

El estado actual de la reputación corporativa de la industria de Rafaela (Argentina). Un análisis a través de las interacciones de sus clientes en las redes sociales

Javier Fornari¹, Mariano Ferrero¹, Constanza Abeille¹, Gonzalo Perez¹, Boglione Sebastián¹

¹ Universidad Católica de Santiago del Estero Departamento Académico Rafaela
Bv. Hipólito Irigoyen 1502, (2300) Rafaela, Argentina

Abstract. *This work aims to analyze the data collected from the interactions of customers in social networks corresponding to the companies of the industrial census of the city of Rafaela, in order to detect their relationships and feelings regarding the reputation of the industry. A detailed study of techniques that are used for sentiment analysis is carried out and an analysis is applied where the results obtained through these techniques are presented to determine based on the comments the approach that clients have regarding the quality of the industrial products and their relationship with companies. The results determine that the reputation of the industry is good, and the level of customer interaction is minimal, where the approach could be used for direct application over another industrial, commercial or service area.*

Keywords: *Corporate Reputation, Industrial Census, Social Networks, Sentiment Analysis.*

Resumen. Este trabajo tiene como objetivo el análisis de los datos recopilados de las interacciones de clientes en redes sociales correspondientes a las empresas del censo industrial de la ciudad de Rafaela, con el fin de detectar sus relaciones y sentimientos respecto a la reputación de la industria. Se realiza un estudio detallado del conjunto de técnicas que se utilizan para el análisis de sentimiento y se aplica un análisis donde se plantean los resultados obtenidos mediante estas técnicas para determinar en base a los comentarios el enfoque que poseen los clientes respecto a la calidad de los productos industriales y su relación con las empresas. Los resultados determinan que la reputación de la industria es buena y el nivel de interacción de los clientes es mínimo, cuyo enfoque podría ser utilizado en otro contexto.

Palabras clave: *Reputación Corporativa, Censo Industrial, Redes Sociales, Análisis de Sentimientos.*

1. Introducción

La estructura económica actual de la ciudad de Rafaela incluye tanto una amplia producción industrial como una significativa producción agropecuaria, que de diversas maneras ha contribuido al desarrollo de actividades manufactureras. Así mismo, Rafaela muestra una estructura industrial con variados sectores que la convierten en un área productiva poli sectorial, por lo que no resultó tan afectada ante las diversas crisis económicas y sociales que se sucedieron en el país. No obstante, de la diversificación de actividades que conforman el tejido industrial local, los sectores de alimentos y bebidas y fabricación de autopartes y de maquinaria y equipos aglutinan el 69% del empleo total (Rafaela, Censo Industrial Rafaela,

2016) (Rafaela, Censo Industrial Rafaela, 2012). El sector industrial, en lo que se refiere a la actividad manufacturera, incluye muchas de las industrias incipientes de la provincia de Santa Fe, que emergieron para responder a la expansión de la demanda producida por una población que crecía rápidamente. Otras, surgieron en relación directa con las explotaciones cerealeras, para proveerlas de los insumos necesarios o para elaborar el producto final. El movimiento cooperativo, motivado por la idea de controlar la propia producción industrial y su posterior comercialización, formó un entramado socio territorial que dio al área ciertas características que se mantuvieron durante mucho tiempo, como ser productores medianos, buena red de caminos, electrificación rural, paisaje agroindustrial, entre otras (Ascúa, 2007). Por otro lado, el complejo escenario competitivo que se presenta a las empresas que actúan en el mercado global, despliega nuevos desafíos para los agentes involucrados. En este marco, el desarrollo de ventajas competitivas depende, cada vez más, de la destreza de las firmas relacionada con las mejoras de productos y procesos, el aseguramiento de la calidad, los cambios organizacionales y la búsqueda de nuevas formas de vinculación con el mercado. Esto indica que, en la actualidad, el logro de la diferenciación depende no sólo de la capacidad de innovación de las empresas en forma individual, sino de una adecuada armonía entre los procesos de innovación que se dan a nivel de la región. En el proceso de creación de ventajas competitivas dinámicas, estos procesos de innovación, sumados a la incorporación de tecnología, al desarrollo de procesos de aprendizaje y a su marca y reputación, juegan un rol clave (Ascúa, 2007) (Albuquerque, 2006). En este contexto cobran fuerza las teorías relacionadas con las ventajas asociadas al desarrollo de clusters y distritos industriales. Los clústeres son concentraciones geográficas de empresas e instituciones interconectadas entre sí por un hilo conductor, el cual puede ser un determinado sector, y todos ligados por externalidades de diversos tipos, mientras que los distritos industriales son entidades socioeconómicas fuertemente identificadas con un territorio geográfico y con una identidad cultural e históricamente determinada (Albuquerque, 2006). Es por ello por lo que en la evolución industrial y comercial de la ciudad de Rafaela varias entidades fueron creadas, a los efectos de representar las necesidades de la región y así poder impulsar el crecimiento y sustentabilidad de la ciudad. Estas instituciones representan el entramado institucional de la ciudad y las competencias principales de cada institución, con sus actividades específicas, sus áreas de trabajo y sus acervos de saberes acumulados, permiten que sean reconocidas en el ámbito local y regional por todo el conjunto de actores involucrados. El grado de cooperación existente entre las instituciones de Rafaela y su región genera una sinergia que permiten apreciar la cooperación interinstitucional en proyectos conjuntos relacionados con la comunidad. Los ejes sobre los cuales se muestra el liderazgo en el tejido institucional son, en primer lugar, la generación de iniciativas y propuestas, la capacidad de articular y generar consenso en segundo lugar, como tercer punto la capacidad de negociación externa y finalmente la visión estratégica de futuro (Comisión Económica para América Latina y el Caribe., 1998). Por otro lado, desde el año 2000, el Instituto de Capacitación y Estudios para el Desarrollo Local (ICEDEL) de Rafaela dependiente de la Municipalidad ha realizado un censo industrial de la ciudad reiterándolo cada 6 años. Actualmente posee el relevamiento de datos de 3 censos junto con sus informes y comparativos. En realidad, las organizaciones públicas y privadas reconocen el valor de los datos como un activo clave para entender profundamente fenómenos empresariales, económicos y sociales que sirven para mejorar la competitividad empresarial en un entorno dinámico (Batini, Cappiello, Francalanci, & Maurino, 2009) (Fox, Levitin, & Redman, 1994) (Madnick, Wang, Lee, & Zhu, 2009). Es por ello que en los últimos años tanto empresas como instituciones dedican importantes esfuerzos a desarrollar sus políticas de marca y visión empresarial de forma de integrar en su estrategia empresarial la consolidación de su imagen y confianza con sus públicos estratégicos.

2. Definición del problema

Hasta hace algunas décadas atrás los medios de intercambio de opiniones en la sociedad eran un tanto acotados. El diálogo cara a cara con otros individuos y los medios gráficos eran dominadores de este tipo de mensajes. No obstante, la instauración de Internet como la herramienta de comunicación que es hoy en día, ha modificado las reglas del juego notablemente. Desde luego que con esto no se pretende transmitir la idea de que aquellos pioneros canales de comunicación han dejado de utilizarse ni mucho menos, pero sí dejar en claro que al día de hoy la facilidad con la que cuentan las personas para "hacerse escuchar" es notoria. Este proceso de cambio que se ha generado en el espacio digital, donde los usuarios no solo consumen contenido sino que además lo generan y de forma continua, ha dado lugar a lo que hoy en día se conoce como Web 2.0.

La aparición de foros, primero, y redes sociales, después, han comenzado a delimitar de alguna forma el escenario en el cual nos encontramos inmersos actualmente. Siendo un tanto precipitados, podríamos decir que la gran mayoría de aplicaciones web que utilizamos diariamente admiten la generación de contenido por parte de los usuarios, principalmente texto. Si bien es difícil dimensionar el tamaño concreto de dicho escenario, solamente Twitter (una de las redes sociales más utilizadas en todo el planeta) reúne aproximadamente 250 millones de usuarios (Hootsuite and We are social, 2019). A continuación, se procede a definir algunos de los factores que se identifican asociados al análisis de opiniones:

- **Negación:** Así como la aparición de determinados términos en el texto puede orientarnos sobre el sentimiento que se expresa, la detección de contraindicaciones de los mismos es igual de importante. Ejemplo: "Este teléfono móvil definitivamente no es el mejor que he tenido en mi vida".

- **Cuantificadores/intensificadores:** Junto con la negación, existen otros factores que actúan como modificadores de otros términos presentes en el texto y deben ser puestos en consideración si lo que se busca es obtener la opinión que se transmite. Ejemplo: "Ha transmitido su mensaje con poca seguridad".

- **Subjetividad:** Tal como se mencionó previamente, no cualquier texto expresa opiniones o sentimientos, por lo que resulta imperioso poder detectar aquellos que si lo hacen frente a los que expresan hechos concretos. Esta tarea, conocida como detección de subjetividad, es considerada incluso más compleja que el propio Análisis de opiniones (Mihalcea, Banea, & Wiebe, 2007). Los ejemplos relacionados con el ámbito del cine mencionados párrafos atrás permiten observar esta diferencia.

- **Ironía:** Los seres humanos hemos sido capaces de desarrollar gradualmente nuestra capacidad de expresión, llegando al punto en que diciendo algo en realidad el mensaje que queremos transmitir es otro distinto. La ironía es una característica compleja de detectar, pero resulta relevante en el ámbito del Análisis de opiniones dado que su interpretación podría derivar en resultados sumamente distintos. Ejemplo: "Gracias por haber chocado el auto, es justo lo que me faltaba para terminar esta semana perfecta".

- **Orden:** Es interesante remarcar que un texto, e incluso una misma oración, pueden denotar opiniones contradictorias que se conjugan de diversas formas para expresar una idea final. Lo que se intenta dejar claro con esto es que no necesariamente una opinión se refiere a favor - o en contra - de algo en su totalidad, sino que pueden producirse matices en su interior. Ejemplo: "Las habitaciones eran amplias y el personal muy respetuoso, pero no volvería a ese antro ni en sueños".

- **Contexto:** Un criterio sumamente importante que clasifica a una opinión de una forma u otra es el contexto en el cual está centrada. A modo de clarificar, una frase u oración - en

principio positiva - relacionada a los beneficios de una empresa privada podría expresar un sentimiento completamente diferente si se relaciona con el ámbito de la política y el bienestar social. Ejemplo: "Los precios de los servicios de luz aumentaron considerablemente".

- Objeto: Otro de los aspectos que no debe pasarse por alto es que una opinión puede referirse a distintos elementos, expresando un sentimiento concreto sobre cada uno de ellos. Esto resulta interesante debido a que, si se desea capturar el mensaje sobre alguno de ellos en particular, debe ser posible identificar aquellas frases que se refieran al mismo y analizarlas de forma aislada. Ejemplo: "Es un móvil muy interesante, la pantalla presenta una calidad increíble y su sensibilidad al tacto es perfecta, aunque el plástico utilizado en la carcasa parece poco resistente y la vida útil de la batería deja mucho que desear".

- Medio/canal: Resulta evidente que el medio sobre el cuál se expresa una opinión condiciona notablemente a la misma. En concreto, el Análisis de opiniones llevado a redes sociales requerirá un tratamiento de emoticones y expresiones informales que caracteriza este tipo de canales de comunicación, mientras que otro distinto deberá ser el tratamiento si se aplica sobre opiniones de películas realizadas por expertos en un sitio dedicado a tal fin.

