistema. Estas tareas constituyen una arquitectura de niveles a través de los cuales y de manera secuencial las oraciones se analizan e interpretan hasta ser comprendidas y asimiladas por el sistema de PLN. A grandes rasgos, existen cuatro componentes principales o niveles de análisis, pero no todos deben ser implementados. Son las funciones a desempeñar por el sistema las que determinan qué niveles de análisis deben ser desarrollados. Estos componentes, ordenados de menor a mayor complejidad, son los siguientes:

- **Nivel de análisis morfológico:** en este componente se examinan las palabras para extraer raíces, rasgos flexivos, sufijos, prefijos y otros elementos. Su objetivo es entender cómo se construyen las palabras a partir de unidades de significado más pequeñas denominadas morfemas.
- **Nivel de análisis sintáctico:** analiza la estructura de las oraciones en base al modelo gramatical empleado con el objetivo de conocer como se unen las palabras para crear oraciones.
- **Nivel de análisis semántico:** proporciona sentido a las oraciones y les otorga un significado, resolviendo además las ambigüedades léxicas y estructurales que pudieran aparecer.
- **Nivel de análisis pragmático:** se encarga del análisis de los textos mas allá del de una oración aislada, teniendo en consideración aquellas inmediatamente anteriores, la relación existente entre ellas y el contexto en el que se producen.

4.2 Aplicaciones del PLN

Es fácil deducir que el PLN es una disciplina que cuenta con un alto potencial y múltiples posibilidades prácticas, tantas como los lenguajes naturales poseen. Su cometido es mejorar y hacer eficaz la comunicación entre personas y computadores. Por esta razón, cualquier área asociada al lenguaje y a las relaciones entre humanos y máquinas se puede ver afectada positivamente por el PLN. Aunque sus aplicaciones son innumerables y el único límite existente es la propia imaginación, algunas de las más populares e importantes son las siguientes:

- **Recuperación de la información:** el objetivo de este tipo de sistemas en la búsqueda y obtención de grupos de documentos electrónicos a partir de un conjunto de palabras clave proporcionadas por el usuario. Los documentos devueltos normalmente se ordenan en base a algún tipo de atributo que mide su relevancia dentro del resultado global. Estos sistemas son en los que basan su funcionamiento los buscadores de contenidos de Internet y representan la primera aplicación implantada masivamente dentro del mundo de las Tecnologías de la Información. Algunos ejemplos populares podrían ser el servicio Google Search o Microsoft Bing.
- **Traducción automática de textos:** una de las aplicaciones paradigmáticas de los sistemas de PLN es la traducción automática entre múltiples lenguajes naturales. Esta labor es tácitamente imposible para un humano debido a la dificultad que existe a la hora de encontrar personas que conozcan decenas de lenguas distintas o la sola combinación de lenguas sobre las que queremos hacer la traducción (por ejemplo, griego y camboyano simultáneamente). Los sistemas actuales de traducción automática utilizan un enfoque basado en mediciones estadísticas y relaciones entre textos a partir de un entrenamiento previo con cientos o miles de textos. Aunque las traducciones no siempre son perfectas y no pueden sustituir a humanos en textos complejos en los que se requiera alta fiabilidad, sí son aceptables para tareas como traducir un mensaje de una red social, una página web o una reseña en una página de alquiler de coches. Uno de los ejemplos más significativos de este tipo de sistemas es el traductor Google Translate.
- **Reconocimiento del habla:** este tipo de sistemas permiten a las personas interactuar con los ordenadores u otros dispositivos electrónicos como teléfonos inteligentes o automóviles mediante el uso de un lenguaje natural y por medio de la voz. En los últimos tiempos se han hecho muy populares los llamados “asistentes virtuales” como Cortana, Siri o Google Assistant. Éstos pueden realizar acciones como enviar un correo electrónico, gestionar el calendario de citas o incluso realizar una compra usando para ello sólo comandos de voz. Otras aplicaciones que hacen uso de estos sistemas de PLN son los servicios telefónicos de atención al cliente. Estos servicios permiten realizar diferentes gestiones sin necesidad de un interlocutor humano, siendo muy habituales en actividades de banca y telecomunicaciones. Virtualmente todo dispositivo electrónico podría en el futuro

poseer un sistema de reconocimiento del habla de manera que sus funciones pudiesen ser controladas mediante comandos de voz. Por tanto, el campo del reconocimiento de la voz humana se postula como uno de los mas importantes debido a la popularización de este tipo de sistemas.

- **Extracción de la información:** este tipo de tareas consiste en analizar textos o mensajes con el objetivo de capturar y extraer automáticamente aquella información considerada de interés. Una aplicación habitual es el escaneo de documentos escritos en algún lenguaje natural y para después volcar la información extraída a una base de datos de manera automática. Estos documentos pueden ser anuncios por palabras, artículos de prensa, informes de carácter científico, etc., y los datos a extraer, nombres de personas, organizaciones, teléfonos, fechas, valores monetarios u otros. El proceso de extracción de la información es básico para poder clasificar documentos, resumirlos o relacionarlos entre sí.
- **Análisis de sentimientos:** como se vera a lo largo de este trabajo, el análisis de sentimientos ofrece la posibilidad de conocer automáticamente cual es la opinión que una persona tiene sobre un determinado tema a partir de las ideas expresadas en un texto. Este sentimiento u opinión es una valoración cualitativa o cuantitativa acerca de un producto, servicio, persona o cualquier otro tipo de entidad. El poder extraer de manera automática esta información permite la creación de poderosas herramientas que facilitaran conocer cual es el sentir de las personas en cada momento en relación a diferentes objetos de estudio.

5. Análisis de sentimientos

Luego de haber hablado de forma general del procesamiento del lenguaje natural, ahora nos enfocaremos en la aplicación que nos compete, el análisis de sentimientos. En este apartado se expondrá el concepto de análisis de sentimientos mediante su definición y un breve repaso por su historia, nombrando también algunos de los autores y trabajos más relevantes. Como se verá a continuación, el análisis de sentimientos es un área de investigación en pleno auge cuyo esplendor se debe a una serie de factores muy determinados. En este capítulo también se indicarán las aplicaciones más importantes que tiene este campo en la vida real, los diferentes niveles de análisis de sentimientos que se pueden ejecutar sobre textos escritos, las tareas necesarias para su realización y una definición formal sobre el concepto de Opinión. Para finalizar esta sección se explicarán cuáles son las dificultades más comunes a las que se enfrenta el análisis de sentimientos.

El Análisis de Sentimientos (AS o SA por sus siglas del inglés Sentiment Analysis) es un campo de investigación dentro del PLN que trata de extraer de manera automática y mediante técnicas computacionales información subjetiva expresada en el texto de un documento dado y acerca de un determinado tema. De esta forma, mediante el análisis de sentimientos podremos saber si un texto presenta connotaciones positivas o negativas. Una definición ampliamente extendida de este concepto es la ofrecida por los investigadores Pang y Lee en (Pang & Lee, 2008) y que define el análisis de sentimientos como:

“Tratamiento computacional de opiniones, sentimientos y subjetividad en textos.”

Esta definición es la más aceptada por la comunidad de investigadores, pero debido a su generalidad otros autores como Cambria y Hussain (Cambria & Hussain, 2012) han definido el análisis de sentimientos de la siguiente manera:

“Conjunto de técnicas computacionales para la extracción, clasificación, comprensión y evaluación de opiniones expresadas en fuentes publicadas en Internet, comentarios en portales web y en otros contenidos generados por usuarios.”

Se puede observar que la segunda definición es mucho más concreta que la primera y sólo hace referencia a las opiniones, dejando fuera del alcance de estudio a los sentimientos y a la subjetividad. Es posible que dejar fuera a los sentimientos sea un error puesto que muchas veces las opiniones están fundamentadas y emanan de los sentimientos de quien las expresa, pero como indica E. Martínez en (Martínez Cámara, 2016), sí es un acierto no hacer referencia a la subjetividad ya que las opiniones se pueden encontrar en oraciones subjetivas y también objetivas. En cualquier caso, ambas definiciones son útiles y válidas para comprender en qué consiste el análisis de sentimientos.

Aunque la historia del análisis de sentimientos pertenece sin ninguna duda al siglo XXI, existen algunos trabajos desarrollados mucho antes considerados como precursores de este campo de investigación. Uno de ellos es (Carbonell, 1979) en donde se propone un modelo computacional que permite representar el pensamiento subjetivo de las personas,

tratando de entender su ideología y su personalidad a través de las subjetividades que contienen sus textos escritos. Pocos años después, en (Wilks & Bien, 1984), se presenta un estudio sobre las creencias que tiene un sujeto sobre otro en base al conocimiento que tienen de ambos por separado. Estos dos estudios, aunque relacionados, no avanzaron hacia lo que hoy se conoce como análisis de sentimientos, sino hacia otros campos de investigación como la interpretación de metáforas, los puntos de vista, el afecto y otras áreas relacionadas.

La verdadera explosión de trabajos de investigación del análisis de sentimientos se produce a partir de 2001 y su número ha ido incrementándose de manera exponencial con el paso de los años. En (Pan & Lee, 2008) se atribuye este progresivo interés a tres factores:

- La popularización de los métodos de aprendizaje automático y su uso dentro de las diferentes áreas del PLN.
- La disponibilidad de datos con los que entrenar a los sistemas de aprendizaje automático provenientes principalmente de Internet y de su capacidad para generar ingentes cantidades de información, en especial a partir de la aparición de la llamada Web 2.0.
- El creciente interés por explotar esta información por parte de organizaciones y empresas debido a las posibilidades que ofrece el poder obtener automáticamente una valoración por parte de las personas acerca de productos, servicios o personas concretas.

