Modelado e implementación de algoritmos inteligentes de análisis de opinión

Tesista

Ing. Juan Pablo TESSORE

Directora

Sandra BALDASSARRI (UNIZAR)

Co director

Hugo RAMÓN (UNNOBA)

TESIS PRESENTADA
PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS INFORMÁTICAS

FACULTAD DE INFORMÁTICA

UNIVERSIDAD NACIONAL DE LA PLATA

Septiembre 2023

Dedico este trabajo a mi familia y amigos, quienes me acompañaron incondicionalmente a lo largo de estos años.

Agradecimientos

Quiero dedicar unas palabras de agradecimiento a quienes colaboraron en el desarrollo de este trabajo.

En primer lugar, a mis directores Sandra Baldassarri y Hugo Ramón, por su inmensa ayuda para llevar adelante esta tesis.

A los investigadores, becarios y colaboradores del Instituto de Investigación y Transferencia en Tecnología de la Universidad Nacional del Noroeste de la Provincia de Buenos Aires, por el enriquecedor intercambio de ideas generado en el trabajo diario.

A la Escuela de Tecnología, el Instituto de Posgrado y la Secretaría de Investigación de la Universidad Nacional del Noroeste de la Provincia de Buenos Aires, por brindar el asesoramiento y apoyo económico para tomar cursos, realizar publicaciones y asistir a congresos.

A la Facultad de Informática de la Universidad Nacional de La Plata por darme la oportunidad de continuar mi formación y por crear un entorno de excelencia que fomenta el intercambio de ideas con profesionales tanto a nivel nacional como internacional.

Al Grupo de Investigación en Interfaces Avanzadas perteneciente a la Universidad de Zaragoza por su apoyo técnico para el desarrollo de este trabajo.

Al Colegio de Psicólogos de la Provincia de Buenos Aires Distrito III, por su colaboración para la realización de parte del trabajo experimental.

Por último y fundamentalmente a mi familia y amigos, quienes estuvieron en todo momento brindándome la fuerza necesaria para recorrer este arduo camino.

A todos les estaré eternamente agradecido.

Resumen

A la par de la amplia adopción que han tenido las redes sociales, ha crecido también la generación contenidos en ellas, en particular en forma de texto. La proliferación de este tipo de contenido ha creado la materia prima necesaria para aplicar técnicas de minería de textos a esos datos con el objetivo de extraer información valiosa.

Numerosos trabajos que intentan categorizar, mediante clasificadores basados en aprendizaje automático, textos provenientes de redes sociales, dependen del etiquetado manual del contenido o de la utilización de *datasets* públicos previamente etiquetados. Dichos abordajes presentan sus inconvenientes, uno de ellos es el tiempo que demanda la clasificación manual de los datos de entrenamiento. Otro problema es que los clasificadores suelen construirse utilizando datos de distinto origen a los que finalmente analizan, esto plantea un desafío debido a que, si el clasificador no fue expuesto, durante la etapa de entrenamiento, a datos similares a los que finalmente debe categorizar, difícilmente pueda hacerlo de manera adecuada. Por otro lado, la cantidad de recursos disponibles (tales como *datasets* etiquetados, corpus o diccionarios afectivos) no es abundante para idiomas distintos del inglés, limitando las posibilidades de construcción de los mencionados clasificadores de texto para otros idiomas, entre ellos el español.

La tarea de recopilación y validación de recursos en el idioma a utilizar se vuelve, en consecuencia, una necesidad para construir clasificadores de texto, basados en aprendizaje automático supervisado. Sin embargo, dichas tareas son extremadamente demandantes en tiempo y recursos humanos. Esta problemática se agrava para los casos en los que el criterio de clasificación no es objetivo, como por ejemplo para la clasificación de emociones en texto. En estas situaciones, se requiere que varios jueces clasifiquen el mismo contenido, de manera de poder validar la veracidad de la etiqueta asignada al mismo.

Con el objetivo de agilizar el desarrollo de clasificadores de emociones en texto para el idioma español basados en aprendizaje automático supervisado, resulta necesario reducir o eliminar la necesidad del etiquetado manual de los *datasets* utilizados para entrenamiento. En esta tesis, a diferencia de otros estudios, las etiquetas que denotan la emoción de cada comentario se obtienen automáticamente de los mismos usuarios que escriben el contenido, en

lugar de clasificarlos de manera manual. Posteriormente, se define un procedimiento para realizar la validación de las etiquetas recopiladas, el cual requiere del etiquetado y validación manual de sólo una pequeña muestra de las mismas y posterior cálculo de métricas para establecer el nivel de consenso. A su vez, durante el proceso de captura de los documentos, se obtiene también información contextual relacionada con los mismos, con el objetivo de utilizarla para medir los cambios, ya sean mejoras o no, en el desempeño de distintos clasificadores basados en aprendizaje automático.

El proceso que se presenta en esta tesis, permite agilizar la construcción de clasificadores de emociones en texto basados en aprendizaje automático y a su vez mejorar su desempeño mediante el uso de información contextual. Estos clasificadores pueden ser utilizados para ofrecer una amplia variedad de propósitos potenciales, como detectar la emoción que surge de la opinión de grandes grupos de personas sobre ciertos productos, servicios o incluso políticas públicas. También podrían utilizarse para identificar demandas o quejas no satisfechas de ciudadanos; o, en seguridad, para la detección automática de factores de riesgo en redes sociales, como amenazas, hostigamiento o acoso.

Los clasificadores construidos a partir del proceso mencionado, alcanzan un desempeño similar al de otros entrenados con *datasets* etiquetados manualmente. Debe resaltarse que, en el trabajo presentado, la necesidad de etiquetado manual en el proceso de recolección y clasificación se reduce significativamente.

El conjunto de datos creado puede ser utilizado en diversas investigaciones que realicen Análisis de Sentimientos en español. Además, el proceso de recopilación y validación presentado en esta tesis puede adaptarse fácilmente para generar nuevos *datasets* en temas o idiomas específicos.

Palabras clave: Análisis de Sentimientos, Procesamiento del Lenguaje Natural, Supervisión Distante, Aprendizaje Automático, Información Contextual.

Abstract

Alongside the widespread adoption of social media, the generation of content on these platforms, particularly in text, has also grown. The proliferation of this type of content has provided the necessary raw material to apply text-mining techniques to extract valuable information from the data.

Numerous studies attempting to categorize texts from social media using machine learning classifiers rely on manual content labeling or using pre-labeled public datasets. These approaches have their drawbacks, including the time-consuming process of manually classifying the training data. Another problem is that classifiers are often built using data from different sources than those they analyze. This poses a challenge because if the classifier hasn't been exposed to similar data during the training phase, it will have difficulty categorizing it correctly. Additionally, the availability of resources such as labeled datasets, corpora, or affective dictionaries is limited for languages other than English, restricting the possibilities of constructing aforementioned text classifiers for other languages, including Spanish.

As a result, the collection and validation of resources in the target language become necessary for building supervised machine learning-based text classifiers. However, these tasks are extremely time-consuming and resource-intensive. This problem is exacerbated in cases where the classification criterion is not objective, such as emotion classification in text. In these situations, multiple judges are required to classify the same content to validate the accuracy of the assigned label.

To expedite the development of supervised machine learning-based emotion classifiers for the Spanish language, reducing or eliminating the need for manual labeling of the datasets used for training is necessary. In this thesis, unlike other studies, the labels denoting the emotion of each comment are automatically obtained from the users who write the content rather than manually classifying them. Subsequently, a procedure is defined to validate the collected labels, which only requires manual labeling and validation of a small sample of them, followed by the calculation of metrics to establish the level of consensus. Furthermore, during the document collection process, contextual information related to the documents is also obtained

and used to measure the changes, whether improvements or not, in the performance of different machine learning-based classifiers.

The process presented in this thesis allows for streamlining the construction of text-based emotion classifiers using machine learning and enhancing their performance using contextual information. These classifiers can be used for a wide variety of potential purposes, such as detecting the sentiment arising from the opinions of large groups of people about specific products, services, or even public policies. They could also be used to identify unmet demands or complaints from citizens or, in security, to automatically detect risk factors in social networks, such as threats, harassment, or bullying.

The classifiers built using the mentioned process perform similarly to others trained with manually labeled datasets. It should be emphasized that in the presented work, the need for manual labeling in the collection and classification process is significantly reduced.

The constructed dataset can be used for various research purposes involving Sentiment Analysis in Spanish. Furthermore, the collection and validation process presented in this thesis can be easily adapted to generate new resources for specific domains or languages.

Keywords: Sentiment Analysis, Natural Language Processing, Distant Supervision, Machine Learning, Contextual Information.

Contenido

i) Lista de tablas

ii) Lista de Figuras

iii) Glosario

- **ANN:** Redes Neuronales Artificiales (Artificial Neural Networks)
- **BoW:** Bolsa de palabras (Bag of Words)
- **DS:** Supervisión Distante (Distant Supervision)
- **GloVe:** Vectores globales (Global Vectors)
- **HTML:** Lenguaje de marcado de hipertexto (HyperText Markup Language)
- **IC:** Información contextual
- **INEGI:** Instituto Nacional de Estadística y Geografía
- **IR:** Recuperación de Información (Information Retrieval)
- **KDD:** Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases)
- **LSA:** Análisis Semántico Latente (Latent Semantic Analysis)
- **LSTM:** Memoria a Largo Plazo de Corto Plazo (Long Short-Term Memory)
- **ME:** Entropía Máxima (Maximum Entropy)
- **ML:** Aprendizaje Automático (Machine Learning)
- **NB:** Naive Bayes
- **NLTK:** Kit de Herramientas de Lenguaje Natural (Natural Language Toolkit)
- **NLP:** Procesamiento del Lenguaje Natural (Natural Language Processing)
- **OOV:** Fuera del Vocabulario (Out of Vocabulary)
- **PAD:** Placer - Activación - Dominancia (Pleasure - Arousal - Dominance)
- **POS tagging:** Etiquetado de Partes del discurso (Part-of-Speech Tagging)
- **RNN:** Redes Neuronales Recurrentes (Recurrent Neural Networks)
- **SVM:** Máquinas de Vectores de Soporte (Support Vector Machines)

- **TASS:** Taller de Análisis de Sentimientos en español
- **TF:** Frecuencia de Término (Term Frequency)
- **TF-IDF:** Frecuencia de Término - Inversa de Frecuencia de Documento (Term Frequency-Inverse Document Frequency)
- **URL:** Localizador Uniforme de Recursos (Uniform Resource Locator)
- **WSD:** Desambiguación del Sentido de las Palabras (Word Sense Disambiguation)

1. Introducción

Contenido

Resumen

La gran cantidad de datos generados a partir la interacción de las personas en las redes sociales y la convicción de que los mismos pueden aportar información valiosa, ha motorizado el desarrollo de la Computación Afectiva y el Análisis de Sentimientos, como áreas de investigación emergentes que tienen como objetivo crear sistemas capaces de reconocer, interpretar, procesar y simular emociones humanas. Sin embargo, cuando se trata de analizar el contenido en formato de texto, estas áreas están estrechamente relacionadas con el procesamiento del lenguaje natural, un subcampo de las Ciencias de la Computación que se dedica a analizar grandes cantidades de datos en lenguaje natural. Aunque cada vez son más frecuentes las investigaciones que utilizan publicaciones en foros, sitios de microblogging y redes sociales para analizar automáticamente el contenido generado por las personas, la mayoría de estos trabajos están enfocados en el idioma inglés, lo que ha generado una abundancia de recursos para ese idioma, pero que no se ha replicado en otros. Tal es el caso del idioma español, que, aunque es el tercer idioma más utilizado en Internet después del inglés y el chino, los recursos de Análisis de Sentimientos específicos para este idioma son limitados. Tal escasez se debe en parte a que la construcción de dichos recursos es una tarea muy demandante en tiempo y recursos humanos, lo que hace necesario buscar alternativas que

vuelvan al proceso más expeditivo. En el presente trabajo de tesis se busca dar respuesta a dicha problemática y su relación con el entrenamiento de clasificadores basados en aprendizaje automático. Para ello, en el presente capítulo, en la sección 1.1 se plantea con mayor nivel de detalle el problema de la escasez de recursos para Análisis de Sentimientos y cuáles son los inconvenientes de las estrategias adoptadas hasta el momento para solucionarlo, en la sección 1.2 se presentan los objetivos generales y específicos, en las secciones 1.3 y 1.4 se detallan respectivamente las contribuciones y publicaciones que son fruto de esta investigación, mientras que en la sección 1.5, titulada Organización de la tesis, se describe brevemente lo tratado por cada capítulo del documento.

1.1 Planteo del problema

Hoy en día, gran parte de la interacción humana se da a través de redes sociales. Éstas permiten inmediatez en las comunicaciones, y son generalmente utilizadas como un medio para que las personas expresen sus opiniones acerca de una gran variedad de temas, participando en discusiones a las cuales antes no tenían acceso. Dichas opiniones e interacciones, si se las evalúa en conjunto, pueden aportar información valiosa acerca de determinados eventos, productos y servicios. Es importante tener en cuenta que las emociones juegan un papel fundamental en la formación de opiniones y que, por lo tanto, el Análisis de Sentimientos y la Computación Afectiva son herramientas clave para comprender mejor las interacciones en las redes sociales y la toma de decisiones en distintos ámbitos.

En consecuencia, el interés que despertó el análisis automático de las opiniones, llevó a las áreas de investigación emergentes llamadas Computación Afectiva y Análisis de Sentimientos que aprovechan la interacción humano-computadora, la recuperación de información y el procesamiento de señales multimodales para intentar interpretar los sentimientos de las personas a partir de la cantidad cada vez mayor de datos presentes de redes sociales (Cambria, 2016). La Computación Afectiva (Picard, 1997) es un campo de la computación cognitiva y la inteligencia artificial cuyo objetivo es desarrollar sistemas capaces de reconocer, interpretar, procesar y simular emociones humanas. Por otro lado, el Análisis de Sentimientos es un campo de estudio que analiza las opiniones, sentimientos, apreciaciones, actitudes y emociones de la gente con respecto a distintas entidades como productos, servicios o incluso otras personas (B. Liu, 2012), según (Serrano-Guerrero, Olivas, Romero, & Herrera-Viedma, 2015) es uno de los temas de investigación más recientes en el campo del Procesamiento de la Información.

Cuando se analiza contenido en formato de texto, las áreas mencionadas están íntimamente relacionadas al procesamiento del lenguaje natural, el cual es un subcampo de las Ciencias de la Computación dedicado a procesar y analizar grandes cantidades de datos en lenguaje natural (Villa Monte, 2019). Según (Cambria, Poria, Gelbukh, & Thelwall, 2017) el Análisis de Sentimientos es un *suitcase problem*, debido a que requiere abordar varias problemáticas de investigación que conciernen al procesamiento del lenguaje natural. Cada vez con mayor frecuencia (Cambria, 2016), se están llevado a cabo investigaciones que utilizan como entrada publicaciones en foros, sitios de *microblogging* y redes sociales, con el objetivo de analizar de

manera automática el contenido generado por las personas (Ravi & Ravi, 2015). Sin embargo, la mayoría de estos trabajos están enfocados en el idioma inglés, lo que ha generado una abundancia de recursos tales como *datasets*, diccionarios afectivos, mapas semánticos y *embeddings* pre entrenados para ese idioma, a pesar de que el 48% de los recursos en Internet están escritos en otros idiomas (Justo, Alcaide, Torres, & Walker, 2018). Según (Acheampong, Wenyu, & Nunoo-Mensah, 2020) existen pocos recursos etiquetados con emociones para idiomas que no sean el inglés. La disponibilidad de estos recursos para el francés, español, hindi, etc., puede en gran medida alentar la investigación y equilibrar el trabajo hecho en los diferentes idiomas.

La situación para el idioma español no es la excepción a esta regla, (Osorio Angel, Peña Pérez Negrón, & Espinoza-Valdez, 2021) remarca que, aunque este idioma es el tercero más utilizado en Internet, después del inglés y el chino, los recursos de Análisis de Sentimientos enfocados al mismo son escasos. También resalta que se han realizado numerosos intentos por traducir recursos de otros idiomas, pero considera que el mejor camino para el mejoramiento del Análisis de Sentimientos en un idioma puntual es la creación de recursos específicos para el mismo (Brooke, Tofiloski, & Taboada, 2009), debido a que cada idioma tiene un nivel distinto de expresividad en lo que concierne a sentimientos.

Sin embargo, la recopilación de tales recursos es una tarea muy demandante en tiempo y recursos humanos (Osorio Angel et al., 2021), lo que hace necesario buscar alternativas que vuelvan al proceso de construcción de recursos de Análisis de Sentimientos en español, así como también en otros idiomas, más expeditivo. Una posibilidad para acelerar el proceso de construcción de tales recursos es hacer uso de la supervisión distante, en este modelo se vincula al contenido a clasificar una etiqueta ruidosa ya existente de manera de construir un juego de datos de manera automática.

Según (Roth, Barth, Wiegand, & Klakow, 2013), la supervisión distante permite crear grandes cantidades de datos de entrenamiento a bajo costo. Sin embargo, debido a que estos son intrínsecamente ruidosos, la calidad de los mismos es un aspecto a considerar. Si bien muchas investigaciones hacen uso de la supervisión distante para la construcción de recursos de Análisis de Sentimientos, no abundan las que tienen algún apartado de validación de la calidad de los mismos. Este aspecto es clave más aún teniendo en cuenta que las emociones

interpretadas de un texto pueden ser subjetivas. En consecuencia, se vuelve necesario, mediante la revisión y comparación de los trabajos presentes en la bibliografía junto con la experimentación propia de este trabajo de tesis, la elaboración de procesos de validación de calidad de juegos de datos de Análisis de Sentimientos recopilados mediante supervisión distante. Un aspecto importante a considerar con respecto a la calidad de los conjuntos de datos es el nivel de consenso que existe entre las etiquetas asignadas y su relación con el desempeño de clasificadores, basados en aprendizaje automático, construidos con dichos *datasets*.

Si bien el Análisis de Sentimientos en textos involucra diversas tareas, tales como detección de polaridad (i.e. clasificación de un texto como positivo o negativo), clasificación de intensidad (i.e. distintos grados de polaridad como en la escala muy negativo, negativo, neutral, positivo, muy positivo) e identificación de emoción básica (i.e. enojo, alegría, tristeza, etc.), no todas ellas han gozado de un desarrollo equitativo. Tal es así que el congreso *Taller de Análisis de Sentimientos en Español*, el cual recopila anualmente investigaciones en el área de distintas partes del mundo, estableció en 2020 una tarea específica de detección de emociones fundamentando que la clasificación de polaridad es una tarea bien establecida con muchos conjuntos de datos estándar y metodologías bien definidas, pero la detección de emociones ha recibido menos atención debido a su complejidad, por lo cual es necesario fomentarla para el idioma español (García-Vega et al., 2020). Un primer paso para la concreción de este objetivo es la creación de juegos de datos etiquetados en este idioma, que permitan el entrenamiento de diversos clasificadores basados en algoritmos de aprendizaje automático. Debido a la escasez de tales recursos para el español, se decidió crear, en el presente trabajo de tesis, una metodología para la construcción y validación de recursos de emoción básica para Análisis de Sentimientos.

La utilización de supervisión distante en la construcción de los recursos mencionados, permite, además de capturar contenido etiquetado de manera automática o semi automática, obtener también información contextual asociada a los mismos. Es una inquietud de esta investigación si dicha información puede utilizarse para mejorar el resultado de la clasificación, por ejemplo, incorporándola como entrada de los clasificadores basados en aprendizaje automático.

1.2 Objetivos

El objetivo general de esta investigación es:

- Desarrollar una metodología que permita implementar clasificadores automáticos de opiniones, a partir de textos de entrada surgidos de la interacción en redes sociales, utilizando algoritmos de aprendizaje automático, y proponiendo mejoras en las etapas del proceso que resulten pertinentes.

Para cumplir con el objetivo general se plantean los siguientes objetivos específicos:

- Construir o adquirir conjuntos de datos, generados a partir de la interacción en redes sociales, que reflejen la opinión del público en diversos sucesos y en lo posible incluyan información contextual acerca de los sucesos previamente mencionados.
- Realizar tareas de limpieza y pre procesamiento sobre los textos de entrada, con el objetivo de verificar su impacto en el desempeño de clasificadores basados en aprendizaje automático.
- Definir una metodología y seleccionar una métrica apropiada para validar las etiquetas asignadas a los textos recopilados.
- Implementar clasificadores de texto basados en aprendizaje automático, que permitan reconocer un conjunto de etiquetas predefinidas según el/los conjunto/s de datos utilizados y medir su desempeño.
- Medir el impacto del uso de la información contextual, presente en el conjunto de datos, en las distintas etapas del proceso de construcción de los clasificadores.

1.3 Contribuciones

Esta tesis contribuye proporcionando una metodología para la creación de recursos de emoción básica para el Análisis de Sentimientos, se enfoca en la construcción de recursos para el idioma español, pero casi la totalidad de la metodología es aplicable sin cambios a otros idiomas. Está orientada a la creación de conjuntos de datos etiquetados con una emoción básica debido que es la tarea de Análisis de Sentimientos que mayores carencias presenta para el idioma mencionado.

Lo anterior se lleva a cabo haciendo uso de técnicas de supervisión distante, pero a su vez incorpora una fase de validación que consiste en un muestreo del conjunto de datos, el etiquetado manual de la muestra y posterior cálculo de métricas de consenso entre las etiquetas asignadas manualmente y la original. Si bien la métrica de consenso utilizada, el Kappa de Fleiss, establece escalas de consenso basadas en los resultados obtenidos, originalmente no fue diseñada para medir el nivel de acuerdo en el contenido de redes sociales. Por lo tanto, resulta necesario establecer el nivel de consenso requerido específicamente para este tipo de *datasets*.

Por otro lado, es importante remarcar que no abundan trabajos para el idioma español que realicen el proceso de construcción y validación del conjunto de datos e implementación, a partir del mismo, de clasificadores de emoción básica basados en aprendizaje automático. Mediante la presentación y ejecución de esta metodología se pretende dar claridad acerca de los niveles de consenso, sobre las etiquetas obtenidas, que resulten adecuados para la creación de recursos de Análisis de Sentimientos de emoción básica.

Como producto de la aplicación de la metodología propuesta esta tesis aporta además un conjunto de datos etiquetado y validado, donde cada texto tiene asignada una emoción básica. Por otro lado, debido a que la metodología minimiza la intervención humana en el proceso, permite la recopilación de conjuntos de mayor tamaño, es decir de unos cientos de miles de muestras en contraposición a sólo miles que se lograrían con etiquetado manual. Esta particularidad es clave si se busca utilizar los datos para la construcción de clasificadores basados en aprendizaje automático, ya que es sabido que dichos clasificadores generalizan mejor al ser expuestos a una mayor cantidad de muestras en la etapa de entrenamiento. Debido a la mencionada escasez de recursos en idioma español, el conjunto de datos recopilado puede ser utilizado como punto de partida de futuras investigaciones.

Por último, la utilización de supervisión distante en el proceso de captura de datos, permite obtener abundante información contextual junto con el conjunto de datos. En los clasificadores, basados en aprendizaje automático, implementados en esta tesis se compara el desempeño obtenido por los mismos en presencia y ausencia de información contextual, de esta manera otro aporte derivado de la metodología utilizada es la mejora del desempeño de los clasificadores implementados mediante el uso de información contextual.