- Multilingüismo: Muchos de los estudios y enfoques planteados para intentar combatir las problemáticas introducidas hasta este momento son propias del lenguaje sobre el que se plantean (principalmente en inglés) y no son extensibles a otros. Si bien esto no resulta extraño debido a las particularidades de cada idioma, presenta otro inconveniente que debe ser tenido en cuenta.

Vale la pena destacar que, si bien se podría pensar en abordar esta problemática como un problema de clasificación temática concretamente (utilizando como categorías de interés: positivo y negativo), la diferencia con ese escenario es notoria. Mientras que el tema que trata un artículo en particular (política o deporte, por ejemplo) puede ser identificado en base a palabras claves (senado o balón, correspondientemente), el sentimiento que se transmite en un texto puede expresarse de formas un tanto más sutiles que dificultan su interpretación de dicha manera (Pang, Lee, & Vaithyanathan, Thumbs up?: Sentiment classification using machine learning techniques., 2002). Un ejemplo de esto son las ironías, tal como ha sido señalado previamente.

Habiendo introducido la temática y las distintas dimensiones que dificultan el llevar a cabo la tarea de automatizar el análisis de opiniones en fuentes de datos no estructuradas como texto, se procede a realizar un repaso de los antecedentes encontrados hasta el momento en relación con este campo de estudio en el plano científico.

3. Revisión del estado del arte

El interés sobre el campo del análisis de opiniones se ha despertado en los últimos años producto de las ventajas que ofrece en el ámbito académico y comercial, no obstante, esta temática ha sido abordada inicialmente desde hace más de 10 años y desde distintos enfoques. A continuación se pretende realizar una revisión de aquellos trabajos más relevantes en este campo, realizando una división de acuerdo a la técnica utilizada.

Debido a que a lo largo de esta sección serán utilizados distintos términos que se consideran transversales a los diversos enfoques presentados, resulta apropiado realizar una primera descripción de dichos términos a modo de facilitar la lectura y el entendimiento.

Dichos términos son:

- N-grama: Un n-grama es una secuencia de n elementos, la cuál puede referirse a letras, palabras u otro tipo de unidades. De este concepto se desprenden otros relacionados, tales como

unigrama, bigrama o trigrama, los cuáles hacen referencia a secuencias de una longitud específica (uno, dos o tres, respectivamente). En este trabajo se hará uso de este término para referirse a secuencias de palabras, salvo que expresamente se mencione lo contrario. De esta forma un unigrama sería "blanco", mientras que un bigrama sería "producción científica".

- **Stemming:** Se denomina "stemming" a la obtención de la raíz de una palabra en particular. Por ejemplo, la raíz de "fabuloso" podría considerarse como "fabul". Esto permite una simplificación cuando se realizan comprobaciones de términos, dado que "fabul" podría utilizarse para reconocer otras palabras que tienen esa misma raíz (como "fabulosos", "fabulosa", etc.) y expresan - a priori - una opinión favorable. Sin embargo, esto también puede constituir una fuente de error, dado que la raíz de "fábula" también es "fabul", y esta última palabra no representa ningún tipo de polaridad evidente. Existen distintos algoritmos de "stemming", siendo el de Porter (Porter, 1997) uno de los más utilizados.

- **Desambiguación:** Un reto para el Procesamiento del Lenguaje Natural ha sido, y continúa siéndolo, la desambiguación de las palabras que presentan más de un significado posible. Un ejemplo evidente de esto es "chica", que bien puede ser utilizado como un adjetivo ("La mesa es chica.") o sustantivo ("La chica de mi barrio."). Si bien las personas podemos identificar el significado concreto en una oración de forma relativamente sencilla, la automatización de este proceso es un desafío. Muchos de los artículos que serán introducidos en este capítulo utilizan distintas técnicas o algoritmos que tratan con este inconveniente, aunque no se mencione la complejidad y los posibles errores que podrían presentarse en este proceso.

- **Stop-words:** Las palabras vacías, denominadas "stop-words" en inglés, son palabras que generalmente se eliminan dado que no aportan información relevante para el análisis que se quiere llevar a cabo. Generalmente son clasificados como "stop-words" determinados pronombres, artículos o preposiciones, entre otros elementos.

4. Léxicos polarizados

Taboada y otros (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011) presentan uno de los trabajos de mayor impacto sobre léxicos polarizados donde se definen como conjuntos finitos de elementos donde cada uno de ellos representa un término junto con su correspondiente polaridad. Dicha polaridad puede ser expresada de diversas formas dependiendo cada caso, por ejemplo: positivo o negativo; con un rango de valores entre -5 y 5; etc.

Este recurso es una herramienta muy útil que permite obtener de alguna forma el resultado de acuerdo con el sentimiento que expresa un texto analizando aquellos términos que se encuentran tanto en el texto como en dicho recurso. La Tabla 1 presenta un ejemplo simplificado de un léxico polarizado con un número sumamente acotado de términos. A partir de esto, a continuación se presentan distintos ejemplos que utilizan dicho léxico para obtener el resultado en distintas oraciones:

1. La película es excelente (+), si bien me he aburrido (-) unos momentos la actuación de cada uno de sus actores principales ha sido sublime (+).

2. Al comenzar a utilizar el servicio me ha abrumado la tristeza (-), me he sentido un torpe (-) al pagar por esto.

3. Si visitas Buenos Aires dicen que debes conocer este estadio, tiene un tamaño imponente (+) aunque suele haber larga espera (-) para ingresar.

Término	Puntaje
Excelente	+
Aburrido	-
Imponente	+
Torpe	-
Tristeza	-
Impredecible	+
Sublime	+
Espera	-

Tabla 1. Ejemplo de un léxico polarizado simplificado

El recuento final para estos ejemplos, en base al léxico presentado, determinaría que el resultado en el ejemplo 1 es positivo, dado que el número de términos de este tipo (2) es mayor al número de términos negativos (1). En el ejemplo 2 sucede lo contrario, dado que el número de términos negativos es 2 y no se encuentran términos positivos, con lo cual el resultado final sería negativo. Por último, se presenta un ejemplo donde la suma de términos positivos y negativos no produce diferencia alguna. En este caso, en principio, no se puede determinar que hay una preponderancia de una polaridad sobre otra, con lo cual el resultado final sería neutral.

Uno de los primeros trabajos, no sólo dentro de este enfoque, sino que, dentro del propia área de análisis de opiniones, ha sido el realizado por Turney (Turney, 2002). Turney se centra en la clasificación de diversas críticas como recomendada o no recomendada, comenzando con un léxico polarizado extremadamente acotado, el cual se basa en un único término por cada categoría polar: excelente ("excellent") y pobre ("poor"). Estos términos han sido escogidos debido a que se consideran de alguna forma los más representativos de las correspondientes categorías. A partir de esto, cada crítica es analizada extrayendo los bigramas que contienen adjetivos o adverbios. Por cada uno de estos bigramas se realizan dos consultas a un motor de búsqueda: bigrama NEAR "excelente" y bigrama NEAR "pobre". El operador NEAR del motor de búsqueda utilizado permite filtrar aquellos resultados donde ambas partes se encuentran como máximo a una distancia de diez palabras, independientemente del orden en el que se encuentren. Una vez obtenidos estos resultados (número de coincidencias entregadas por el motor de búsqueda), las cantidades de cada uno se utilizan para obtener las medidas de IM y restarlas entre sí. De esta forma cada frase identificada en la crítica es ponderada, y finalmente el cálculo final se obtiene operando con estos valores. Una de las grandes bondades de esta propuesta en particular es que no requiere de la creación manual de un léxico polarizado, lo cual suele ser una tarea que demanda recursos considerables, principalmente tiempo y dinero. Sin embargo, la utilización de un motor de búsqueda se presenta como una gran limitante debido a que presupone una carga computacional extra en cada análisis que se quiera realizar.

Nasukawa y Yi (Nasukawa & Yi, 2003), por su parte, presentan un método que se diferencia con el anterior principalmente en que el léxico polarizado que utilizan ha sido creado manualmente (incluyendo verbos, sustantivos, adjetivos y adverbios), un número acotado de situaciones que podrían expresar negaciones son tenidas en cuenta y la clasificación se realiza sobre un tema en concreto más que sobre el texto en general. Esto permite filtrar aquellas opiniones que no están relacionadas directamente con lo que se pretende analizar, pero podrían estar presentes en el texto de interés. Si bien en este último artículo se evidencia la importancia de contemplar los cambios de polaridad que podrían presentarse producto de negaciones, existen otros trabajos que se han centrado de forma más profunda en este tipo de cuestiones (Lapponi, Read, & Ovreid, 2012) (Li, Wang, Lee, & Huang, Sentiment classification with polarity shifting detection., 2013). Por su parte, Polanyi y Zaenen (Polanyi & Zaenen, 2006) mencionan que hay dos grandes tipos de modificadores de polaridad: contextuales (donde se incluyen negaciones e intensificaciones, entre otras) y basados en el discurso (como por ejemplo la utilización de determinados tipos de conectores - "a pesar de", "pero", etc.).

Uno de los trabajos más relevantes basado en léxicos polarizados concretamente es el desarrollado por Taboada y otros (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011), donde los autores llevan adelante la construcción de un diccionario compuesto de adjetivos, sustantivos, verbos y adverbios, junto con la ponderación de cada término dentro de un rango de -5 (extremadamente negativo) y 5 (extremadamente positivo). El cálculo final se realiza teniendo en cuenta negaciones (las cuales invierten el puntaje original de los términos involucrados), intensificadores (los cuáles modifican en un porcentaje específico arbitrariamente definido los términos relacionados, ya sea aumentando o disminuyendo su valor original) e "irrealis blocking" (lo cual se asemeja al concepto de modificadores basados en discurso que mencionan en su artículo Polanyi y Zaenen (Polanyi & Zaenen, 2006)). Vale la pena destacar que los resultados obtenidos son interesantes, incluso sobre distintos dominios. Esta última característica es destacable, dado que un gran problema que existe con la creación de léxicos polarizados es su dependencia al dominio sobre el cuál se desarrollan.

Con el correr de los años, como se ha sido mencionado previamente en este trabajo, el interés por analizar la información presente en redes sociales ha aumentado cuantiosamente. (Hutto & Gilbert, 2014) se centran en este contexto, realizando en primer lugar una revisión de los recursos léxicos disponibles para esta tarea hasta el momento. Al concluir que ninguno de éstos se encontraba correctamente desarrollado para su utilización en este tipo de herramientas, proceden con la creación de un léxico polarizado personalizado que nombran como "gold standard".

Por último, Khan y otros (Khan, Bashir, & Qamar, 2014) proponen un modelo que consta principalmente de la utilización de tres léxicos polarizados independientes para clasificar mensajes de la red social Twitter en positivos, negativos o neutros. El primero de ellos utiliza un conjunto de 145 emoticones (70 positivos y 75 negativos), el segundo emplea un total de 9493 palabras que se obtienen combinando dos léxicos existentes en la literatura (2360 positivas y 7133 negativas), mientras que el último de los tres se apoya en SentiWordNet (Baccianella et al. (Baccianella, Esuli, & Sebastiani, 2010)) para determinar la polaridad de cada palabra. El modelo se completa con una primera fase de pre-procesado de los mensajes, realizando tareas de lematización, eliminación de enlaces encontrados, entre otras cosas. Se introduce en la Figura 1 el detalle completo del modelo presentado. Si bien en esta propuesta no son considerados aspectos tales como negaciones, los resultados obtenidos son interesantes.