Aunque la lista de trabajos de investigación es interminable, es importante destacar algunos considerados como los verdaderos creadores de los métodos que más se utilizan a la hora de capturar el sentimiento global de textos y documentos. (Pang et al., 2002) es el primer trabajo conocido que hace uso de algoritmos de aprendizaje automático para la clasificación de textos; en este caso, críticas de películas extraídas de un sitio Web especializado en esa temática. Otro precursor es Turney que en su estudio (Turney, 2002) muestra un sistema que es capaz de clasificar opiniones de usuarios sobre diversos productos y servicios como automóviles, viajes o películas, mediante un análisis gramatical de las oraciones y una serie de consultas en el motor de búsquedas AltaVista.

Además de estos trabajos de investigación, también es necesario destacar el estudio de divulgación de Pang y Lee (Pang & Lee, 2008) en donde se muestran diversas técnicas y procedimientos para la construcción de sistemas de análisis de sentimientos, así como el libro (Liu, 2012) en donde se desarrollan en detalle todos los conceptos relacionados con el análisis de sentimientos y cuyos planteamientos siguen vigentes a fecha de hoy.

5.1 Aplicaciones del análisis de sentimientos

Los beneficios del análisis de sentimientos son múltiples y sustanciales. Tanto es así, que empresas, organizaciones y gobiernos de todo el mundo son los principales interesados en el avance de este campo de investigación. Poder saber qué es lo que piensa la gente sobre sus productos, medidas y políticas en cada momento es una herramienta muy valiosa y bien utilizada puede ofrecer ventajas competitivas impensables de conseguir hasta hace pocos años. De la misma forma, la monitorización de las redes sociales y la extracción del sentir global sobre determinados temas puede ayudar a detectar la gestación de determinados acontecimientos sociales como huelgas, sediciones, revueltas, etc.

En concreto, algunas de las aplicaciones del análisis de sentimientos podrían ser las siguientes:

- **Valoración de opinión de productos y servicios:** probablemente esta sea la aplicación más práctica y directa del análisis de sentimientos. Mediante esta técnica es posible que las empresas puedan conocer la opinión de los usuarios acerca de sus productos sin necesidad de llevar a cabo estudios tradicionales como encuestas de satisfacción. Así, mediante las opiniones vertidas en foros, blogs y especialmente redes sociales, será posible conocer a los usuarios les gusta o no un determinado producto. De esta forma, las empresas pueden conocer en cualquier momento si sus productos son del agrado de los usuarios y, en caso negativo, poder replantear estrategias en el menor tiempo posible otorgando así ventajas competitivas.
- **Posicionamiento de publicidad on-line:** los anunciantes de determinados productos podrían requerir que sus anuncios fuesen publicados solo en sitios web en donde se expresen conceptos positivos, huyendo de aquellas páginas en donde los textos expresen sentimientos negativos.
- **Corrección de opinión:** es habitual que los usuarios expresen su opinión en sitios de compras online indicando, además de una reseña, una puntuación. Puede ocurrir que, por error, el usuario no indique correctamente dicha puntuación. Así, un sistema de análisis de sentimientos podría analizar las palabras del usuario y corregir automáticamente dicha puntuación.
- **Mejora de los sistemas de recomendación de productos:** en base a las opiniones de los usuarios, una tienda online podrá priorizar los productos que ofrece en base a dichas opiniones o no recomendar aquellos cuya opinión general sea negativa.
- **Reputación política:** el análisis de sentimientos demuestra un fuerte potencial para conocer la opinión de la gente sobre un determinado partido político o un candidato.
- **Análisis del mercado financiero:** a partir de la información contenida en páginas Web, foros y redes sociales sobre una empresa concreta, es posible prever cual

será su evolución en el mercado financiero a partir del valor agregado de la polaridad de todas las opiniones encontradas.

5.2 Niveles de análisis de sentimientos

El análisis de sentimientos de un documento se puede llevar a cabo a tres niveles distintos en base a la granularidad, profundidad y detalle requeridos. Estos niveles son:

- **Análisis a nivel de documento:** en este nivel se analiza el sentimiento global de un documento como un todo indivisible, clasificándolo como positivo, negativo o neutro o usando otro sistema de calificación. En estos casos, se asume que dicho documento expresa una valoración sobre una única entidad por lo que no es aplicable en aquellos que hablen sobre varias entidades simultáneamente.
- **Análisis a nivel de oración:** en este caso, se divide el documento en oraciones individuales para extraer posteriormente la opinión que contiene cada una de ellas. La opinión de cada oración puede ser, de nuevo, positiva, negativa o neutra o bien tomar un valor en base a cualquier otro tipo de medida.
- **Análisis a nivel de aspecto y entidad:** este es el nivel de análisis con mayor detalle posible, en donde una entidad esta formada por distintos elementos o aspectos y sobre cada uno de ellos se expresa una opinión cuya polaridad puede ser distinta en cada caso.

6. Análisis de documentos

Aquí se explicará en mayores detalles en que consiste y cuales son las particularidades del análisis de sentimientos a nivel de documento, que es en lo que se va a enfocar este trabajo. Además, se presentarán los dos grandes grupos en los que se encuadran la mayor parte de los métodos conocidos para clasificar los textos en base a su polaridad: mediante aprendizaje supervisado y no supervisado.

Los documentos son considerados las unidades básicas de información y estos pueden ser opiniones en blogs, en tiendas online, sitios web o mensajes en redes sociales. La opinión generalmente toma un valor de entre tres posibles: sentimiento positivo, negativo o neutro, pero también puede haber otras escalas.

Para llevar a cabo la clasificación de un documento en base a su sentimiento existen diversos métodos y técnicas que se van refinando y mejorando a medida que avanzan las investigaciones sobre esta materia y aparecen en escena nuevos estudios y trabajos. A pesar de la multitud de artículos y publicaciones presentados cada año, no parece existir un consenso claro sobre que técnicas se deben usar para obtener los mejores resultados en el proceso de clasificación de textos. Y es debido a este gran numero de publicaciones y a un campo de investigación sumido en un proceso de fuerte expansión por lo que no es sencillo fijar una división clara de los métodos existentes en la actualidad. Aun así, se establecen dos grandes grupos, métodos supervisados y no supervisados y estos últimos a su vez basados en diccionarios o en relaciones lingüísticas.

6.1 Tipos de clasificación

6.1.1 Clasificación mediante aprendizaje supervisado

La clasificación mediante técnicas de aprendizaje supervisado está basada en el uso de algoritmos de aprendizaje automático, conocidos también como machine learning. Su tipificación de “supervisados” se debe a que estos métodos necesitan de un grupo de documentos de ejemplo previamente etiquetados para generar un modelo que será usado posteriormente para clasificar nuevos textos y que en este contexto es conocido como “corpus”. Su funcionamiento se basa en la relación matemática creada entre los elementos de ejemplo durante un proceso conocido como entrenamiento y en donde se genera un modelo estadístico que agrupa dichos elementos en tantos conjuntos como diferentes etiquetas o clases existan en el grupo de documentos de entrenamiento. Posteriormente, el modelo generado se utiliza con un ejemplo no etiquetado para determinar de cuál de los grupos existentes formaría parte, realizando así una predicción en base a los ejemplos aportados durante la fase de entrenamiento.

El éxito y efectividad de los sistemas de aprendizaje automático a la hora de clasificar nuevos elementos depende principalmente de dos factores: del algoritmo de clasificación seleccionado y de las características o features elegidas para representar los elementos de

ejemplo y con los que entrenar dicho algoritmo. Existen decenas de algoritmos de aprendizaje automático distintos cuyos resultados además pueden ser mejorados mediante la configuración de sus diferentes parámetros; pero son las características elegidas para el entrenamiento las verdaderamente importantes y de ellas depende en gran medida el éxito de este tipo de métodos de clasificación.

Los métodos de clasificación mediante aprendizaje automático son sistemas que ofrecen buenos resultados en el trabajo de clasificación por sentimiento, pero cuentan con dos desventajas.

Por una parte, necesitan un corpus o juego de datos inicial con sus ejemplos previamente clasificados y no siempre es posible contar con él debido al coste que supone tener que categorizar, muchas veces a mano, dichos ejemplos. Además, el tamaño del corpus es fundamental para poder obtener resultados aceptables.

Por otra parte, los modelos resultantes son muy dependientes del dominio. Esto significa que un algoritmo entrenado con un corpus sobre comentarios de películas puede no ofrecer el mismo rendimiento a la hora de clasificar reseñas sobre automóviles, ya que una misma palabra no siempre posee el mismo sentimiento en contextos distintos. Por ejemplo, si hablamos de restaurantes, en la frase “La porción era grande”, en este contexto la palabra “grande” tiene una connotación positiva. En cambio, si el tema son notebooks, el comentario “La notebook es grande” puede indicar que el tamaño de la notebook no es atractivo y, por tanto, en este contexto “grande” tiene una connotación negativa. Por tanto, se hace necesario un juego de pruebas diferentes para cada uno de los dominios con el coste de tiempo y esfuerzo que esto conlleva.

6.1.1.a Algoritmos de clasificación

Los algoritmos de aprendizaje supervisado se pueden dividir principalmente en dos grandes grupos: de regresión y de clasificación. Los primeros permiten inferir un valor numérico a partir de una serie de datos de entrada, por ejemplo, las ventas que tendrá una determinada empresa. En cambio, los de clasificación se utilizan para deducir a qué grupo pertenece un ejemplo dado de entre los grupos disponibles. Aunque ambos tipos de algoritmos pueden ser usados en el análisis de sentimientos, nos centraremos en tres algoritmos de clasificación muy populares y que ya han sido utilizados en múltiples ocasiones para esta tarea: Naive Bayes, máquinas de vectores de soporte y regresión logística.