1.4 Publicaciones

Las siguientes publicaciones, en las cuales el tesista es autor o coautor, respaldan el trabajo realizado en esta tesis doctoral:

- Russo, C., Ramón, H., Alonso, N., Cicerchia, B., Esnaola, L., & Tessore, J. P. (2016). Tratamiento masivo de datos utilizando técnicas de Machine Learning. Actas del XVIII Workshop de Investigadores en Ciencias de la Computación. XVIII WICC. Concordia, Argentina. Abril 14- 15, 2016. Páginas 131-134.
- Tessore, J. P., Esnaola, L. M., Russo, C. C., Ramón, H. D., & Pompei, S. (2018). Análisis automático de grandes volúmenes de datos en redes sociales mediante minería de textos combinado con algoritmos inteligentes. Actas del XX Workshop de Investigadores en Ciencias de la Computación. XX WICC. Corrientes, Argentina. Abril 26-27, 2018. Páginas 60-64.
- Esnaola, L., Tessore, J. P., Ramón, H. D., & Russo, C. C. (2019). Análisis comparativo de tareas de pre procesamiento de textos sobre contenido extraído de redes sociales. Actas del XXI Workshop de Investigadores en Ciencias de la Computación. XXI WICC. San Juan, Argentina. Abril 25-26 de, 2019. Páginas 164-168.
- Tessore, J. P., Esnaola, L. M., Russo, C. C., & Baldassarri, S. (2019). Comparative analysis of preprocessing tasks over social media texts in Spanish. Proceedings of the XX International Conference on Human Computer Interaction, Interacción 2019. Donostia - San Sebastián, País Vasco, España. Junio 25-28, 2019. ACM, 27:1 - 27:8. https://doi.org/10.1145/3335595.3335632
- Esnaola, L., Tessore, J. P., Ramón, H., & Russo, C. (2019). Effectiveness of preprocessing techniques over social media texts for the improvement of machine learning based classifiers. Proceedings of the XLV Latin American Computing Conference, CLEI 2019. Ciudad de Panamá, Panamá. Septiembre 30 - Octubre 4, 2019. IEEE, 10 páginas. https://doi.org/10.1109/CLEI47609.2019.235076
- Tessore, J.P., Esnaola, L.M., Lanzarini, L. Baldassarri, S. (2021) Distant Supervised Construction and Evaluation of a Novel Dataset of Emotion-Tagged Social Media Comments in Spanish. Cognitive Computation 14, 407–424. Springer. ISSN: 1866-9964; 1866-9956. DOI: https://doi.org/10.1007/s12559-020-09800-x.

- Tessore, J.P.; Esnaola, L.M.; Ramón, H.D.; Lanzarini, L.; Baldassarri, S. (2022) Contextual information usage for the enhancement of basic emotion classification in a weakly labelled social network dataset in Spanish. Multimedia Tools and Applications 82, 9871–9890. Springer. ISSN: 1573-7721; 1380-7501. DOI: https://doi.org/10.1007/s11042-022-13750-x.

1.5 Organización de la tesis

Este documento se encuentra estructurado en seis capítulos, cuyos contenidos más importantes se detallan a continuación.

- **Capítulo 1:** Se hace una introducción al tema tratado en esta tesis. En primer lugar, se resalta la escasez existente de recursos de Análisis de Sentimientos en español sobre todo para la tarea de clasificación de emoción básica. El costo asociado al etiquetado manual lleva a plantear un enfoque basado en supervisión distante incorporando una etapa de validación de manera de no dejar de lado la calidad. Por otro lado, la utilización de supervisión distante permite la captura de información contextual, la cual puede ser aprovechada para la mejora del desempeño de los clasificadores basados en aprendizaje automático. A partir de lo anterior, se definen los objetivos y contribuciones del presente trabajo de tesis junto con las publicaciones que lo respaldan.
- **Capítulo 2:** En primer lugar, se presenta una fundamentación teórica de los diversos temas relacionados con esta tesis. Se comienza con definiciones de Computación Afectiva, Análisis de Sentimientos y de los modelos de representación de emociones existentes. Luego se desarrollan los temas relacionados a la minería de textos y el procesamiento del lenguaje natural, en lo que respecta al preprocesamiento y formatos de representación del contenido. Posteriormente, se revisan los algoritmos de aprendizaje automático más utilizados para la clasificación de textos.
- **Capítulo 3:** En este capítulo se realiza un estado del arte de los temas vinculados a los aportes de la presente tesis. Primero se aborda la construcción de conjuntos de datos mediante etiquetado manual para luego enfocarse en los realizados mediante supervisión distante. Por último, se trata la utilización de información contextual para la mejora de clasificadores basados en aprendizaje automático supervisado.

- **Capítulo 4:** Se presenta el proceso desarrollado para la construcción de un conjunto de datos de emoción básica utilizando supervisión distante. Dicho proceso comprende las etapas de recopilación de los datos y la de aplicación de técnicas de preprocesamiento para la normalización de los mismos. Posteriormente, también se presentan estadísticas del conjunto de datos recopilado. Por otro lado, para asegurar la calidad de los datos se discuten y establecen los requisitos de consenso necesario sobre las etiquetas del conjunto de datos, para luego medir y validar el mismo mediante cuestionarios a especialistas.
- **Capítulo 5:** En primer lugar, se selecciona un formato de representación basado en *embeddings* neurales para el idioma español, posteriormente se establece una línea base de la cual se replica la arquitectura para el diseño del clasificador, la misma está basada en redes neuronales recurrentes. Luego se diseñan y ejecutan dos experimentos, el primero de ellos tiene como objetivo mostrar el impacto del entrenamiento con un conjunto de datos más numeroso, fruto de la recopilación utilizando la metodología descrita en los capítulos anteriores, mientras que el segundo experimento deja en evidencia el impacto del uso de la información contextual en el desempeño del clasificador construido.
- **Capítulo 6:** Se presentan las conclusiones del presente trabajo de tesis junto con las posibles líneas de investigación futuras.

2. Fundamentación teórica

Contenido

Resumen

La construcción de sistemas capaces de realizar Análisis de Sentimientos y/o Clasificación de Emociones requiere la combinación de conocimientos provenientes principalmente de los ámbitos de la Psicología y de las Ciencias de la Computación. En el presente capítulo se presenta la fundamentación teórica, es decir se revisan los modelos, procesos, esquemas y algoritmos necesarios para la construcción de los sistemas detallados, se inicia con lo referente a Computación Afectiva y Análisis de Sentimientos en la sección 2.2, luego en la sección 2.3 se aborda lo referido a la minería de textos y el procesamiento del lenguaje natural, en la sección 2.4 se tratan los diferentes esquemas de representación de contenido en formato de texto que se utilizan como entrada para diversos algoritmos de aprendizaje automático, y finalmente, en la sección 2.5 se mencionan los algoritmos de aprendizaje automático relevantes para el presente trabajo de tesis.

2.1 Introducción

Con el auge de las redes sociales y las comunicaciones, las personas están cada vez más involucradas en muchos aspectos en los que antes solían ser sólo consumidores pasivos (Hu & Liu, 2012). Las redes sociales permiten a las personas interactuar, expresándose rápida y libremente acerca de una amplia variedad de temas. Además de ser ampliamente utilizadas por empresas u otras organizaciones para influir en los usuarios y realizar campañas de marketing dirigido (Aggarwal & Zhai, 2012).

Por otro lado, el contenido generado a partir de las interacciones mencionadas contiene información valiosa acerca de la opinión del público con respecto a distintos productos, servicios, personas, sucesos, etc. El análisis adecuado de dicha información puede ofrecer una amplia gama de posibles propósitos (Bollen & Mao, 2011) (Nasukawa & Yi, 2003) (Ortigosa, Martín, & Carro, 2014) como detectar la emoción que surge de la opinión de grandes grupos de personas sobre determinados productos, servicios, políticas públicas, o de los aspectos que componen estas entidades. También podría utilizarse para identificar demandas no satisfechas o quejas de los ciudadanos; en seguridad, para la detección automática de factores de riesgo en las redes sociales como amenazas o ciberacoso (Sintaha, Bin Satter, Zawad, Swarnaker, & Hassan, 2016).

Si bien las principales redes sociales, ya sea Twitter, Facebook o Instagram, permiten o incluso algunas de ellas están orientadas a contenido multimedia, la mayoría de las interacciones o el contenido generado es en forma de texto. Este fenómeno no es exclusivo de las redes sociales, según (Villa Monte, 2019) además los correos electrónicos, reclamos de clientes, noticias periodísticas, páginas web, entre otros están repletos de texto. Este autor también resalta que, según estimaciones, el 80 % de los datos disponibles en el mundo son almacenados en formato texto, por lo cual su procesamiento automático se vuelve una tarea crucial (Miner, Elder, Hill, & Delen, 2012).

La proliferación de contenido en formato texto, motorizado por el desarrollo de la Web y posteriormente de las redes sociales, y el deseo de obtener información valiosa a partir del mismo, ha motivado el desarrollo de nuevas disciplinas cuyo objetivo es sacar provecho de estas nuevas fuentes de información. El análisis de opiniones/emociones en texto se enmarca

en la disciplina llamada Análisis de Sentimientos, cuyo objetivo es el análisis de las opiniones, sentimientos, evaluaciones, impresiones, actitudes y emociones de las personas con respecto a diversas entidades ya sea productos, servicios, organizaciones, otros individuos, etc. (B. Liu, 2012).

Si bien este tipo de análisis en textos puede a priori parecer más sencillo, debido a que no se debe lidiar con la complejidad de procesar contenido multimedia, presenta otros inconvenientes que es preciso abordar para llevar adelante la tarea. En primer lugar, el texto carece de marcadores muy útiles para detectar opiniones, sentimientos o emociones, tales como el tono de voz o las expresiones faciales. Por otro lado, los textos surgidos de la interacción en redes sociales se presentan en modo no estructurado y además contienen errores no presentes en otros tipos de textos formales.

Las problemáticas inherentes a los textos mencionados hacen que el Análisis de Sentimientos se relacione estrechamente con el procesamiento del lenguaje natural y comprenda además muchas de sus tareas. Los sistemas de Análisis de Sentimientos suelen incluir, como primer paso, una etapa de pre procesamiento que abarca la limpieza y normalización de los textos a analizar. Algunas de las tareas comprendidas en esta primera etapa son: la normalización léxica de los micro textos (i.e. textos informales), la delimitación de las oraciones, lematización, truncado, resolución de ambigüedades, representación del texto en un formato numérico (ya sea con matrices densas o ralas), etc. Una vez realizado lo anterior es posible avanzar con la extracción o reconocimiento de emociones, sentimientos u opiniones de los textos analizados, para ello debe definirse en primer lugar el modelo de representación de emociones a utilizar, estos modelos se dividen en dos grandes grupos, categóricos o dimensionales. Por último, debe seleccionarse un algoritmo de aprendizaje automático para la construcción de un clasificador a partir de entrenamiento con parte de los datos recopilados y posteriores pruebas del mismo con las muestras restantes.

2.2 Computación Afectiva y Análisis de Sentimientos

2.2.1 Definición y tareas involucradas

La Computación Afectiva (Picard, 1997) es un campo de la computación cognitiva y la inteligencia artificial cuyo objetivo es desarrollar sistemas capaces de reconocer, interpretar, procesar y simular las emociones humanas. Por otro lado, el Análisis de Sentimientos es un área de la Computación Afectiva que se enfoca en la tarea específica de analizar y comprender las emociones expresadas en el lenguaje natural, como textos, comentarios, reseñas y redes sociales.

Las definiciones anteriores son generales, al profundizar en la cuestión se evidencia que existen matices en la bibliografía acerca de las áreas en las que se divide tanto la Computación Afectiva como el Análisis de Sentimientos, en particular a lo que respecta a la detección y/o clasificación de emociones. Por tal motivo, resulta necesario discutir brevemente las definiciones de los autores más relevantes para luego determinar cuales se adoptarán para el resto del presente trabajo. Varias de las definiciones que se presentan a continuación fueron recopiladas por (Osorio Angel et al., 2021).

Según (Cambria et al., 2017) el Análisis de Sentimientos es un área de la Computación Afectiva que contiene tres capas. La primera es una capa sintáctica que tiene como objetivo el pre procesamiento de textos e incluye tareas como etiquetado de parte del discurso, lematización y normalización de micro texto. La segunda es una capa semántica que tiene como objetivo deconstruir el texto normalizado de la capa anterior en conceptos, resolver entidades y filtrar contenido neutral para mejorar la precisión de la clasificación de sentimientos. Las tareas de esta capa son, entre otras, la extracción de conceptos, la desambiguación del sentido de las palabras y la detección de subjetividad (Chaturvedi, Cambria, & Vilares, 2016). La última es la capa pragmática, enfocada en extraer significado tanto de la estructura de la oración como de la semántica obtenida de capas anteriores, e incluye tareas como detección de polaridad, reconocimiento de aspecto, detección de sarcasmo (Majumder et al., 2019), reconocimiento de personalidad (Majumder, Poria, Gelbukh, & Cambria, 2017) y clasificación de emociones.

Para (Poria, Cambria, Bajpai, & Hussain, 2017), por otro lado, la Computación Afectiva se divide en dos grandes áreas: el Análisis de Sentimientos, encargada principalmente de la detección de polaridad; y la detección de emociones, encargada de la clasificación de emoción básica.

(Medhat, Hassan, & Korashy, 2014) propuso otra definición para Análisis de Sentimientos, estableciendo que puede ser considerado como un proceso con tres niveles de clasificación primarios: nivel de documento, nivel de oración y nivel de aspecto, en el cual el objetivo es detectar cuando se expresa una opinión positiva o negativa, generalmente conocido como detección de polaridad. Por otro lado, dicho trabajo menciona que existe cierta ambigüedad en las definiciones de sentimiento y emoción que de alguna manera impactan en la delimitación de las tareas del Análisis de Sentimientos. Más allá de esto, los autores de dicho trabajo mencionan que la detección de emociones puede ser considerada como una tarea dentro del Análisis de Sentimientos.

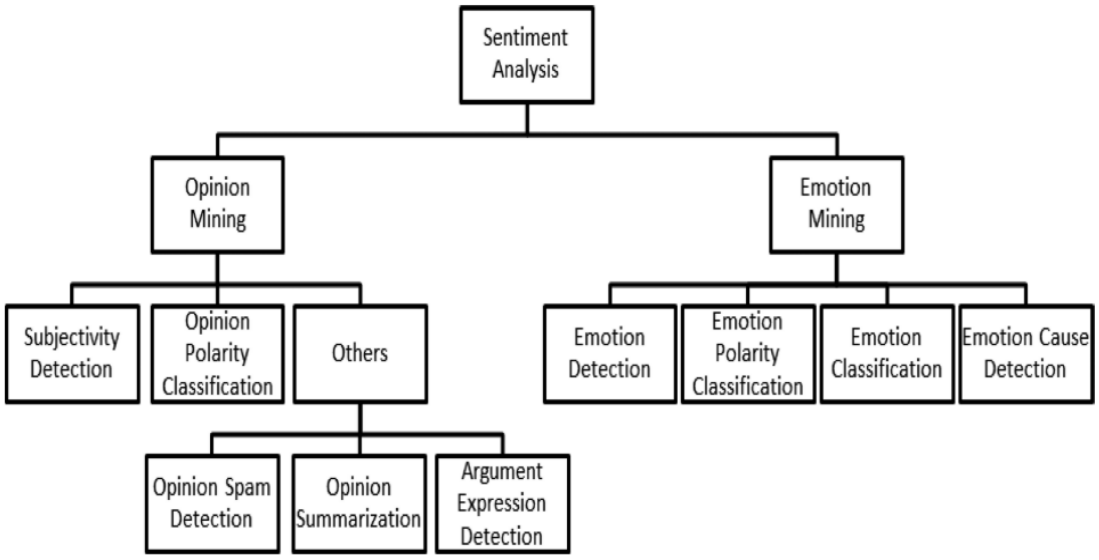


Figura 1: Tareas del Análisis de Sentimientos (Yadollahi, Shahraki, & Zaiane, 2017).

Por otro lado, para (B. Liu, 2012) el Análisis de Sentimientos incluye las tareas de análisis de subjetividad, análisis de emociones, extracción de opiniones, minería de sentimientos, entre otras. (De Albornoz, Plaza, & Gervas, 2012) divide el Análisis de Sentimientos en: clasificación de polaridad, clasificación de intensidad e identificación de emociones. (Yadollahi et al., 2017) coincide en incluir la detección de emociones como tarea del Análisis

de Sentimientos y presenta un diagrama que incluye una subclasificación de las tareas del Análisis de Sentimientos, este puede verse en la Figura 1.

Según lo discutido en la sección, la mayoría de los referentes en el área de Análisis de Sentimientos mencionados coinciden en que la detección de emociones constituye una de sus tareas fundamentales. En consecuencia, esta definición es la que se considera a lo largo del presente trabajo de tesis.

2.2.2 Modelos de representación de emociones existentes

Los términos sentimiento y emoción se utilizan ampliamente, pero generalmente se confunden o se malinterpretan y a menudo se han utilizado indistintamente; sin embargo, los sentimientos se diferencian de las emociones por la duración en que se experimentan (Munezero, Montero, Sutinen, & Pajunen, 2014). (Z. Wang, Ho, & Cambria, 2020) afirmó que, si bien los sentimientos reflejan sensaciones y actitudes, las emociones proporcionan una caracterización más refinada de los sentimientos involucrados. A menudo el Análisis de Sentimientos suele asociarse con la detección de polaridad, es decir clasificar con mayor o menor nivel de granularidad si el contenido tiene una connotación positiva o negativa. Sin embargo, la detección de emociones va más allá para intentar revelar las emociones exactas expresadas en el texto. En su estudio, los autores sostienen que cualquiera que sea la metodología de detección de emociones utilizada, siempre es muy importante tener un modelo de categorización adecuado para las emociones. En este sentido, el estudio revisa los modelos emocionales existentes considerando la visión de los psicólogos, así como las perspectivas de las ciencias sociales, las ciencias de la computación y la ingeniería. Los diferentes modelos relevados en el estudio varían en el número de emociones que reconocen; algunos constan de seis emociones primarias, mientras que otros identifican hasta 24.

Por otro lado, la manera más utilizada para clasificar los modelos de representación de emociones es dividirlos principalmente entre categóricos y dimensionales. A su vez, esta última categoría, puede subdividirse según la cantidad de dimensiones que contengan los modelos que pertenecen a ella.

En primer lugar, se discuten los ***modelos categóricos***. Estos se caracterizan por definir un conjunto finito y discreto de emociones básicas o universales, si bien existe un solapamiento

de algunas de ellas entre los distintos enfoques no hay un acuerdo acerca de las mismas. Entre los modelos categóricos, se destaca el modelo de emociones de Ekman (Ekman & Friesen, 1971), que distingue seis expresiones faciales universales, es decir que pueden ser reconocidas por cualquier ser humano sin importar la influencia cultural a la que estuvo expuesto. Estas expresiones son: ira, miedo, disgusto, alegría, tristeza y sorpresa, más la expresión neutral (ver Figura 2). Dichas expresiones se identifican con las emociones básicas de los seres humanos, cualquier otra emoción es, en consecuencia, una combinación de las anteriores. Según varias fuentes (Yadollahi et al., 2017) (Susanto, Livingstone, Ng, & Cambria, 2020) (Alswaidan & Menai, 2020) (Poria, Cambria, Bajpai, et al., 2017) (Zeng, Pantic, Roisman, & Huang, 2014) el modelo de Ekman es el más ampliamente utilizado para el Análisis de Sentimientos.

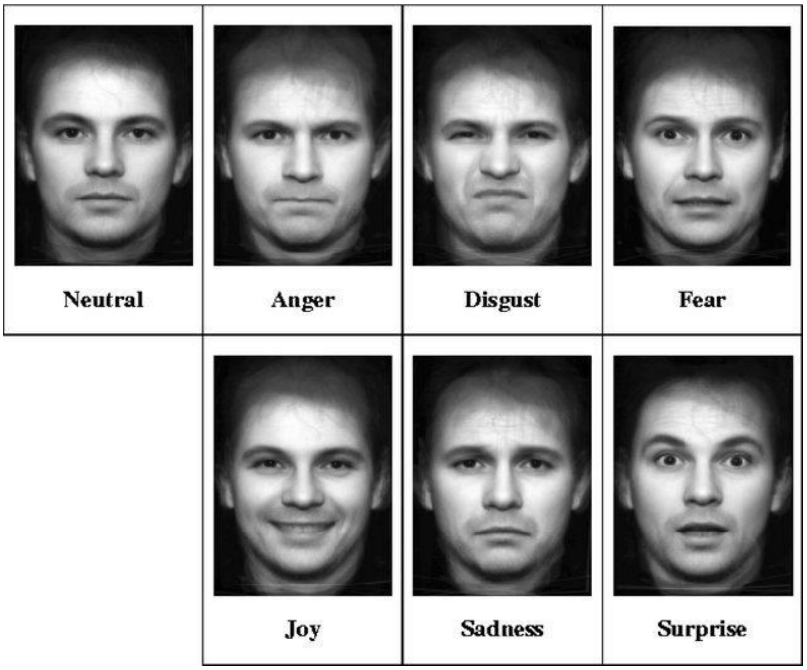
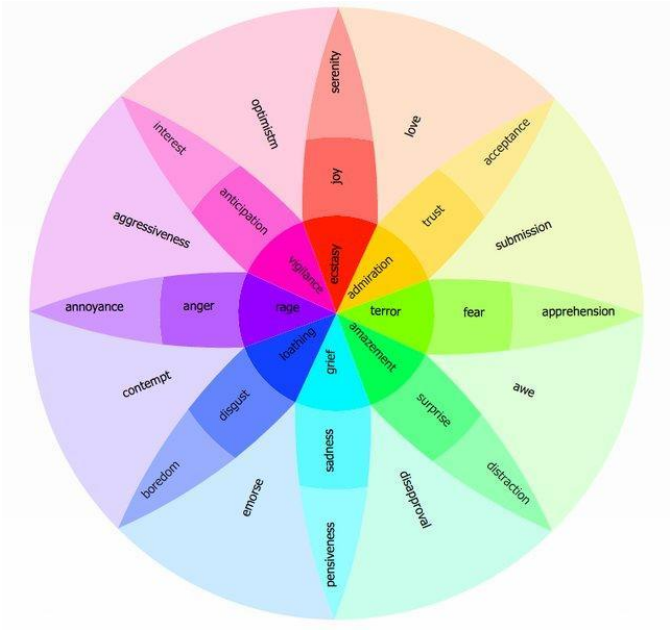


Figura 2: Emociones básicas de Ekman asociadas a expresiones faciales. Fuente: (Mizgajski & Morzy, 2019).

Otro de los modelos más influyentes es la rueda de emociones de (Plutchik, 1980), el cual surge de estudios acerca de la reacción psicológica producida por las diversas emociones en humanos y animales, producto de estos experimentos determinó la existencia de ocho emociones básicas: alegría, confianza, miedo, sorpresa, tristeza, asco, enojo y anticipación, que se muestran en la Figura 3-a. Si bien, a partir de dichas emociones, este modelo es catalogado como categórico por algunos autores (Mizgajski & Morzy, 2019), en otros casos es considerado tridimensional (Vaughan, 2011) dado que incorpora las dimensiones de polaridad,

similaridad e intensidad, como se observa en la Figura 3-b. Las emociones mencionadas corresponden a un nivel de intensidad medio. Para el nivel bajo se convierten en: serenidad, aceptación, aprehensión, distracción, persistencia, aburrimiento, molestia e interés, mientras que para el nivel alto son: éxtasis, admiración, terror, asombro, dolor, aversión, ira y vigilancia, en ambos casos respectivamente. Por otro lado, existen también emociones de segundo nivel, por ejemplo, la alegría y la confianza forman el amor.

a)



b)

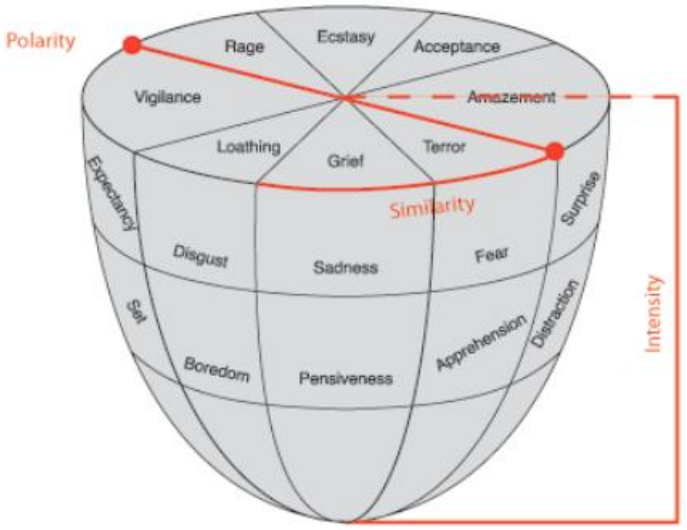


Figura 3: a) Rueda de emociones de Plutchik. Fuente: (Nielek, Ciastek, & Kopeć, 2017). b) Dimensiones de la rueda en 3D. Fuente: (Vaughan, 2011).