A modo de resumen, se considera que este enfoque tiene ventajas interesantes, como por ejemplo la simplicidad en la interpretación de los resultados que entrega, tal como se mencionó previamente. No obstante, la creación de léxicos polarizados conlleva una ardua

tarea que no puede ser pasada por alto. Con respecto a los artículos analizados concretamente, si bien se presentan algunos intentos por automatizar la creación o ayudar a la creación de estos recursos, sería interesante contar con herramientas que permitan llevar adelante esta tarea, así como también un mantenimiento y actualización de los léxicos incluidos.

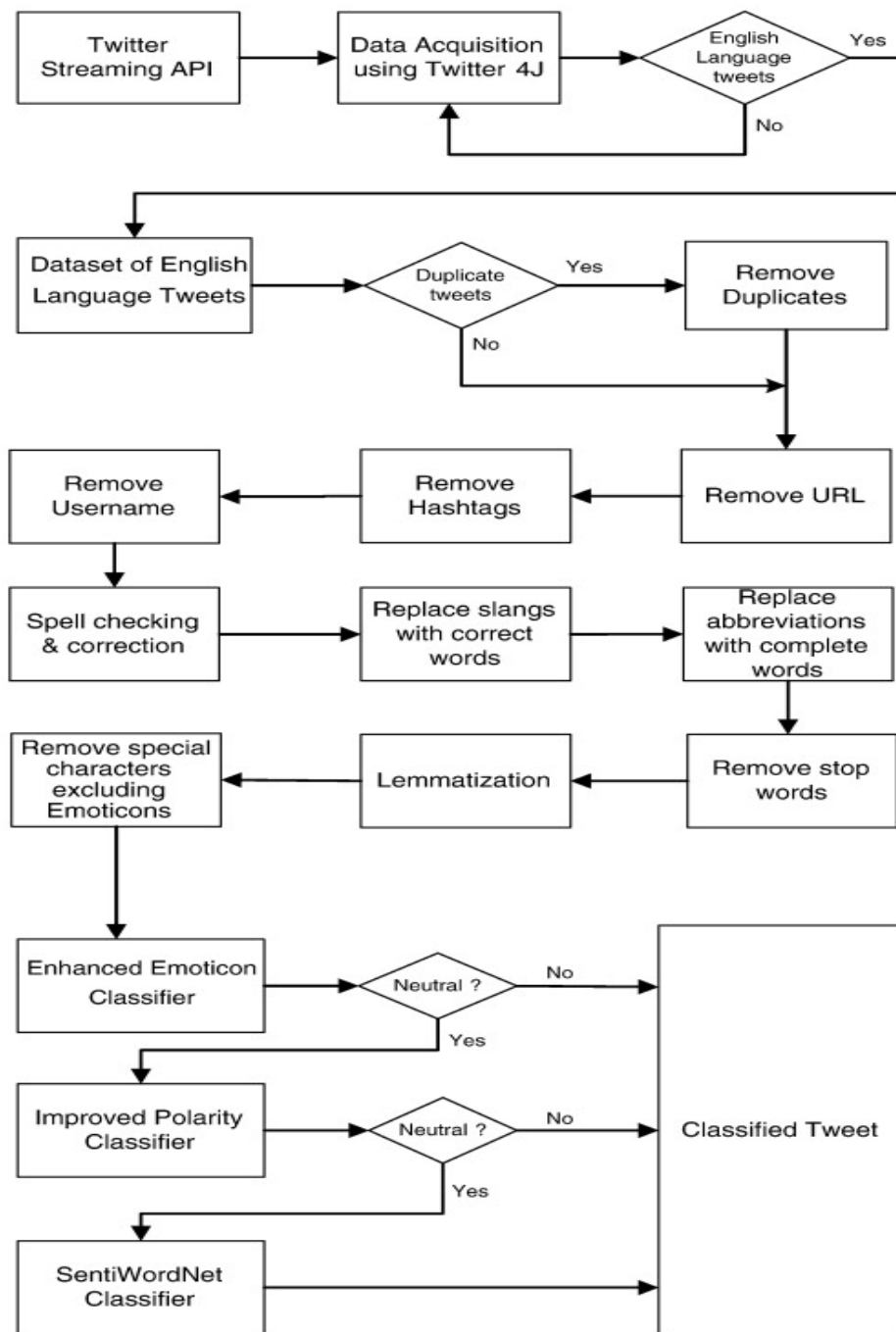


Figura 1. Descripción del modelo propuesto por Khan y otros (Khan, Bashir, & Qamar, 2014).

5. Aprendizaje automático

Se define al concepto de aprendizaje automático o aprendizaje de máquina ("machine learning", en inglés), como la programación de computadoras para optimizar un determinado criterio utilizando datos de ejemplo o experiencia pasada (Alpaydin (Alpaydin, 2009)). Dicho de otra manera, lo que se persigue es buscar aproximar mediante un modelo una función determinada que ante la presencia de determinados datos produzca determinadas salidas. La

búsqueda de dicha optimización es lo que le confiere el término de aprendizaje, y dicho proceso suele conocerse técnicamente como entrenamiento del modelo. Algunas de las tareas, aunque no todas, en las cuáles se han aplicado técnicas de aprendizaje de máquina son: traducción automática (Güvenir and Cicekli (Güvenir & Cicekli, 1998)), detección de correos electrónicos basura, clasificación de textos (Sebastiani (Sebastiani, 2002)), predicción meteorológica (Chen et al. (Chen, Chang, & lin, 2004)), detección de enfermedades (Sajda (Sajda, 2006)), clasificación de imágenes (Chapelle et al. (Chapelle, Haffner, & Vapnik, 1999)), detección de fraude (Aleskerov et al. (Aleskerov, Freisleben, & Rao, 1997)) y análisis de riesgo crediticio (Galindo and Tamayo (Galindo & Tamayo, 2000)).

Actualmente se cuenta con una gran cantidad de técnicas y modelos disponibles, los cuáles dependiendo el problema sobre el que se apliquen son mejores o peores candidatos para encontrar buenas soluciones. Una primera clasificación que surge de este gran conjunto de modelos es en base a cómo se realiza el aprendizaje: supervisado, semi-supervisado o no supervisado. A continuación, se introducen cada una de estas categorías en detalle, junto con los distintos trabajos encontrados en el área del Análisis de opiniones que utilizan este tipo de técnicas.

6. Aprendizaje supervisado

Lo que caracteriza a este tipo de métodos en particular, el más estudiado dentro de los tres, es que el algoritmo aprende con ejemplos conociendo la salida esperable para cada uno de ellos. Esto quiere decir que por cada caso que es proporcionado, se especifica el resultado que se debería entregar. Una vez que el modelo ha sido entrenado con estos ejemplos, ante la llegada de nuevas instancias de las cuáles se desconoce si son positivas o negativas, el mismo podría clasificarlas en una u otra categoría (con un margen de error asociado). En lo que respecta al Análisis de opiniones, esta tarea de clasificación suele denominarse Clasificación de opiniones o Clasificación de sentimientos. Vale la pena destacar que si bien en este sencillo ejemplo se han utilizado solo dos etiquetas (positivo o negativo), lo cual se conoce a su vez como Clasificación polar de opiniones, existen otras posibilidades. Por otra parte, tal como mencionan Pang y Lee (Pang & Lee, Opinion mining and sentiment analysis., 2008), es posible enfocar el problema utilizando otro tipo de etiquetas, como "a favor" y "en contra" en el ámbito político.

Una de las principales ventajas que presenta la utilización de estos métodos es que adquieren niveles altos de exactitud si se entrenan con volúmenes de datos suficientes, a pesar de que su capacidad de generalización es menor comparada con la utilización de léxicos polarizados. Otra de las ventajas que presentan es su capacidad de adaptación a diversos escenarios modificando las variables que se utilizan como entrada para el algoritmo en cuestión. Uno de los primeros trabajos que ha abordado el problema de Análisis de opiniones mediante la utilización de algoritmos de aprendizaje supervisado es el de Pang y otros (Pang, Lee, & Vaithyanathan, Thumbs up?: Sentiment classification using machine learning techniques., 2002). En su caso, se han centrado en críticas de películas obtenidas de la "Internet Movie Database" ¹. Los algoritmos que han sido tenidos en cuenta son tres: Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés), Naive Bayes (NB) y Máxima Entropía (ME). La justificación sobre la utilización de estos modelos particularmente se basa en los buenos resultados obtenidos en problemas de clasificación temática de textos en trabajos previos. Aquel lector que no se encuentre familiarizado con estos modelos puede encontrar una explicación sencilla introducida por los mismos autores en dicho artículo, o bien acudir al libro de Russell y Norvig (Russell & Norvig, 2009). La estructura que se ha utilizado para modelar los datos en dicho artículo se conoce como bolsa de palabras, o "bag-of-words" en inglés, la cuál es simple y permite obtener buenos resultados, pero tiene algunos inconvenientes dado

que rompe el orden en el que las palabras se presentan en el texto. Si se considera $[f_1, \dots, f_m]$ como un vector predefinido de m términos que podrían encontrarse en un texto y $n_i(d)$ el número de veces que el término i aparece en el documento d , cada instancia que representa un documento se expresa de la siguiente forma: $(n_1(d), n_2(d), \dots, n_m(d))$, junto con su etiqueta correspondiente. A su vez, también se ha experimentado en este estudio reemplazando la frecuencia de cada término por valores binarios, indicando con un 1 que dicho término se encuentra presente y con un 0 lo contrario. Por último, en la fase de experimentación también se han realizado pruebas modificando las variables contempladas (unigramas, unigramas y bigramas, adjetivos, etc.). Es posible encontrar en la literatura enfoques similares a este que utilizan estos algoritmos (Ye et al. (Ye, Zhang, & Law, 2009), Narayanan et al. (Narayanan, Arora, & Bhatia, 2013) y Tripathy et al. (Tripathy, Agrawal, & Rath, 2015)). Dentro de dicho conjunto, vale la pena destacar el trabajo presentado por Basari y otros (Basari, Hussin, Ananta, & Zeniarja, 2013) donde se busca adaptar los parámetros del método SVM mediante la utilización de Optimización por enjambre de partículas (PSO, por sus siglas en inglés), una técnica proveniente de la rama de las metaheurísticas.

Si bien los distintos modelos mencionados hasta este momento, y en concreto NB y SVM, han sido los mayormente utilizados para abordar la problemática, también existen otras propuestas. Moraes y otros (Moraes, Valiati, & Neto, 2013) comparan la utilización de SVM con Redes Neuronales Artificiales (RNA), concluyendo que a pesar de que estas últimas no han sido tenidas muy en cuenta. Orimaye y otros (Orimaye, Alhashmi, & Siew, 2012) se centran en diversas críticas provenientes de Amazon 2 y destacan la imposibilidad de dicho enfoque ("bag-of-words") de modelar características propias de los discursos utilizados en este contexto, donde el autor puede variar su opinión a lo largo del texto (es decir, puede comenzar hablando a favor, luego en contra y finalizar con comentarios favorables). Para lidiar con esto introducen distintas variables que luego son utilizadas por el clasificador para predecir la crítica como positiva o negativa. Otro enfoque que presenta una variante al clásico "bag-of-words" utiliza una primera identificación de términos que expresan polaridad, pero luego a partir de estos generan distintas variables (como por ejemplo, la cantidad de términos - positivos y negativos - presentes) que son las que finalmente utiliza el algoritmo de aprendizaje automático Katz et al. (Katz, Ofek, & Shapira, 2015).