Naive Bayes:

Este método se basa en el teorema de Bayes sobre la probabilidad condicional en la que se quiere calcular la probabilidad de que ocurra el evento C sabiendo que ha ocurrido X. La fórmula es la siguiente:

$$P(c|x) = \frac{P(x|c)P(c)}{P_x}$$

1. $P(c|x)$ es la probabilidad posterior que deseamos calcular. Esta nos indica que la probabilidad de que se tenga una clase c dados los datos en x .
2. $P(c)$ es la probabilidad posterior de la clase. Qué tan probable que se obtenga una clase c .
3. $P(x|c)$ es la esperanza, que es la probabilidad de los datos dada una cierta clase.
4. $P(x)$ es la probabilidad a priori de los datos o predictor.

Los datos 2,3 y 4 es posible obtenerlos de una tabla de frecuencias. También se asume que el efecto del predictor x dada una clase c es independiente de los valores del predictor. Esta condición es llamada la independencia condicional de clases.

La tabla de probabilidad puede ser calculada construyendo una tabla de frecuencias para cada atributo con respecto a su objetivo. Con estas tablas de frecuencias se construyen tablas de probabilidad que se utilizarán para calcular la probabilidad de Naive Bayes. La clase con la probabilidad posterior más alta es a la cual se le asigna el vector de datos.

En el caso concreto de la clasificación de textos, los sucesos excluyentes y exhaustivos son las diferentes clases que se pueden asignar a un mensaje, de manera que no es posible asignar más de una simultáneamente (excluyentes) y esas clases son todos los tipos que existen (exhaustivos). Los algoritmos Naive Bayes suelen recibir el apelativo de “ingenuos” debido a que en sus cálculos las características seleccionadas para representar a los ejemplos de entrenamiento son estadísticamente independientes y contribuyen por igual en el proceso de clasificación. Dicho de otro modo y en el caso concreto de la clasificación de textos, se considera que las palabras de un mismo mensaje no mantienen ningún tipo de relación entre sí y es indiferente la posición que tienen dentro del texto al que pertenecen.

Regresión logística:

Se trata de calcular la probabilidad en la que una de las opciones de la variable dicotómica dependiente sucederá en función de cómo puntúa en una serie de variables dependientes que pueden estar en diferentes escalas de medida.

Si tenemos un conjunto de variables independientes X_1, X_2, \dots, X_p que nos clasifican a los n sujetos trataremos de saber a cuál de las dos categorías de la variable Y pertenece. La probabilidad de que un sujeto « i » pertenezca a una de ellas será la combinación lineal

$$Z = b_1 x_1 + b_2 x_2 + \dots + b_p x_p + b_0$$

y será igual a:

$$p_i = \frac{e^z}{1 + e^z} \quad \text{Que es igual a:} \quad p_i = \frac{1}{1 + e^{-z}}$$

Por lo que para el sujeto «i»:

$$p_i = \frac{1}{1 + e^{-(b_1x_1 + \dots + b_px_p + b_o)}}$$

Si la probabilidad p_i de que el sujeto esté encuadrado en esa categoría es mayor que 0,5 se le asigna, si es menor se le asignará la otra categoría

SVM(Support vector machine) / Máquinas de vectores de soporte:

Las máquinas de vectores de soporte son un grupo de algoritmos de aprendizaje supervisado desarrollados por Vapnik en 1982 en los laboratorios AT&T. De manera visual, podemos pensar en este tipo de algoritmos como la representación gráfica de un espacio multidimensional en donde se sitúan los puntos que simbolizan los ejemplos de entrenamiento. Un hiperplano, denominado vector de soporte, los separa la mayor distancia posible en base a su clase. De esta forma, el vector determina la frontera que sirve para clasificar un nuevo elemento, por lo que dependiendo a qué parte del espacio pertenezca, se le asignará una clase u otra.

Este tipo de algoritmos cuenta con una serie de parámetros que permiten ajustar su configuración interna y así optimizar los resultados durante el proceso de clasificación. Uno de estos parámetros es el *kernel* y se utiliza cuando no es posible separar las muestras mediante una línea recta, plano o hiperplano de N dimensiones, permitiendo tal separación mediante otro tipo de funciones matemáticas como polinomios, funciones de base radial Gaussiana, Sigmoid u otras. Otro de estos parámetros es *regularization* (también conocido como “C”) que permite crear un margen blando de manera que se consientan ciertos errores en la clasificación y se evite el sobre entrenamiento. Y, para terminar, el parámetro *gamma* determina la distancia máxima a partir de la cual una muestra pierde su influencia en la configuración del vector de soporte, y *margin*, qué es la separación entre el vector y las muestras de cada clase más cercanas al mismo.

6.1.1.b Opciones de muestreo para los algoritmos

Existen diversas técnicas para determinar qué datos del conjunto formarán parte del conjunto de entrenamiento del modelo y cuáles serán usados para validarlo y saber si es realmente efectivo. Se trata con estos métodos de lograr que no se produzca "overfitting", es decir, que el modelo se ajuste muy bien a nuestros datos pero que luego no se pueda generalizar. Algunas opciones que se utilizaron son las siguientes:

- Cross-validation: se utilizó el modelo de K-iteraciones, donde los datos se dividen en K subconjuntos. Uno de esos subconjuntos se utiliza como datos de prueba y el resto como datos de entrenamiento. Esto se repite K veces, con cada uno de los posibles subconjuntos de datos de prueba. Por último, se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. Los valores para K que se utilizaron fueron

- Muestreo aleatorio: en este caso se utilizan dos argumentos para definir el muestreo. El primero es el porcentaje que se va a tomar del total de los datos para usar como datos de prueba. El segundo es la cantidad de iteraciones que se van a realizar para entrenar al algoritmo.
- Leave one out: implica separar los datos de forma que para cada iteración tengamos una sola muestra para los datos de prueba y todo el resto conformando los datos de entrenamiento. La evaluación viene dada por el error, y en este tipo de validación cruzada el error es muy bajo, pero en cambio, a nivel computacional es muy costoso, puesto que se tienen que realizar un elevado número de iteraciones, tantas como N muestras tengamos y para cada una analizar los datos tanto de entrenamiento como de prueba.

6.1.2 Métodos basados en diccionarios

Los métodos basados en diccionarios (o lexicones, del inglés lexicon) hacen uso de listados de palabras y frases previamente etiquetadas con la polaridad de sentimiento que expresan y, en ocasiones, además con su intensidad o la fuerza de dicho sentimiento. Los textos por clasificar se dividen en unidades más pequeñas, como palabras o frases, y se buscan en los diccionarios de sentimiento. Así, el sentimiento global del texto vendrá dado por algún tipo de función matemática que tenga en cuenta el sentimiento individual de las unidades de trabajo y en base a lo indicado para ellas en el diccionario.

El uso de técnicas basadas en diccionarios facilita la obtención de sistemas de clasificación independientes del dominio, pero esto presenta también algunos inconvenientes. Uno de estos problemas es que en ocasiones se pierde precisión ya que las palabras pueden poseer diferente polaridad dependiendo del contexto en el que se usen. Por ejemplo, el adjetivo “silencioso” tiene una connotación positiva si se aplica al ruido que hace un lavarropas durante su funcionamiento, pero es negativa si se utiliza en referencia al sistema de sonido de un televisor. Este tipo de problemas se pueden resolver mediante diccionarios contruidos a partir de las palabras de un corpus centrado en el dominio que se desea estudiar.

También dentro del mismo dominio pueden existir palabras que no siempre tienen la misma polaridad de sentimiento. Por ejemplo, si hablamos de teléfonos móviles, el adjetivo “largo/a” es positivo en “la batería tiene una autonomía larga”. En cambio, en “las aplicaciones necesitan un tiempo largo para arrancar” esa misma palabra es negativa.

Otro inconveniente evidente de este tipo de técnicas es la necesidad de contar con un diccionario de palabras etiquetadas con su sentimiento. Estos diccionarios pueden ser contruidos a mano, mediante técnicas automatizadas partiendo de diccionarios ya existentes o extrayendo las palabras que forman parte de un corpus concreto. No obstante, ya existen varios recursos disponibles para ser usados, pero la gran mayoría de estos son en lengua inglesa, como por ejemplo SentiWordNet o BLOL. Aun contando con estos elementos, uno de los problemas de este tipo de sistemas es la dificultad para encontrar

diccionarios en cualquier idioma siendo necesario muchas veces hacer traducciones a partir de los ya disponibles.

6.1.3 Métodos basados en relaciones lingüísticas

Además de los métodos basados en diccionarios, en los sistemas de clasificación no supervisada existe otro tipo de modelos basados en relaciones lingüísticas. Estos métodos buscan ciertos patrones en los textos que puedan expresar opiniones y sentimientos con mayor probabilidad, extrayendo las palabras que lo forman para luego ser usadas en la categorización del texto global. Para ello, se obtiene la categoría gramatical de las palabras, llamada parts-of-speech o POS, y se determina si dichos patrones expresan una opinión positiva o negativa. Finalmente, el sentimiento global texto se calcula mediante algún tipo de función matemática.

El trabajo precursor de este tipo de métodos es (Turney, 2002). El método de Turney consistía en extraer frases de los textos a clasificar formadas por adjetivos y adverbios y que respondiesen a una serie de patrones preestablecidos que suelen expresar opiniones y sentimientos. Posteriormente, estimaba el sentimiento de dichos patrones con ayuda del buscador web ya desaparecido AltaVista y dos palabras a modo de semillas: 'excellent' y 'poor'. Para ello, obtenía el número de resultados de cada frase en combinación con cada una de las semillas y mediante el punto de información mutua (pointwise mutual information o PMI, en inglés) podía saber si la frase era positiva o negativa. Para finalizar, se usaba el sentimiento parcial de cada frase para calcular el sentimiento global del texto.