A continuación, se mencionan los principales ***modelos dimensionales***. Estos definen regiones continuas en las cuales las distintas emociones se sitúan según los valores que tomen para cada una de las dimensiones consideradas. Los modelos bidimensionales se caracterizan por la valencia que indica si una emoción es positiva o negativa y por la activación que refleja la relajación o tensión. El principal exponente de esta categoría es el modelo circunflejo de Russell, en la Figura 4 puede verse que la valencia se encuentra en el eje de las abscisas mientras que la activación en el de las ordenadas, a su vez en los distintos cuadrantes pueden verse las diferentes emociones según los valores que adoptan para cada dimensión. Otros exponentes destacados de los modelos bidimensionales son el de (Arnold, 1960), (Lazarus, 1991) y (Hekkert & Desmet, 2002).

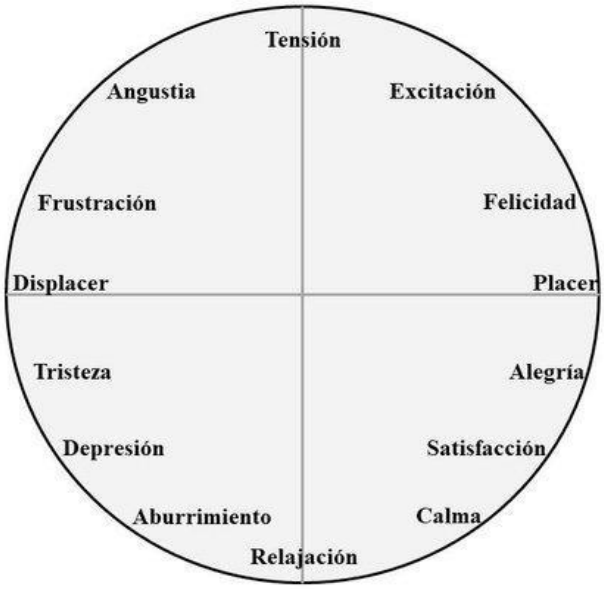


Figura 4: Modelo Circunflejo de las Emociones de Russell. Fuente: (Scandar, 2019).

Los modelos tridimensionales incorporan una dimensión adicional, la cual varía según el caso en cuestión. Entre estos últimos, uno de los más utilizados es el de estado emocional PAD (Pleasure - Arousal - Dominance, por sus siglas en inglés) (Bakker, van der Voordt, Vink, & de Boon, 2014), este modelo además de la valencia y la activación incorpora el control como tercera dimensión. Por último, existen modelos multidimensionales, como el reloj de arena de emociones (Susanto et al., 2020), que considera la sensibilidad, la aptitud, la simpatía y la

atención como dimensiones, en la Figura 5, pueden verse un conjunto de emociones consideradas según este esquema. Este es un modelo de categorización afectiva, inspirado principalmente en los estudios de (Plutchik, 2001) sobre las emociones humanas.

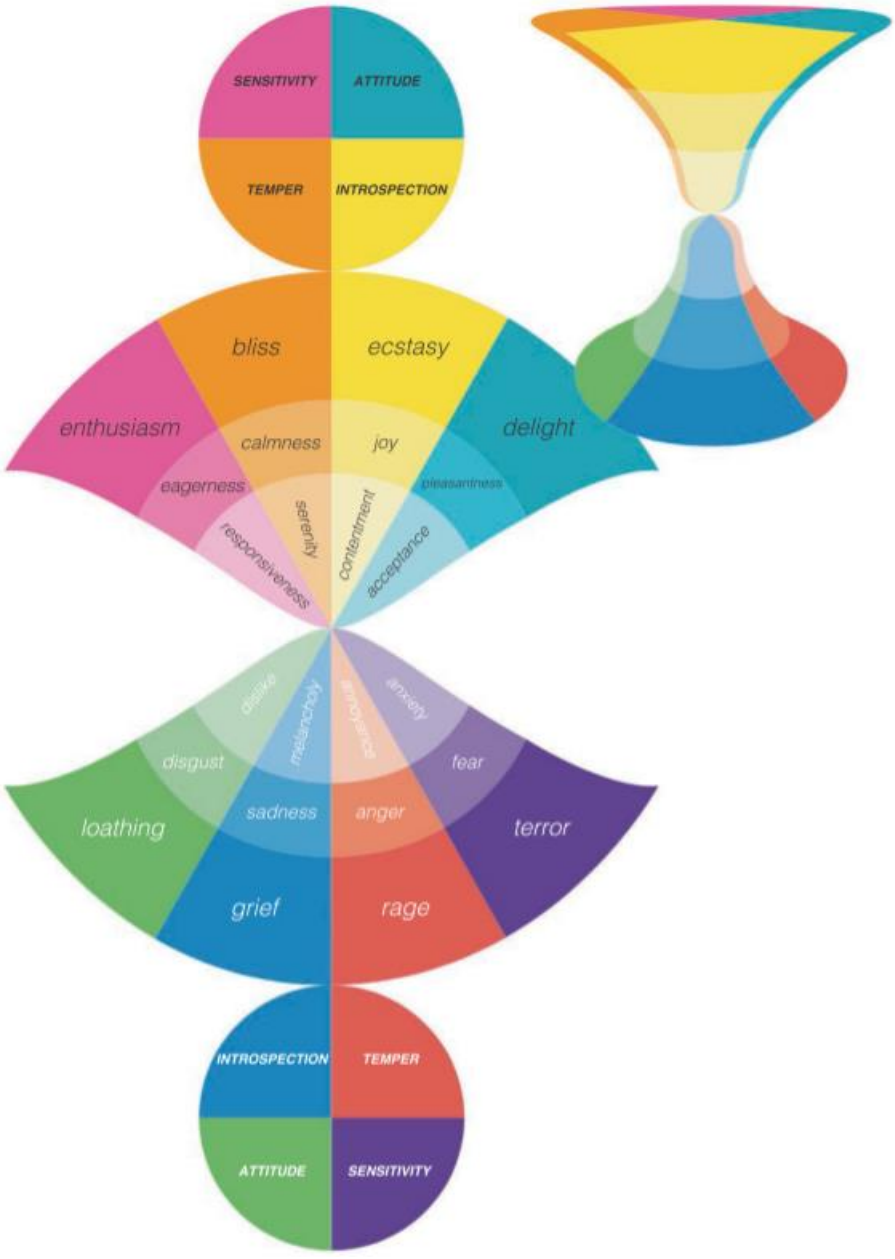


Figura 5: Reloj de arena de las emociones. Fuente: (Susanto et al., 2020).

Según (Yadollahi et al., 2017), entre otros referentes a destacar se encuentran el modelo categórico de (Shaver, Schwartz, Kirson, & O'Connor, 1987) este tiene estructura de árbol en donde las ramas principales corresponden a las emociones básicas (enojo, miedo, alegría,

amor, tristeza y sorpresa) y cada una tiene sus propias subcategorías. También se encuentra el modelo dimensional de (Lövheim, 2012) que utiliza tres hormonas (serotonina, dopamina y noradrenalina) como dimensiones y ubica sus emociones básicas (enojo, asco, angustia, miedo, alegría, interés, vergüenza y sorpresa) en este espacio.

2.3 Minería de textos y procesamiento del lenguaje natural

Con el auge y desarrollo de la web y posteriormente de las redes sociales se dio también un incremento considerable en la generación de datos en general y en formato de texto en particular. Estos datos en formato texto se han vuelto un recurso muy valioso para las distintas organizaciones deseosas de saber las preferencias del público al cual intentan acceder.

La lingüística es la ciencia que se dedica al estudio del lenguaje humano y de su intersección con la computación surge, entre otras ramas, la minería de textos. Según (Aggarwal, 2018) la minería de texto se ha vuelto cada vez más popular en los últimos años debido a la ubicuidad de los datos en forma de texto en la Web, las redes sociales, los correos electrónicos, las bibliotecas digitales y los sitios de chat. Sin embargo, el aprovechamiento de dichos datos requiere de la aplicación de métodos y algoritmos específicos que permitan la extracción de información valiosa a partir de ellos debido a su naturaleza no estructurada. Para ello, suele ser necesaria una adaptación de los algoritmos de minería de datos de manera que sea posible aplicarlos al formato texto (Hotho, Nürnberger, & Paaß, 2005).

La minería de textos se origina de la intersección de tres áreas fundamentales: el procesamiento del lenguaje natural (NLP *natural language processing* por sus siglas en inglés), la recuperación de información (conocida como IR *information retrieval* por sus siglas en inglés) y el aprendizaje automático (ML *machine learning* por sus siglas en inglés).

IR se refiere a encontrar documentos relevantes en respuesta a una solicitud. Para ello debe medir la similitud entre distintos documentos. La minería de texto, puede hacer uso de los sistemas de IR, por ejemplo, para clasificación, en dicho caso se examinan las propiedades de los documentos retornados por los sistemas de IR, como pueden ser las etiquetas de clase, y se determina la categoría de un documento nuevo a partir de las etiquetas de los que son similares (Weiss, Indurkhya, Zhang, & Damerau, 2005).

Por otro lado, la minería de textos se vale del NLP para representar de manera más precisa el significado de los textos no estructurados (Kao & Poteet, 2007), en otras palabras, intenta identificar los actores involucrados, lugares, tiempos y motivaciones de los hechos que se describen en el texto. Debe lidiar con ambigüedades tanto a nivel palabra, por ejemplo, en caso de palabras homógrafas, como en la estructura gramatical, que puede darse entre otros casos en las anáforas. Para ello utiliza tareas de soporte, como la segmentación y normalización del texto, para determinar las características relevantes a ser analizadas por el sistema de minería de textos, el etiquetado de categorías gramaticales (POS *part-of-speech tagging* por sus siglas en inglés) que consiste en etiquetar las palabras según su tipo (verbo, sustantivo, adjetivo, adverbio, etc.), el *parsing* de dependencias, que analiza la estructura gramatical y establece relaciones entre las palabras, la desambiguación del sentido de las palabras y el reconocimiento de entidades nombradas. En la implementación de algunas de las tareas mencionadas anteriormente necesita recursos como *lexicones*, que contienen el significado y las propiedades gramaticales de las palabras, ontologías de entidades y acciones, junto con *thesaurus* y redes semánticas de sinónimos y abreviaturas como es el caso de WordNet (Kilgarriff & Fellbaum, 1998).

El ML es un área de la inteligencia artificial cuyo objetivo es el desarrollo de técnicas que permitan generalizar el conocimiento existente, típicamente utilizando un conjunto de muestras durante una fase de entrenamiento, para luego aplicarlo a muestras nuevas de la manera más precisa posible. La minería de textos lo utiliza para crear modelos que sean capaces de clasificar o extraer información específica de los documentos analizados. Los sistemas de IR, que miden la similitud de un documento de consulta contra una colección de manera de retornar resultados similares, pueden valerse de algoritmos de ML para lograr este propósito.

Muchas de las investigaciones en minería de textos derivan de otras análogas en minería de datos, en consecuencia, las arquitecturas de alto nivel de los sistemas de ambas disciplinas comparten muchas similitudes en cuanto a las rutinas de preprocesamiento, algoritmos de descubrimiento de patrones y herramientas de visualización/presentación. Sin embargo, la diferencia fundamental entre ellas radica en que, la minería de datos asume que los datos ya han sido almacenados en un formato estructurado, por el contrario, la entrada de los sistemas de minería de textos generalmente se compone de datos no estructurados, como pueden ser

correos electrónicos, publicaciones en redes sociales, páginas web, fragmentos de chats, etc. Esta situación hace necesario empezar la labor en un nivel inferior, realizando un pre procesamiento que identifique las características más representativas de los documentos en lenguaje natural a analizar. A partir de ellas, se construye una vista minable sobre la cual se aplican distintos algoritmos con el objetivo de extraer patrones significativos (Feldman & Sanger, 2006).

En tal sentido, el proceso llevado a cabo por los sistemas de minería de textos se asemeja al de descubrimiento de conocimiento en bases de datos descrito por (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) (KDD por sus siglas en inglés *Knowledge Discovery in Databases*) que puede verse en la Figura 6. Sin embargo, es relevante mencionar que, si bien a grandes rasgos el proceso contiene las mismas etapas, las tareas a realizar dentro de cada una de ellas deben adaptarse a las necesidades de los sistemas de minería de textos, principalmente como consecuencia de la naturaleza de los datos de entrada y con el objetivo de darles una estructura que permita su análisis posterior.

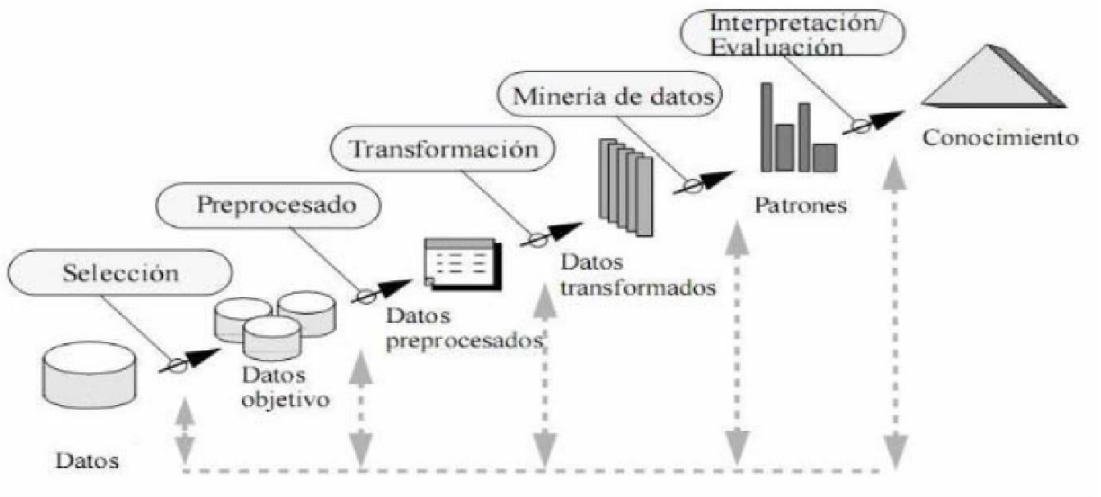


Figura 6: Proceso de extracción de conocimiento en bases de datos. Fuente: (Batista, 2015).

Como ejemplo de lo anterior, en los sistemas tradicionales de minería de datos, las fases de Selección, Preprocesado y Transformación de la Figura 6 generalmente involucran la recopilación de diversas bases de datos heterogéneas y su posterior consolidación mediante el tratamiento de datos anómalos o faltantes, la selección de atributos o datos relevantes o incluso la adición de nuevos atributos producto de alguna operación específica sobre los originales.

En cambio, en los sistemas de minería de textos, se llevan a cabo tareas como la partición o tokenización del texto, filtrado de palabras no relevantes, normalización de las palabras (mediante unificación de variantes, *stemming* o lematización) y la clasificación de las mismas (por medio de POS *tagging*, reconocimiento de entidades nombradas y desambiguación del significado). También se deben transformar los textos a una representación numérica, para ello suelen utilizarse matrices, que puede ser dispersas, como en el caso de las bolsas de palabras (o n-gramas de palabras/caracteres), o densas como pueden ser Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) o GloVe (*Global Vectors* por sus siglas en inglés) (Pennington, Socher, & Manning, 2014) entre otras.

En la fase de Minería de datos (Figura 6), en primer lugar, se realiza el análisis propiamente dicho, para lo cual se selecciona el algoritmo de ML apropiado para la tarea en cuestión. Dentro de las tareas más habituales se pueden mencionar la clasificación de textos, el agrupamiento o *clustering* y la extracción automática de información. Finalmente, en la fase de Interpretación y Evaluación (Figura 6), se comprueba el sistema mediante diversas métricas de desempeño como la exactitud, precisión, exhaustividad o F1 score, como así también de manera gráfica mediante matrices de confusión, gráficos de silueta, proyecciones vectoriales de los *embeddings*, etc.

El proceso de KDD junto con los ajustes mencionados para sistemas de minería de textos se pueden ver en la arquitectura genérica enunciada por (Feldman & Sanger, 2006) que se presenta en la Figura 7, donde se consideran, además, elementos correspondientes a la interfaz de usuario y un conjunto de técnicas de refinamiento, pudiendo estas últimas asociarse a la etapa de evaluación e interpretación del proceso KDD.

Hasta aquí se definió minería de textos, las tareas que involucra y su relación con la minería de datos. En las próximas subsecciones se abordan técnicas, herramientas y algoritmos específicos, utilizados generalmente en sistemas de minería de textos, que contribuyeron en parte al desarrollo del presente trabajo de tesis. En particular se discuten: técnicas de preprocesamiento, esquemas de representación de contenido y algoritmos de ML.

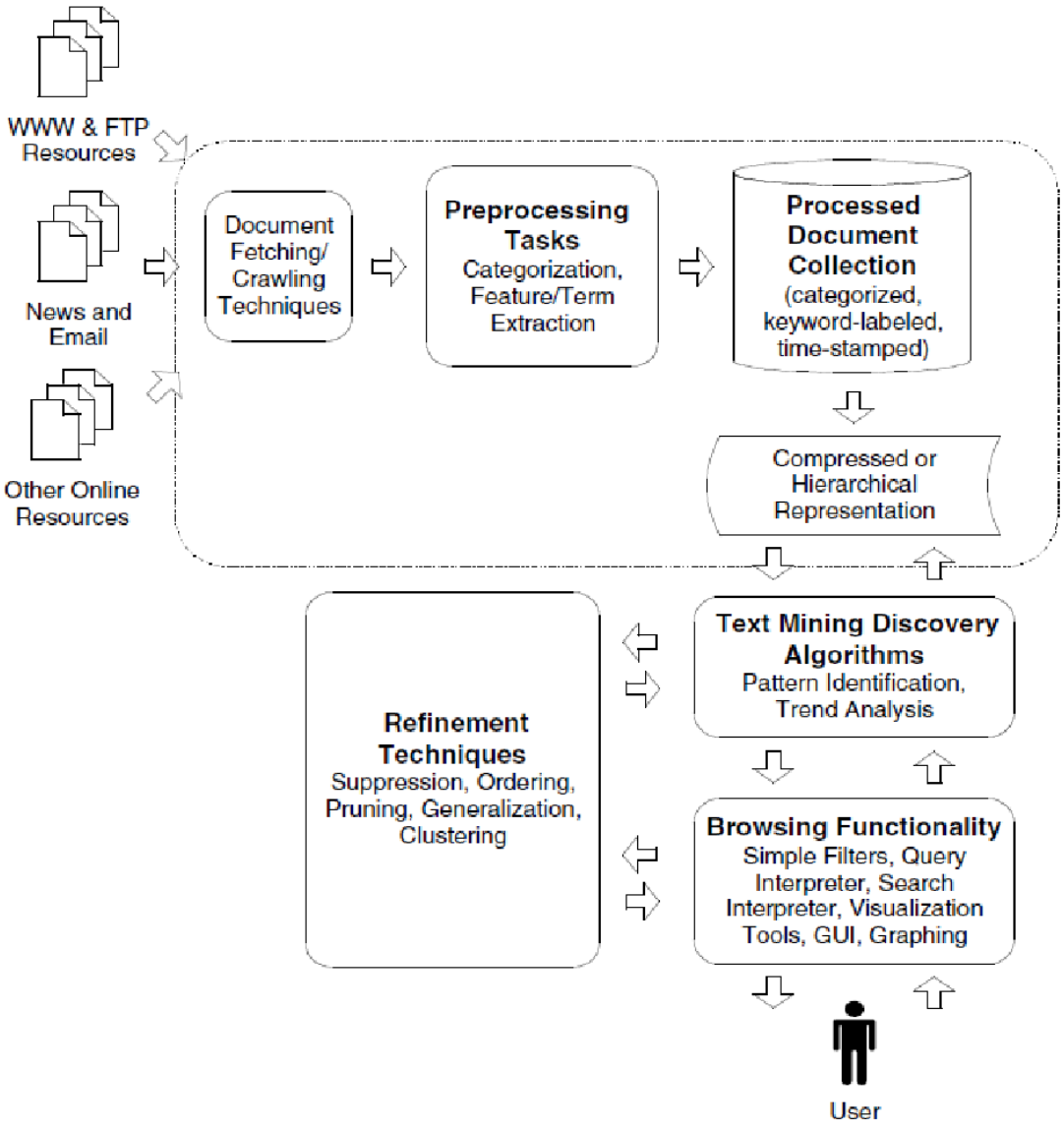


Figura 7: Arquitectura genérica para sistemas de minería de textos (Feldman & Sanger, 2006).

2.4 Preprocesamiento del contenido recopilado

En los últimos años con el surgimiento y expansión de las redes sociales, también ha crecido el interés por analizar lo que se publica allí. En particular, para el contenido en formato texto, uno de los aspectos clave a considerar es si se puede analizar con técnicas tradicionales de minería de texto, es decir, las utilizadas para analizar contenido de texto formal. Ese problema se abordó en el trabajo de (Baldwin, Cook, Lui, Mackinlay, & Wang, 2013), en el que se concluyó que el contenido de texto procedente de las redes sociales es generalmente más

ruidoso que el texto tradicional. Sin embargo, este "ruido" se puede abordar aplicando técnicas de preprocesamiento como, el filtrado de las palabras, la normalización léxica, la delimitación del texto y/o el etiquetado de las palabras. Una de las conclusiones más importantes obtenidas en ese trabajo es que, una vez aplicadas, las técnicas de preprocesamiento reducen la variabilidad de los datos de tal manera que se pueden analizar mediante técnicas tradicionales de minería de textos. En esta subsección se mencionan las técnicas de preprocesamiento más utilizadas, algunas de las cuales son luego aplicadas durante el desarrollo de la presente tesis.

2.4.1 Filtrado del texto

En primer lugar, como se mencionó anteriormente, los textos provenientes de la web y las redes sociales contienen características no presentes en los textos formales que dependiendo de la tarea a realizar deben ser tratados de manera correspondiente. Entre estos elementos podemos encontrar los enlaces, hashtags, emojis y emoticones, nombres de usuario, tags HTML (*HyperText Markup Language* por sus siglas en inglés), entre otras. En particular, el presente trabajo de tesis se enfoca en la tarea de Análisis de Sentimientos por lo cual generalmente se deben remover los elementos anteriores mediante el uso de heurísticas y/o expresiones regulares. Dicho proceso es descrito en el Capítulo 4.

En segundo lugar, otro elemento a considerar durante el filtrado de texto son las palabras de paro (*stop words* por su nombre en inglés), estas son palabras que no aportan demasiado sentido a la oración y que generalmente pueden quitarse sin alterar el significado de la misma. Son dependientes del idioma y existen listas pre compiladas de las mismas las cuales en general son independientes del contexto. Usualmente se utilizan para filtrar los términos del conjunto de datos a analizar. Si bien esta técnica se encuentra muy difundida y es ampliamente utilizada, genera controversias en cuanto a que tan efectiva es para la tarea de Análisis de Sentimientos, ya que algunas fuentes sugieren que es más recomendable remover los términos poco frecuentes utilizando umbrales de frecuencia mínimos de manera de reducir el espacio de características (Saif, Fernandez, He, & Alani, 2014).

2.4.2 Corrección de errores ortográficos

Otra característica presente en los textos provenientes de las redes sociales es la utilización, ya sea de manera involuntaria o adrede, del lenguaje informal. Dicha característica incrementa el tamaño del vocabulario allí utilizado debido a la introducción de términos nuevos o de variantes léxicas diferentes a las ya presentes en el lenguaje formal. Según (Alegria et al., 2015), la normalización del lenguaje no estándar puede ser crucial para facilitar el procesamiento textual subsiguiente y, en consecuencia, ayudar a mejorar el rendimiento de las herramientas de NLP aplicadas al texto de las redes sociales. En el mismo artículo se comparan herramientas diversas, algunas muy efectivas, para la reducción de tokens fuera de vocabulario (OOV *out-of-vocabulary* por sus siglas en inglés). Sin embargo, no se dan detalles ni se evalúa cómo esa reducción ayuda a mejorar el rendimiento de clasificadores basados en ML.