7. Conjuntos de clasificadores

Tal como se mencionó anteriormente, existen distintos tipos de algoritmos de aprendizaje automático. A pesar de que muchas veces existen "candidatos" a ser utilizados de acuerdo a la problemática con la cual se esté tratando (como en este caso podrían ser NB o SVM), no es posible conocer con anterioridad cuál de ellos obtendrá mejores resultados. Por otra parte, una de las ideas que ha surgido en la comunidad científica es combinar distintos clasificadores con el objetivo de intentar mejorar los resultados que se obtendrían utilizando solo uno de ellos. Esta técnica es conocida en inglés como "ensemble learning". Dos de los enfoques más reconocidos de este tipo son Bagging (Breiman, 1996) y Boosting (Freund & Schapire, 1996).

La ventaja de la utilización de este tipo de métodos reside principalmente en una disminución del número de errores que se producen, aunque como desventajas se presentan la carga computacional elevada que suponen junto con la dificultad asociada para comprender los resultados de estos modelos. Wang y otros (Li, Wang, Lee, & Huang, Sentiment classification with polarity shifting detection., 2013) trabajan sobre la problemática de Análisis de opiniones comparando distintas técnicas de "ensemble" (entre ellas Bagging y Boosting), utilizando diversos clasificadores base (por ejemplo, NB, SVM y ME). Al igual que este artículo, existen otras propuestas que utilizan métodos de "ensemble" para trabajar sobre la problemática de

Análisis de opiniones (Fersini, Messina, & Pozzi, 2014) (da Silva, Hruschka, & Hruschka, 2014).

8. Aprendizaje profundo

Los grandes volúmenes de información disponibles junto con las mejoras en las capacidades de cómputo han dado paso a la creación de Redes Neuronales Artificiales de gran tamaño, es decir, con numerosas capas y neuronas ocultas. La utilización de este tipo de estructuras se conoce como aprendizaje profundo, o "deep learning" en inglés. Se puede encontrar una revisión detallada de los distintos modelos que constituyen esta área en el artículo de Schmidhuber (Schmidhuber, 2015). Por su parte, Socher y otros (Socher, y otros, 2013) proponen un método denominado "Recursive Neural Tensor Networks", el cuál lo utilizan para clasificar oraciones mediante el análisis de pares de unigramas. Zhang y LeCun (Zhang & LeCun) han presentado un estudio siguiendo este enfoque, utilizando "Temporal Convolutional Networks" (conocidas como "ConvNets"), donde entre otras propuestas se ha llevado a cabo un Análisis de opiniones en reseñas de Amazon.

La mejora que ofrecen estos modelos es su alta capacidad de generalización, algo que se ve contrastado con la demanda computacional que requieren. Por otra parte, una crítica frecuente que se oprime sobre las RNA es que son vistas como "cajas negras", debido a que no resulta posible encontrar explicaciones a las salidas que estos modelos producen.

9. Aprendizaje semi-supervisado

Como un escenario hipotético podría plantearse la situación en la que el equipo de marketing de una compañía que cuenta con su sitio de comercio electrónico desea conocer si las críticas que los usuarios dejan hacia sus productos en dicho sitio son favorables o no. Si bien la aplicación permite a los usuarios acompañar los mensajes con una puntuación, esto no es obligatorio. Al momento de comenzar el Análisis de opiniones sobre estos mensajes, se descubre que un bajo porcentaje de ellos cuenta con dicha puntuación. Llevar adelante la lectura y clasificación de cada uno de ellos supone un esfuerzo ingente si se plantea hacerlo manualmente. Las técnicas de aprendizaje semi-supervisado surgen para trabajar en este tipo de escenarios, donde solo se cuenta con un número reducido de datos etiquetados y un gran número de ellos sin etiquetar. Dasgupta y Ng (Dasgupta & Ng, 2009) proponen un mecanismo que combina una serie de técnicas para trabajar en estas situaciones, enfocado en la problemática de Análisis de opiniones. Lo que se busca, en términos generales, es clasificar de forma manual aquellos datos en los cuáles resulta complejo determinar su etiqueta, y utilizar un algoritmo que clasifique automáticamente aquellos que son más sencillos. Otros métodos similares que utilizan un enfoque semi-supervisado aplicado a la problemática de interés del presente trabajo son los de Aue y Gamon (Aue & Gamon, 2005), Goldberg y Zhu (Goldberg & Zhu, 2006)

10. Aprendizaje no supervisado

Existe un tercer tipo de métodos donde el entrenamiento se realiza sin conocer la clase o etiqueta a la que pertenece cada instancia. Este otro tipo de métodos se denominan no supervisados, y una de las técnicas más conocidas dentro de esta rama es el "clustering": dado un determinado conjunto de datos se intenta encontrar posibles grupos bien definidos, buscando la máxima homogeneidad posible dentro de los elementos de un mismo grupo y la máxima heterogeneidad entre los elementos de distintos grupos. Nuevamente, la ventaja de este tipo de técnicas es que no requieren grandes volúmenes de datos etiquetados.

En el proceso de revisión de antecedentes se percibe que la utilización de este tipo de técnicas aplicadas a la problemática de Análisis de opiniones ha sido menos abordada que las técnicas supervisadas. Li y Liu (Li & Liu, A clustering-based approach on sentiment analysis., 2010) explican que esto puede deberse, entre otros factores, a que:

- La precisión obtenida (asumiendo que los grupos son conocidos y se puede ubicar cada dato en uno de ellos) suele ser menor.
- Si bien entregan grupos definidos no se especifica cada uno de ellos con una etiqueta en concreto. Esto suele ser una fase posterior que requiere la intervención humana, donde en algunos casos esos posibles grupos se conocían o intuían con antelación, mientras que en otros casos no.
- Los resultados pueden ser inestables, es decir, podrían variar notablemente en diversas ejecuciones del algoritmo.

En su artículo, Li y Liu proponen la utilización de uno de los algoritmos de "clustering" más reconocidos, denominado k-means, para el análisis de críticas de películas de cine. Este algoritmo, explicado de forma sencilla, dado un número candidato de grupos intenta agrupar los datos para conseguir que la diferencia entre las instancias de cada grupo sea lo menor posible. De esta forma, los autores utilizan "bag-of-words" (teniendo en cuenta únicamente adjetivos y adverbios) para representar cada instancia y aplicar luego el algoritmo.

11. Reglas

Si bien los dos grandes enfoques introducidos hasta este momento han sido los más abordados por la comunidad científica para trabajar con el Análisis de opiniones, existen otro tipo de propuestas que resultan interesantes de destacar. Una de ellas es la utilización de reglas que denotan una polaridad en particular (positiva o negativa), las cuáles pueden estar compuestas por una única palabra o un conjunto de expresiones.

Un ejemplo de una de dichas reglas podría ser el siguiente:

"ninguna" + {palabra_polar_positiva} = negativo

Uno de los trabajos que aplican este enfoque es el de Poria y otros (Poria, Cambria, Winterstein, & Huang, 2014), en el cuál utilizan también razonamiento del sentido común para trabajar a un nivel de conceptos más que de palabras únicamente ("bag-of-concepts"). Esto les permite de alguna forma, mediante una base de conocimientos (SenticNet), detectar conceptos y clasificarlos polarmente como un todo más que de manera individual.

12. Ontologías

Una ontología es una "especificación explícita de una conceptualización" (Gruber, 1993), o, dicho de otra forma, una representación de los términos y las relaciones que existen entre ellos, asociadas a un vocabulario en particular. Dentro del ámbito de interés de este estudio, una ontología constituye una herramienta que permite reconocer los distintos aspectos que están relacionados con un objeto en particular, de modo que puedan ser reconocidos en el texto. De esta forma, Zhao y Li (Zhao & Li, 2009) han desarrollado una ontología para trabajar con críticas de cine y reconocer aspectos concretos (musicalización, director, actor, escritor, entre otros). Por otra parte, Kontopoulos y otros (Kontopoulos, Berberidis, Dergiades, & Bassiliades, 2013) utilizan una ontología con una finalidad diferente. En este artículo se centran en la clasificación de tuits 3 para un dominio en particular: los teléfonos móviles. En su caso, la ontología constituye una herramienta que les permite filtrar aquellos mensajes que contengan aspectos relacionados con dichos dispositivos, tales como: pantalla, teclado, antena, interfaz,

etc. Esto les permite ampliar el número de mensajes que son analizados, en su caso, mediante la utilización de un servicio web. La utilización de dicha interfaz representa una gran limitación en este artículo, dado que se desconoce con exactitud el funcionamiento de ella.

13. Híbridos

Uno de los caminos más transitados, debido tal vez a que han sido los enfoques más abordados, es la combinación de léxicos polarizados con algoritmos de aprendizaje de máquina. Kennedy e Inkpen (Kennedy & Inkpen, 2006), describen por un lado la utilización de un léxico polarizado (construido a partir de la combinación de recursos existentes), y por otro la utilización de SVM con unigramas como variables binarias de entrada (1 o 0, dependiendo si se encuentra presente o no en el mensaje a analizar). Luego, combinan ambos de dos formas distintas: sumando las predicciones ponderadas de cada uno o entregando ambos resultados a un nuevo clasificador (SVM) que los utiliza como variables de entrada, generando este otro modelo la clasificación final. Por otra parte, Mudinas y otros (Mudinas, Zhang, & Levene, 2012) combinan estos mismos enfoques de una manera diferente. En primer lugar, se lleva adelante la creación de un léxico polarizado, y luego a partir del mismo se obtienen distintas variables que son utilizadas por un modelo SVM.

Ortigosa y otros (Ortigosa, Martín, & Carro, 2014) se han centrado en el contexto de las redes sociales, más específicamente en Facebook 6, aplicado al campo de "e-learning" y en concreto para el idioma español. Tal como se introdujo anteriormente, los enfoques híbridos que intentan combinar la capacidad de generalización de los léxicos polarizados con los altos niveles de precisión de los algoritmos de aprendizaje de máquina (entre ellos la utilización de técnicas de "ensemble" y aprendizaje profundo) han sido ampliamente abordados. Un número interesante de artículos de este tipo han sido encontrados en el proceso de revisión de antecedentes: aprendizaje supervisado y léxicos polarizados (Ikeda, Takamura, arie Ratinov, & Okumura, 2008) (Dang, Zhang, & Chen, 2010) (Jiang, Yu, Zhou, Liu, & Zhao, 2011), aprendizaje semi-supervisado y léxicos polarizados (Sindhwani & Melville, 2008) (Ortigosa-Hernández, Alzate, Lucania, Inza, & Lozano, 2012), aprendizaje no supervisado y léxicos polarizados (Hatzivassiloglou & McKeown, 1997), "ensembles" y léxicos subjetivos (Wilson, Wiebe, & Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis., 2005) y aprendizaje profundo junto con léxicos polarizados (Zhou, Wu, & Tang, 2002).

14. Recursos disponibles

El esfuerzo invertido por la comunidad científica y el mundo empresarial sobre el área de análisis de opiniones han dado paso a la creación de numerosos recursos que se encuentran disponibles actualmente. Se pretende introducir aquellos recursos que han sido encontrados en el proceso de revisión de antecedentes en el momento que se lleva adelante este trabajo, los cuales podrían ser útiles para el problema planteado en esta investigación.