Otro ejemplo que hace uso de este tipo de enfoque para clasificar textos es (Hatzivassiloglou & McKeown, 1997). Este trabajo afirma que, dependiendo del conector usado para unir las palabras, éstas tendrán una orientación semántica igual u opuesta. Así, si dos adjetivos están relacionados mediante la conjunción "y" ambas tendrán la misma polaridad de sentimiento. Ocurre así lo contrario con la conjunción "pero". Haciendo uso de un grupo de conectores y de varias palabras semilla es posible obtener una relación de palabras etiquetadas por su sentimiento y en base a ellas deducir el sentimiento global del texto en el que se encuentran.

Este tipo de sistemas resuelven el problema de la clasificación de una manera vistosa y elegante, pero de alguna manera necesitan apoyarse en algún tipo de recurso que verdaderamente aporte orientación semántica como por ejemplo un grupo de palabras semilla o alguna base de conocimiento. Si éste no existe o no cuenta con la calidad suficiente, el sistema completo puede no ofrecer buenos resultados.

6.2 Métricas y métodos de evaluación de resultados

Para determinar el rendimiento de los algoritmos y de su configuración, es necesario contar con una serie de medidas que permitan evaluar de manera objetiva su eficacia a la hora de clasificar los ejemplos que se le proporcionen. Para ello, es importante no tener solo en cuenta las muestras clasificadas correcta e incorrectamente, sino también las que habiéndose clasificado de manera errónea podrían haberse etiquetado bien.

Para entender los seis posibles estados de un ejemplo a clasificar, pensemos en tres clases A, B y C, y en un algoritmo que determina si dicho ejemplo pertenece o no a alguna de esas clases:

- **True Positives (Verdaderos Positivos o TP):** son los ejemplos que han sido marcados de manera correcta como pertenecientes a la clase A.
- **False Positives (Falsos Positivos o FP):** serán los ejemplos marcados como de clase A, pero en realidad no pertenecen a ella, es decir, han sido clasificados de manera incorrecta.
- **True Neutrals (Verdaderos Neutros o TNeu):** son los ejemplos que han sido marcados de manera correcta como pertenecientes a la clase B.
- **False Neutrals (Falsos Neutros o FNeu):** estos ejemplos son marcados que son de clase A o C, pero de forma incorrecta ya que pertenecen a la B.
- **True Negatives (Verdaderos Negativos o TN):** en este caso, los ejemplos no son de la clase A ni B y han sido clasificados correctamente.
- **False Negatives (Falsos Negativos o FN):** en este grupo estarán los ejemplos marcados como no pertenecientes a la clase A ni B, pero en realidad sí lo son y, por tanto, no se han clasificado correctamente.

Teniendo en cuenta los estados anteriores, podemos definir las siguientes medidas que serán usadas para evaluar nuestros modelos:

- **Exactitud (CA / Classification accuracy) :** esta es la medida de rendimiento más simple e intuitiva y representa la razón entre las predicciones correctas sobre el total de predicciones realizadas. Dicho de otra manera, es el número de elementos clasificados correctamente entre el número total de clasificaciones llevadas a cabo.

$$Accuracy = \frac{TP + TN + TNeu}{TP + FP + TN + FN + TNeu + Fneu}$$

Es habitual pensar que el modelo que ofrezca una mayor exactitud es el mejor modelo. En realidad, esta medida es adecuada en el caso de que el número de

elementos de cada clase sea aproximadamente el mismo y el corpus esté balanceado. En caso contrario, es necesario hacer uso de otro tipo de medidas como la precisión, la exhaustividad y el valor-F. Al contrario que la exactitud, estas medidas no valoran el rendimiento del modelo teniendo en cuenta todas las clases del sistema, sino que lo hacen sobre clases individuales.

- **Precisión (del inglés Precision):** es la razón entre el número de documentos clasificados correctamente como pertenecientes a la clase A y el número total de documentos de que han sido clasificados por el modelo como de clase A.

$$Precision = \frac{TP}{TP + FP}$$

La precisión mide la proporción de identificadores positivas que son realmente correctas. Nótese que su valor aumenta a medida que el número de falsos positivos disminuye.

- **Exhaustividad (del inglés Recall):** es la relación entre los documentos clasificados correctamente como pertenecientes a la clase A y la suma de todos los documentos de la clase A.

$$Recall = \frac{TP}{TP + FN + FNeu}$$

Donde asumimos que FN y FNeu son los falsos que corresponden al grupo de A. La cobertura es la proporción de elementos positivos reales identificados acertadamente. También se puede ver como la capacidad que tiene el modelo de construir de manera correcta las clases.

Cuanto más cercano a 1, mejor estarán definidas las distintas clases existentes ya que su valor aumenta a medida que disminuye el número de falsos.

- **Valor-F:** es habitual que para medir la eficiencia de un modelo de clasificación se haga uso de los valores de cobertura y exhaustividad. Para ello, el valor-F se presenta como la media armónica entre ambas medidas y suele utilizarse como referencia para comparar el rendimiento entre varios modelos. La fórmula del valor-F combina las dos medidas anteriores de manera ponderada a través de un parámetro β lo que permite otorgar una mayor importancia a una que a otra:

$$F_{\beta} = (1 + \beta^2) * \frac{Precision * Recall}{(\beta^2 * Precision) + Recall}$$

Es frecuente que la precisión y la exhaustividad tengan el mismo peso en la fórmula, es decir, con un valor β igual a 1. A esta configuración se la conoce como Valor-F₁.

6.3 Proceso de entrenamiento de los algoritmos

El proceso habitual de la construcción de un clasificador de textos basado en un sistema de aprendizaje automático consta de varias etapas secuenciales. En Primer lugar, es necesario preparar los datos del corpus para entrenar los algoritmos. Para ello, se debe limpiar y normalizar su información con el objetivo de reducir o eliminar aquellos datos que pueda influir de manera negativa en el resultado final. A continuación, cada uno de los textos de ejemplo se somete a un proceso denominado *tokenización*, el cual los divide en unidades más pequeñas o tokens y que habitualmente son las palabras de los mensajes. A partir de los tokens se extraen las características que representen a los mensajes originales. Para finalizar, estas características se ponderan en función de la importancia que se les quiera dar y con ellas se entrenan los clasificadores.



Figura 1 Fases para el entrenamiento de los algoritmos de aprendizaje supervisado

6.4 Preprocesamiento

En todo método que se haga uso de algoritmos de aprendizaje automático es necesario tratar previamente los datos con los que serán entrenados. El objetivo de esta fase es limpiar y normalizar la información para evitar que determinados datos puedan influir de manera negativa en el resultado final.

Esta cuestión es crucial cuando hablamos, en este caso, de mensajes extraídos de reseñas de sitios web ya que es muy habitual encontrar mensajes con faltas de ortografía, repeticiones de caracteres, mezcla de letras mayúsculas y minúsculas, entre otras.

Se han seleccionado un conjunto de reglas para aplicar que suelen ser comunes en la construcción de este tipo de clasificadores. El objetivo que persiguen todos ellos es la normalización de los mensajes, pero evitando en todo momento que los cambios aplicados provoquen la pérdida de la polaridad de sentimiento.

Las reglas son las siguientes:

- **Normalización de mayúsculas y minúsculas:** para los algoritmos de aprendizaje no es lo mismo la palabra “algo” que “ALGO”. Estas palabras son tratadas como dos totalmente distinto, sin ningún tipo de relación entre ellas. Para evitar que esto suceda y mantener el significado de las palabras sin tener en cuenta la forma de sus

caracteres, todos los mensajes serán convertidos a su equivalente en letras minúsculas.

- **Eliminación de tildes:** ya sea en comentarios de Facebook, reseñas de hoteles o en tweets, los usuarios de internet no acostumbran a hacer buen uso de las tildes. Por esta razón, las palabras “relación” y “relacion” serían consideradas por los algoritmos como distintas. Para evitar esta pérdida de relación semántica, serán eliminadas todas las tildes de las vocales de los mensajes de entrenamiento.
- **Eliminación de números:** por norma general, las cifras numéricas no suelen contener información que ayude al proceso de clasificación de polaridad de sentimiento por lo que serán removidas de los textos y así ayudar a reducir la cantidad de características del corpus.
- **Eliminación de stopwords:** existe un conjunto de palabras que, aunque son necesarias para construir oraciones con sentido, carecen de información que ayude a determinar la polaridad de los textos en los que se encuentran. En español, estas palabras son las preposiciones, los pronombres, las conjunciones y las distintas formas del verbo haber, entre otras.
- **Lematización:** este es un proceso de normalización morfológica que transforma cada palabra en su lema mediante el uso de diccionarios y de un proceso de análisis morfológico. Por ejemplo, la lematización convertiría la palabra “lindas” a su lema “guapo”. De esta forma, muchas características tomarían la misma forma, reduciendo así su variabilidad.
- **Stemming:** se trata de otro método de normalización morfológica, pero más agresivo que la lematización. En este caso, una palabra se transforma a su raíz por medio de la supresión de sus sufijos e inflexiones.
- **Tokenización:** Una vez completado el proceso de normalización de los mensajes del corpus, la siguiente etapa es la denominada *tokenización*. En esta fase los textos se dividen en unidades más pequeñas llamadas tokens y que normalmente se corresponden con las palabras de cada texto. Para este proceso se separa cada palabra por los espacios blancos entre las mismas y los caracteres de puntuación.
- **Extracción de las características:** a partir de los tokens obtenidos, se definirá la manera de representar con ellos los mensajes de los que proceden, creando así las llamadas características. Aquí se hará uso del modelo de bolsa de palabras (bag of words) en donde cada mensaje se representa mediante sus tokens sin tener en cuenta ningún orden concreto entre ellos.