2.4.3 Stemming y lematización

En general, por cuestiones de redacción, los documentos utilizan diferentes derivaciones de una misma palabra base. Por ejemplo, las palabras *cantamos*, *canto* y *cantas* son conjugaciones del verbo *cantar*. En muchas ocasiones la palabra base y sus derivados tienen un significado similar, por tal motivo suele ser útil agrupar las distintas variantes léxicas para el posterior análisis del texto. Un caso de uso es el de los motores de búsqueda, donde al buscar una palabra se deberían retornar los documentos que contengan la misma y también los que presenten sus derivaciones.

Tanto el *stemming* y la lematización tienen por objetivo transformar las derivaciones de grupos de palabras con el mismo significado a una forma base común. Sin embargo, varían en el método utilizado para lograr ese objetivo.

Los algoritmos de *stemming* cortan el final o el principio de la palabra valiéndose de procesos heurísticos, también de un conjunto de prefijos y sufijos comúnmente utilizados en las palabras derivadas de un lenguaje particular. Este corte puede resultar efectivo en muchos casos, sin embargo, pueden producirse problemas en los casos en los que el principio o el fin de una palabra base contenga los prefijos o sufijos del idioma objetivo. Algunos ejemplos de *stemming* pueden verse en la Tabla 1, obtenidos mediante el SnowballStemmer (Porter, 2001) de NLTK (Natural Language Processing Toolkit por sus siglas en inglés) (Bird, Klein, &

Loper, 2009). Entre los algoritmos más famosos y efectivos para el idioma inglés se encuentran el algoritmo de Porter (Porter, 1980), el de Lovins (Lovins, 1968) y el de Paice/Husk (Paice, 1990).

Por otro lado, la lematización generalmente se refiere a no sólo remover los sufijos comunes que se añaden a las palabras para un determinado lenguaje sino además devolver las mismas a su forma base conocida como *lema*, para ello se valen de un vocabulario y del análisis morfológico de las palabras. Esta diferencia se puede observar en la Tabla 1, para las para las palabras *ser*, *es* y *fue*, mientras que el *stemming* deja las mismas inalteradas el lematizador las convierte al lema *ser*. Los dos también pueden diferir en que el *stemming* comúnmente colapsa palabras relacionadas deliberadamente, mientras que la lematización comúnmente sólo colapsa las diferentes formas flexivas de un lema (Manning, Raghavan, & Schütze, 2009).

Palabra base	**Stem**	**Lema**
nacer	nac	nacer
nació	nac	nacer
naca	nac	naco
corona	coron	corona
coronar	coron	coronar
coronación	coron	coronación
vacaciones	vacacion	vacación
vacacionar	vacacion	vacacionar
vacacionó	vacacion	vacacionar
ser	ser	ser
seres	ser	seres
es	es	ser
fue	fue	ser
arrollo	arroll	arrollo
arrollar	arroll	arrollar

Tabla 1: Ejemplos de stems y lemas para algunas palabras en español.

Debe evaluarse la conveniencia de aplicar tanto lematización como *stemming* en los sistemas de minería de textos y no utilizarlos ciegamente, ya que ambos son procesos que no

están exentos de costos ni tampoco garantizan obtener mejoras en el desempeño. Para el caso de la lematización el costo computacional de aplicarla para todos los datos suele ser muy alto. En cambio, el *stemming* es mucho más rápido en dicho aspecto debido a la simpleza de sus algoritmos, sin embargo, puede relacionar palabras inadecuadamente al recortarlas por demás (conocido como *overstemming*) o por el contrario no detectar relaciones debido a que el recorte no fue suficiente (conocido como *understemming*).

2.4.4 Delimitación del texto

Otro paso fundamental es la delimitación de los textos, esto es la partición del texto crudo en componentes que resulten significativos, como pueden ser capítulos, secciones, párrafos, oraciones, palabras, o secuencias de caracteres. La forma más usual de representar al texto es mediante secuencias de caracteres llamadas tokens, que generalmente se corresponden con las palabras del texto. Este proceso es conocido como tokenización.

Oración original	'Este es un ejemplo de algunas de las distintas formas de tokenización'
bi-gramas de palabras	'Este es', 'es un', 'un ejemplo', 'ejemplo de', 'de algunas', 'algunas de', 'de las', 'las distintas', 'distintas formas', 'formas de', 'de tokenización'
tri-gramas de palabras	'Este es un', 'es un ejemplo', 'un ejemplo de', 'ejemplo de algunas', 'de algunas de', 'algunas de las', 'de las distintas', 'las distintas formas', 'distintas formas de', 'formas de tokenización'
bi-gramas de caracteres	'Es', 'st', 'te', 'e ', ' e', 'es', 's ', ' u', 'un', 'n ', ' e', 'ej', 'je', 'em', 'mp', 'pl', 'lo', 'o ', ' d', 'de', 'e ', ' a', 'al', 'lg', 'gu', 'un', 'na', 'as', 's ', ' d', 'de', 'e ', ' l', 'la', 'as', 's ', ' d', 'di', 'is', 'st', 'ti', 'in', 'nt', 'ta', 'as', 's ', ' f', 'fo', 'or', 'rm', 'ma', 'as', 's ', ' d', 'de', 'e ', ' t', 'to', 'ok', 'ke', 'en', 'ni', 'iz', 'za', 'ac', 'ci', 'ió', 'ón'
tri-gramas de caracteres	'Est', 'ste', 'te ', 'e e', ' es', 'es ', 's u', ' un', 'un ', 'n e', ' ej', 'eje', 'jem', 'emp', 'mpl', 'plo', 'lo ', 'o d', ' de', 'de ', 'e a', ' al', 'alg', 'lgu', 'gun', 'una', 'nas', 'as ', 's d', ' de', 'de ', 'e l', ' la', 'las', 'as ', 's d', ' di', 'dis', 'ist', 'sti', 'tin', 'int', 'nta', 'tas', 'as ', 's f', ' fo', 'for', 'orm', 'rma', 'mas', 'as ', 's d', ' de', 'de ', 'e t', ' to', 'tok', 'oke', 'ken', 'eni', 'niz', 'iza', 'zac', 'aci', 'ció', 'ión'

Tabla 2: Ejemplo de distintas variantes de tokenización generados con el módulo n-gramas de la herramienta NLTK.

Si bien la manera más natural que se puede pensar para tokenizar el texto es que cada token se corresponda con una palabra, existen otras formas de agrupar el texto de entrada. Esto es utilizando *n-gramas* de palabras o *q-gramas* de caracteres con una ventana deslizante para delimitar cada uno de los tokens generados, donde *n* y *q* indican la cantidad de palabras o caracteres que componen cada token respectivamente. Así por ejemplo los *bi-gramas* están compuestos por secuencias de dos palabras o caracteres, los *tri-gramas* por tres y así sucesivamente. En la Tabla 2 puede verse un ejemplo del proceso de tokenización para palabras, *bi-gramas* y *tri-gramas* de palabras y caracteres generados utilizando el módulo *ngrams* de la herramienta NLTK (Bird et al., 2009).

La regla más sencilla para tokenizar un texto en el caso de los *n-gramas* de palabras es, en primer lugar, reemplazar los signos de puntuación por espacios en blanco y, luego, utilizar estos últimos (junto con las tabulaciones y los saltos de línea) como separadores. Sin embargo, muchas veces los textos, sobre todo los generados mediante la interacción en redes sociales, contienen características que hacen que dicha regla no funcione de manera adecuada en todos los casos, algunas de ellas pueden ser:

- Omisiones voluntarias o involuntarias de los espacios en blanco.
- Usos particulares de los símbolos de puntuación, como por ejemplo para delimitar números, fechas, horas, indicar abreviaturas, etc. O también símbolos especiales como puede ser el apóstrofo en el idioma inglés.
- Características propias del texto de la web y las redes sociales como enlaces, direcciones de correo electrónico, nombres de usuario y/o emoticones.
- Expresiones de varias palabras que generalmente se corresponden a entidades nombradas por ejemplo “Ciudad Autónoma de Buenos Aires” o “Catedral de la Inmaculada Concepción de La Plata”.

Estos casos puntuales y otros descritos anteriormente como los errores ortográficos hacen que sea necesario el uso de expresiones regulares en el proceso de tokenización. Las expresiones regulares son un lenguaje que permite caracterizar un conjunto de cadenas y a partir de ello detectar determinados patrones en el texto para darles un tratamiento especial. Durante el proceso de tokenización, algunas de estas expresiones fueron utilizadas en (Tessore, Esnaola, Russo, & Baldassarri, 2019) y (Esnaola, Tessore, Ramon, & Russo, 2019) en primer

lugar para medir su impacto en la reducción de tokens OOV para un lenguaje puntual y en segundo lugar para determinar la incidencia de lo anterior en el desempeño de clasificadores basados en ML, dicho proceso se detalla en los siguientes capítulos.

2.4.5 Etiquetado de las palabras

Por último, dentro de esta subsección de preprocesamiento, es necesario mencionar el etiquetado de las palabras. Si bien estos métodos no han sido incorporados en gran medida para el desarrollo de esta tesis, resulta pertinente mencionarlos debido a su adopción por una parte de los trabajos que se detallan en la bibliografía. Existen varios tipos de etiquetado, los más relevantes son el reconocimiento de entidades nombradas, la desambiguación del sentido de las palabras y el etiquetado de categorías gramaticales o POS *tagging*.

La desambiguación del sentido de las palabras WSD (*word sense disambiguation* por sus siglas en inglés), es la tarea que permite determinar el sentido que debe adoptar una palabra en el contexto de una oración, generalmente este proceso es llevado a cabo por los seres humanos de manera automática sin embargo debe ser abordado de manera específica para algunas aplicaciones del NLP, tales como traducción automática, la recuperación de información o la búsqueda de respuestas entre otras. Para llevar a cabo esta tarea existen tres enfoques principales:

- Basados en conocimiento: en este enfoque se utilizan diccionarios, tesauros y ontologías (siendo la más conocida WordNet) para desambiguar un término valiéndose de las palabras que aparecen en su contexto y de los mencionados recursos.
- Basados en ML supervisado: en los cuales se construye un clasificador, típicamente bayes o una máquina de soporte vectorial, utilizando como etiqueta el sentido del término a desambiguar y como atributos diversas características de las palabras de su contexto (categoría gramatical, lemas, posiciones, entre otros).
- Basados en ML no supervisado: utilizan un algoritmo de *clustering* para agrupar palabras basadas en las similitudes de sus contextos. La idea detrás de esto es que palabras similares ocurren en contextos similares y, en consecuencia, al finalizar el algoritmo, cada grupo representa una categoría distinta, aunque en ausencia de etiquetas.

Un caso particular de desambiguación es el reconocimiento de entidades nombradas. Esto se refiere a detectar palabras o secuencias de palabras y clasificarlas en caso de ser nombres de sitios, personajes, empresas, organizaciones gubernamentales o no gubernamentales, etc. Un buen proceso de reconocimiento de entidades nombradas debe:

- Agrupar todas las palabras que corresponden a una entidad, ya sea por ejemplo "Laguna natural de Gómez" o "Teatro de la Ranchería", ambas compuestas por secuencias de palabras.
- Ser capaz de lidiar con las ambigüedades que surgen en los casos que distintas entidades compartan el mismo lexema. Por ejemplo, "Según el último censo Junín tiene 90.305 habitantes" y "El ejército unido de los patriotas obtuvo una importante victoria en Junín posibilitando la independencia del Perú". En el primer caso, el lexema Junín, se refiere a una ciudad en el noroeste de la Provincia de Buenos Aires, mientras que en el segundo a una batalla desarrollada en el primer cuarto del siglo XIX.

El último tipo de etiquetado a mencionar es el de categoría gramatical o POS *tagging*, este permite determinar qué rol tiene una palabra en el contexto de una oración, esto es, si actúa como verbo, sustantivo, adjetivo, adverbio, etc. Las etiquetas a asignar dependen del *tagset* adoptado, existen varios con mayor o menor nivel de detalle, siendo el más conocido el de Penn Treebank que cuenta con 45 etiquetas. Existen distintos tipos de algoritmos para llevar a cabo esta tarea, los más antiguos son los basados en reglas, dentro de los cuales se encuentra el *tagger* de Brill, también se encuentran los enfoques estocásticos que se basan en la frecuencia, probabilidades y estadísticas de aparición de las palabras. Una clasificación detallada de los algoritmos de POS *tagging* puede encontrarse en (Kumawat & Jain, 2015).

2.5 Esquemas de representación de contenido

Una vez realizado el pre procesamiento correspondiente sobre los documentos, se deben seleccionar las características relevantes de los mismos que sirven como entrada a los algoritmos utilizados en los sistemas de minería de textos. Estas características dependen de la tarea a resolver y del algoritmo seleccionado y se dividen en dos grandes grupos estáticas y dinámicas (Layton, Watters, & Dazeley, 2012). Las primeras de ellas son establecidas a priori y se basan, generalmente, en estadísticas acerca de las categorías gramaticales, longitudes de

las palabras, interacciones entre los usuarios (en caso de datos provenientes de redes sociales), etc. Por otro lado, las características dinámicas se encuentran de manera literal en el contenido y surgen como consecuencia de su procesamiento, a partir de ellas surgen representaciones como la bolsa de palabras (BoW *Bag of Words* por sus siglas en inglés), de n-gramas de palabras y q-gramas de caracteres, entre otras. En el resto de la subsección se abordan los esquemas de representación que surgen a partir de las características dinámicas, estos son ampliamente utilizados para la tarea de Análisis de Sentimientos.

2.5.1 Esquemas basados en bolsas de palabras/caracteres

La representación basada en bolsas de palabras o caracteres utiliza como características los tokens generados en la sección 2.4.4 de delimitación del texto, estos pueden ser palabras, n-gramas de palabras o q-gramas de caracteres. A partir de ellos construye una representación matricial dispersa en donde las filas representan cada uno de los documentos del conjunto de datos a analizar, mientras que las columnas son los tokens, una por cada token de la colección de documentos. Dentro de cada celda se encuentra el peso del token para el documento en cuestión, tal y como se presenta gráficamente en la Figura 8. El peso asignado a cada término depende de la métrica utilizada para tal fin.

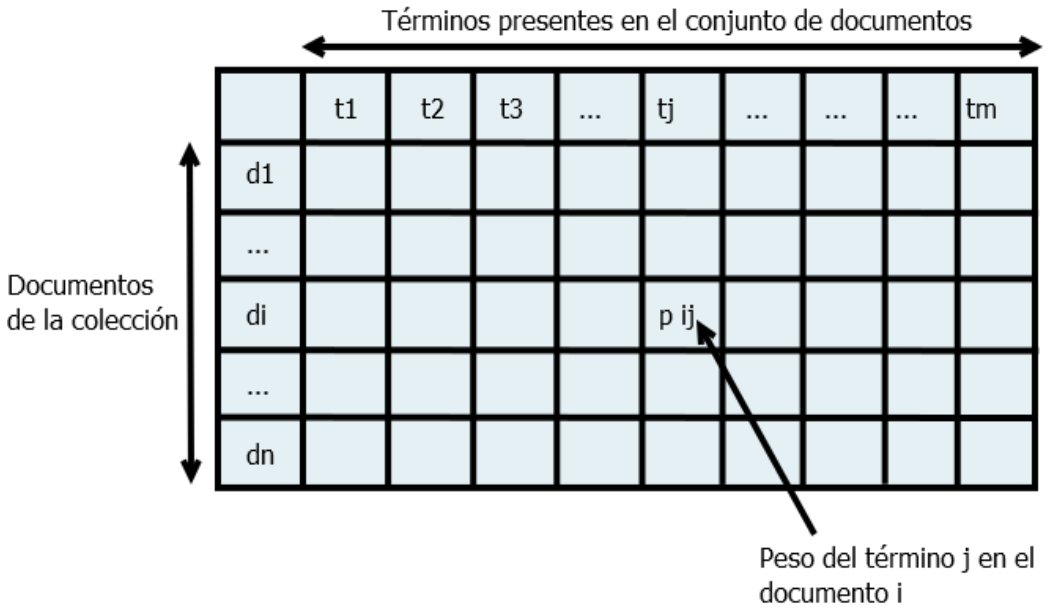


Figura 8: Representación de bolsa de palabras para un conjunto de datos de n documentos y m términos.

2.5.2 Esquemas de ponderación del contenido

La manera más básica de asignar los pesos a la matriz es considerar la presencia o ausencia de un término dentro de un determinado documento, también conocida como pesado binario. Como ejemplo de lo anterior se puede ver la bolsa de palabras de la Tabla 3 que se genera a partir del siguiente conjunto de documentos (texto convertido previamente a minúsculas):

1. "Ayer fue el día de lluvia. De lluvia intensa."
2. "La hizo bajo el sol de ayer."
3. "El día de ayer viajé de hecho."

	D1	**D2**	**D3**
ayer	1	1	1
fue	1	0	0
el	1	1	1
día	1	0	1
de	1	1	1
lluvia	1	0	0
intensa	1	0	0
la	0	1	0
hizo	0	1	0
bajo	0	1	0
sol	0	1	0
viajé	0	0	1
hecho	0	0	1

Tabla 3: Ejemplo de bolsa de palabras con pesado binario para un conjunto de tres documentos y trece términos.

El pesado binario, sin embargo, no captura la frecuencia con la que aparecen los términos en el documento ni en la colección de los mismos, lo cual podría representar un problema si se supone que la cantidad de veces que se encuentra un término en un documento refleja su importancia en él. Por ello, otra métrica a considerar es construir la matriz utilizando como peso la cantidad de veces que un término aparece en un determinado documento (TF *term frecuency* por sus siglas en inglés). El resultado para la misma colección de documentos puede

verse en la Tabla 4. En la práctica se suele usar la máxima frecuencia de un término en cada documento como factor de normalización, es decir para evitar favorecer a los documentos más largos, por motivos de simpleza esto se dejó fuera del ejemplo.

	D1	D2	D3
ayer	1	1	1
fue	1	0	0
el	1	1	1
día	1	0	1
de	2	1	1
lluvia	2	0	0
intensa	1	0	0
la	0	1	0
hizo	0	1	0
bajo	0	1	0
sol	0	1	0
viajé	0	0	2
hecho	0	0	1

Tabla 4: Ejemplo de bolsa de palabras con ponderación frecuencia de términos para un conjunto de tres documentos y trece términos.

Si bien esta ponderación resalta la importancia del término “lluvia” para el documento 1, también lo hace con el término “de” el cual seguramente aparece frecuentemente en una colección más grande y a su vez es de escasa importancia para sus respectivos documentos. Para solucionar este problema, surge la idea de que la importancia de un término para un determinado documento disminuye a medida que aumenta el número de documentos en la colección que contienen dicho término, en otras palabras, cuanto más común es un término en una colección de documentos menos útil es para discriminar las clases de esos documentos. La ponderación que surge a partir de tal razonamiento es llamada frecuencia de término frecuencia inversa de documento (TF-IDF *term frecuency inverse document frecuency* por sus siglas en inglés). Y se calcula de la siguiente manera:

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D)$$

donde t es un término, d es un documento, D es la colección de documentos, $tf(t,d)$ es la frecuencia de un término para un determinado documento e $idf(t,D)$ es la frecuencia inversa de un término para la colección de documentos. Esta última se calcula de la siguiente forma:

$$idf(t,D) = log\frac{|D|}{|\{d \in D : t \in d\}|}$$

donde $|D|$ es la cantidad de documentos en la colección y $|\{d \in D : t \in d\}|$ es el número de documentos que contienen el término t. Nuevamente, en la práctica suelen utilizarse factores de normalización que no se detallan por motivos de sencillez.

	D1	**D2**	**D3**
ayer	0,21	0,26	0,23
fue	0,34	0	0
el	0,2	0,26	0,23
día	0,26	0	0,30
de	0,40	0,26	0,23
lluvia	0,68	0	0
intensa	0,34	0	0
la	0	0,45	0
hizo	0	0,45	0
bajo	0	0,45	0
sol	0	0,45	0
viajé	0	0	0,78
hecho	0	0	0,39

Tabla 5: Ejemplo de bolsa de palabras con ponderación frecuencia de término frecuencia inversa de documento para un conjunto de tres documentos y trece términos.

De esta manera se privilegia a los términos que aparecen muchas veces en pocos documentos debido a que el logaritmo del cociente se aproxima a 0 a medida que un término aparece en más documentos de la colección. El resultado puede verse en la Tabla 5. Esta matriz fue calculada con la herramienta scikit-learn (Buitinck et al., 2013), las diferencias entre los valores que se obtendrían con la fórmula presentada anteriormente y los finalmente expuestos, se deben a pequeñas variaciones en la forma de cálculo y la utilización de factores de

normalización. En la matriz resultante, se puede observar que, para el documento uno, el término “lluvia” tiene una mayor ponderación que “de” a pesar de que ambos tienen dos ocurrencias para ese documento, esto se debe a que “de” aparece en todos los documentos de la colección y es penalizado por ello. Por otro lado, los términos “sol” y “viajé” se encuentran entre los más relevantes para los documentos dos y tres respectivamente, debido a que solo se aparecen en ellos.

Hasta aquí se revisó el esquema de BoW y las ponderaciones más ampliamente utilizadas para el mismo. Esta representación contiene algunos inconvenientes, entre los cuales se pueden mencionar:

- Al aumentar la cantidad de documentos así lo hace el vocabulario de la colección resultando en vectores más extensos.
- Los vectores que se utilizan para representar los documentos contienen una gran cantidad de ceros, lo cual hace que la matriz que representa la colección sea extremadamente dispersa, esto no es ideal si se tiene la intención de utilizarla para entrenar algoritmos de ML.
- No almacena información acerca del orden, coocurrencia y otras relaciones semánticas entre los términos.

Para intentar abordar estos inconvenientes se han desarrollado variantes de representación de texto densas, siendo las más difundidas son Word2Vec y GloVe, que se abordan en la siguiente subsección.