15. Código y aplicaciones

En la fase de preprocesamiento, dependiendo el enfoque y el alcance de este, es factible llevar a cabo diversas tareas. Algunas de ellas han sido mencionadas anteriormente, tales como desambiguación o lematización. Otras tareas factibles de llevar adelante son el reconocimiento de nombres de entidades ("Named-Entity Recognition" en inglés, NER) o etiquetado gramatical ("Part-Of-Speech tagging", o POS tagging), las cuales permiten determinar y clasificar nombres de personas, organizaciones, etc., o etiquetar gramaticalmente las palabras de una oración (sustantivo, verbo, etc.), respectivamente. La Tabla 2 presenta un resumen con

los distintos recursos específicos encontrados actualmente, acompañados con una breve descripción y el tipo de tarea con la que tratan.

Tarea	Nombre	Descripción
Corrección ortográfica	GNU Aspell	Libre y de código abierto. Soporta múltiples idiomas.
Corrección ortográfica	Jazzy	API para entornos Java e idioma inglés únicamente
Corrección ortográfica	Jspell	Múltiples entornos con 5 idiomas (incluido español).
Lematización	XIP 10	Servicio web disponible para inglés, francés y alemán.
Stemming	Snowball Stemmer 11	Implementación del algoritmo de Porter (1997)

Tabla 2. Código y aplicaciones disponibles para preprocesamiento de textos.

Por otra parte, existen distintas librerías o paquetes que integran diversas tareas relacionadas con el Procesamiento del Lenguaje Natural. La Tabla 3 introduce estas herramientas junto con una breve descripción de cada una de ellas.

Por último, existen distintas implementaciones de algoritmos que se encuentran relacionadas más específicamente con el análisis de opiniones. En concreto, dichas implementaciones son métodos o enfoques que han sido desarrollados y se encuentran disponibles para ser utilizados. La Tabla 4 introduce estos recursos.

En síntesis, se puede mencionar que, si bien existen herramientas e implementaciones de distintos enfoques para trabajar con Análisis de opiniones, el número es considerablemente menor en comparación con la cantidad de investigación desarrollada al día de la fecha. Esto se encuentra directamente relacionado con la gran cantidad de artículos que no dejan disponible el código utilizado en la fase de experimentación, lo que restringe llevar a cabo una replicación de los resultados obtenidos. Por otra parte, un grupo acotado de los proyectos que se han encontrado y parecieran presentar un mayor grado de madurez son aquellos en los cuáles existe una comunidad de gente trabajando detrás de ellos, siendo algunos ejemplos de esto Weka y NLTK. A su vez, y tal como será precisado más adelante en este artículo, de las herramientas encontradas ninguna de ellas resulta factible de ser utilizada en el contexto que se enfoca esta investigación (mensajes provenientes de redes sociales sobre páginas de Facebook en Argentina).

Nombre	Descripción
FreeLing	<p>Librería escrita en C++ que implementa numerosas tareas (NER, POS, etc.) en diversos idiomas (español, catalán, francés, entre otros).</p>
LingPipe	<p>Conjunto de herramientas que ofrece versiones de pago y gratuitas.</p> <p>Algunas de las funcionalidades que incluye son: detección de idioma, "POS tagging", NER y corrección ortográfica.</p>
NLTK	<p>Plataforma escrita en Python que contiene numerosas librerías ("stemming", NER, POS, entre otras), al mismo tiempo que provee interfaces de acceso a diversos recursos como WordNet.</p>
OpenNLP	<p>Colección de componentes Java que incluye "POS tagging", NER y detección de oraciones, entre otros.</p>
Stanford Core NLP	<p>Conjunto de herramientas de código abierto desarrollado en Java que incluye, entre otras funcionalidades, "POS tagging", NER, lematización y división de oraciones. Permite trabajar con 6 idiomas, incluido español.</p>

Tabla 3. Librerías y paquetes disponibles para preprocesamiento de textos.

16. Léxicos

Muchos de los métodos introducidos anteriormente se basan en la utilización de léxicos polarizados para funcionar. Actualmente, y producto del interés detrás de esta área del conocimiento, existen diversos léxicos que han sido creados y puestos a disposición para trabajar con ellos. En este apartado se pretende introducir el listado de recursos encontrados en el proceso de revisión de antecedentes de esta memoria, describiendo particularidades de cada uno de ellos.

Nombre	Descripción
Narayanan y otros (2013)	Implementación en Python del método presentado en el artículo, el cuál utiliza técnicas de aprendizaje supervisado.
NLTK	Esta plataforma mencionada anteriormente contiene también implementados métodos para clasificar textos de acuerdo a su polaridad.
Pattern.en	Implementación abierta en Python para el idioma inglés utilizando léxicos polarizados, la cual forma parte de un recurso completo denominado Pattern.
Stanford Core NLP	Código abierto en Java del algoritmo basado en aprendizaje profundo desarrollado por la universidad de Stanford.
VADER	Código Python basado en léxicos polarizados diseñado para análisis de textos en redes sociales en idioma inglés.

Tabla 4. Algoritmos y métodos existentes para el análisis de opiniones.

A continuación, se procede a introducir estos recursos:

- Loughran McDonald Master Dictionary (Loughran & McDonald, 2011): Estos autores han desarrollado un diccionario de términos en inglés centrado en el ámbito de las finanzas, el cual han utilizado para demostrar que un gran número de palabras clasificadas como negativas en otros dominios no representan dicha polaridad en su área. Este recurso cuenta con más de 80.000 palabras, de las cuales 2355 han sido clasificadas como negativas y 354 como positivas.

- EmotiNet (Balahur, Hermida, Montoyo, & Muñoz, 2011): base de conocimiento desarrollada con el objetivo de poder detectar emociones en datos no estructurados (texto), teniendo presente la idea de que dichas emociones se transmiten asociadas a diversas situaciones y no de forma aislada. La idea subyacente en la utilización de este recurso es poder captar las interacciones que pueden producirse en el texto a través de los sentimientos que se transmiten, junto con el resultado final que se refleja.

- General Inquirer (Stone, Dunphy, & Smith, 1996): Uno de los recursos pioneros de este tipo ha sido el desarrollado por Stone y otros en el año 1966 (nótese la antigüedad del mismo). Este recurso, el cuál es una aplicación que se apoya en un léxico polarizado para funcionar, fue concebido con el objetivo de analizar el contenido de diversos textos para encontrar características particulares en los mismos. A pesar de haber sido desarrollado hace medio siglo aproximadamente, continúa en desarrollo. El léxico utilizado fue desarrollado manualmente a partir de distintas fuentes en idioma inglés, y actualmente contiene un total de 1915 palabras positivas y 2291 palabras negativas, generando un total de 4206 unigramas.

- Hu y Liu (Hu & Liu, 2004): Este recurso ha sido desarrollado originalmente en el año 2004 pero ha continuado su evolución. Actualmente cuenta con 4783 unigramas en idioma inglés clasificados como negativos y 2006 como positivos, lo que hacen un total de 6789 palabras. El origen de este ha sido en el contexto de las críticas de películas, aunque diversos autores lo han utilizado en otros escenarios.

- Linguistic Inquiry and Word Count (LIWC) (Tausczik & Pennebaker, 2010): LIWC es una aplicación de software que ha sido diseñada para estudiar los rasgos cognitivos, estructurales y emocionales que pueden presentarse en un texto. Este recurso ha sido desarrollado por más de 10 años y continúa en vigencia, reuniendo a diversos profesionales (psicólogos, sociólogos y lingüistas). La incorporación de dichos profesionales es una ventaja interesante para este recurso ya que desde el ámbito de la psicología se incorpora la interpretación de los significados, de la sociología la aplicación al ámbito del contexto en particular y desde el ámbito de la lengua para comprender la sintaxis y semántica estructural del lenguaje. La herramienta trabaja con un diccionario en idioma inglés que contiene alrededor de 4500 palabras anotadas con al menos una categoría dentro de un extenso conjunto.

- Multi-Perspective Question Answering (MPQA) Subjectivity Lexicon (Wilson, Wiebe, & Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis., 2005): Este recurso constituye uno de los léxicos polarizados disponibles más ampliamente utilizados por la comunidad científica para tratar con la problemática de Análisis de opiniones. Consta de un total de 8221 palabras, las cuáles se encuentran clasificadas en distintas categorías: 4912 negativas, 2718 positivas, 570 neutral y 21 que pertenecen a más de una al mismo tiempo (clasificadas como "both"). A su vez, cada registro es acompañado por otros datos que son interesantes: tipo (indica el grado en el que el unigrama expresa subjetividad: "strongsubj" o "weaksbj"), categoría gramatical (sustantivo, adjetivo, verbo, etc.) y un valor que indica si el término expresa una raíz o la palabra original. Existe un valor que indica la longitud de palabras que contiene el término.

- NRC Emotion Lexicon (Mohammad & Turney, 2010): Un léxico presentado más recientemente es el desarrollado por Mohammad y Turney, conocido también como EmoLex. El mismo representa una lista de 14182 palabras clasificadas de acuerdo a distintas emociones (ira, miedo, expectación, confianza, sorpresa, tristeza, alegría y disgusto) y sentimientos (positivo o negativo). El mismo puede ser utilizado gratuitamente con propósitos académicos. Si bien es un recurso interesante por el método utilizado para crearlo, en una opinión considero que utilizar este tipo de servicios de traducción no puede ser considerado altamente confiable. De esta forma, si el idioma con el que se pretende trabajar difiere al inglés, utilizar este léxico podría conllevar a resultados imprecisos.

- NRC Hashtag Sentiment Lexicon (Kiritchenko, Zhu, & Mohammad, 2014): A diferencia del recurso anterior, en este caso el léxico ha sido creado de forma automática. El mismo cuenta con 54129 unigramas ("#fabulous", "#nice", "elegant"), 316531 bigramas ("perfect combination", "perfect hangover", "loved it") y 480010 pares de elementos (donde cada uno de dichos elementos pueden ser un unigrama o un bigrama, "it—happening", "i—happy today", "excellent—from"), en inglés y disponible gratuitamente para usos académicos.

- SenticNet: Este recurso forma parte de un proyecto que está en continuo desarrollo. En su versión actual, la 3.0, la base de conocimiento cuenta con aproximadamente 30000 conceptos en idioma inglés. Cada uno de los registros que componen este almacén de datos, cuenta con información semántica (específicamente los cinco conceptos que se encuentran más relacionados), de sentimientos (con cuatro valores que indican su relación con distintas emociones: "Pleasantness", "Attention", "Sensitivity" y "Aptitude") y polar (representado con un número decimal que se ubica en el rango -1 y 1).

- Sentiment140 Lexicon (Kiritchenko, Zhu, & Mohammad, 2014): Este léxico es similar a "NRC Hashtag Sentiment Lexicon", mencionado previamente, con la diferencia de que ha sido creado mediante un corpus distinto, utilizando 1,6 millones de tuits. A su vez, en lugar de utilizar "hashtags" como base para determinar la polaridad del mensaje, en este caso se utilizaron emoticones. Contiene 62.468 unigramas, 677.698 bigramas y 480010 pares de términos, todos ellos en inglés.

- SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010): El último de los recursos que se incluye dentro de esta lista, pero no menos importante, es SentiWordNet. Para describirlo, es apropiado introducir antes WordNet ©, dado que está basado en éste último. WordNet © es un listado extenso de sustantivos, verbos, adjetivos y adverbios en idioma inglés que se encuentran agrupados en conjuntos de sinónimos cognitivos denominados "Synsets", los cuáles se refieren cada uno a un concepto en particular. Estos "Synsets" se relacionan entre ellos por medio de asociaciones semánticas-conceptuales. De esta forma, conceptos con significados similares se encuentran cercanos entre ellos. Una de las principales virtudes que podría ser explotada son las relaciones de sinonimia que provee, las cuáles podrían servir para ampliar los léxicos polarizados disponibles. SentiWordNet es una extensión de WordNet © creada concretamente para trabajar en el área de Análisis de opiniones. Asigna a cada "Synsets" tres puntajes: positividad, negatividad y objetividad.

Una vez presentados los distintos recursos encontrados, se puede mencionar que existe cierta diversidad entre ellos. Algunos pueden ser considerados más sencillos, mientras que otros constituyen herramientas sumamente ricas para trabajar en el área del Análisis de opiniones. Vale la pena destacar que, si bien algunos recursos han sido creados automatizando cierta parte del proceso, la intervención humana es en la mayoría de los casos ineludible. Un aspecto interesante de estos recursos es el número de n-gramas que son tenidos en cuenta. Haciendo referencia nuevamente a los recursos listados, es importante destacar que el idioma inglés se presenta como un común denominador en la gran mayoría de ellos, lo cual se presupone como una desventaja al momento de centrarse en otras lenguas.

17. Propuesta de solución

En las secciones previas se ha realizado una definición del problema que conlleva la automatización del análisis de opiniones, junto con las diversas problemáticas subyacentes, y posteriormente una revisión de antecedentes junto con un repaso de los recursos disponibles. Si bien es factible utilizar cualquiera de las técnicas descriptas debido al contexto en el que se sitúa esta investigación resulta poco factible utilizar dichas herramientas de forma directa debido a que el idioma y localización resultan un impedimento para esto, así como su proceso de implementación no resultaría adecuado para el análisis correspondiente ya que el conjunto de API para conectar con las redes sociales no funcionaría por el contexto del idioma y su funcionalidad básica se vería disminuida repercutiendo en los resultados. De todo el listado de recursos disponibles la gran mayoría de ellos ha sido creado y funciona únicamente con el idioma inglés. A su vez, aquellos que funcionan con español han sido adaptados para trabajar con textos provenientes de alguna región de España concretamente, lo cual dificulta su utilización en otras regiones como la que se plantea en este proyecto. Es por esto por lo que en el presente trabajo para llevar a cabo una evaluación automática de los mensajes recopilados se ha optado para desarrollar un modelo siguiendo como base la estructura propuesta por Khan y otros (Khan, Bashir, & Qamar, 2014) ya que la misma dispone de los elementos asociados a la API para poder conectar adecuadamente a las redes sociales y extraer los comentarios necesarios y por otro lado, el uso de léxicos resulta independientes del idioma elegido, lo que pondera por encima del resto en la elección realizada. Antes de comenzar con la definición del

modelo de solución propuesto, se procede a definir el proceso de recolección de información realizado.

18. Fuentes de información y proceso de recolección utilizado

Para poder continuar con el presente proyecto se necesitó definir cuál sería la fuente principal de información, para ello se realizó un relevamiento para definir cuál era el medio de comunicación en la web preferida por las empresas rafaquinas delimitadas por esta investigación. De las más de 400 empresas seleccionadas, se notó que la mayoría no disponía de una red social que sirviera para obtener información referida a sus comentarios o interacciones con usuarios. Si bien este problema fue un gran inconveniente en cuanto a la obtención de la información necesaria, un total de 132 empresas disponían de fanpages de Facebook en donde realizan la interacción con diversos usuarios en su día a día, por este motivo, se eligió avanzar con la extracción de datos desde Facebook, tomando como fuente principal de información a esta red social. Una vez seleccionada la fuente, se debió indagar sobre las diversas formas de extracción de datos con las que se contaba o las herramientas disponibles en la actualidad para realizar la extracción de información necesaria. En este punto y luego de indagar sobre varios métodos de extracción y herramientas, se dió con una herramienta de código abierto denominada Facepager.

Facepager permite realizar la extracción de datos públicos disponibles en diversas redes sociales como Facebook, Twitter y otras APIs que sean basadas en JSON. Este programa utiliza una base SQLite en donde se almacena la información obtenida, aunque además este software permite exportar estos datos a archivos CSV (Comma Separated Values), lo cual nos permitió poder manipular esta información obtenida según necesitemos. De cada una de estas 132 empresas que disponían de fanpage en Facebook, se realizó la obtención con Facepager de cada una de las publicaciones de estas páginas y con ellas también se obtuvieron sus comentarios. Por cada página procesada, se exportó la información obtenida a un archivo CSV que contenía todos los comentarios que fueron realizados en dicha página y una vez generados los archivos para cada página, culminó este punto obteniendo como resultado la información que necesitábamos para realizar la evaluación automática que se propone en este proyecto.

19. Modelo propuesto

Para trabajar en el contexto en el que se ubica la presente investigación se ha optado por utilizar léxicos polarizados. Esta decisión se debe a la falta de recursos (conjuntos de datos etiquetados) que permitan entrenar modelos de aprendizaje automático, siendo ésta una de las técnicas que mejor resultados entrega de acuerdo con el proceso de revisión de antecedentes. Para llevar a cabo la clasificación de cada mensaje recopilado, se aplica un modelo que consta de 3 partes: preprocesamiento, clasificación por emoticones y clasificación por léxicos polarizados. Es necesario destacar que la estructura de la solución propuesta ha sido inspirada en el trabajo de Khan y otros (Khan, Bashir, & Qamar, 2014), donde cada mensaje a ser evaluado se somete en primera instancia a una serie de tareas que facilitan su procesamiento posterior, y luego distintos modelos son consultados de forma consecutiva hasta que alguno de ellos emita una predicción. La estructura de la solución planteada se puede visualizar en la Figura 2, donde a partir del preprocesamiento de los datos relevados en las redes sociales se procede a realizar la clasificación por los diversos emoticones utilizados, en este caso en particular las “caras sonrientes o enojadas”. En caso de que no se pudiera realizar la clasificación por emoticones se procede a la clasificación por léxicos polarizados donde vocablos que representen “me encantó” o “me gusta” se clasifican como positivos mientras que otros como “nunca responde” o “es incomible el producto” se clasifican como negativos,

continuando luego con el resto del procedimiento. A continuación se detallan las distintas partes de esta.

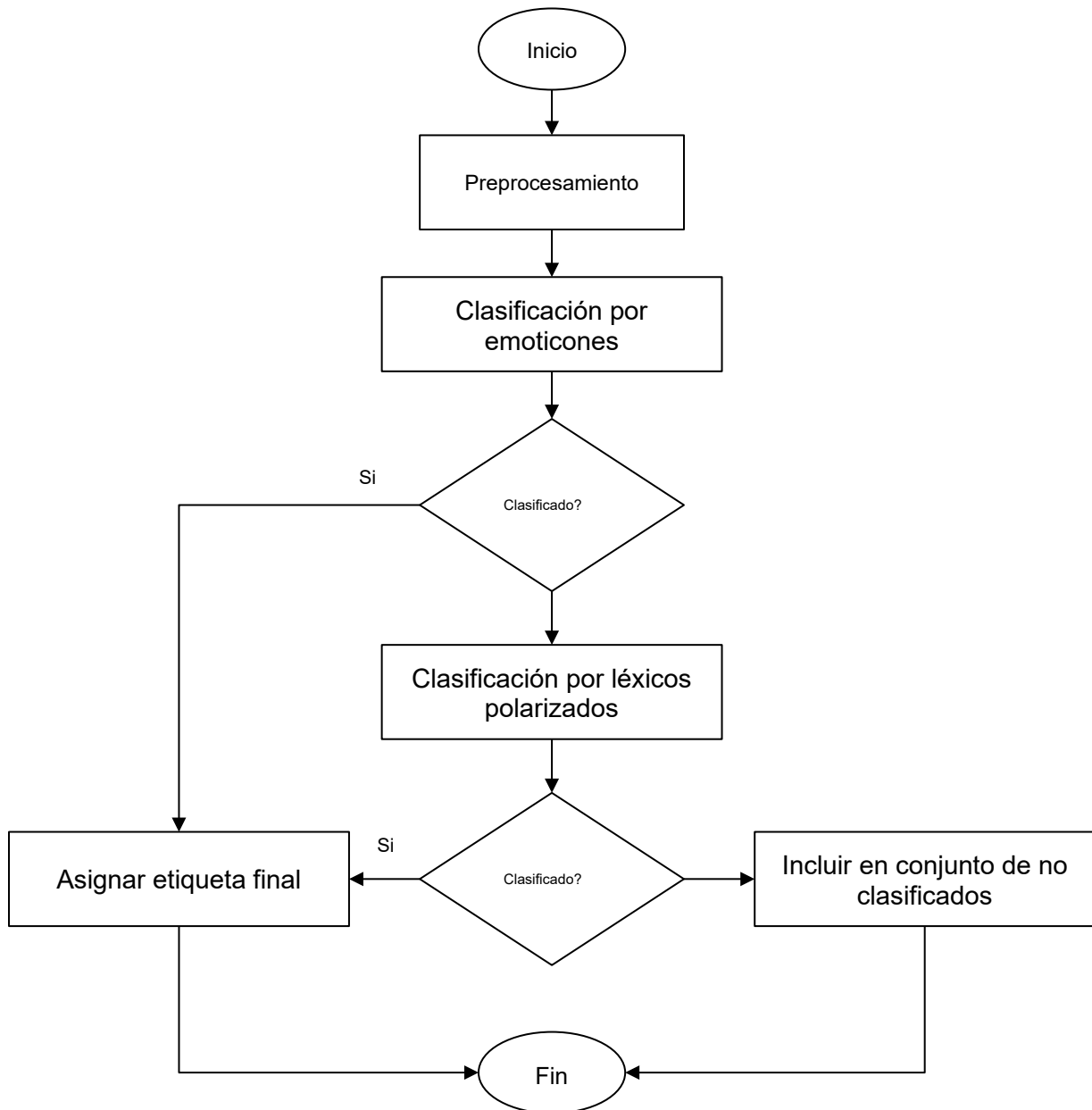


Figura 2. Diagrama de flujo del funcionamiento del esquema de solución propuesto.

20. Preprocesamiento

Debido a que la fuente de datos con la que se trata en este problema es desestructurada resulta necesario llevar a cabo un preprocesamiento que permita estandarizarla para su posterior tratamiento, así como también filtrar aquellos casos que por alguna razón deban eliminarse. Las tareas definidas concretamente para esta fase se pueden visualizar en la Figura 3 y se describen a continuación:

Eliminación de mensajes carentes de contenido: Como primer criterio de exclusión se eliminaron aquellos registros cuyo contenido era nulo o exclusivamente números. Esta última condición es no menor dado que se encontró una gran cantidad de casos en los que el contenido

del mensaje hacía referencia a 3 números decimales. Realizando un análisis más detallado de esta situación se pudo observar que se hacía referencia a mensajes para participar en sorteos organizados por la página, en la cual aquellos usuarios que quisieran participar debían ingresar los últimos 3 dígitos de su documento de identidad. Habiendo realizado este análisis se llegó a la conclusión de que dichos mensajes no tenían un significado susceptible de ser ponderado.