7. Análisis del Data set

Se analizará la clasificación de textos por su sentimiento a nivel de documento. Estos documentos serán comentarios que han sido publicados en el sitio web *Restorando*(www.restorando.com). Para ello, será necesario contar con un corpus de entrenamiento cuyos ejemplos deberán haber sido etiquetados previamente con la categoría del sentimiento al que pertenecen.

7.1 Corpus de entrenamiento

Uno de los problemas de los métodos supervisados es la necesidad de contar con un juego de pruebas representativo y previamente etiquetado para entrenar los algoritmos de aprendizaje automático, es decir, un corpus. En el caso concreto de la clasificación de reseñas de restaurantes no es una tarea sencilla de conseguirlos. La creación de este tipo de elementos a menudo resulta complicada debido al enorme coste en términos de tiempo y esfuerzo necesarios para completarla.

Para las pruebas de este trabajo se hará uso de un corpus en español creado a mano ya que no se dispone de un corpus ya existente que cuente con la información que necesitamos. Para el mismo se evaluaron distintos sitios como *PedidosYa*, *Guia Oleo*, *Rappi*, *Glovo* y *Restorando* para tomar datos de los comentarios y realizar las pruebas. Teniendo en cuenta los factores que se explican a continuación, *Restorando* fue la elección más favorable.

El primer punto que se tuvo en cuenta a la hora de elegir el sitio de donde tomar los datos era el volumen comentarios que recibía el sitio y la cantidad de usuarios que lo usaba. *Restorando* es una de las plataformas de restaurantes mas utilizadas y cuenta con una gran cantidad de usuarios registrados, por lo que era una gran opción de donde obtener la información. Pero además de elegir la plataforma de donde obtener la información, también habría que elegir un restaurante específico en donde enfocar el estudio ya que no tiene sentido usar comentarios de distintos lugares con el fin de entrenar los algoritmos. Aquí se consultaron varios rankings y blogs de opiniones de restaurantes para obtener una primera lista de restaurantes candidatos. Luego fue cuestión de ir evaluando uno por uno y ver cuál de ellos iba a brindar la data mas útil. El restaurante elegido fue Cabaña Las Lilas ubicado en Puerto Madero. El mismo al estar ubicado en una zona turística, recibe una gran cantidad de comensales todos los días y también al ser bastante popular tenia una buena combinación de reseñas positivas y negativas lo que iba a generar un corpus más variado y rico en información.

No solo alcanzaba que contengan un gran volumen de datos. También era necesario que esos datos se puedan extraer sin inconvenientes y de la manera más eficiente. Esto fue posible utilizando la tecnica Web Scraping, que es el proceso de recopilar información de forma automática de la Web, ya que si esto se quiere hacer de forma manual se tardaría demasiado tiempo por la cantidad de datos que se quieren obtener.

Para esto se utilizó el plug-in de Web Scraper para Google Chrome que permite la extracción de datos de manera simple y muy fácil de configurar. El plug-in agrega una opción al Developer tools de Chrome en donde se permiten crear Site Maps para realizar el Scraping. Por medio del uso de Selectors se puede definir qué información del sitio se desea extraer. Una vez definido, se realiza el Scraping y se extrae la información en formato CSV.

La herramienta que se utilizó para correr los algoritmos de aprendizaje supervisado acepta archivos CSV para procesar los datos, pero no con el formato que te brinda el Web Scraper. Por este motivo se realizó un código en Java para parsear el CSV y dejarlo en el formato correcto. Todo el código desarrollado se encuentra en la sección de anexos y además se encuentra en el repositorio del proyecto.

La clasificación de los datos se realizó de la siguiente forma. Cada comentario obtenido del sitio de Restorando viene con un puntaje que se encuentra entre el 1 y el 10. Se plantea como hipótesis que una alta calificación se corresponde con una connotación positiva y una calificación baja corresponde a una negativa. Si su puntaje se encontraba por debajo de un 5, se lo clasifico como Negativo. Si está por encima de un 6, se lo clasifico como positivo. Los que quedaron comprendidos entre 5 y 6 se los clasifico como Neutro.

7.2 Análisis de datos obtenidos

Los mensajes del corpus se encuentran clasificados en tres clases: Positivo (P), Neutro (NEU) y Negativo (N), como se mencionó en la sección anterior. Esta será la clasificación en la que se basaran las pruebas de este trabajo.

A continuación, se puede ver en mas detalle los datos obtenidos luego de la clasificación, junto con gráficos de mapa de palabras y tablas con los pesos de las palabras que mas aparecen en cada caso.

7.2.a) Negativos

La siguiente tabla muestra que los valores catalogados como negativos se encuentran entre el puntaje 2.5 y 4.5.

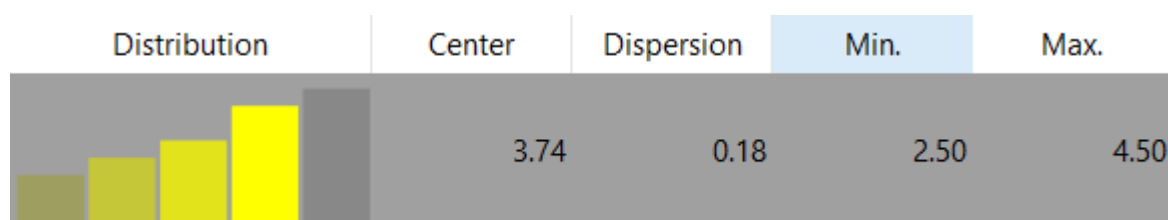


Tabla 1 Estadísticas comentarios negativos

Tabla 2 - Tabla de pesos de las palabras de comentarios negativos

7.2.b) Neutros

En el caso de los neutros, mucho no se puede observar en la tabla siguiente, ya que se impuso un intervalo chico entre 5 y 6 para comprender estos valores.


Distribution	Center	Dispersion	Min.	Max.
	5.52	0.07	5.00	6.00

Tabla 3 - Tabla de estadísticas de comentarios neutros

En estos comentarios aparece una mezcla de palabras positivas y negativas, dando una distribución mas pareja entre todas las palabras.



Figura 3 Nube de palabras de comentarios neutros

Weight	Word
0.383	calidad
0.381	carne
0.269	atencion
0.247	servicio
0.243	comida
0.221	bien
0.203	lugar
0.156	tener
0.151	mejor
0.140	restaurante
0.133	pedir
0.129	veces
0.123	mesa

Tabla 4 Tabla de pesos de las palabras de comentarios neutros

7.2.c) Positivos

Algo que se puede percibir de los comentarios positivos, es que cuando a la gente le gusta el servicio o la comida del lugar, los puntajes tienden a ser la mayoría de las veces el puntaje máximo. Esto se observa en la distribución de puntajes, marcando una gran diferencia con el puntaje máximo 10 respecto de los otros.

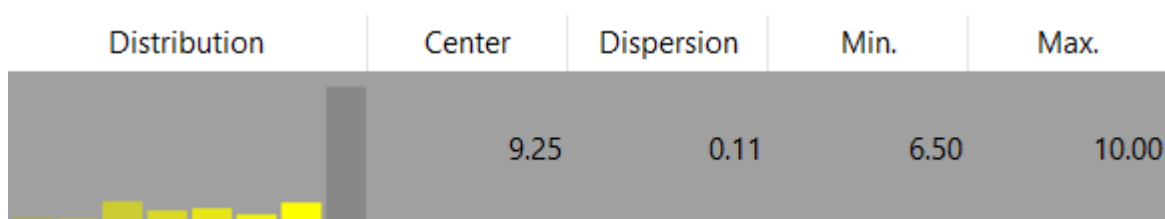


Tabla 5 Tabla de estadísticas de comentarios positivos

Observando la nube de palabras y las tablas de peso, se destacan varios adjetivos con connotaciones positivas como “excelente”, “bueno/a”, “bien”.

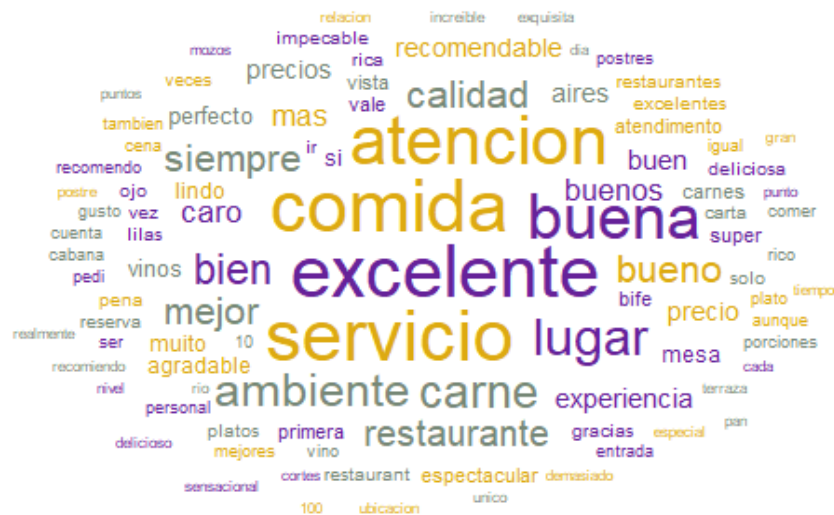


Figura 4 Nube de palabras de comentarios positivos

Weight	Word	Weight	Word
0.619	excelente	0.062	buenos
0.495	comida	0.060	precio
0.300	atencion	0.057	recomendable
0.259	servicio	0.056	precios
0.205	buena	0.049	vinos
0.180	lugar	0.049	si
0.133	ambiente	0.046	aires
0.127	carne	0.039	caro
0.099	calidad	0.037	experiencia
0.092	bien	0.036	vale
0.082	mejor	0.035	primera
0.075	bueno	0.034	super
0.074	siempre	0.034	mesa
0.067	mas	0.033	agradable
0.067	restaurant	0.032	atendimento
0.063	buen	0.031	lindo

Tabla 6 Tabla de pesos de las palabras de comentarios positivos

8. Resultados

En esta sección, la performance de distintos algoritmos y configuraciones va a ser comparada. Los algoritmos utilizados son los que se describieron previamente: Regresión Logística, SVM y Naive Bayes. En todos se utilizó la configuración que muestre los mejores resultados usando la misma data de prueba para todos. Además, se va a evaluar la performance de un algoritmo que utiliza un lexicon en español.