2.5.3 Esquemas basados en neural embeddings

El modelo **Word2Vec** (Mikolov et al., 2013) se utiliza para producir *embeddings* de palabras densos utilizando un procedimiento novedoso. En lugar de contar las ocurrencias como hacen los enfoques revisados en la sección anterior o ratios de coocurrencias como lo hace GloVe (Pennington et al., 2014), este modelo entrena una red neuronal con la intención de predecir palabras a partir de su contexto, este clasificador posteriormente se descarta y se utilizan los pesos de la capa oculta de la red como *embeddings* de las palabras. Esta particular forma de producir los *embeddings* permite capturar las relaciones sintácticas y semánticas entre las palabras, una consecuencia de ello es la posibilidad de realizar operaciones aritméticas entre

los vectores para resolver analogías de palabras de la forma "A es a B lo que C es a X" usando aritmética simple o como "Rey - Hombre + Mujer = Reina". Estas relaciones pueden verse en las proyecciones en el plano de los *embeddings* correspondientes a las distintas palabras, como se muestra en la Figura 9-a y la Figura 9-b.

a)

b)

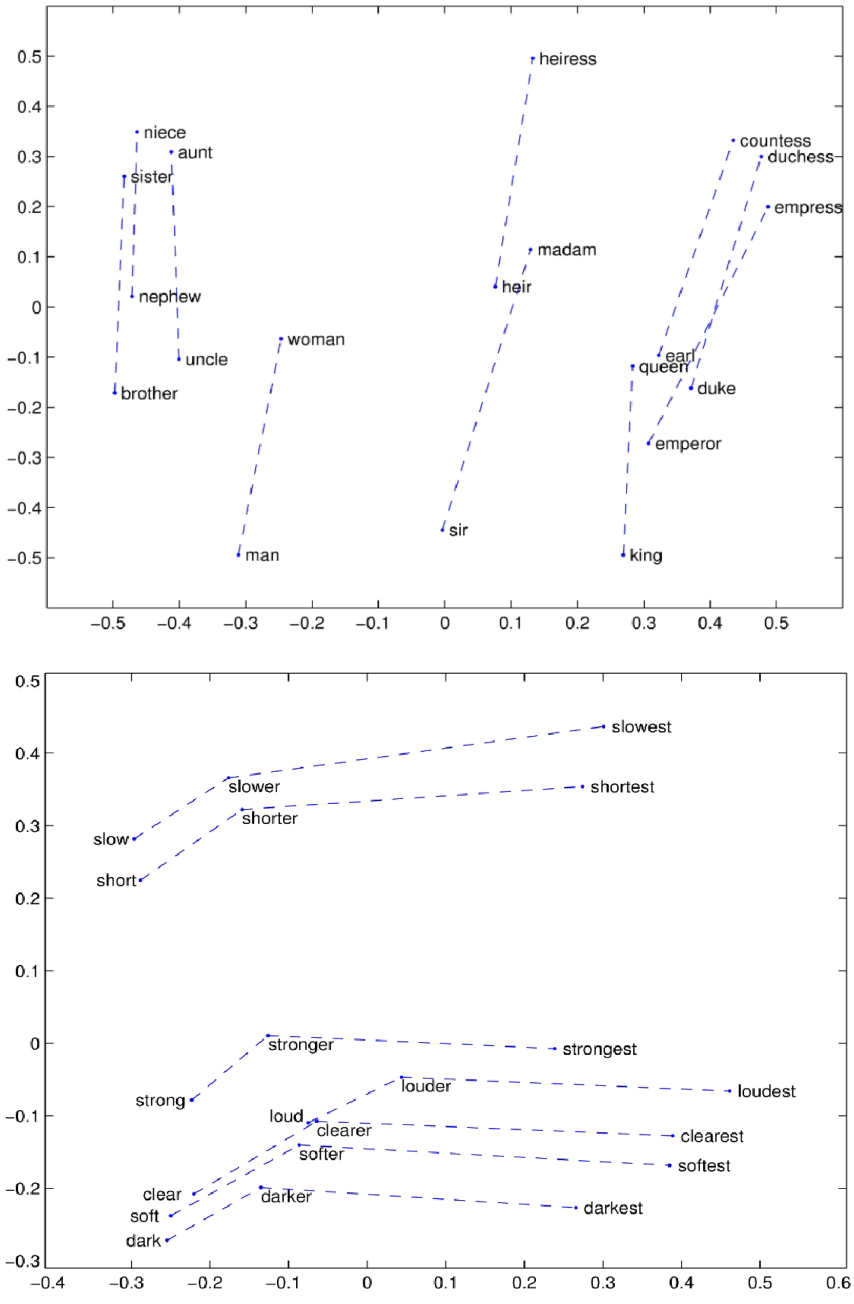


Figura 9: Proyección al plano de distintos vectores de embeddings de Word2Vec. Extraído de: (Jurafsky & Martin, 2023).

Por otro lado, la dimensión de los vectores de *embeddings*, que es un hiper parámetro igual a la cantidad de neuronas de la capa oculta de la red, tiene siempre un tamaño relativamente pequeño y flexible con respecto a los esquemas vistos en la sección anterior en los cuales el tamaño de los vectores aumentaba linealmente con respecto al tamaño del vocabulario.

Cabe aclarar también que la generación de los datos de entrenamiento no requiere etiquetado manual, dado que las palabras y sus contextos son extraídos automáticamente a partir de los textos de entrada, concretamente analizando la vecindad de cada palabra delimitada por otro hiper parámetro llamado ventana.

Existen dos arquitecturas para obtener los *embeddings* de Word2Vec:

- *Continuous bag-of-words* (CBOW): en esta variante se construye un clasificador para predecir una palabra a partir de su contexto.
- Skip-Gram: en este caso, a diferencia del anterior se intenta predecir el contexto a partir de una determinada palabra clave.

En ambos modelos el clasificador se entrena generalmente no con la intención de utilizarlo sino la de extraer la matriz de pesos de su capa oculta de la cual se obtienen los vectores que sirven como *embeddings* de las palabras. En general, se utiliza Skip-Gram para corpus de gran tamaño, aunque el entrenamiento es más lento y CBOW se usa para los más pequeños, pero es más rápido. A continuación, se describe brevemente la arquitectura de Skip-Gram, la cual puede verse en la Figura 10.

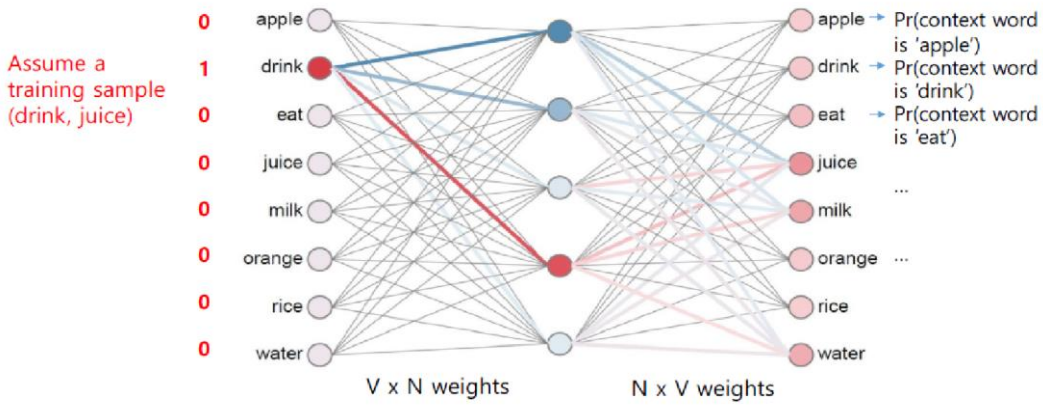


Figura 10: Arquitectura de un clasificador SkipGram utilizado para generar los embeddings de Word2Vec, Fuente: (Ghosh, 2020).

Para construir un clasificador SkipGram, en primer lugar, se generan los datos de entrenamiento a partir de un corpus de texto, del cual se extraen tuplas de la forma (P, C) donde P es una palabra y C es su contexto, compuesto por una o más palabras cuya cantidad está delimitada por un hiper parámetro llamado ventana. Las palabras se codifican en vectores *one-hot*, es decir vectores de longitud V (tamaño del vocabulario) con solo un número uno y luego ceros, siendo uno de estos vectores lo que recibe la capa de entrada de la red. El algoritmo utilizado para realizar el entrenamiento es *back propagation*.

Luego se encuentra la capa oculta, de la cual se extraen los *embeddings* una vez entrenado el clasificador. Está compuesta por N neuronas (lo cual determina el tamaño de los vectores de *embeddings*), la matriz de pesos es en consecuencia de dimensiones V x N. La capa no cuenta con función de activación con lo cual su salida es lineal. Para extraer el *embedding* de una palabra puntual, se realiza el producto vectorial entre el vector *one-hot* correspondiente a dicha palabra (que funciona como un índice) y la matriz de pesos mencionada.

Por último, la capa de salida tiene una dimensión de N x V. Cada neurona de esta capa recibe un vector de tamaño N correspondiente al *embedding* de la palabra contextual (en el ejemplo el contexto está compuesto por solo una palabra) y lo multiplica contra su propio vector de pesos produciendo un escalar. Finalmente, a las salidas de las neuronas de esta capa se les aplica la función de activación *softmax* que convierte los valores en probabilidades. Idealmente, al finalizar el entrenamiento de la red, la salida de una única neurona tiene una probabilidad alta y se corresponde con la palabra contextual correspondiente.

Una vez finalizado el entrenamiento, se descarta prácticamente todo el clasificador y se extrae la matriz de pesos de la capa oculta que contiene los vectores que se utiliza como *embeddings* de las palabras del vocabulario V.

GloVe fue desarrollado por (Pennington et al., 2014) argumentando que el enfoque utilizado por Word2Vec es subóptimo ya que no explota completamente la información estadística con respecto a las coocurrencias de palabras. Este enfoque combina la utilización de contextos locales mediante una ventana, de la misma manera que Word2Vec, y una matriz de coocurrencia entre las palabras. Los autores argumentan que los vectores globales combinan lo mejor de ambos mundos, es decir, hacer uso de las estadísticas de coocurrencia de todo el corpus mientras se capturan relaciones lingüísticas complejas.

La novedad de GloVe con respecto a otros métodos preexistentes de factorización de matrices, como por ejemplo Análisis de Semántica Latente (LSA *Latent Semantic Analysis* por sus siglas en inglés), es que utiliza ratios de probabilidades de coocurrencia en lugar de probabilidades de coocurrencia propiamente dichas. Un ejemplo de esto puede verse en la Figura 11.

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Figura 11: Ejemplo de probabilidades y ratios en GloVe. Extraído de (Pennington et al., 2014).

Como puede verse en dicha figura, si P(*k* | *w*) es la probabilidad de que la palabra *k* aparezca en el contexto de la palabra *w*, pueden darse las siguientes situaciones:

- Una palabra *k* fuertemente relacionada con *w1* pero no con *w2* en cuyo caso la proporción P(*k* | *w1*) / P (*k* | *w2*) es grande, como es el caso para *k=solid*, *w1=ice* y *w2=steam*.
- Una palabra *k* fuertemente relacionada con *w2* pero no con *w1*, en cuyo caso la proporción P(*k* | *w1*) / P (*k* | *w2*) es pequeña, como es el caso para *k=gas*, *w1=ice* y *w2=steam*.
- Una palabra *k* fuertemente relacionada tanto con *w1* como con *w2*, en cuyo caso la proporción P(*k* | *w1*) / P (*k* | *w2*) es cercana a 1, como es el caso para *k=water*, *w1=ice* y *w2=steam*.
- Una palabra *k* débilmente relacionada tanto con *w1* como con *w2*, en cuyo caso la proporción P(*k* | *w1*) / P (*k* | *w2*) es cercana a 1, como es el caso para *k=fashion*, *w1=ice* y *w2=steam*.

Según (Pennington et al., 2014), debido a que estos ratios pueden codificar alguna especie de significado, esta información también se codifica como diferencias entre los vectores. Por esta razón, los vectores de palabras resultantes funcionan muy bien en tareas de analogías de palabras. Sin embargo, según (Agrawal & Suri, 2019), este tipo de características posee

algunas limitaciones, entre ellas se puede mencionar la falta de robustez ante la presencia de errores ortográficos y que no permiten capturar el significado de los tokens OOV de manera efectiva. Con el objetivo de superar dichos inconvenientes, los autores proponen utilizar una combinación de características neurales y léxicas. Un estudio de ablación realizado en dicho trabajo, demostró que los n-gramas de caracteres son las características que, individualmente, más ganancia aportan al desempeño del modelo propuesto.

2.6 Algoritmos de clasificación de textos basados en ML

En la presente sección se describen tres algoritmos para la clasificación de textos basados en ML que se utilizan en el trabajo experimental correspondiente a la tesis. En particular, se presentan las máquinas de soporte vectorial (SVM *support vector machines* por sus siglas en inglés), un algoritmo que busca maximizar la distancia entre el hiperplano de separación y las muestras más cercanas a él; el clasificador *Naïve Bayes* (NB), que utiliza el teorema de Bayes con la suposición "ingenua" de independencia condicional entre cada par de características dado el valor de la variable de clase; y las redes neuronales artificiales (ANN *artificial neural networks* por sus siglas en inglés), que se basan en la estructura de las neuronas del cerebro humano y pueden tener una o varias capas de neuronas. Estos algoritmos son útiles para clasificar y categorizar grandes cantidades de datos de texto, lo que los hace aplicables en áreas como la minería de textos.

2.6.1 Máquinas de soporte vectorial

Las SVM tienen su origen en los trabajos de V. Vapnik (V. N. Vapnik, 1999) (V. Vapnik, 2006). Es un algoritmo de ML supervisado que se aplica para resolver problemas de clasificación, regresión y detección de valores fuera de rango.

Este algoritmo no sólo intenta encontrar un hiperplano que permita separar las clases a categorizar, sino que también busca maximizar la distancia entre el hiperplano que las separa y las muestras más cercanas a él, estas últimas llamadas vectores de soporte. En la Figura 12 puede verse un diagrama de esto para R^2, al tratarse del plano la separación en este caso está dada por una recta. La intuición de detrás de las SVM es que mientras mayor sea la distancia

entre el hiperplano y los vectores de soporte menos probable es que una futura muestra a clasificar caiga del lado equivocado del hiperplano trazado.

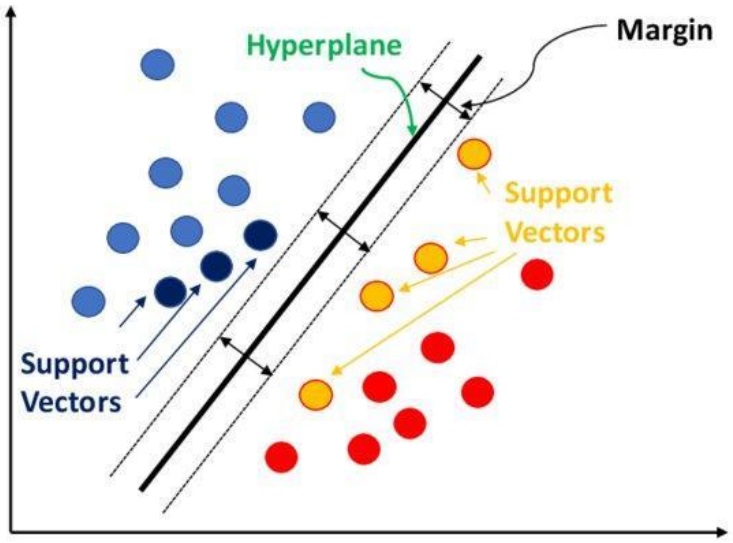


Figura 12: Hiperplano de separación (recta) y muestras utilizadas como vectores de soporte. Fuente: (Manjrekar & Dudukovic, 2019).

Si bien las SVM fueron originalmente concebidas para problemas de clasificación lineales y binarios puede adaptarse para ser utilizadas en problemas no lineales (se logra en este caso aumentando la dimensionalidad del espacio de características) y multiclase (descomponiendo el problema multiclase en múltiples problemas de clasificación binaria, llamado enfoque uno a uno, o construyendo un clasificador para cada clase que la separe de las demás llamado enfoque uno contra el resto).

2.6.2 Naïve Bayes

El clasificador NB es un algoritmo de ML supervisado basado en la aplicación del teorema de Bayes con la suposición "ingenua" de independencia condicional entre cada par de características dado el valor de la variable de clase. A partir de esto, si se tienen k clases $y_1, y_2, \dots, y_k$, junto un vector de n características $x = (x_1, x_2, \dots, x_n)$, se quiere encontrar la clase y_i que maximiza la siguiente ecuación:

$$P(y_i | \, x_1, x_2 \dots, x_n) = \frac{P(x_1, x_2, \dots, x_n, y_i)}{P(x_1, x_2, \dots, x_n)}$$

Siendo $P(y_i \mid x_1, x_2 \dots, x_n)$ la probabilidad de y_i dados $x_1, x_2 \dots, x_n$. Mientras que $P(x_1, x_2, \dots, x_n, y_i)$ es la probabilidad de que $x_1, x_2, \dots, x_n, y_i$ ocurran al mismo tiempo y $P(x_1, x_2, \dots, x_n)$ de que $x_1, x_2, \dots, x_n$ ocurran al mismo tiempo. Si se supone que las características de entrada $x_1, x_2, \dots, x_n$ son independientes la ecuación anterior puede transformarse en la siguiente:

$$P(y_i \mid x_1, x_2 \dots, x_n) = \frac{P(y_i)P(x_1|y_i)P(x_2|y_i) \dots P(x_n|y_i)}{P(x_1, x_2, \dots, x_n)}$$

En este caso $P(x_n|y_i)$ es la probabilidad de x_n dado y_i. Para calcular estas probabilidades es necesario contar todas las entradas que tienen la característica x_n en las cuales se da y_i y se divide sobre el total de entradas en las que se da y_i. Luego de calcular las probabilidades para cada valor de i se selecciona el y_i que maximiza $P(y_i \mid x_1, x_2 \dots, x_n)$.

2.6.3 Redes neuronales

Las ANN se basan en las estructuras neuronales análogas que se encuentran en el cerebro de los seres vivos. El modelo de neurona o nodo actualmente utilizado se basa la siguiente fórmula:

$$y = g(\sum_{i=1}^{n} w_i a_i + b)$$

En donde g actúa como función de activación, b es un sesgo que se incorpora llamado *bias*, a_i es una señal de entrada, w_i es el peso asociado a la señal anterior, y n es el número de entradas de la neurona.

Las neuronas o unidades descritas se agrupan en capas para formar la red neuronal, cada capa puede tener una o varias neuronas. Las ANN tradicionales cuentan con tres capas: una capa de entrada, una oculta y finalmente una capa de salida. En el caso que la ANN posea más de una capa oculta, entonces se trata de una red neuronal profunda.

Un tipo especial son las redes neuronales recurrentes (RNN *recurrent neural networks* por sus siglas en inglés) (Rumelhart, Hinton, & Williams, 1986). Las RNN permiten procesar y obtener información de datos secuenciales, en particular son útiles para el NLP, esto se debe a que permiten capturar las dependencias secuenciales y temporales de los datos de entrada.

Las RNN agregan ciclos que conectan nodos adyacentes y actúan como una especie de memoria de la red que se utiliza para incorporar datos del pasado en la evaluación de las propiedades del dato actual. Un diagrama de estas redes puede verse en la Figura 13.

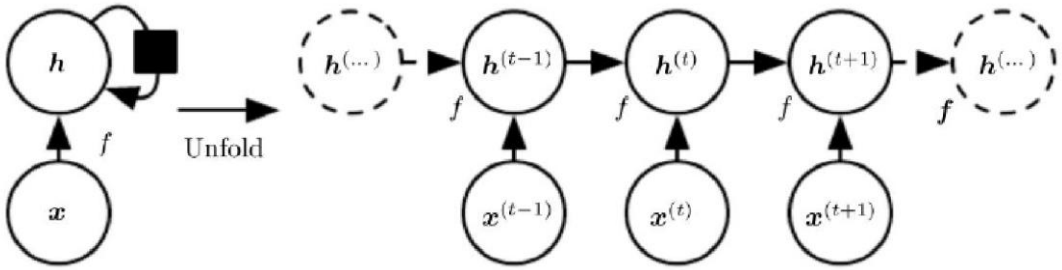


Figura 13: Diagrama de una RNN. Fuente: (Bengio, Goodfellow, & Courville, 2015).

La salida un nodo de la red es función de su entrada x_t y de los datos históricos que recibe h_{t-1}, la fórmula sería la siguiente:

$$h_t = f(h_{t-1}, x_t)$$

Un problema que afecta en general a las RNN es el del desvanecimiento del gradiente, esto hace que este tipo de redes tengan problemas para recordar patrones muy extendidos en el tiempo debido a que a medida que la red procesa más elementos en la cadena se le dificulta recordar información pasada. Para solucionar este inconveniente, se desarrollaron las redes neuronales recurrentes de memoria de corto y largo plazo (LSTM *long short term memory* por sus siglas en inglés). La novedad de estos nodos es que incorporan tres compuertas: una de entrada, una de olvido y una de salida. Un diagrama de una celda LSTM y sus compuertas puede verse en la Figura 14.

La compuerta de olvido se encarga de recordar solo algunas partes de la memoria a largo plazo y decide qué recordar en función de la entrada actual y memoria que se recibió del paso anterior. Por otro lado, la compuerta de entrada recuerda solo algunas partes de la entrada actual y memoria de trabajo anterior y decide qué recordar en función de la entrada actual y memoria que se recibió del paso anterior. En función de las dos compuertas anteriores se actualiza la memoria de largo plazo. Finalmente, la compuerta de salida decide que partes de la memoria de corto plazo se recuerdan y se pasan a la próxima iteración.

Las LSTM son ampliamente utilizadas en diversos problemas de clasificación de texto como es el caso del Análisis de Sentimientos y se utilizan en parte del desarrollo del presente trabajo de tesis.

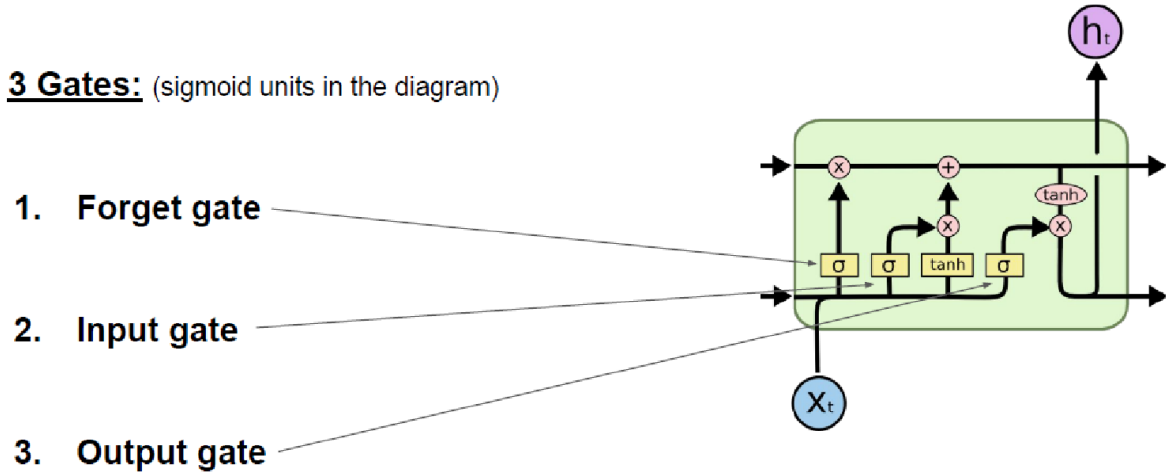


Figura 14: Celda LSTM indicando sus compuertas de entrada, olvido y salida. Fuente: (Voleti, 2017).

3. Estado de la cuestión

Contenido

Resumen

Las tareas de Análisis de Sentimientos y/o Clasificación de Emociones necesitan de diversos recursos para ser llevadas a cabo, entre los más importantes se encuentran los conjuntos de datos etiquetados. En el presente capítulo se revisa el estado del arte en lo que respecta a la construcción de conjuntos de datos para el Análisis de Sentimientos y/o Clasificación de Emociones para dos estrategias, la primera de ellas es el etiquetado manual y la segunda para el automático y/o semi automático. Para el segundo caso, se analizan las metodologías utilizadas para la recopilación de los recursos junto con la presencia o ausencia de una etapa de validación del contenido recopilado. Por otro lado, la aplicación de métodos automáticos o semi automáticos suele permitir capturar abundante información contextual, debido a esto, también se analiza la utilización de la misma para las tareas de Análisis de Sentimientos y/o Clasificación de Emociones. El capítulo se organiza de la siguiente manera, en primer lugar, se revisan diversos conjuntos de datos etiquetados para el Análisis de Sentimientos en la sección 3.2, ya sea los recopilados mediante etiquetado manual (sección 3.2.1) o los construidos mediante etiquetado automático (sección 3.2.2). En la sección 3.3 se presentan diversas métricas para la comparación de diversos conjuntos de datos. Luego, en la

sección 3.4, se revisan métricas de comparación y medición del nivel de consenso. Posteriormente, en la sección 3.5, se aborda el uso de la información contextual capturada para la mejora en el desempeño de clasificadores basados en ML. Por último, en la sección 3.6, se presentan las conclusiones del capítulo.

3.1 Introducción

La aplicación del ML para diversas tareas ha tenido también su impacto en el Análisis de Sentimientos. Muchas de las tareas de clasificación incluidas en el Análisis de Sentimientos basadas en ML utilizan aprendizaje supervisado, dicha variante requiere de recursos etiquetados para el entrenamiento de los algoritmos. Para el caso de los sistemas que intentan realizar Análisis de Sentimientos y/o Clasificación de Emociones, utilizando dichos algoritmos, la disponibilidad de los conjuntos de datos requeridos es dispar, dependiendo fuertemente del idioma. Ante la escasez de recursos para determinados idiomas, entre ellos el español, distintos investigadores los han construido de manera ad-hoc utilizando dos estrategias. La primera de ellas es recurrir al etiquetado manual de los textos, mientras que la segunda es realizar un etiquetado automático o semiautomático utilizando marcadores ya presentes en los datos.