Eliminación por tipo de contenido: El corpus de mensajes extraído se compone de 2 tipos de contenido, publicaciones de página (aquellas que se realizan desde la misma página por aquellos usuarios creadores o autorizados) y comentarios (respuestas que esos mismos usuarios u otros dentro de la red realizaron sobre una publicación de página). Debido a que el objetivo perseguido en esta investigación trata de evaluar las interacciones de una empresa (en este caso a través de su perfil de Facebook) con otros usuarios, se consideró relevante analizar únicamente aquellos mensajes que se refieran a comentarios realizados por otros.

Filtro por idioma: Tal como se mencionó anteriormente en este trabajo, la multilingüedad es un criterio que dificulta la automatización del análisis de un determinado mensaje. En este caso, y por el contexto planteado, el estudio se enfoca en la lengua española, localizada concretamente en Argentina. Es por esto que aquellos mensajes que han sido escritos en una lengua distinta deben ser filtrados. Para llevar a cabo la detección de la lengua asociada a un mensaje se utilizó la Librería de Detección de Lenguaje desarrollada por Google, por intermedio de la interfaz proporcionada por el paquete python langdetect.

Eliminación de duplicados: En menor medida a otras situaciones comentadas anteriormente, se detectó la presencia de mensajes duplicados en el conjunto de datos recolectados. A partir de esto se produjo un filtrado de los mismos para que cada mensaje sea analizado una única vez. Vale la pena destacar que para determinar que dos mensajes son exactamente iguales el contenido de cada uno de ellos debe coincidir exactamente, sin tener en cuenta otros factores.

Tokenización: En esta fase cada mensaje se desagrega en las distintas partes que lo componen (palabras, signos de puntuación, emoticones, etc.), denominadas "tokens", dando lugar a una lista finita de elementos.

Eliminación de puntuaciones: Debido al carácter informal que presentan las redes sociales, se considera pertinente realizar una limpieza de aquellos símbolos de puntuación que podrían estar presentes o ausentes (dependiendo el caso) para aumentar la cobertura del análisis realizado. Se presupone que los beneficios obtenidos mediante esta determinación compensan aquellos casos en los que la eliminación de dichos símbolos arroje un resultado diferente al original.

21. Clasificación por emoticones

Este primer modelo intenta inferir la polaridad del mensaje haciendo uso de emoticones. Esto quiere decir que si el mensaje contiene el emoticón ":)" (o alguno incluido dentro de la lista de positivos), el mismo será clasificado como positivo. En consecuencia, si un emoticón del tipo ":(" o similar es encontrado, el mensaje es clasificado como negativo. Existen situaciones donde la clasificación no puede llevarse a cabo, las cuáles son: mensajes que no contienen emoticón alguno, o, por el contrario, mensajes que contienen emoticones positivos y negativos al mismo tiempo. En ambas situaciones se considera que la polaridad no es evidente para este modelo, y por lo tanto, se debe delegar en otro criterio. Vale la pena mencionar que aquellos mensajes que han sido clasificados en esta instancia, ya sea por la presencia de uno o más emoticones de una polaridad en concreto, no continúan el flujo principal y son excluidos de las fases siguientes, recibiendo la etiqueta que este modelo ha asignado. Con respecto a la utilización de emoticones, se debe mencionar que uno de los trabajos

pioneros en introducir esta idea para abordar el Análisis de opiniones es el de Read (Read, 2005), donde se menciona entre otras cosas que estos elementos son completamente independientes del dominio y tema sobre el que se apliquen, así como también permanecen invariantes a lo largo del tiempo en la mayoría de los casos. El listado de emoticones utilizado ha sido construido a partir de los que introducen en Wikipedia (https://en.wikipedia.org/wiki/List_of_emoticons). Si bien dicho listado es sumamente extenso, se realizó un filtrado por aquellos que se consideraron más apropiados para este estudio. A su vez, esta misma fuente ha sido utilizada también en el estudio de Ortigosa y otros (2014). Tal como se describe en la figura 2 con un ejemplo, es el mismo a utilizar en la figura 3 focalizando en este caso particular en los emoticones. En total, se consideraron 26 emoticones positivos y 19 negativos.

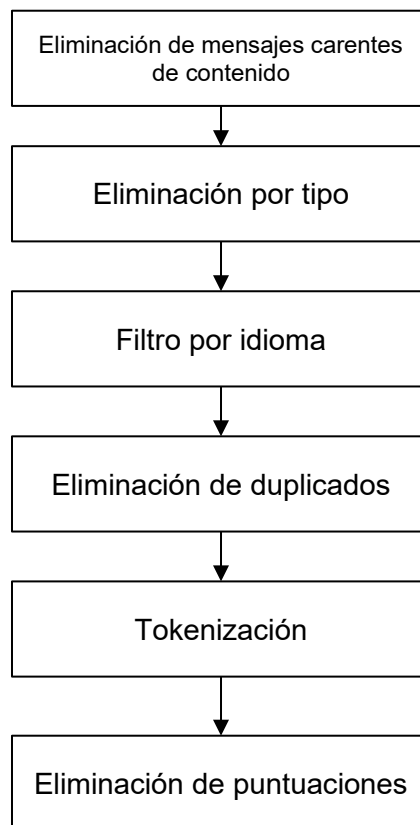


Figura 3. Fases que componen el preprocesamiento de cada mensaje.

22. Clasificación por léxicos polarizados

La clasificación realizada por este modelo, tal como se describió anteriormente en este trabajo, se basa en la utilización de un conjunto de unigramas anotados con una polaridad determinada (positiva o negativa). Los léxicos de base que se han utilizado han sido adaptados a partir de los disponibles en la herramienta SOCAL (<https://github.com/sfu-discourse-lab/SO-CAL>) y los generados por Pérez-Rosas y otros (Pérez-Rosas, 2012). A partir de la utilización de estos recursos, para clasificar un mensaje se realiza un recuento del número de unigramas que aparecen dentro de cada polaridad en el texto y se etiqueta como positivo en caso de que exista mayoría de dichos unigramas, o como negativo en caso contrario. En cualquier otra situación se considera nuevamente que el modelo no es apto para realizar un juicio sobre dicho

caso y se delega la clasificación. Vale la pena destacar que este modelo se compone por dos iteraciones aplicando la misma lógica desarrollada anteriormente, con la diferencia de que en la primera de ellas no se realiza un stemming de los tokens contenidos en el mensaje ni en los léxicos polarizados, mientras que en la segunda iteración sí se lleva a cabo esa tarea. La decisión de implementarlo de esta manera responde a que, si bien la utilización de la técnica de stemming permite incrementar la cobertura del análisis realizado, también da lugar a la aparición de ciertos errores. Para llevar a cabo esta tarea de stemming en la segunda iteración de este modelo se ha optado por la implementación del algoritmo Snowball Stemmer (https://www.nltk.org/_modules/nltk/stem/snowball.html) disponible en la librería NLTK.

23. Resultados obtenidos

Del total de 324.638 mensajes obtenidos a partir de las páginas seleccionadas, detallado en la sección “Fuentes de información y proceso de recolección utilizado” del presente documento, luego de aplicar los filtros y tareas de preprocesado comentadas en la sección previa el número final descendió a 61.702. Las tareas de clasificación aplicadas dieron lugar a los valores descritos en la Tabla 5.

Modelo	Corpus de entrada	Casos etiquetados
Basado en emoticones	61.702	1.621
Basado en léxicos polarizados	60.081	24.240
Basado en léxicos polarizados + stemming	35.841	17.171

Tabla 5. Cantidad de casos analizados por cada modelo.

El proceso de validación de los resultados se realizó sobre una muestra de 200 mensajes obtenidos de forma aleatoria. Dicha muestra fue etiquetada de forma manual y completamente independiente por tres revisores distintos, donde cada uno de ellos solo disponía del mensaje en cuestión (sin ningún tipo de contexto o información extra). Esto se dispuso de esta manera puesto que esa misma limitación se aplica para los modelos utilizados. Al igual que el proceso de clasificación automático, los revisores debían indicar un único valor posible indicando la polaridad del mensaje de acuerdo con su interpretación del mismo, siendo los únicos valores posibles -1 (negativa), 0 (neutra) y 1 (positiva). Realizando una primer comparativa entre los resultados del proceso de etiquetado manual se observa que existen diferencias en las apreciaciones realizadas por los revisores. La Tabla 6 introduce el nivel de concordancia entre los distintos revisores.

	Revisor 1	Revisor 2	Revisor 3
Revisor 1	1		
Revisor 2	0.79	1	
Revisor 3	0.87	0.79	1

Total	0.725
--------------	-------

Tabla 6. Comparativa de resultados obtenidos en la generación manual de etiquetas.

Para llevar a cabo una unificación de estos criterios y poder obtener una única etiqueta por cada caso contra la cual comparar los resultados obtenidos en el proceso automático se ha determinado utilizar la técnica de voto por mayoría, es decir, que aquella etiqueta que recibe el mayor número de votos entre los revisores es la que se utiliza. En caso de que más de una etiqueta reciba la misma cantidad de votos entonces se determina que dicho mensaje no es debe ser contrastado. La Tabla 7 introduce el resultado al aplicar el proceso de unificación de etiquetas y la Tabla 8 indica el balanceo de cada clase.

Número de votos para la clase con mayoría	Cantidad de casos
1	0
2	55
3	145

Tabla 7. Detalle del proceso de unificación de etiquetas manuales.

Clase	Cantidad de casos
-1	4
0	121
1	75

Tabla 8. Cantidad de casos por cada etiqueta luego del proceso de unificación de etiquetas manuales.

		etiquetas automáticas		
		-1	0	1
etiquetas manuales	-1	2	0	2
	0	22	50	49
	1	10	9	56

Tabla 9. Matriz de confusión para los resultados entre las etiquetas creadas automática y manualmente.

Una vez obtenidas las etiquetas finales se realizó una comparación con las salidas del proceso automático. La Tabla 9 introduce los resultados obtenidos.

A su vez, si se calcula la métrica de accuracy (porcentaje de aciertos sobre el porcentaje total de casos), la misma arroja un valor general de 0.54. A partir de esto, se logra una estimación del proceso de etiquetado automático realizado por el modelo propuesto. La Tabla 10 introduce el detalle completo del balanceo generado por este proceso.

Clase	Cantidad de casos
-1	9.609
0	18.670
1	33.423

Tabla 10. Cantidad de casos por cada etiqueta luego del proceso de etiquetado automático.

Como resultado final, se introduce en la Tabla 11 el detalle de resultados obtenidos para las 20 páginas con mayor cantidad de mensajes, detallando cantidad de mensajes en el período analizado y la proporción dentro de cada clase.

Ranking	Página	Cantidad de mensajes	Porcentaje negativos	Porcentaje neutros	Porcentaje positivos
1	Lario	25.697	0.20	0.287	0.513
2	La Espiga de Oro	14.478	0.094	0.346	0.55
3	Agofa Chocolates	7.730	0.108	0.279	0.612
4	RR Escapes	3.322	0.269	0.356	0.373
5	Noemí Vallone Alta Moda	936	0.111	0.221	0.667
6	Analucia Dulces	835	0.101	0.136	0.761
7	Madame Pasteliere	736	0.183	0.298	0.517
8	Bici Peretti	716	0.11	0.248	0.641
9	Artecoc	698	0.11	0.21	0.679
10	Super Family	614	0.084	0.436	0.478
11	Spaggiari Alta Costura	565	0.092	0.184	0.723
12	IMG Rafaela	488	0.135	0.26	0.604
13	Panificación San	476	0.14	0.388	0.47

	Rafael				
14	Carnicería MUMU	473	0.105	0.397	0.496
15	La Flor del Lapacho	447	0.17	0.237	0.592
16	Granja el Ceibo	390	0.089	0.1	0.81
17	Sigma Gráfica y Papelería	302	0.102	0.211	0.685
18	Imprenta y Diario La Opinión	260	0.1307	0.323	0.546
19	Rotisería Rafaela	246	0.052	0.402	0.544
20	Adorni Muebles	244	0.102	0.43	0.467

Tabla 11. Resultados obtenidos del modelo para cada una de las 20 páginas con mayor cantidad de mensajes.