8.1 Pruebas de algoritmos de machine learning

Se realizaron muchas pruebas para los distintos algoritmos y se decidió mostrar con que configuración y cuál fue el mejor resultado por algoritmo. De esta forma, se evita mostrar los resultados de cada una de las pruebas y así lograr una mayor claridad a la hora de observar y comparar resultados.

Para estas pruebas, los datos que se utilizaron poseen las siguientes características:

- 1926 comentarios
- 358 comentarios neutros
- 623 comentarios negativos
- 945 comentarios positivos
- Comentarios en español
- Valores positivos mayores a 6, negativos menores a 5

NAIVE BAYES

En las pruebas con el algoritmo de Naive Bayes, no hay configuración particular del algoritmo que se pueda modificar. Por lo tanto, de la única forma que se podrían obtener distintos resultados fue variando la configuración del muestreo.

Para cross-validation se tomaron valores K comprendidos entre 5 y 20. Para obtener más resultados aún, se volvieron a hacer evaluaciones con esos valores de K, pero usando data estratificada. Esto Consiste en la división previa de la población de estudio en grupos o estratos que se suponen homogéneos respecto a característica a estudiar y que no se solapen.

En el muestreo aleatorio, el porcentaje de datos de prueba a tomar vario entre 30% y 95%. Con respecto al numero de iteraciones a realizar para entrenar al algoritmo, este tomo los valores 10, 20, 50 y 100.

Con el Leave One Out no había parámetros para modificar por la forma en que se realiza el muestreo.

Tomando los mejores resultados de cada una de las pruebas recién mencionadas, se formó la siguiente tabla:

	Exactitud	Precisión	F1	Exhaustividad
Cross-validation	70,2%	74,7%	69,4%	70,2%
Muestreo aleatorio	71,1%	75,1%	70,3%	71,1%
Leave one out	70,6%	74,7%	69,8%	70,6%

Tabla 7 Mejores resultados para las pruebas de Naive Bayes

REGRESIÓN LOGÍSTICA

Para el caso de regresión logística, se utilizaron dos tipos de regularización para el análisis de regresión: Lasso y Ridge. La diferencia entre ellos es que Ridge aproxima a cero los coeficientes de los predictores, pero sin llegar a excluir ninguno. En cambio, Lasso aproxima a cero los coeficientes, excluyendo predictores.

Además, se cuenta con un parámetro C que es una variable de control que retiene la fuerza de modificación de la regularización. Para valores bajos de C, la fuerza de regularización va a incrementarse con lo que se crearon modelos mas simples que se adaptaran a los datos. En cambio, con valores de C altos, el poder de regularización va a disminuirse lo que le permite al modelo incrementar su complejidad y por lo tanto se ajustan en exceso a los datos.

Se realizaron las pruebas utilizando los dos tipos de regularización y valores de 1 y 200 para C. También se vario la configuración de muestreo de la misma forma que se hizo para Regresión Logística y se eligió el mejor resultado de todas las pruebas.

A continuación, los resultados:

	Exactitud	Precisión	F1	Exhaustividad
Lasso – C=200	81,4%	81,7%	81,4%	81,4%
Lasso – C=1	80,9%	75,9%	79,3%	83%
Ridge – C=200	82,9%	81%	80,5%	79,9%
Ridge – C=1	81%	74,3%	80,3%	87,3%

Tabla 8 Mejores resultados para las pruebas de Regresión Logística

SVM

Con SVM se hace uso de funciones kernel que son las que le permiten convertir lo que sería un problema de clasificación no lineal en el espacio dimensional original, a un sencillo problema de clasificación lineal en un espacio dimensional mayor.

Dentro de las distintas funciones kernel que existen, se realizaron pruebas con las de tipo “Lineal”, “RBF” y “Sigmoid”. La diferencia entre ellos es el tipo de curva que utiliza el algoritmo para realizar el mapeo de los datos.

Como se realizo para los algoritmos anteriores, en este caso también se fue modificando la configuración de muestreo y de todos los resultados evaluados, se eligió el mejor de cada caso.

	Exactitud	Precisión	F1	Exhaustividad
Lineal	72,9%	72,9%	72,9%	72,9%
RBF	73,8%	73,6%	73,8%	73,7%
Sigmoid	72,9%	72,9%	72,9%	72,9%

Tabla 9 Mejores resultados para las pruebas de SVM

Conclusión de resultados

Tomando en cuenta todos los resultados obtenidos y haciendo un promedio de ellos, se concluye en la siguiente tabla:

	Exactitud	Precisión positivos	Precisión negativos
Naive Bayes	70,64%	70,6%	79,6%
Regresión Logística	81.55%	81%	74,3%
SVM	72,3%	75,5%	71%

Tabla 10 Resultados finales de los algoritmos utilizados

Se puede ver que utilizando el método de Regresión Logística se obtuvo un mejor porcentaje de exactitud comparado a los demás, que, aunque tengan menor valor tampoco dejan de ser buenos resultados.

También se puede observar que, aunque el método Naive Bayes fue el que obtuvo menor puntaje en términos generales de exactitud, el mismo fue el mas preciso a la hora de clasificar los comentarios negativos.

8.2 Pruebas con lexicon

Realizar pruebas con lexicon no fue tarea sencilla ya que casi todos los lexicons disponibles se encuentran en inglés. De los pocos que hay en español, muchos no están a disposición de forma pública y se necesita hacer un pedido privado para que lo compartan.

Lamentablemente, no se obtuvo respuesta a todos los pedidos realizados a distintas instituciones y universidades que desarrollaron lexicon para distintos usos.

De todas formas, se pudo encontrar un lexicon en español que contenía una gran cantidad de palabras junto a su valor de polaridad (positivas o negativas) que fue utilizado en este trabajo. El mismo fue creado por los autores Fermín L. Cruz, José A. Troyano, Beatriz Pontes y F. Javier Ortega. Fue obtenido desde su sitio web de la Red Temática en Tratamiento de la Información Multilingüe y Multimodal (<http://timm.ujaen.es/recursos/ml-senticon/>).

La idea detrás de estas pruebas es evaluar todos los comentarios utilizando los valores de polaridad del lexicon y ver si esos comentarios son categorizados correctamente en el grupo correspondiente.

Para las pruebas con el lexicon se utilizó una menor cantidad de datos para analizar. De esta manera iba a ser mas fácil de identificar puntajes de comentarios que llamen la atención o no coincidan con la polaridad original del comentario.

Primera versión del código

En esta primera versión, se conto con una lista de comentarios que contenía calificaciones positivas, negativas y neutras. Para analizarlos con el lexicon, se creó un algoritmo en donde se recorría cada comentario y se armaba una lista con las palabras que contenía. A cada una de esas palabras, se le asignaba el valor de polaridad correspondiente (si es que tenía uno) del lexicon. Luego se sumaba cada uno de estos valores y se clasificaba el resultado como positivo o negativo si ese valor es mayor o menor a 0.

Por último, se comparaba con calificación real del comentario.

Estadísticas datos previos al análisis:

Total: 586

Positivos: 415

Negativos: 143

Neutros: 28

Resultados:

Positivos:209

Negativos:25

Falsos Positivos:194

Falsos Negativos:158

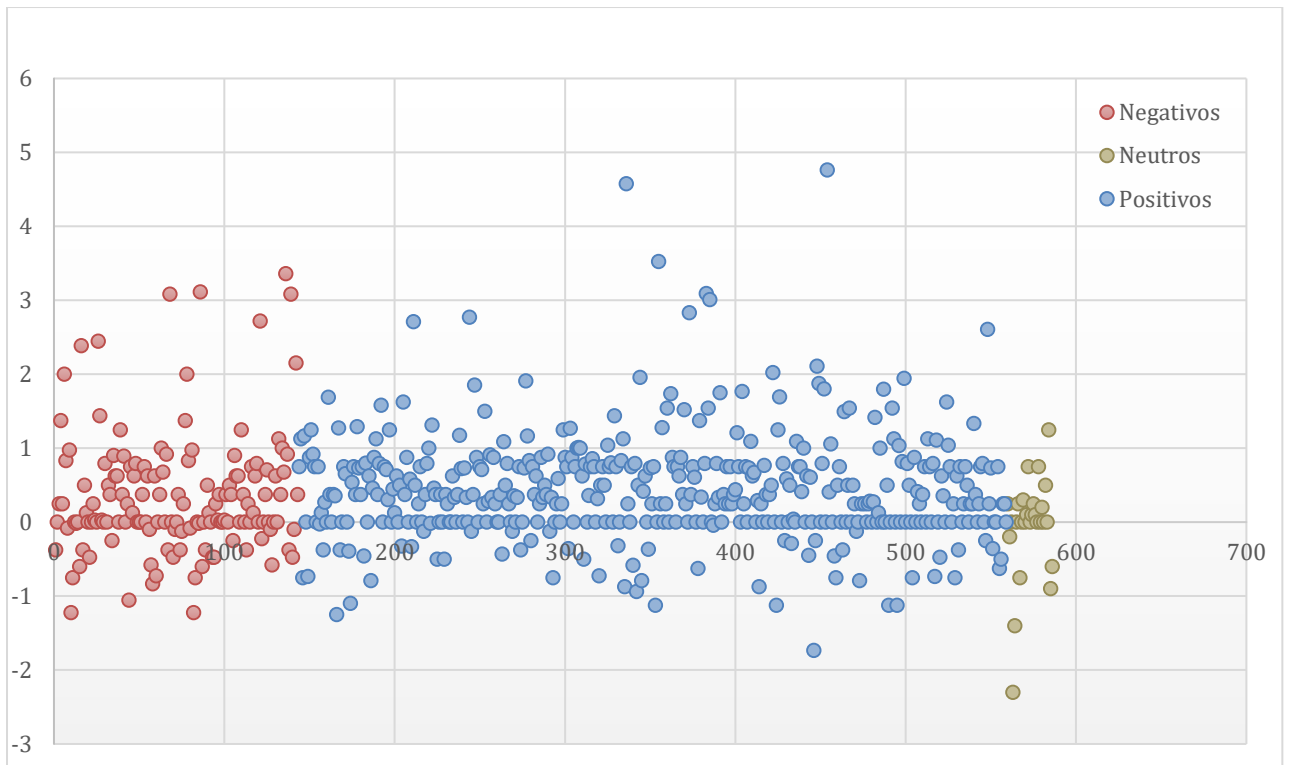


Figura 5 Grafico de dispersión con los resultados de la primera versión

Segunda versión del código

Para esta segunda versión, se realizaron varios cambios en el algoritmo del lexicon y los datos que se usaron.

En principio, se eliminaron los datos neutros del conjunto de información que había que analizar. Como solo se estaba evaluando por positivos o negativos en esta instancia, no tenía sentido incluir los datos neutros ya que generaban ruido en los resultados.

Por otro lado, se hicieron algunas mejoras en el algoritmo para analizar los comentarios, en donde se incluyó un mejor preprocesamiento de los datos. Con esta mejora hubo más palabras que pudieron obtener un puntaje correcto del lexicon.

Por último, al revisar muchos de los resultados obtenidos en la versión 1, se notó que había muchas palabras que no tenían su referencia en el lexicon, por lo que no sumaban ni restaban puntaje y tendían a dar un puntaje de 0 en la sumatoria del comentario. Esto resultaba en que no se clasificaran en ningún grupo y terminaban en los falsos positivos o falsos negativos. Para resolver este problema, incluí nuevas palabras al lexicon tratando de asignarles un valor de polaridad que concuerde, teniendo en cuenta otras palabras que ya se encontraban en el lexicon.

Aquí se puede observar un ejemplo de una de las palabras que se agregó al lexicon:

```
<lemma pos="r" pol="-0.525" std="0.137"> malo </lemma>
```

Estadísticas datos previos al análisis:

Total: 558

Positivos: 415

Negativos: 143

Resultados:

Positives:371

Negatives:47

False Positives:96

False Negatives:44

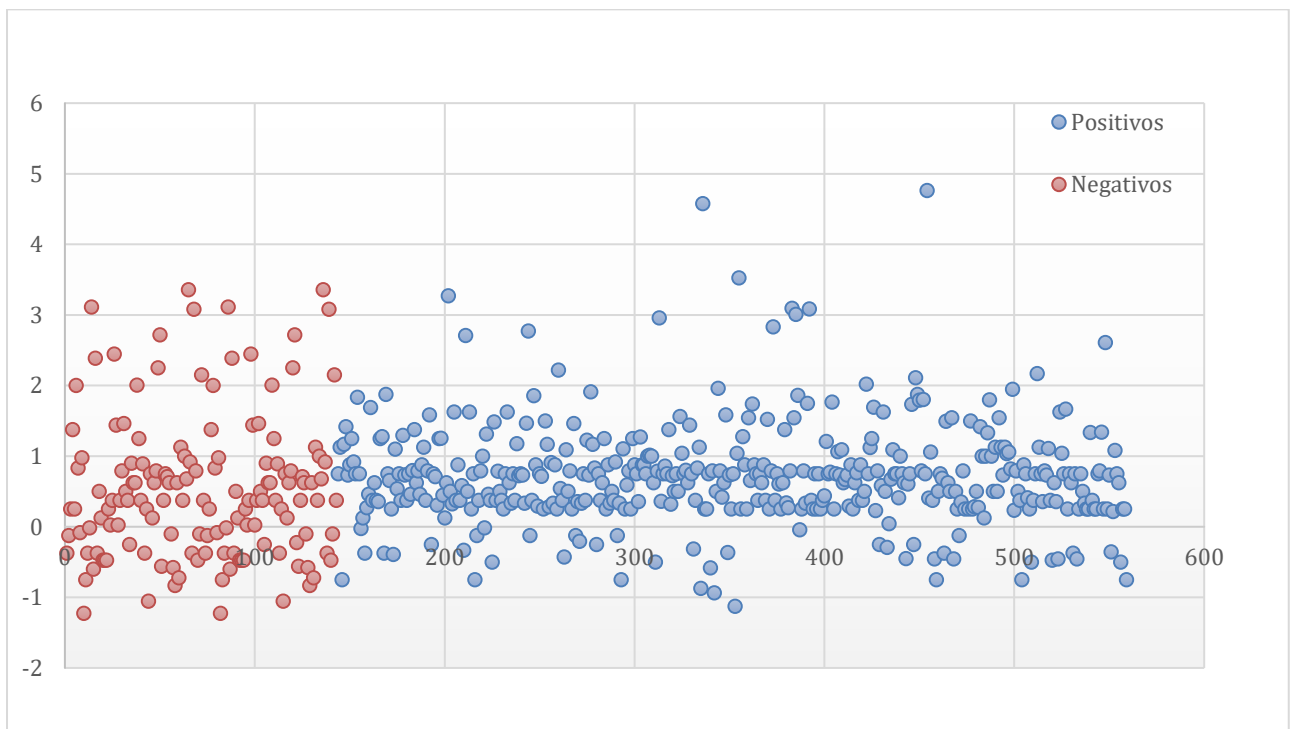


Figura 6 Grafico de dispersión con los resultados de la segunda versión

Conclusión de resultados

Observando los resultados de ambas versiones se pudo ver un gran progreso en los porcentajes obtenidos. Si bien los resultados para los comentarios negativos siguen siendo malos, hubo un incremento de más del doble del valor previo que tenían.

Desde el punto de vista general del algoritmo, se puede observar un buen valor de exactitud comparable con lo que dieron las pruebas de aprendizaje supervisado.

	Exactitud	Precision Positivos	Precision Negativos	Exhaustividad Positivos	Exhaustividad Negativos
Version 1	39,9%	51,8%	13,6%	56,9%	11,4%
Version 2	74,9%	79,4%	51,6%	89,3%	32,8%

Tabla 11 Comparación de resultados de las distintas versiones

9. Conclusiones

En esta sección del trabajo se expondrán algunas ideas y conclusiones sobre los diversos temas tratados, sobre su realización y sobre los resultados obtenidos en las pruebas prácticas realizadas.

9.1 Conclusiones generales sobre el análisis de sentimiento y su futuro

En el transcurso de este trabajo hemos podido conocer las bases teóricas en las que se fundamenta el análisis de sentimientos y de qué diferentes maneras es posible crear un sistema automatizado para evaluar textos y clasificarlos en base a su polaridad. Así mismo, se ha mostrado la comparación de distintos clasificadores basados en un sistema de aprendizaje supervisado, una de las formas más extendidas y efectivas de construir este tipo de sistemas. Estas soluciones evolucionarán hacia desarrollos más ambiciosos y eficaces, en donde no sólo las palabras escritas serán usadas para evaluar los sentimientos.

En el futuro, problemas clásicos del análisis de sentimientos y del PLN como la detención del sarcasmo y de la ironía, la ambigüedad o la dependencia del contexto, se irán mejorando hasta alcanzar, quizás no una solución definitiva, pero sí tolerable. Sin embargo, al mismo tiempo empezarán a aparecer con fuerza otros problemas para los que habrá que buscar soluciones. Bajo mi punto de vista, uno de los más preocupantes es el spam de opiniones, mensajes y publicaciones que tratan de manipular el sentimiento global que las personas tienen sobre un producto, servicio o cualquier tema general. Es posible que cada vez sospechemos más de la autenticidad de ciertas reseñas publicadas en sitios web de restaurantes o e compras online. Tener control sobre este tipo de comportamientos es fundamental para que el análisis de sentimientos pueda ser considerado una herramienta de verdadero valor.

9.2 Problemas hallados, posibles mejoras y resultados obtenidos

En primer lugar, se puede decir que los objetivos marcados fueron logrados en mayor o menor medida. Si bien algunos resultados no fueron los esperados, se entendió el motivo y por qué esto estaba sucediendo, y se trató de llegar a una explicación de este.

A continuación, se explicarán algunas conclusiones obtenidas sobre los distintos problemas que fue apareciendo en el desarrollo del trabajo, como así también mejoras o posibles soluciones a situaciones que no se han podido resolver y también entender un poco los datos obtenidos y tratar de explicar el porque se llegó a eso.