En este capítulo se presentan y analizan distintos conjuntos de datos para el Análisis de Sentimientos presentes en la literatura. En primer lugar, se discuten los construidos mediante etiquetado manual, para luego detenerse en los recopilados mediante etiquetado automático o semi automático. Se incluyen distintas métricas relevantes de los conjuntos de datos informadas por sus autores, tales como cantidad de muestras, tamaño del vocabulario, número de clases, necesidad de etiquetadores en el proceso de construcción junto con la presencia o ausencia de una etapa de validación adecuada, lo último es particularmente relevante para los *datasets* construidos mediante supervisión distante, debido la necesidad de establecer la calidad de los conjuntos de datos generados. Para el análisis anterior no solo se tienen en cuenta conjuntos de datos con textos en español, sino también en otros idiomas que puedan contener características relevantes o novedosas con respecto al contenido o al proceso de construcción. También se pone énfasis en los tipos de información contextual (IC) recopilada en los distintos conjuntos de datos presentados y su utilización para obtener mejoras en los algoritmos de clasificación de Análisis de Sentimientos, principalmente en los basados en ML.

El objetivo del capítulo es identificar la disponibilidad de recursos de Análisis de Sentimientos de emoción básica, principalmente en español, como así también las metodologías adoptadas por los investigadores para el proceso de construcción y validación de

los mismos. Asimismo, verificar el impacto del uso de la IC en los distintos trabajos presentes en la literatura.

3.2 Conjuntos de datos para el Análisis de Sentimientos

Los algoritmos de ML supervisado requieren de datos etiquetados para su entrenamiento y posterior evaluación. Según (Yadollahi et al., 2017) un problema en general que se presenta con el etiquetado de conjuntos de datos para el Análisis de Sentimientos, ya sea polar o de emoción básica, es que suelen estar condicionados por las apreciaciones subjetivas de los etiquetadores. Por otro lado, el etiquetado manual suele ser una tarea lenta y costosa motivo que ha entorpecido la creación de este tipo de recursos para Análisis de Sentimientos en lenguajes distintos del inglés. Por ello, los investigadores enfocados en la construcción de recursos para Análisis de Sentimientos han adoptado estrategias que permitan etiquetar el contenido de manera automática o semiautomática, utilizando marcadores ya presentes en los datos, de manera de acelerar el proceso y reducir el impacto de las apreciaciones subjetivas de los etiquetadores humanos externos, este tipo de etiquetado es conocido como supervisión distante (DS *distant supervision* por sus siglas en inglés).

En las próximas subsecciones se revisan algunos de los conjuntos de datos etiquetados más relevantes, ya sea de polaridad o emoción básica.

3.2.1 Construidos mediante etiquetado manual

Los textos de la mayoría de los exponentes de esta subsección son en idioma inglés ya sea polares o de emoción básica, esto se debe a que se le ha dedicado mayor interés y recursos a lo largo de los años. En (Yadollahi et al., 2017), (Sailunaz, Dhaliwal, Rokne, & Alhajj, 2018) y (Yadav & Vishwakarma, 2020) se presentan varios de los principales referentes.

A continuación, se detallan los principales conjuntos de datos que utilizan ***etiquetas polares***, estas indican si el contenido tiene una connotación positiva o negativa. Existen también casos de estos conjuntos de datos donde se aporta una mayor granularidad, por ejemplo, sumando las categorías *muy positivo*, *muy negativo*, *neutral*, etc.

(Melville, Gryc, & Lawrence, 2009) recopiló un conjunto de publicaciones en blogs en inglés de dos temas, el primero acerca de revisiones de productos y el segundo corresponde a

comentarios políticos. En total contiene 252 posts y fue clasificado manualmente con las etiquetas positivo, negativo, neutral e irrelevante.

El conjunto de datos utilizado en la competencia SemEval-2013 (Nakov et al., 2013) está compuesto por un conjunto de tweets y mensajes de textos en inglés. Los mismos fueron etiquetados manualmente por cinco colaboradores utilizando la herramienta *Amazon Mechanical Turk*. Cada elemento fue clasificado de manera redundante por los cinco etiquetadores, asignándose finalmente la clase más votada resultante. Está compuesto por 6.302 elementos positivos, 2.801 negativos y 8.187 neutrales.

El *Stanford Sentiment Treebank* (Socher et al., 2013) es un *dataset* de árboles de análisis sintáctico completamente etiquetados. Este permite analizar los efectos composicionales del sentimiento en el lenguaje. Incluye un total de 215.154 oraciones en inglés, cada una anotada por tres jueces humanos como positivo, negativo o neutral.

Gran parte de los trabajos que realizan Análisis de Sentimientos en español se basan en conjuntos de datos construidos con etiquetado manual. Tal es el caso de la mayoría de los conjuntos de datos proporcionados o presentados en el Taller de Análisis de Sentimientos en español (TASS) ("Taller de Análisis de sentimientos en Español (TASS)," n.d.), su conjunto de datos original, contiene más de 70.000 tweets, en español de distintas personalidades y celebridades del mundo de la política, la economía, la comunicación, los medios de comunicación y la cultura, entre noviembre de 2011 y marzo de 2012. Cada mensaje se encuentra etiquetado con su polaridad global en cinco niveles: fuerte positivo (P+), positivo (P), neutral (NEU), negativo (N), fuerte negativo (N+) y una etiqueta adicional sin sentimiento (NONE).

También el conjunto de datos utilizado en la competencia IberLef 2019 (Díaz-Galiano et al., 2019), que recopiló tweets con diferentes variantes del español. En total, este último conjunto de datos contiene un total de 7.294 comentarios pertenecientes a las variantes uruguaya, mexicana, española, costarricense y peruana del idioma. Cada texto fue etiquetado por expertos en cuatro categorías: Positivo, Negativo, Neutral o Ninguno.

Los siguientes conjuntos de datos a mencionar, a diferencia de los discutidos hasta este punto, se encuentran ***etiquetados con diversas emociones básicas***.

Uno de los exponentes más antiguos entre *datasets* de emoción básica en idioma inglés es ISEAR (Scherer & Wallbott, 1994), para el mismo se solicitó a los participantes que contaran experiencias en las que habían vivido alguna de las emociones básicas pre seleccionadas (alegría, miedo, enojo, tristeza, asco, vergüenza y culpa). Cada muestra, por tanto, se compone de un párrafo etiquetado con su correspondiente emoción básica, y contiene un total de 7.666 textos etiquetados. Alrededor de 3.000 personas participaron en la clasificación de los textos. Se lo considera un *dataset* altamente confiable debido a que el contenido fue etiquetado por su mismo autor (Yadollahi et al., 2017).

Otro referente importante en este campo es el *dataset Fairy Tales* (Alm & Sproat, 2005), compuesto por 185 cuentos para niños, en el cual cada una de las 15.000 oraciones de los cuentos fueron etiquetadas con alguna de las emociones básicas seleccionadas (enojo, asco, miedo, felicidad, tristeza, sorpresa positiva, sorpresa negativa y neutral). Las emociones fueron asignadas por un grupo de seis personas, todos hablantes de idioma inglés.

La competencia Semeval, en sus distintas ediciones, presentó diversos *datasets* de emoción básica. Uno de ellos, introducido en 2007, (Strapparava & Mihalcea, 2007) compuesto por títulos de noticias en inglés extraídas de *Google news* y otros portales relevantes, en total consiste de 1.250 títulos, 250 para entrenamiento y otros 1.000 para pruebas. Cada titular fue etiquetado de manera manual por una de un total de seis personas y puede reflejar una o varias de las emociones seleccionadas para el estudio (enojo, asco, miedo, alegría, tristeza y sorpresa). La novedad de este *dataset* es que incorpora una escala de 0 a 100 para medir la intensidad de cada una de las emociones censadas en lugar de sólo reflejar la presencia o ausencia y otra escala de polaridad que va desde -100 (negativo) hasta 100 (positivo). También en la Tarea 4 de la edición 2017 de la competencia Semeval, se presentó un *dataset* de 1.250 textos en inglés, entre ellos tweets y títulos de noticias, anotados manualmente en una de las seis emociones básicas de Ekman (Rosenthal, Farra, & Nakov, 2017). Mientras que en la Tarea 3 de la edición 2019 (Chatterjee, Narahari, Joshi, & Agrawal, 2019), se utilizó un conjunto de datos de diálogos textuales en inglés para una de las tareas del congreso, compuesto por 38.424 diálogos, donde cada uno fue etiquetado manualmente en cuatro clases diferentes (enojo, alegría, tristeza y otros) por siete evaluadores humanos.

En (Buechel & Hahn, 2017) se creó un conjunto de 10.000 textos llamado EmoBank, utilizando un modelo tridimensional de valencia, activación y dominancia. La particularidad de este estudio es que cada elemento fue anotado teniendo en cuenta la perspectiva del lector y la del autor del contenido. Otro aspecto relevante es que parte del contenido también está etiquetado utilizando el modelo categórico de Ekman, lo que puede ser útil para entrenar clasificadores que se nutran de ambos conjuntos de características (Acheampong et al., 2020).

En el trabajo de (Gambino & Calvo, 2019), se compiló un conjunto de datos de 3.572 mensajes de Twitter en español. Luego, cada tweet fue clasificado en una de las seis emociones básicas (amor, alegría, sorpresa, ira, tristeza y miedo) por cuatro anotadores. Si bien el etiquetado de este *dataset* de emoción básica fue hecho de manera manual se incorporó una etapa de validación al proceso el cual se comentará más en detalle en la sección 3.4.

Tanto para realizar clasificación de emoción básica en TASS 2020, como así también en la tarea EmoEvalEs de IberLEF 2021 (Plaza-Del-Arco et al., 2021), se utilizó el *dataset EmoEvent* (Plaza-Del-Arco, Strapparava, Alfonso Ureña-López, & Teresa Martín-Valdivia, 2020), un corpus de emociones multilingüe de tweets basado en eventos. Estos fueron etiquetados de manera redundante con la principal emoción expresada en el tweet por tres anotadores según las siguientes categorías: ira, asco, miedo, alegría, tristeza, sorpresa y "neutro o sin emoción". El etiquetado se realizó mediante crowdsourcing utilizando la herramienta *Amazon Mechanical Turk*. La versión en español del *dataset* contiene un total de 8.409 tweets. Para la validación del consenso en este *dataset* se utilizó la Kappa de Cohen, no se informó un nivel de consenso global, sin embargo, este varía entre 0,17 y 0,55, para el contenido en español, dependiendo de la emoción básica analizada.

En (Y. Li & Su, 2017) se recopiló un *dataset* de 13.118 diálogos cada uno de ellos con un promedio de ocho intervenciones en total. Posteriormente se procedió a etiquetarlo en varios aspectos, siendo uno de ellos las emociones de cada intervención de un determinado diálogo. Las categorías adoptadas fueron las emociones Ekman agregando además la clase "otros". El proceso fue realizado por expertos en comunicación. Según los autores se logró un acuerdo entre los anotadores de un 78,9%, sin embargo, no se aclara la métrica utilizada, adicionalmente la categoría "otros" pudo haber contribuido a elevar el valor de acuerdo debido

a que el contenido desafiante o dudoso generalmente es catalogado de esa forma, alrededor del 83% del contenido fue clasificado como “otros”.

En (Ghazi, Inkpen, & Szpakowicz, 2015) se construyó un conjunto de unidades léxicas (pares palabra-emoción) utilizando FrameNet (Fillmore, Petruck, Ruppenhofer, & Wright, 2003), el diccionario de sinónimos de Oxford y etiquetado manual. Las palabras se etiquetaron en alguna de las seis emociones básicas de Ekman junto con la vergüenza. Se obtuvieron 102 unidades léxicas con un alto nivel de consenso entre las etiquetas. Estas unidades léxicas luego fueron utilizadas para etiquetar automáticamente un conjunto de 2.414 oraciones. La particularidad de este trabajo es que a 820 oraciones también se las etiquetó con el estímulo que produjo la emoción.

(Preoţiuc-Pietro et al., 2016) recolectó un conjunto de 2.895 publicaciones de Facebook las cuales fueron etiquetadas por dos especialistas en psicología utilizando el modelo circunflejo de Russell, es decir en las dimensiones de valencia y activación. Si bien los autores resaltan que el acuerdo entre anotadores es de 0,768 para valencia 0,827 para activación, no se especifica la métrica utilizada.

En (Chen, Hsu, Kuo, Huang, & Ku, 2019) se recopiló un conjunto de 2.000 diálogos compuestos por 1.000 de la serie *Friends* y otros 1.000 de chats privados de Facebook, posteriormente se realizó etiquetado manual de cada intervención de los participantes en el diálogo mediante crowdsourcing. En total, se etiquetaron 14.503 intervenciones para el *dataset* de *Friends* y 14.742 para el de los chats privados de Facebook, un total de 29.245. Las emociones utilizadas fueron las de Ekman junto con la etiqueta neutral. Cada elemento fue etiquetado por cinco personas, en cada uno de ellos se eligió la emoción más votada y en caso de empate se asignó la etiqueta “no neutral”. Para la medición del consenso entre las etiquetas asignadas se utilizó Kappa de Fleiss cuyo valor fue 0,33.

El fragmento correspondiente a la serie *Friends* del conjunto de datos de (Chen et al., 2019) fue tomado por (Poria et al., 2020) y enriquecido con información multimedia (visual y audio) para cada elemento del diálogo. El *dataset* fue re etiquetado, incorporando los datos multimedia, por tres especialistas obteniendo un consenso, medido por Kappa de Fleiss de 0.43. El conjunto de datos de (Chen et al., 2019) contenía 14.503 elementos etiquetados,

mientras que en (Poria et al., 2020) este número se redujo a 13.708 debido a que se detectaron algunas inconsistencias en los diálogos del conjunto de datos original.

En (B. Wang, Liakata, Zubiaga, Procter, & Jensen, 2016) se recopiló un *dataset* de 3.085 tweets en inglés correspondientes a museos asociados al proyecto SMILES, los mismos fueron etiquetados manualmente por estudiantes de doctorado en sociología, utilizando las emociones básicas de Ekman (exceptuando miedo) junto con las etiquetas "*not relevant*" y "*no code*". La medición del consenso se realizó mediante acuerdo simple reportando que el 82,1% de los tweets fue clasificado de la misma forma por al menos dos de los anotadores humanos. Este conjunto de datos presenta la particularidad de que algunos tweets fueron clasificados con más de una emoción. Sin embargo, presenta un gran desbalance dado que más del 70% de los datos están etiquetados como "*no code*" o "*happy*".

En el trabajo de (C. Liu, Osama, & de Andrade, 2019) se recopiló un conjunto de historias de la plataforma Wattpad, de cada historia se seleccionaron los pasajes susceptibles de ser anotados con una emoción básica. Para la anotación se eligieron algunas de las emociones del modelo de Plutchik (Plutchik, 2001), estas son alegría, tristeza, enojo, miedo, expectativa, sorpresa, amor, asco y neutral. El proceso de etiquetado se realizó utilizando *Amazon Mechanical Turk*, es decir *crowdsoursing*, con tres anotadores independientes. Se etiquetaron un total de 9.710 pasajes, el Kappa de Fleiss obtenido luego de dicho proceso fue del 0,4. Posteriormente los autores realizaron una revisión de los pasajes que tuvieron menos consenso, re etiquetando o descartando según su criterio. El *dataset* fue utilizado para entrenar varios algoritmos de ML, logrando desempeños de hasta 0,604 medido en micro F1.

En (Aman & Szpakowicz, 2007) se recopiló un conjunto de 5.205 oraciones de un total de 173 publicaciones en blogs. Las oraciones fueron etiquetadas por dos personas cada una, utilizando las categorías de Ekman agregando las etiquetas "varias emociones" y "sin emoción". Debido a que cada elemento fue etiquetado por dos personas, la medición del consenso se realizó utilizando la Kappa de Cohen, la cual resultó en un promedio de 0,76. Posteriormente se dividieron las etiquetas entre "emoción" y "no emoción" y se entrenaron algoritmos NB y SVM para su clasificación, siendo esta última la de mayor exactitud, con un 73,89%.

De lo discutido hasta aquí, pueden observarse las limitaciones del etiquetado manual en cuanto al tamaño de los conjuntos de datos producidos. En general la cantidad de muestras en cada *dataset* va desde los cientos hasta las decenas de miles, rara vez supera las 20.000, algunas excepciones para el caso de etiquetado polar son el Stanford Sentiment Treebank (Socher et al., 2013) con 215.154 en inglés y el conjunto de datos del TASS ("Taller de Análisis de sentimientos en Español (TASS)," n.d.) con alrededor de 70.000 comentarios en español, mientras para emoción básica se encuentra el *dataset* de SemEval 2019 con 38.324 elementos. Cabe resaltar que estas excepciones corresponden a organizaciones relevantes, lo que les permite coordinar esfuerzos de diversos grupos de investigación para lograr dichos resultados. En general, los conjuntos de datos producidos por investigadores individuales suelen ser mucho más acotados. En la próxima sección se abordan trabajos que hicieron uso de etiquetado automático o semi automático mediante DS lo que permite, entre otras ventajas, recopilar un mayor número de muestras a menor costo.

3.2.2 Construidos con etiquetado automático o semi automático

El proceso de construcción de recursos para el Análisis de Sentimientos puede ser extenso y costoso de realizar, y esto impacta directamente en la disponibilidad de los mismos. En el idioma inglés esto se ve compensado en parte por los recursos volcados a la investigación en esta área por distintos grupos alrededor del mundo, sin embargo, en otros idiomas, como por ejemplo el español, este no es el caso. Según (Elnagar, Khalifa, & Einea, 2018) la mayoría de la atención en lo que respecta a investigación en Análisis de Sentimientos se dedica al idioma inglés, esto supone un problema a la hora de construir clasificadores basados en ML robustos, dado que el entrenamiento con grandes conjuntos de datos es clave para su desarrollo.

Dentro de los principales problemas en la construcción de conjuntos de datos, se encuentra la necesidad del etiquetado manual. No solo por el tiempo que conlleva dicha tarea, sino porque, además, para el caso del Análisis de Sentimientos, las etiquetas asignadas al contenido pueden ser subjetivas, debido a que la persona que escribe el texto generalmente no es la misma que lo clasifica (Alm, 2008). Esto obliga a realizar la clasificación de manera redundante y adoptar la etiqueta que posea mayor consenso, por consiguiente, el trabajo necesario se vuelve aún más extenso y costoso.

En consecuencia, los investigadores enfocados en la construcción de recursos para Análisis de Sentimientos han adoptado estrategias que permiten etiquetar el contenido de manera automática o semiautomática, utilizando marcadores ya presentes en los datos, de manera de acelerar el proceso y reducir el impacto de las apreciaciones subjetivas de los etiquetadores humanos. Este tipo de etiquetado es conocido como supervisión distante (DS *distant supervision* por sus siglas en inglés) y permite crear grandes cantidades de datos de entrenamiento a bajo costo (Roth et al., 2013). Sin embargo, debido a que los datos obtenidos son inherentemente ruidosos, el problema más desafiante es mejorar su calidad al reducir la cantidad de ruido.

Por otro lado, el etiquetado automático o semi automático del contenido permite crear conjuntos de datos más numerosos, los cuales han hecho posible construir clasificadores basados en ML supervisado de texto que generalizan mejor al haber sido entrenados con un mayor número de muestras (Ahmad Refaee, 2016) (Banko & Brill, 2001). Esto último es particularmente importante si se intenta analizar textos provenientes de redes sociales ya que el entrenamiento con un mayor número de muestras permite una mejor cobertura de las variantes léxicas presentes en el idioma analizado (Ahmad Refaee, 2016).

Otro aspecto a resaltar con respecto a la utilización de DS es que la eficacia de los distintos marcadores de DS puede ser dependiente del idioma, al respecto (Ahmad Refaee, 2016) señala que por ejemplo los hashtags y emoticones son muy útiles para el Análisis de Sentimientos en inglés, pero son altamente ruidosos para el árabe. En consecuencia, el desempeño de las etiquetas seleccionadas debe ser evaluada para cada lenguaje.

Las redes sociales ofrecen un conjunto de marcadores con los que los usuarios pueden etiquetar fragmentos de contenido multimedia, entre ellos texto. Dentro de estos marcadores podemos mencionar, hashtags, emoticones, puntuaciones y reacciones. Este tipo de etiquetas se han empezado a utilizar para construir recursos para Análisis de Sentimientos en diversos idiomas. A continuación, se mencionan los casos más destacados.

En (Maas et al., 2011) se recopiló un conjunto de datos de evaluación de películas del sitio IMDB. Está compuesto por 25.000 muestras de entrenamiento y otras 25.000 de pruebas, cada uno de los subconjuntos esta balanceado entre muestras positivas y negativas. Para el

etiquetado automático del contenido se utilizaron las calificaciones anexas al texto en cada una de las revisiones de películas.

El conjunto de datos recopilado por la plataforma YELP (Yelp, 2014), contiene aproximadamente siete millones de revisiones en inglés sobre negocios producidas por alrededor de dos millones de usuarios. Cada revisión consta de un rating de una a cinco estrellas y una breve reseña, ambos creados por el mismo usuario. A su vez las mismas revisiones pueden ser calificadas por otros usuarios. También cuenta con IC acerca de los negocios (ubicación, horarios, puntaje global), revisiones (fecha, interacciones de los demás usuarios) y de los usuarios (cantidad de revisiones, interacciones con otros usuarios, puntaje promedio dado, etc.).

(Blitzer, Dredze, & Pereira, 2007) recopiló un *dataset* de opiniones de productos vendidos en el sitio de comercio electrónico Amazon. Está compuesto por 8.000 comentarios de los cuales la mitad corresponden a revisiones positivas y el resto negativas. Los autores mapearon los comentarios con cuatro estrellas o más como positivos y los que tienen dos estrellas o menos como negativos. Además, contiene 9.630 revisiones sin etiquetar.

(Pang & Lee, 2004) (Pang & Lee, 2005) recopiló conjuntos de comentarios, extraídos de los sitios IMBD y *rottentomatoes* acerca de películas. Estos fueron etiquetados utilizando las puntuaciones de los autores de los comentarios en ambos sitios. El *dataset* de IMBD contiene 2.000 comentarios, con 1.000 de cada clase y está etiquetado a nivel de documento. Por otro lado, el correspondiente a *rottentomatoes* está clasificado a nivel oración y tiene 10.662 oraciones, mitad positivas y mitad negativas.

El *Twitter Emotional Corpus* (TEC) (Mohammad, 2012) fue recopilado utilizando la API de la red social Twitter. Se recolectaron los tweets que contenían como hashtag alguna de las emociones básicas de Ekman en inglés (*#anger*, *#disgust*, *#fear*, *#joy*, *#sadness* y *#surprise*). Se recopilaron un total de 21.051 tweets, sin embargo, una de las críticas que se le hace a este *dataset* es que es muy desbalanceado, como ejemplo #joy representa casi el 40% de las muestras del *dataset*. Tampoco presentó una etapa de validación sobre las etiquetas recopiladas.

(W. Wang, Chen, Thirunarayan, & Sheth, 2012), recopiló un conjunto de datos de 2.488.982 tweets anotados con emociones básicas (amor, alegría, sorpresa, enojo, tristeza, miedo y

gratitud) mediante DS. Para ello capturaron tweets que incluyeran como hashtags palabras específicas asociadas extraídas de (Shaver et al., 1987) para cada una de las emociones consideradas. Este estudio incluyó una etapa de validación del conjunto de datos, en la cual se tomó una muestra de 400 tweets y se le aplicó una serie de heurísticas. Siendo estas últimas la verificación de la posición de hashtag al final de tweet, el descartado de tweets con menos de cinco palabras, el descartado de tweets que incluyan enlaces URL (*Uniform Resource Locator* por sus siglas en inglés) o citas y un re etiquetado manual. El resultado del proceso de validación logró una precisión del 93,16% con respecto a las etiquetas originales de la muestra.