24. Agradecimientos

Este artículo fue desarrollado bajo el proyecto de investigación “Caracterización de la reputación corporativa de la industria de Rafaela mediante el análisis de la interacción de sus usuarios en redes sociales y la calidad de sus sitios web” y financiado por la Universidad Católica de Santiago del Estero.

25. Conclusiones

El presente estudio constituye un caso de aplicación de técnicas de análisis de opiniones a una problemática local de la ciudad de Rafaela. Como primera conclusión luego de haber realizado una revisión de antecedentes y habiendo obtenido resultados experimentales a partir del proceso introducido en este trabajo se puede apreciar el faltante de recursos abiertos y de fácil acceso para trabajar con esta problemática en el contexto de nuestro país. Tal como se mencionó previamente, la gran mayoría de recursos existentes por la comunidad científica han sido desarrollados para trabajar con el idioma inglés, y aquellos que funcionan con español están preparados para trabajar con mensajes provenientes por hispanohablantes no americanos.

Los resultados presentados en la Tabla 9 de la fase de clasificación hacen pensar que existen múltiples mejoras a ser aplicadas en el modelo propuesto, como trabajar con léxicos mejor adaptados y extensos, modelar los cambios de polaridad (negaciones, amplificadores, etc.) o construir un corpus de mensajes etiquetados que permita utilizar técnicas de aprendizaje automático, aproximación que de acuerdo a lo mostrado en la fase de revisión de antecedentes ofrece mejores resultados que otras técnicas pero que por faltante de un corpus anotado no se pudo aplicar en este proyecto.

Tal como se pudo apreciar, la construcción de una herramienta que permita analizar automáticamente la totalidad de mensajes disponibles acarrea un número de imprecisiones debido a que resulta una simplificación de la problemática expuesta a lo largo de este trabajo.

No obstante, es importante destacar que mediante esta herramienta se ha podido llevar a cabo un análisis que de otra manera hubiera resultado muy difícil de afrontar con los recursos disponibles en las redes sociales ya que las empresas actuales se mostraron muy renuentes a compartir su información interna para mejorar la propuesta presentada.

A partir de los resultados presentados en la Tabla 11 se concluye que no existe una correlación entre el número de mensajes o interacciones y una opinión positiva con respecto a otros usuarios en las redes. A su vez, es importante mencionar que todas las páginas analizadas presentan una mayoría de mensajes positivos, lo cual hace pensar que la reputación de las empresas en este medio es buena.

26. Referencias

Alburquerque, F. (11 de 07 de 2006). Clusters, territorio y desarrollo empresarial: Diferentes modelos de organización productiva. Recuperado el 09 de 07 de 2019, de <https://publications.iadb.org/es/publicacion/14444/clusters-territorio-y-desarrollo-empresarial-diferentes-modelos-de-organizacion>

Aleskerov, E., Freisleben, B., & Rao, B. (1997). Cardwatch: a neural network based database mining system for credit card fraud detection. *Computational Intelligence for Financial Engineering (CIFEr)*, 220–226.

Alpaydin, E. (2009). *Introduction to Machine Learning*. The MIT Press.

Ascúa, R. (2007). *Rafaela 125 años desorientando a Adam Smith. Evolución de su Rafaela: Comisión de Industrias del Centro Comercial e Industrial de Rafaela*.

Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: a case study. *International Conference on Recent Advances in Natural Language Processing*. Borovets, BG.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Language Resources and Evaluation Conference (LREC 2010)*, 7, 2200–2204.

Balahur, A., Hermida, J. M., Montoyo, A., & Muñoz, R. (2011). EmotiNet: A Knowledge Base for Emotion Detection in Text Built on the Appraisal Theories. *Natural Language Processing and Information Systems: 16th International Conference on Applications of Natural Language to Information Systems*, 27–39.

Basari, A. S., Hussin, B., Ananta, I. G., & Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. *Procedia Engineering*(53), 453 – 462.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*(41), 16:1–16:52.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 2(24), 123–140.

Chapelle, O., Haffner, P., & Vapnik, V. N. (1999). Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5), 1055–1064.

Chen, B.-J., Chang, M.-W., & lin, C.-J. (2004). Load forecasting using support vector machines: a study on eunite competition. *IEEE Transactions on Power Systems*, 19(4), 1821–1830.

Comisión Económica para América Latina y el Caribe. (1998). Entorno Institucional y Desarrollo. La articulación y las interacciones entre instituciones de soporte técnico al desarrollo empresarial. El caso de Rafaela y su Región. Ambiente.

da Silva, N. F., Hruschka, E. R., & Hruschka, E. R. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*(66), 170 – 179.

Dang, Y., Zhang, Y., & Chen, H. (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 4(25), 46-53.

Dasgupta, S., & Ng, V. (2009). Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification. *Proceedings of the Joint Conference of the 47th Annual REFERENCIAS 101 Meeting of the ACL*, 2(4), 701–709.

Fersini, E., Messina, E., & Pozzi, F. (2014). Sentiment analysis: Bayesian ensemble learning. *Decision Support Systems*(68), 26 – 38.

Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. *Information Processing & Management*(30), 9-19.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *ICML*(96), 148–156.

Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics*, 15(1), 107-143.

Goldberg, A. B., & Zhu, X. (2006). Seeing stars when there aren't many stars: Graphbased semi-supervised learning for sentiment categorization. *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, 45-52.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*(5), 199–220.

Güvenir, H. A., & Cicekli, I. (1998). Learning translation templates from examples. *Information Systems*, 23(6), 353 – 363.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, ACL*, 174–181.

Hootsuite and We are social. (01 de 2019). *Data Reportal*. Recuperado el 09 de 07 de 2019, de <https://datareportal.com/reports/digital-2019-global-digital-overview>

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*(10), 168–177.

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM*, 8, 1-4.

Ikeda, D., Takamura, H., arie Ratinov, L., & Okumura, M. (2008). Learning to shift the polarity of words for sentiment classification. *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)*.

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent twitter sentiment classification. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1, 151-160.

Katz, G., Ofek, N., & Shapira, B. (2015). Consent: Context-based sentiment analysis. *Knowledge-Based Systems*(84), 162 – 178.

Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*(22), 110–125.

Khan, F. H., Bashir, S., & Qamar, U. (2014). Tom: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, 57, 245–257.

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*(50), 723–762.

Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*, 10(40), 4065 – 4074.

Lapponi, E., Read, J., & Ovrelid, L. (2012). Representing and resolving negation for sentiment analysis. *12th International Conference on Data Mining Workshops (ICDMW)*.

Li, G., & Liu, F. (2010). A clustering-based approach on sentiment analysis. *International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, 331-337.

Li, G., & Liu, F. (2010). A clustering-based approach on sentiment analysis. *International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, 331-337.

Li, S., Wang, Z., Lee, S. Y., & Huang, C. R. (2013). Sentiment classification with polarity shifting detection. *International Conference on Asian Language Processing (IALP)*, (págs. 129-132).

Li, S., Wang, Z., Lee, S. Y., & Huang, C. R. (2013). Sentiment classification with polarity shifting detection. *International Conference on Asian Language Processing*.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*(66), 35–65.

Madnick, S. E., Wang, R. Y., Lee, Y. W., & Zhu, H. (2009). Overview and framework for data and information quality research. *Journal of Data and Information Quality*, 2(1), 1-22.

Mihalcea, R., Banea, C., & Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. *En Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (págs. 976–983). Prague, Czech Republic.

Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 26–34.

Moraes, R., Valiati, J. F., & Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, 2(40), 621-633.

Mudinas, A., Zhang, D., & Levene, M. (2012). Combining lexicon and learning based. *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM '12*, 5(1), 5:1–5:8.

Narayanan, V., Arora, I., & Bhatia, A. (2013). *Intelligent Data Engineering and Automated Learning. IDEAL*. Hefei, China.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. *Proceedings of the 2nd International Conference on Knowledge Capture*, 2, 70-77.

Orimaye, S. O., Alhashmi, S. M., & Siew, E.-G. (2012). Buy it - don't buy it: Sentiment classification on amazon reviews using sentence polarity shift. *En Lecture Notes in Computer Science* (págs. 386–399). Springer.

Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*(31), 527 – 541.

Ortigosa-Hernández, J. R., Alzate, L., Lucania, M., Inza, I., & Lozano, J. A. (2012). Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers. *Neurocomputing*(92), 98 – 115.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2, 1-135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, 10(2), 79-86.

Pérez-Rosas, V. (2012). Learning sentiment lexicons in Spanish. *International Conference on Language Resources and Evaluation (LREC-2012)*. Istanbul, Turkey.

Polanyi, L., & Zaenen, A. (2006). Contextual Valence Shifters. *En Computing Attitude and Affect in Text: Theory and Applications* (págs. 1-10). Netherlands, Dordrecht.: Springer.

Poria, S., Cambria, E., Winterstein, G., & Huang, G.-B. (2014). Sentic patterns: Dependencybased rules for concept-level sentiment analysis. *Knowledge-Based Systems*(69), 45-63.

Porter, M. F. (1997). *Readings in information retrieval. An Algorithm for Suffix Stripping*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Rafaela, M. d. (2012). Censo Industrial Rafaela. Recuperado el 09 de 07 de 2019, de <https://www.rafaela.gob.ar/GobiernoAbierto/BibliotecaVirtual-detalle.aspx?c=17&t=vrAbdCzOGSx4im9FhAx43Q&f=18&index=&txt=&i=2148&v=lista>

Rafaela, M. d. (2016). Censo Industrial Rafaela. Recuperado el 09 de 07 de 2019, de https://www.rafaela.gob.ar/nuevo/Files/Archivos/arc_93.pdf

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *Proceedings of the ACL Student Research Workshop*, 43–48.

Russell, S. J., & Norvig, P. (2009). *Artificial intelligence: a modern approach* (3rd edition). Prentice Hall.

Sajda, P. (2006). Machine learning for detection and diagnosis of disease. *Annual Review of Biomedical Engineering*(8), 537–565.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*(61), 85-117.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47.

Sindhwani, V., & Melville, P. (2008). Document-word co-regularization for semi-supervised sentiment analysis. *Eighth IEEE International Conference on DataMining, 2008. ICDM*(8), 1025–1030.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, 1631–1642.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1996). *The General Inquirer: A Computer Approach to Content Analysis*. MIT press.

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

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 1(29), 24-54.

Tripathy, A., Agrawal, A., & Rath, S. K. (2015). Classification of sentimental reviews using machine learning techniques. *Procedia Computer Science*(57), 821 – 829.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. En *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (págs. 417–424). Stroudsburg, PA, USA: Association for Computational Linguistics.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 347–354.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 347–354.

Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 3(36), 6527–6535.

Zhang, X., & LeCun, Y. (s.f.). Text understanding from scratch. CoRR.

Zhao, L., & Li, C. (2009). Ontology Based Opinion Mining for Movie Reviews. *Knowledge Science, Engineering and Management: Third International Conference, KSEM 2009*. Vienna, Austria.

Zhou, Z.-H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 1(137), 239 – 263.