- Como se pudo observar en los resultados de la última versión del código del lexicon, los resultados para los comentarios positivos fueron mejores que los resultados de los comentarios negativos. Al notar esto, se volvió a analizar esos comentarios y tratar de entender porque esto estaba sucediendo. Se pudo identificar que una gran cantidad comentarios negativos contenían muchas palabras con connotación positiva pero las mismas venían negadas o de forma de deseo o de lo que se esperaba del servicio/comida brindada. Por la forma en que se diseñó el algoritmo, esto ya no era algo que se pudiera analizar, ya que requeriría de un algoritmo que evalué las distintas combinaciones de palabras que pueden existir y no de evaluar las palabras individualmente como se hizo.
- El lexicon termino siendo una muy buena forma para el análisis de sentimiento, superando a algunos casos de machine learning en cuestiones de exactitud. Su problema principal, como era de esperarse, fue la cantidad de información provista por el lexicon. Ya de por sí, fue difícil conseguir un lexicon español para utilizar y además a ese lexicon hubo que hacerles modificaciones a lo largo del trabajo para poder obtener mejores resultados. Pero consideramos que este diccionario puede ser un aporte a futuros trabajos de AS.
- Por el otro lado, los algoritmos de aprendizaje supervisado fueron mas confiables en sus resultados sin necesitar de con demasiada preparación previa de los datos. Se considero que su mayor punto en contra es el gran procesamiento que requieren algunos de estos métodos. A la hora de desarrollar las pruebas, hubo una gran diferencia de tiempo computacional comparándolo con el tiempo que llevaba correr el algoritmo del lexicon. Esto se debe principalmente a que los algoritmos de machine learning necesitan todo un preproceso de aprendizaje y entrenamiento.
- En ambos métodos utilizados, muchos comentarios que contenían sarcasmo, ironía, ambigüedad o dependían del contexto no fueron evaluados correctamente. Este es un problema clásico del análisis de sentimiento y todavía no existe una solución definitiva.
- Uno de los problemas encontrados es que la principal fuente de información se encuentra en estudios publicados por universidades. Es cierto que cada vez aparecen mas libros sobre el análisis de sentimientos y que existen diversas paginas web que tratan este tema, pero no es aun tan popular y accesible como lo pueden ser otras ramas de investigación. Encontrar información organizada y que no se contradiga con ella, no ha sido sencillo. Esto se ha notado mas a la hora de encontrar información sobre el lexicon, especialmente al buscar información en castellano.
- Hubo técnicas de reducción de características como *stemming* y lematización, que en el caso de los datos para usar con los algoritmos de machine learning, se utilizó las herramientas provistas por el Orange. Pero para los datos del lexicon no se utilizo ninguna herramienta especializada en esto, sino que se hizo un

preprocesamiento manual. En algunos casos, esto pudo haber interferido en la evaluación de los comentarios usando el lexicon y ciertas palabras no ser ponderadas correctamente. Este fue uno de los motivos por el que muchas palabras fueron agregadas al lexicon, ya que palabras como “belleza” o “bello” contaban con su *stem* “bell”.

- Para concluir, observando los resultados finales obtenidos, se puede decir que los resultados fueron bastante positivos. Tras leer varios papers e informes sobre distintas investigaciones realizadas, se vio que los resultados de exactitud en estos quedaban comprendidos entre un 70-80%. Viendo los resultados de este trabajo, la mayoría de los resultados termino quedando en ese intervalo o superándolo, salvo por algunos casos particulares.

9.3 Trabajo futuro

Para obtener mejores resultados y que este trabajo sea de más utilidad para futuras investigaciones se podría trabajar los siguientes puntos.

En principio sería ideal seguir incrementando el dataset de entrenamiento que utilizan los algoritmos de machine learning. No solo incrementar su tamaño, sino también conseguir información mas confiable en donde por ejemplo los comentarios de las personas sean más fieles a los puntajes que se les asignan.

Otro punto importante es la mejora del lexicon, especialmente para este dominio en particular. Es necesario seguir analizando comentarios y así poder ir incluyendo mas palabras al mismo y enriquecer lo mas posible el diccionario. Si bien es algo que se fue haciendo durante el desarrollo del trabajo, es una tarea en la que se pueden seguir realizando mejoras constantemente.

10. Bibliografía

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

Martínez Cámara, E. (2016). Análisis de opiniones en español. Tesis Doctoral. Universidad de Jaén.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.

Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M. T., & Perea-Ortega, J. M. (2013). Semantic orientation for polarity classification in Spanish reviews. Expert Systems with Applications, 40(18), 7250-7257.

Perez-Rosas, V., Banea, C., & Mihalcea, R. (2012, May). Learning Sentiment Lexicons in Spanish. In LREC (Vol. 12, p. 73).

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

Pedro Larranaga, Inaki Inza, Abdelmalik Moujahid. Clasificadores Bayesianos

Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences, 181(6), 1138-1152

Martin-Valdivia, M. T., Perea-Ortega, J. M., Martinez-Camara, E., Ureña Lopez, L. (2013). Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches.

Khan, A., Khan, K., Baharudin, B. (2011). Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure.

Perez-Rosas, V., Banea, C. Mihalcea, R. (2011). Learning sentiment lexicons in Spanish.

Collomb, A., Costea, C., Joyeux, D., Hasan, O., Brunie, L. (2012). A study and comparison of sentiment analysis methods for reputation evaluation.

Henriquez, C., Guzman, J., (2017). A review of sentiment analysis in Spanish, TECCIENCIA, Vol. 12 No. 22, 35-48

11. Enlaces consultados

Análisis de sentimiento - Wikipedia

https://es.wikipedia.org/wiki/Análisis_de_sentimiento

Sentiment Analysis - Wikipedia

https://en.m.wikipedia.org/wiki/Sentiment_analysis

The importance of Neutral Class in Sentiment Analysis - Datumbox

<http://blog.datumbox.com/the-importance-of-neutral-class-in-sentiment-analysis/>

Text Classification and Sentiment Analysis - Ahmet Taspinar

<http://ataspinar.com/2015/11/16/text-classification-and-sentiment-analysis/>

Multiclass classification - Wikipedia

https://en.m.wikipedia.org/wiki/Multiclass_classification

Precision and recall - Wikipedia

https://en.m.wikipedia.org/wiki/Precision_and_recall

Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures - Exsilio Blog

<http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures>

Text Classification with NLTK and Scikit-Learn - Libelliinar

<https://bbengfort.github.io/tutorials/2016/05/19/text-classification-nltk-sckit-learn.html>

Dive Into NLTK, Part IV: Stemming and Lemmatization - Text Mining Online

<http://textminingonline.com/dive-into-nltk-part-iv-stemming-and-lemmatization>

Text Classification for Sentiment Analysis – Precision and Recall - StreamHacker

<https://streamhacker.com/2010/05/17/text-classification-sentiment-analysis-precision-recall>

Como hacer Análisis de Sentimiento en español - Pybonacci

<https://www.pybonacci.org/2015/11/24/como-hacer-analisis-de-sentimiento-en-espanol-2>

A Tour of Machine Learning Algorithms - Machine learning mastery

<https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms>

Webscraper – Web scrapping tool

<https://www.webscraper.io/>

Orange – Data mining

<https://orange.biolab.si/>

Portal Estadística Aplicada

<http://www.estadistica.net/>

Sentiment analysis in Spanish

<http://blog.manugarri.com/sentiment-analysis-in-spanish>

Clasificador bayesiano ingenuo - Wikipedia

https://es.wikipedia.org/wiki/Clasificador_bayesiano_ingenuo

Regresión logística – Wikipedia

https://es.wikipedia.org/wiki/Regresi%C3%B3n_log%C3%ADstica

Análisis de sentimiento, ¿Qué es, como funciona y para qué sirve?

<http://www.itelligent.es/es/analisis-de-sentimiento>

SVM (Support Vector Machine) – Theory – Machine Learning 101 - Medium

<https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72>

Validación cruzada – Wikipedia

https://es.wikipedia.org/wiki/Validacion_cruzada

The future of Sentiment Analysis – Terrific data

<http://terrificdata.com/2017/04/07/future-sentiment-analysis>

Red Temática en Tratamiento de la Información Multilingüe y Multimodal

<http://timm.ujaen.es/>

12. Anexo

Este trabajo se complementa con los siguientes recursos adicionales: el código de fuente del procesamiento de datos, el código fuente para generar el algoritmo del lexicon, hojas de cálculo con los datos y resultados obtenidos a partir de las pruebas realizadas y el lexicon utilizado. Ambos recursos se explican con mayor detalle en esta sección.

Código fuente

La mayor parte del procesamiento de datos ha sido escrito en lenguaje de programación Java 1.8. Se destacan dos clases principales:

- *HTMLparserOrange*: clase utilizada para realizar el procesamiento de los datos obtenidos por el Webscrapper y que puedan ser utilizados en la herramienta del Orange.
- *LexiconParser*: clase utilizada para correr el algoritmo que analiza los comentarios utilizando la información provista por el lexicon.

El proyecto está adjunto como anexo a este documento de memoria y además puede ser consultado en el siguiente proyecto alojado en:

- <https://pabloagustinpauli@bitbucket.org/itba/bpm-institucional.git>

Lexicon

Se trata de varias listas de lemas positivos y negativos para español. Cada lema viene acompañado de una estimación numérica de su polaridad (entre -1.0 y 1.0) así como de un valor de desviación típica de dicha polaridad. Las listas están organizadas en varias capas, de manera que las primeras capas contienen estimaciones más precisas de los valores anteriores, aunque contienen menos elementos que las capas posteriores.

El mismo fue obtenido de la Red Temática en Tratamiento de la Información Multilingüe y Multimodal y se le aplicaron varios cambios a lo largo del trabajo.

Orange

Es un programa informático para realizar minería de datos y análisis predictivo desarrollado en la facultad de informática de la Universidad de Ljubljana. Consta de una serie de componentes desarrollados en C++ que implementan algoritmos de minería de datos, así como operaciones de preprocesamiento y representación gráfica de datos.

Los componentes de Orange pueden ser manipulados desde programas desarrollados en Python o a través de un entorno gráfico.

Por medio de este programa, se armó todo el modelo para poder operar con el corpus de datos, procesar esa información y luego aplicar los algoritmos de aprendizaje supervisado y analizar los datos obtenidos.