Otro de los *datasets* recopilados en Twitter es CBET (*Clean Balanced Emotional Tweets* por sus siglas en inglés) (Gholipour Shahraki & Zaïane, 2017), como su nombre lo indica, está perfectamente balanceado entre las nueve emociones que lo componen (enojo, miedo, alegría, amor, tristeza, sorpresa, agradecimiento, asco y culpa). Al igual que en el caso anterior se utilizaron hashtags para minar cada una de las emociones básicas recolectadas. En total está compuesto por 81.163 tweets, de los cuales 76.860 están marcados con una única emoción (8.540 cada una). Los 4.303 tweets restantes están marcados con dos emociones, este último fragmento del *dataset* no se encuentra balanceado.

DS se aplicó también en el trabajo de (Go, Bhayani, & Huang, 2009), en el que los emoticones se utilizaron como etiquetas para clasificar automáticamente un conjunto de datos de tweets en una de tres categorías: positivo, negativo y neutral. El conjunto de datos recopilado contiene más de 1.600.000 tweets (entre positivos y negativos). Luego se entrenaron clasificadores basados en SVM, *Maximum Entropy* (ME) y NB utilizando los tweets positivos y negativos. La mejor precisión reportada por este estudio, 82,7%, se logró usando una combinación de características de unigramas y bigramas de palabras con ME y NB.

(Pool & Nissim, 2016) utilizaron las reacciones de Facebook para aplicar DS sobre un *dataset* (no se informa el tamaño) y luego para entrenar una SVM para la detección de emociones. No obstante, vincularon las reacciones a la publicación original, que es la asociación más adoptada en los estudios que utilizan las reacciones de Facebook (Kaur, Balakrishnan, Rana, & Sinniah, 2019), en lugar de asociarlas al comentario. Además, tampoco midieron la confiabilidad de las etiquetas automáticas.

DS también es útil cuando se trabaja con idiomas de poco estudiados, como se presenta en el trabajo de (Ahmad Refaee, 2016), en el que se realizaron varios experimentos con conjuntos de datos en árabe, algunos categorizados manualmente y otros con DS. Para medir el acuerdo en el etiquetado manual se utilizó Kappa de Cohen. El κ promedio fue de 0,786, lo que indica un acuerdo sustancial. Las clases utilizadas fueron positiva, negativa, neutral, mixta, indefinida y omitir. Las dos últimas clases, si bien son útiles, tienden a mejorar los resultados de las medidas de consenso, ya que el contenido más difícil de interpretar suele caer en alguna de ellas. El conjunto de datos más numeroso, entre los etiquetados de manera manual, contiene 6.894 tweets. Mientras que el del mayor tamaño categorizado utilizando DS cuenta con 1.173.432 tweets.

En el trabajo de (Suttles & Ide, 2013), se recopiló un conjunto de tweets en inglés en el que se usaron los hashtags, emoticones y emojis para el etiquetado con ocho emociones básicas (alegría, tristeza, expectativa, sorpresa, enojo, miedo, confianza y asco) utilizando DS. Se recopilaron aproximadamente tres millones de tweets. Tener múltiples formas de categorizar el conjunto de datos permitió realizar una validación cruzada del proceso de etiquetado. Los autores concluyeron que, con excepciones menores, hubo consenso entre las etiquetas. Posteriormente, el conjunto de datos construido se utilizó para entrenar clasificadores basados en ML que obtuvieron precisiones de entre el 75 y el 91%.

En la competencia *Emotion Intensities* (EmoInt) del Workshop de Enfoques Computacionales a Subjetividad, Sentimientos y Análisis de Social Media 2017 (WASSA 2017 por sus siglas en inglés) (Mohammad & Bravo-Marquez, 2017b) se utilizó un conjunto de datos recopilado en (Mohammad & Bravo-Marquez, 2017a). Para su construcción se preseleccionaron términos asociados a las emociones de enojo, miedo, alegría y tristeza según el *Roget Thesaurus*, luego se utilizó la API de Twitter para capturar tweets asociados a esos términos. Si bien el etiquetado de las emociones se realizó utilizando DS, la posterior medición de las intensidades se hizo de manera manual mediante etiquetadores humanos empleados a través de sitios de *crowdsourcing*. El *dataset* cuenta con 7.097 tweets divididos en conjuntos de entrenamiento, prueba y validación.

(V. Liu, Banea, & Mihalcea, 2017) utilizó DS para construir un conjunto de datos de tweets con las etiquetas "*happy*" y "*sad*". Para automatizar el etiquetado se capturaron tweets que

incluyeran los hashtags #happy y #sad en el texto. En total se recopilaron 2.557 comentarios, 1.525 pertenecientes a la categoría "*happy*" mientras que 1.032 pertenecían a la clase "*sad*". Si bien no se realizó una validación externa, lo interesante de este estudio es que se midió la relación entre la emoción manifestada y factores externos como el clima, noticias, el timeline previo del autor, etc.

(Vosoughi, Zhou, & Roy, 2015) también utilizó un enfoque con DS para recopilar y etiquetar un conjunto de datos de 18 millones de tweets, que fueron clasificados como positivos o negativos según la presencia de emoticones específicos. Para validar las etiquetas asignadas al contenido se utilizó la métrica Kappa de Fleiss, se tomó una muestra de 3.000 tweets y se la clasificó de manera manual por tres etiquetadores. El Kappa resultante fue de 0.82, mostrando un consenso sustancial.

Dado que el conjunto de datos compilado y analizado en el presente trabajo de tesis está en español, se revisó la literatura en busca de artículos que trabajen con este lenguaje específico utilizando DS para el etiquetado automático de emociones básicas. Sin embargo, la mayoría de los artículos usan etiquetas polares (o una variante similar que incluye neutral y varios grados de positivo y negativo). De esta forma, en la investigación de (Moctezuma et al., 2017), se clasificó un conjunto de datos de 18 millones de tweets en español en positivos y negativos utilizando léxicos afectivos en español. (Martín, Aguilar, Torres, & Díaz, 2020) utilizó la calificación adjunta a los comentarios de un sitio web turístico también en una clasificación polar. (Sandoval-Almazan & Valle-Cruz, 2018) midió el impacto de las publicaciones de Facebook en las campañas políticas recopilando publicaciones en español junto con algunas estadísticas, como el número de comentarios, compartidos y reacciones. Sin embargo, en ese trabajo, el análisis se realizó únicamente en base a los emoticonos incluidos en el texto de los comentarios, sin analizar la emoción que el propio texto pudiera reflejar.

Ante la escasez de recursos en determinados idiomas como el español (Osorio Angel et al., 2021) argumenta que muchos investigadores han optado por la alternativa traducir, ya sea el recurso o el texto analizado, a un idioma con mayor nivel de desarrollo en lo que respecta a investigación del Análisis de Sentimientos, típicamente se traducen al inglés. Si bien esta primera aproximación permite llevar a cabo investigaciones de manera rápida, no es sustituta de la creación de recursos en el idioma destino, esto se evidencia en los resultados obtenidos

mediante cada uno de los enfoques, en la misma investigación se detalla que ha habido muchos intentos de utilizar recursos de Análisis de Sentimientos en inglés para el español y otros idiomas, sin embargo argumenta, basándose en los resultados de (Brooke et al., 2009) que el mejor camino para obtener mejoras a largo plazo en Análisis de Sentimientos es la creación de recursos específicos para cada idioma, ya que en el caso de traducción de español a inglés se produce una perdida semántica notable debido a la mayor riqueza del primer idioma con respecto al segundo. En el mismo sentido, (Balahur, Turchi, & Steinberger, 2011) concluyó que el desempeño de los clasificadores de sentimientos basados en datos traducidos automáticamente puede ser optimizado mediante la incorporación de otros datos en el idioma destino, los cuales incluso en pequeñas cantidades puede producir una mejora notable.

La regla general es la escasez de recursos para el idioma español, sobre todo en lo que respecta a la tarea de detección de emoción básica. Según (Yadollahi et al., 2017) el Análisis de Sentimientos para emoción básica es un tema que no se encuentra adecuadamente explotado en español y aún tiene mucho camino por delante, dado que la mayoría del trabajo se ha enfocado en la detección de polaridad. (Plaza-del-arco, Molina-González, Jiménez-Zafra, & Martín-Valdivia, 2018) aclara que, si bien algunas investigaciones han comenzado a descubrir el potencial de la clasificación de emoción básica, la mayoría se centran en el idioma inglés. Para el caso del español, concluye, el desarrollo de aplicaciones de minería de emociones depende de la generación e integración de recursos específicos en este idioma.

La escasa cantidad de *datasets* de emoción básica en español ha motivado la creación de tareas específicas para incentivar investigaciones en este sentido. Tal es el caso de TASS 2020, donde se incluyó una tarea de detección de emociones dando como fundamentación que, si bien la clasificación de polaridad es una tarea bien establecida con muchos *datasets* y metodologías bien definidas, la detección de emociones ha recibido menos atención debido a su complejidad (García-Vega et al., 2020) (Plaza-Del-Arco et al., 2021) (Rosá & Chiruzzo, 2021). Para dicha competencia se utilizó el conjunto de datos de *EmoEvent* (Plaza-Del-Arco et al., 2020), el equipo que logró mejor desempeño reportó valores de *precision*, *recall* y F1 de alrededor de 0,45.

Sin embargo, el desarrollo del Análisis de Sentimientos y la detección de emociones en español dista mucho del observado para el idioma inglés. Esto se debe a que, si bien es deseable

la utilización de recursos diseñados específicamente para el idioma a analizar, la creación de los mismos puede resultar una tarea ardua e interdisciplinaria, que involucra el trabajo de estadísticos, informáticos, especialistas en psicología, lingüistas, entre otros, para la creación de diccionarios afectivos, conjuntos de datos etiquetados (cualquiera sea la etiqueta), POS *taggers*, *lematizadores*, truncadores, desambiguadores, mapas semánticos y otras numerosas herramientas o recursos. Además, cabe resaltar que la creación de conjuntos de datos para el Análisis de Sentimientos y la detección de emociones presenta la particularidad de que las etiquetas asignadas al contenido no son objetivas y pueden depender de las apreciaciones de la persona que las asigna.

3.3 Métricas para comparación de conjuntos de datos

Los conjuntos de datos mencionados en las secciones anteriores tienen la particularidad que tratan temas diversos, provienen de distintas fuentes ya sea formales o informales o incluso en idiomas diferentes.

Para la realización de tareas de Análisis de Sentimientos o detección de emociones es útil comparar conjuntos de datos con características diversas, por ejemplo, para establecer a priori qué algoritmos o técnicas pueden tener un buen desempeño en base a resultados experimentales de otros estudios. Más allá de las características diversas mencionadas, inherentes a los conjuntos de datos, existen métricas que permiten realizar comparaciones útiles a la hora de la toma de decisiones. Una métrica de comparación universal entre *datasets* de texto es el ***tamaño del vocabulario***, esto último impacta no solo en la representación de los datos de entrada sino también en la dificultad, de un modelo basado en ML, para establecer generalizaciones a medida que el tamaño del vocabulario crece.

El ***crecimiento del vocabulario*** en un *dataset* con respecto a la cantidad de documentos que el mismo contiene es otra métrica universal de comparación. Experimentalmente se determinó que sigue la ***Ley de Heaps*** (Baeza-Yates & Riberio-Neto, 2011), la cual respeta la siguiente fórmula:

$$V_P(n) = Kn^{\beta}$$

Donde $V_P \subseteq V$ siendo V el vocabulario total, n es la cantidad de documentos, mientras que K y β son parámetros libres determinados de manera experimental y dependientes del idioma. Generalmente, el tamaño del vocabulario con respecto a la cantidad de documentos crece con una escala logarítmica, es decir con un incremento importante inicial para luego hacerse asintótico a un valor determinado.

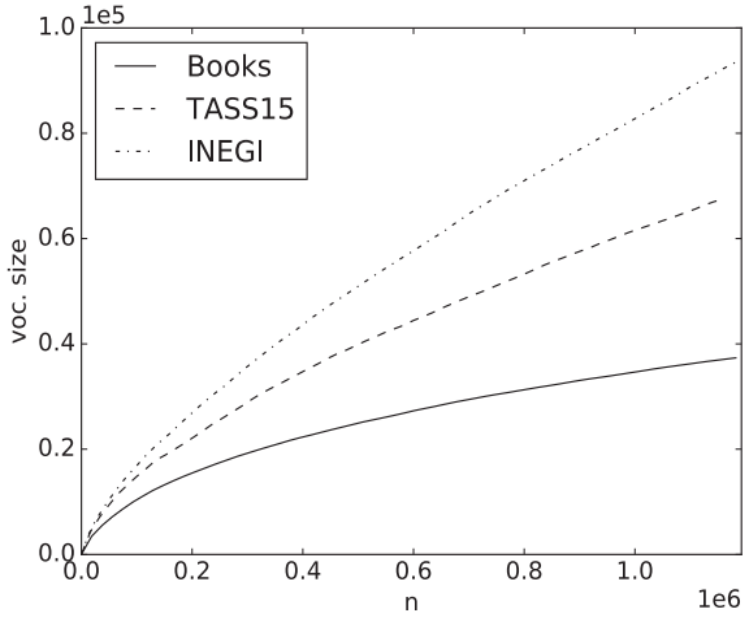


Figura 15: Crecimiento del vocabulario en función de la cantidad de documentos en el conjunto de datos. Fuente: (Tellez et al., 2017).

Como ejemplo de lo anterior, en (Tellez et al., 2017) se determinó experimentalmente la Ley de Heaps para tres *datasets* en español. El primero, compuesto por textos de libros en lenguaje formal, recopilados por el proyecto Gutenberg, el segundo compuesto por tweets utilizado en el workshop TASS'15 (Martínez-Cámara, Díaz-Galiano, García-Cumbreras, García-Vega, & Villena-Román, 2015) y el último también de tweets georreferenciados, elaborados por la población general, recopilados por el Instituto Nacional de Estadísticas y Geografía de México (INEGI). En la Figura 15 puede visualizarse el crecimiento del vocabulario a medida que se aumenta la cantidad de documentos del *dataset*. Naturalmente mientras más informal sea el texto, mayor es su crecimiento y el valor al cual se vuelve asintótico.

El tamaño del vocabulario y su crecimiento son métricas interesantes que permiten comparar conjuntos de datos de texto de distinta naturaleza, sin embargo, no brindan

información acerca de la distribución de las palabras o los tópicos tratados en los documentos, dado que, por ejemplo, una palabra que aparece una vez contribuye de la misma manera que otra que se encuentra en casi todos los documentos.

Debido a lo anterior, también es útil contar con otro tipo de métricas que brindan más información acerca de la distribución de las palabras o los temas en el conjunto de documentos. Una de ellas es la ***distancia del coseno***, métrica que calcula el coseno del ángulo entre los vectores que representan a cada documento. Cuando dicho valor se acerca a 1, los documentos analizados son similares, mientras que cuanto más próximo sea a 0 más distinto es el contenido comparado. Naturalmente, este valor puede promediarse para obtener una noción de qué tan homogéneo es el conjunto de documentos. La distancia del coseno puede aplicarse tanto a vectores basados en bolsas de palabras, como a formatos densos como Word2Vec o GloVe. Posteriormente, puede utilizarse la distancia de coseno para realizar un análisis de los componentes principales (PCA) sobre el conjunto de datos. Esta métrica es comúnmente utilizada en sistemas de recomendación, por ejemplo, para el caso de catálogos de películas (Chiny, Chihab, Bencharef, & Chihab, 2022).

3.4 Métricas de medición del nivel de consenso sobre las categorías

Un aspecto importante cuando se trabaja con conjuntos de textos para el Análisis de Sentimientos y/o la detección de emociones, es la validación de las etiquetas asignadas al contenido. Este tipo de etiquetas presenta una dificultad adicional debido a su carácter subjetivo, frente a, por ejemplo, clasificar textos en un conjunto predefinido de temas (economía, política, espectáculos, deportes, etc.), tarea que puede ser realizada con un mayor grado de objetividad. Por tal motivo, una buena cantidad de estudios que han recopilado conjuntos de datos con este propósito, realizan validaciones sobre las etiquetas asignadas, principalmente utilizando Kappa de Cohen o en su defecto Kappa de Fleiss. En particular, los conjuntos de datos recopilados mediante DS se categorizan mediante etiquetas ya presentes en el contenido, por lo cual prescinden de la intervención humana en el proceso de clasificación. Esto tiene como ventaja que permite crear colecciones particularmente numerosas, sin embargo, este tamaño considerable puede ser un problema a la hora de validar los datos, debido

a que, si fuera necesario revisarlos por completo, el ahorro en tiempo se vería contrarrestado por tal proceso.

La métrica ***Kappa de Cohen*** (Cohen, 1960), predecesora de Kappa de Fleiss, pero limitada a dos evaluadores, se utilizó en algunos de los estudios mencionados previamente para medir el consenso entre los etiquetadores (Ahmad Refaee, 2016) (Plaza-Del-Arco et al., 2021) (Aman & Szpakowicz, 2007). La fórmula para su cálculo es la siguiente:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Siendo p_o el acuerdo observado y p_e la probabilidad de acuerdo por azar. En el caso de Kappa de Cohen estos se calculan de la siguiente manera:

$$p_o = \frac{1}{N} \sum_{1}^{k} m_{i,i}$$

$$p_e = \frac{1}{N^2} \sum_{1}^{k} n_{k\,1}\, n_{k\,2}$$

En este caso k es el número de categorías, m es una matriz cuadrada donde se registran los resultados de la clasificación y en cuya diagonal se encuentran las coincidencias entre los dos etiquetadores, N es la cantidad de muestras a clasificar. Mientras que $n_{k\,1}$ y $n_{k\,2}$ son las muestras clasificadas como clase k por el etiquetador uno y dos respectivamente.

La segunda estrategia es la medida de consenso entre evaluadores por ***Kappa de Fleiss*** (Fleiss, 1971). Esta es útil para medir el consenso entre un número fijo de evaluadores sobre datos categóricos. La fórmula para calcular esta medida se puede ver en la siguiente ecuación, donde $\bar{P}$ es la probabilidad de acuerdo entre evaluadores y $\bar{P}_e$ es la probabilidad de acuerdo por azar.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Si bien la ecuación general está relacionada y es similar a la Kappa de Cohen, en realidad es una generalización de la *pi* de Scott. Para calcularlo, si se considera N etiquetadores, n evaluaciones por etiquetador y k categorías, primero hay que obtener los p_j esto es la

proporción de evaluaciones asignadas a la categoría *j* (siendo n_{ij} el número de etiquetadores que asignaron el elemento i a la categoría *j*):

$$p_j = \frac{1}{N\,n}\sum_{i=1}^{N} n_{ij}$$

Luego es necesario calcular P_i esto es la cantidad de evaluadores que están de acuerdo para la muestra i:

$$P_i = \frac{1}{n\,(n-1)}\sum_{j=1}^{k} n_{ij}\,(n_{ij}-1)$$

Finalmente se calculan $\bar{P}$ y $\bar{P}_e$ (previo a obtener κ), de la siguiente manera:

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i$$

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2$$

El valor de κ se mueve entre −1 (perfecto desacuerdo) y 1 (perfecto acuerdo). Carletta (Carletta, 1996) estableció $\kappa > 0.80$ como una buena medida de confiabilidad, y valores de $0.67 < \kappa < 0{,}80$ que permitirían extraer conclusiones provisionales. Sin embargo, el autor de este último trabajo también insinúa que los fenómenos del discurso y el diálogo pueden ser más complicados que otros tipos de análisis (como la clasificación temática de artículos periodísticos). (Hearst, 1997) sugirió que esto implicaría que la confiabilidad requerida para este tipo de estudios puede llegar a ser menor. Además, también se debe señalar que estas conclusiones se enunciaron antes de la era de las redes sociales, y los estudios posteriores fueron más permisivos con los requisitos de fiabilidad.

En el trabajo de (Gambino & Calvo, 2019) mencionado en la sección 3.2.1 se compiló un conjunto de datos de 3.572 mensajes de Twitter en español. Luego, cada tweet del conjunto de datos fue clasificado en una de seis emociones básicas (amor, alegría, sorpresa, ira, tristeza y miedo) por cuatro anotadores. Una vez completado el proceso de etiquetado, la medida de consenso de Kappa resultante fue de 0,49, lo que indica una concordancia moderada.

Esta misma métrica también se utilizó para medir el consenso en el *dataset* utilizado en (Chatterjee et al., 2019). La puntuación Kappa de Fleiss obtenida para los datos etiquetados fue de 0,59, indicando también un acuerdo moderado. En este caso, la clase "otros" también puede haber ayudado a mejorar el puntaje de consenso final.

Como puede concluirse de los estudios mencionados, si bien (Carletta, 1996) planteó requisitos de consenso desafiantes, la mayoría de los investigadores que realizan clasificación de emociones en datos en formato texto de redes sociales consideran que una concordancia moderada, es decir entre 0,4 y 0,6, en la escala de Kappa de Fleiss es aceptable. Debido al consenso existente en tomar como aceptables datos con un acuerdo moderado, este será también el caso en el presente trabajo.

Por otro lado, aunque muchos estudios miden la calidad del etiquetado manual usando la métrica Kappa de Fleiss, ningún estudio, al menos en español, compara la confiabilidad de los conjuntos de datos etiquetados usando DS versus los etiquetados manualmente, lo cual resulta novedoso y se realiza en la presente tesis.

3.5 Información contextual para el Análisis de Sentimientos

Clasificar un texto según su connotación positiva o negativa, o incluso en una emoción básica, es una tarea desafiante debido a que el mismo carece de información que está presente en otros formatos (Chatterjee et al., 2019) como pueden ser los gestos, el tono de voz, el lenguaje corporal, etc. Por ejemplo, la frase "casi lloro" puede interpretarse como tristeza (o con polaridad negativa), pero si el texto va precedido de "recibí el regalo que tanto deseaba", entonces la emoción básica puede ser de felicidad (o con polaridad positiva).

Dicha información adicional es conocida como contexto, según (Dabrowski & De Villiers, 2015), el contexto se puede definir como cualquier información que se puede utilizar para caracterizar la situación de un objeto o entidad. En tal sentido, en el marco de esta tesis, se llama IC a los distintos marcadores o información anexa a un fragmento de texto, que pueda ser útil para la correcta interpretación del mismo. Esta definición es amplia en varios aspectos. En primer lugar, porque permite incluir palabras claves, emoticones, ratings, otros textos anexos, etc. Y, en segundo lugar, porque la IC puede utilizarse para tareas de clasificación de

textos que van más allá del Análisis de Sentimientos o la detección de emociones, como por ejemplo la clasificación de tópicos, el perfilado de autor, predicción de la próxima palabra, etc.

Según (Yusof, Mohamed, & Abdul-Rahman, 2018), la orientación del sentimiento, es decir si es positivo o negativo, es altamente dependiente del texto a su alrededor, en consecuencia, es muy importante considerar el contexto para realizar una clasificación adecuada, ya que, en ausencia del mismo, el contenido en sí puede ser engañoso. Según la revisión del estado del arte presentada en dicho trabajo, al utilizar al menos uno o dos tipos de IC, la mayoría de los investigadores han mejorado el desempeño con respecto a la línea base (la cual generalmente no utiliza IC). Además, según (Poria, Cambria, Winterstein, & Huang, 2014) (Cambria, Schuller, Xia, & Havasi, 2013) el Análisis de Sentimientos y la minería de emociones avanzan hacia el análisis del contenido, concepto y contexto de los textos en lenguaje natural.

A continuación, en el resto de la sección, se detallan diversos trabajos enfocados al Análisis de Sentimientos o la detección de emociones en texto, junto con los resultados obtenidos, a partir del uso de diversos tipos de IC, para mejorar el desempeño de clasificadores basados en algoritmos de ML.

La utilización de algoritmos de ML, entre otros, en combinación con tareas de clasificación de textos llevó a los investigadores a buscar características que permitieran mejorar el desempeño de los clasificadores construidos, como consecuencia pusieron su atención en la IC. En primer lugar, se detallan los trabajos enfocados en detección de polaridad.

Como ejemplo de lo anterior, en el trabajo (Poria, Cambria, Hazarika, et al., 2017) se comparó el desempeño de una red LSTM con y sin IC para contenido multimodal (videos). En primer lugar, se dividió cada video en fragmentos y luego extrajeron características textuales, auditivas y visuales de los fragmentos. Primero se utilizaron las características de cada fragmento para clasificarlo, y luego hacerlo también sumando las del resto de los fragmentos como IC. El modelo que incorpora contexto mostró una mejora de alrededor del 6-8 % (81,3 % de métrica F1) con respecto a la línea base, el conjunto de datos utilizado en este estudio fue el CMU-MOSI (Zadeh, Zellers, Pincus, & Morency, 2016). En (Ghosal et al., 2018) se presentó un modelo refinado de este enfoque, que mejora el desempeño levemente al 82,31%.

En (Agarwal, Mittal, Bansal, & Garg, 2015), los autores midieron el impacto en la precisión de clasificación utilizando tres conjuntos de datos de texto polares (software, películas y

reseñas de restaurantes) mediante: una ontología específica del dominio; ponderación de características; e IC para determinar la polaridad de los términos ambiguos. Para esto último se construyó un lexicón que permitía determinar la polaridad de los términos ambiguos según la polaridad de los términos que aparecían en su contexto. Esta tarea se realizó utilizando SenticNet (Cambria, Havasi, & Hussain, 2012), SentiWordNet (Baccianella, Esuli, & Sebastiani, 2008) y General Inquirer (Stone & Hunt, 1963). Los resultados mostraron que la IC fue individualmente la adición más relevante, ya que logró la mejora de exactitud más significativa sobre los tres conjuntos de datos utilizados (+14,01% para el dataset de software, +6,1% para el de películas y +15,9% para el de restaurantes).

(Muhammad, Wiratunga, & Lothian, 2016) construyó lexicones para dominios específicos utilizando un enfoque de DS que, combinados con SentiWordNet (Baccianella et al., 2008), produjeron puntajes de sentimiento ajustados (valencia + o -) para los términos analizados, a lo que llamó contexto global. Esto se integró junto con un enfoque basado en ventanas de palabras en el que se utilizan modificadores léxicos y no léxicos para determinar la valencia del término dentro de un intervalo de texto específico, referido como contexto local. El sistema construido superó la línea de base (implementada con SVM), medido en F1, en dos de los tres conjuntos de datos utilizados, las mejoras fueron de entre un 5% y un 12%, mientras que en el restante fue un 7% inferior. En un enfoque similar, (Saif, He, Fernandez, & Alani, 2016) utilizó la polaridad anterior, asignada usando SentiWordNet, MPQA (Wilson, Wiebe, & Hoffmann, 2010) y los lexicones de (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010) (Thelwall, Buckley, & Paltoglou, 2012) de términos co-ocurrentes (contexto) para construir lo que llamaron "círculos sénticos", que se utilizaron posteriormente para determinar la valencia y la polaridad de las palabras para luego aplicarlo a documentos, como tweets. Los resultados reflejaron que la utilización del contexto permitió superar a la línea base seleccionada (llamada SentiStrength) en dos de los tres conjuntos de datos utilizados.

En el trabajo de (Vosoughi et al., 2015) mencionado previamente, también se utilizó IC (como geolocalización, tiempo de publicación y autor del contenido) para calcular las probabilidades previas de sentimientos negativos y positivos utilizando un modelo bayesiano. La precisión del modelo que utilizó todas las características contextuales fue de 0,862, una mejora de 0,077 con respecto a la línea de base (0,785).

Además, en (Vanzo, Croce, & Basili, 2014) se utilizó un conjunto de tweets como contexto, en primer lugar, obteniendo los *n* tweets anteriores en una conversación y, en segundo lugar, obteniendo los *n* tweets previos pero que compartían un hashtag específico con el tweet objetivo. El conjunto de datos utilizado fue el proporcionado en SemEval-2013 Task 2 (Nakov et al., 2013), y el clasificador adoptado fue una SVM. El estudio logró una mejora de la precisión en casi todos los experimentos realizados; el mayor incremento de precisión fue de alrededor del 5,2%. Los autores concluyeron que las mejoras logradas con el uso de IC son sorprendentes ya que la aplicabilidad de su enfoque no requiere recursos adicionales etiquetados manualmente.

Como puede observarse, la mayoría de los trabajos analizados se enfocan en la detección de polaridad. Entre los trabajos que utilizaron IC para mejorar la tarea de detección de emoción básica se encuentran los presentados en (Chatterjee et al., 2019). En dicha tarea, se compiló un conjunto de datos de 38.424 diálogos de texto (entre entrenamiento y prueba) el cual fue etiquetado manualmente por un grupo de jueces en cuatro clases distintas (felicidad, tristeza, enojo y otros). Cada contenido fue etiquetado por siete personas. Cada diálogo está compuesto de tres fragmentos de texto, dos iniciales de contexto y el tercero que se intenta clasificar. La línea base utilizada en este trabajo logró una puntuación micro F1 de 0,5861 para tres clases (Felicidad, Tristeza, Enojo) utilizando una red LSTM. Varios equipos (Agrawal & Suri, 2019) (Bae, Choi, & Lee, 2019) (Basile et al., 2019) (Huang, Trabelsi, & Zaïane, 2019) (Winata et al., 2019) (Xiao, 2019) participaron en la tarea propuesta logrando una puntuación máxima micro F1 de 0,7959. Sin embargo, la puntuación más alta para el equipo que también presentó un documento describiendo la arquitectura fue de 0,7765 (Agrawal & Suri, 2019). Una particularidad de esta competencia es que no se reporta el desempeño de los clasificadores prescindiendo de la IC, lo cual sería útil para determinar el diferencial de desempeño.

En (Chen et al., 2019), los autores expresaron que modelar la emoción básica de una sola frase sin IC tiene el problema de que una misma oración puede expresar diferentes cuestiones dependiendo de su contexto. En consecuencia, en el etiquetado del conjunto de datos de emoción básica construido en dicho trabajo, se proporcionó un diálogo completo a los etiquetadores (en vez de una sola frase) de manera que clasificaran cada una de las intervenciones del mismo. La utilización de los diálogos completos como IC en combinación

con redes LSTM mostró una mejora de exactitud del 59,2% al 63,9% y del 71,5% al 77,4% en la tarea de clasificación de emoción básica para los dos conjuntos de datos utilizados (*Friends* y *EmotionPush* respectivamente). En (Poria et al., 2020) se depuró y anotó nuevamente el conjunto de datos del estudio anterior, para posteriormente comparar el desempeño de clasificadores de emoción básica con y sin IC. Los resultados fueron consistentes con los de (Chen et al., 2019) en el sentido de que los clasificadores que utilizaron IC lograron una mejora en el desempeño.

Por otro lado, en (V. Liu et al., 2017) se utilizaron factores externos como IC para ayudar a predecir las etiquetas *#Happy* o *#Sad* de un conjunto de datos de tweets. Los factores mencionados fueron el clima, noticias relevantes en una ventana de tiempo, tiempo de publicación, tweets de personas seguidas en un intervalo de tiempo y tweets previos del usuario en un intervalo de tiempo. La utilización en conjunto de las características mencionadas (sin tener en cuenta el contenido del tweet) permitió lograr una exactitud del 66,9%, por lo cual los autores concluyen que estos factores externos tienen un impacto significativo en las respuestas emocionales.

Entre los trabajos específicos para el idioma español para clasificación de emoción básica, en (Plaza-Del-Arco et al., 2021) se propuso una tarea de clasificación de emoción básica utilizando el conjunto de datos *EmoEvent* (Plaza-Del-Arco et al., 2020). En esta tarea se inscribieron 70 equipos, 15 presentaron resultados y 11 presentaron trabajos describiendo sus sistemas. Además, este concurso proporcionó IC para cada muestra del conjunto de datos utilizado, compuesto por tweets en español, añadiendo el dominio asociado y si el tweet expresaba ofensividad. Los equipos que se presentaron a la competencia construyeron clasificadores que lograron exactitudes de hasta el 72%, sin embargo, no se reporta la ganancia en desempeño producto del uso de la IC.

3.6 Conclusiones del capítulo

En este capítulo se han presentado en detalle los distintos conjuntos de datos para el Análisis de Sentimientos disponibles en la literatura, junto con diversas métricas que permiten extraer conclusiones acerca de los mismos. Por otro lado, también se analizó el impacto del uso de la IC en sistemas para el Análisis de Sentimientos.

Con respecto a lo analizado acerca de los conjuntos de datos, se identificaron dos técnicas principales para la realización del etiquetado. La primera de ellas consiste en categorizar el contenido de manera manual por una o más personas. Este proceso puede ser largo y complejo dado que la interpretación del contenido suele ser subjetiva sobre todo cuando se intentan construir conjuntos de datos para el Análisis de Sentimientos, lo que obliga a un etiquetado redundante y revisión de los casos controvertidos si se quiere asegurar la calidad del conjunto de datos. Como consecuencia de lo anterior los conjuntos de datos construidos de esta manera suelen tener un tamaño reducido, las excepciones a esta regla se dan en casos de trabajos coordinados por diversas organizaciones o grupos de investigación que posiblemente contaron con cuantiosos recursos para tal fin.

La segunda técnica consiste en el etiquetado automático o semi automático de los datos utilizando marcadores ya presentes en los mismos, lo que suele conocerse como supervisión distante. Las ventajas son el ahorro en tiempo y recursos para el etiquetado (lo que se suele traducir en conjuntos de datos de mayor tamaño) y también, en muchos casos, la etiqueta asignada al contenido suele ser ingresada por el autor del mismo. Sin embargo, la necesidad de una etapa de validación de calidad para este tipo de conjuntos de datos resulta más imperiosa, dado que las etiquetas pueden no ser confiables. Si bien varios de los trabajos que construyeron *datasets* mediante DS incluyeron una etapa de validación, lo realizado en los mismos dista mucho de ser un proceso estandarizado, es decir que cada investigador validó sus datos de distinta manera, por ejemplo, utilizando diversas métricas de consenso, variando el tamaño de la muestra sobre la cual se calculan tales métricas y, en su mayoría, sin especificar los detalles del proceso de re etiquetado de la muestra. Esto dificulta la comparación de la calidad de los conjuntos de datos recopilados mediante DS, debido a que se estaría intentando cotejar resultados con distintas unidades de medida. Tampoco existe claridad acerca del criterio de selección de las métricas de consenso ni determinación del tamaño muestral. En tal sentido, para estimular la creación de conjuntos de datos para el Análisis de Sentimientos recopilados mediante DS, sería interesante contar con una metodología estandarizada para la validación de los mismos, que se base en procesos claramente definidos, métricas confiables y muestreos debidamente justificados. Por otro lado, es importante resaltar que los investigadores que utilizaron Kappa de Fleiss como métrica de consenso, tomaron como aceptable un nivel de acuerdo moderado, es decir entre 0,4 y 0,6.

También cabe señalar que muchas de las tareas relacionadas al Análisis de Sentimientos y detección de emociones, ya sea de pre procesamiento o posterior clasificación son dependientes del idioma subyacente que se pretende analizar, es decir, por ejemplo, el proceso llevado a cabo para realizar una de ellas con textos en inglés es diferente que con textos en español. La calidad del output retornado por las tareas de Análisis de Sentimientos mencionadas depende en consecuencia de los recursos o herramientas disponibles para la realización de las mismas en el idioma objetivo. Esto genera un inconveniente, al respecto (Justo et al., 2018) señala que la mayoría de las investigaciones en disciplinas como el Análisis de Sentimientos se abordan el inglés, aunque el 48% de los recursos de Internet están escritos en otros idiomas.

El idioma español sufre de los mismos inconvenientes con respecto a la escasez de recursos etiquetados para el Análisis de Sentimientos, sobre todo en lo que respecta a conjuntos de datos de emoción básica. Las excepciones a esto son, utilizando etiquetado manual, los recursos del TASS, *EmoEvent* utilizado en la competencia IberLef 2019 (Díaz-Galiano et al., 2019) (Plaza-Del-Arco et al., 2020), el recopilado por (Gambino & Calvo, 2019). Utilizando DS encontramos el conjunto de datos de (Moctezuma et al., 2017), (Martín et al., 2020) y (Sandoval-Almazan & Valle-Cruz, 2018).

Teniendo en cuenta que, por lo analizado en este capítulo, la mayoría de los conjuntos de datos de texto etiquetados para las tareas de Análisis de Sentimientos y detección de emociones se encuentran disponibles para el idioma inglés, el contar con una metodología para la creación y validación de conjuntos de datos recopilados mediante DS sería un paso para suplir, en parte, la necesidad de recursos para el Análisis de Sentimientos en español, pero también aplicable a otros idiomas realizando los ajustes pertinentes.

Por otro lado, otra ventaja que se desprende del proceso de construcción de recursos de Análisis de Sentimientos mediante DS es que permite capturar además IC asociada al contenido específico a ser etiquetado, esta información puede presentarse en diversas formas, y algunas investigaciones en otros lenguajes han intentado hacer uso del mismo para mejorar el desempeño conseguido en distintas tareas del Análisis de Sentimientos y detección de emociones. Los distintos trabajos mencionados en la sección 3.5 lograron mejoras en el desempeño de los clasificadores construidos mediante la utilización de IC.

La gran mayoría de los trabajos referenciados correspondientes a DS con su posterior validación, junto con la utilización de IC, se enfocaron en idiomas distintos del español. Por tal motivo, en los próximos capítulos se adoptan y adaptan técnicas mencionadas en los trabajos referenciados en este estado del arte, con el objetivo de definir una metodología que permita crear y validar conjuntos de datos etiquetados para el Análisis de Sentimientos mediante DS. A su vez, se busca medir el impacto del uso de parte de la IC capturada en el desempeño de clasificadores basados en ML para verificar si se producen mejoras consistentes con el uso de otros tipos de IC para distintos idiomas.

4. Proceso de construcción y validación de conjuntos de datos

Contenido

Resumen

El presente capítulo se inicia con una descripción general del proceso desarrollado para recopilación, preprocesado y validación de conjuntos de datos para la construcción de clasificadores basados en ML para Análisis de Sentimientos, estableciendo el contexto para el resto de la tesis. Debido a la extensión del proceso, se decidió dividir su descripción detallada en dos partes, este capítulo se enfoca en las etapas de recopilación, preprocesamiento y validación de etiquetas del proceso desarrollado.

El capítulo se estructura de la siguiente manera, en primer lugar, en la sección 4.1 se realiza una introducción, donde se detallan algunas de las problemáticas detectadas en el capítulo previo y se las relaciona con el proceso a desarrollar. En la sección 4.2 se describe de forma general el proceso para el Análisis de Sentimientos desarrollado, este incluye las etapas recopilación, preprocesado y validación de conjuntos de datos junto con la selección de formatos de representación y algoritmos de clasificación. Posteriormente, en la sección 4.3, se aborda la recopilación del conjunto de datos, explicando los métodos y fuentes utilizados para

obtener la información necesaria. El siguiente paso es el preprocesamiento del conjunto de datos recopilado, descrito en la sección 4.4, en la cual se valida la efectividad de diversas tareas de preprocesamiento aplicadas, midiendo su incidencia en la reducción de tokens OOV y en el desempeño de clasificadores basados en ML. Luego, la sección 4.5 se centra en la selección y validación de las etiquetas del conjunto de datos. Se describen los criterios utilizados para seleccionar y filtrar los comentarios relevantes, así como el proceso de etiquetado y la medición del consenso entre los etiquetadores. También se proporciona una descripción detallada del conjunto de datos utilizado, incluyendo su tamaño, características y distribución de etiquetas. Finalmente, se presentan las conclusiones del capítulo en la sección 4.6, destacando los hallazgos de las etapas del proceso analizadas en el capítulo.

4.1 Introducción

Según lo establecido en el Capítulo 3, la disponibilidad de recursos de emoción básica para el Análisis de Sentimientos en español es limitada. En parte, esto se debe a lo costoso del etiquetado manual en lo que respecta a tiempo y recursos humanos. En este capítulo se presenta un método alternativo para la construcción de conjuntos de datos de emoción básica utilizando DS, es decir etiquetar automáticamente el contenido a partir de marcadores ya presentes en los datos. Para este método se consideran las experiencias recopiladas en el capítulo de estado del arte. Pero, además, a diferencia de otros trabajos, también se incluye una etapa de validación que en general no se aplica en otros trabajos que utilizan DS debido al gran tamaño del conjunto de datos. Para ello se realiza una validación estadística en la cual se extrae una muestra de los datos, para luego, re etiquetarla con la ayuda de un grupo de especialistas, y finalmente, medir el nivel de consenso sobre las etiquetas seleccionadas.

El método que se presenta permite construir conjuntos de datos de emoción básica de un tamaño importante con respecto a lo que hay hoy presente en la bibliografía para el idioma español, pero también podría aplicarse a otros lenguajes. Los resultados obtenidos a partir del proceso mencionado, no son un fin en sí mismo, sino que contribuyen al Análisis de Sentimientos y a la detección de emociones en lo que respecta a clasificadores construidos mediante algoritmos de ML supervisado. Es sabido que dichos algoritmos, en este caso aplicados a clasificación de textos de redes sociales, pueden generalizar mejor en presencia de mayor cantidad de datos de entrenamiento. Esto se debe a que la abundancia de muestras de texto permite tener una mejor cobertura de las variantes léxicas para el lenguaje analizado.

Si bien la DS agiliza el proceso de construcción de conjuntos de datos, este no está exento de desafíos. En primer lugar, si se pretende posteriormente construir clasificadores de propósito general, el *dataset* debe cubrir una amplia variedad de temas y vocabulario. Por otro lado, se debe seleccionar una fuente de datos que permita asociar fácilmente etiquetas ya presentes a alguna emoción básica de los modelos emocionales previamente desarrollados. Por último, se debe utilizar una métrica para establecer qué tan fiables son las etiquetas seleccionadas como así también para relacionarla con el desempeño de los clasificadores construidos a partir de los datos recopilados.

Con el objetivo de construir un conjunto de datos de tales características y poder realizar las validaciones correspondientes, se buscaron fuentes de datos con etiquetas que se asemejen a los modelos emocionales discutidos en secciones previas, principalmente el modelo categórico de (Ekman & Friesen, 1971) ya que es uno de los modelos más utilizados para el Análisis de Sentimientos (Yadollahi et al., 2017). Dicha característica permite la comparación de los resultados de la presente tesis con otros trabajos encontrados en la bibliografía. Dentro de las fuentes de datos analizadas al momento de iniciar este trabajo, se encontró que las reacciones de Facebook expresan emociones básicas similares a las detalladas en el modelo de Ekman y además se las utiliza ampliamente para interactuar con noticias de temas variados publicadas en los perfiles de los diversos portales presentes en dicha red social.

Las experiencias presentadas en este capítulo fueron publicadas parcialmente en (Tessore et al., 2019) (Esnaola et al., 2019) (Tessore, Esnaola, Lanzarini, & Baldassarri, 2021).

4.2 Descripción general de proceso

En la Figura 16 se presentan las distintas etapas del proceso para el Análisis de Sentimientos desarrollado a lo largo de esta tesis, el mismo se basa en la arquitectura general para sistemas de minería de texto (Feldman & Sanger, 2006) descrita en el Capítulo 2, e incluye las etapas recopilación, preprocesado y validación de conjuntos de datos junto con la selección de formatos de representación y algoritmos de clasificación.

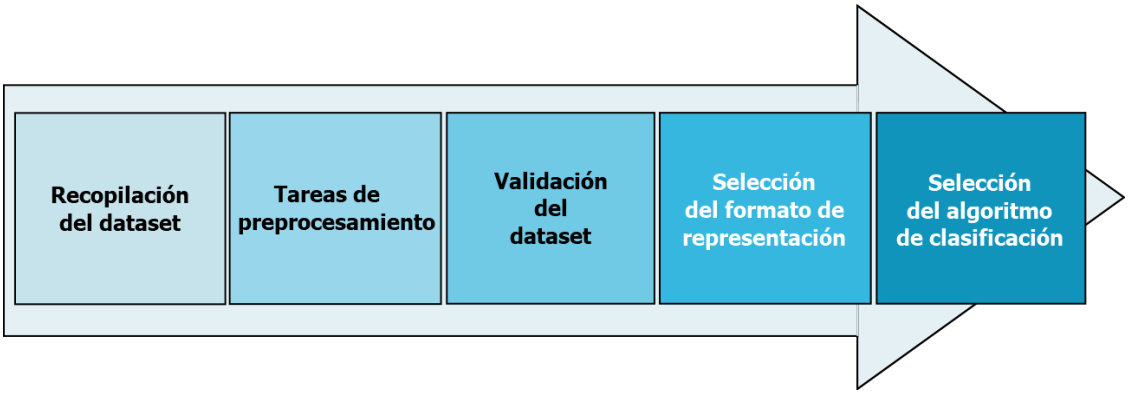


Figura 16: Descripción general del proceso para el Análisis de Sentimientos desarrollado.

Debido a la extensión del proceso, las etapas de recopilación, preprocesamiento y validación del conjunto de datos son descritas en el presente capítulo, mientras que las de selección del

formato de representación y algoritmo de clasificación junto con el uso de la IC se detallan en el Capítulo 5.

En la primera etapa del proceso, se realiza la recopilación de los datos, seleccionando una fuente que contenga mayoritariamente textos en español en la variante de Argentina, y que permita asociarlos de manera sencilla a un marcador asimilable a una emoción básica. De esta manera, se busca construir un conjunto de datos con un tamaño de muestras considerable (en la escala de los cientos de miles), utilizando el enfoque conocido como supervisión distante. Para la selección de la fuente de datos, también se considera que la misma incluya IC asociada al contenido, de manera de utilizarla posteriormente para intentar mejorar el desempeño del clasificador basado en ML a construir.

A continuación, se realizan las distintas tareas de preprocesamiento sobre el conjunto de datos: tokenización, filtrado de muestras por longitud, filtrado y descarte de muestras en otro idioma y filtrado de comentarios pertenecientes a *trolls*. Una vez realizadas dichas tareas, se compara el desempeño de clasificadores basados en ML utilizando el conjunto de datos original (esto es, sin tareas de preprocesamiento) y el revisado, de manera de establecer la efectividad de las tareas desarrolladas para los datos recopilados. Este paso resulta necesario debido a que no necesariamente todas las tareas de procesamiento resultan efectivas independientemente del conjunto de datos.

Posteriormente, debido a que el conjunto de datos fue construido mediante un enfoque de DS, se encara una validación sobre las etiquetas asignadas al contenido. En esta etapa, no resulta posible validar la totalidad de las muestras dado que el tiempo que insumiría tal proceso diluiría las ventajas de la DS. Por el contrario, se realiza una validación estadística tomando un conjunto reducido de muestras, las cuales se clasifican de manera manual por un grupo de psicólogos, para luego calcular la métrica Kappa de Fleiss (Fleiss, 1971) para las etiquetas asignadas por los psicólogos entre sí y entre las anteriores respecto a la etiqueta original de contenido.

Por último, se selecciona un formato de representación de textos, que sirva como entrada para el clasificador basado en ML para el cual también debe seleccionarse un algoritmo. En ambos casos, se pondera un formato y algoritmo que sean susceptibles de ser mejorados con

el uso de IC, pero a su vez que sea posible comparar la mejora en el desempeño con otros estudios presentes en la bibliografía.

4.3 Recopilación de el conjunto de datos

Para realizar la recolección de datos se buscó una fuente que permitiera construir un conjunto de datos de textos en español etiquetados con una emoción básica mediante el enfoque de DS. En distintos trabajos mencionados en el Capítulo 3, las etiquetas asignadas se dedujeron del contenido en sí, o fueron asignadas de manera automática o semi automática a partir de un proceso externo. Según (Yadollahi et al., 2017) los *datasets* altamente confiables son los que tienen la particularidad que los autores del contenido son los mismos que lo etiquetan, un ejemplo de esto es el conjunto de datos ISEAR (Scherer & Wallbott, 1994). Por otro lado, sería deseable que las etiquetas asignadas al contenido puedan ser mapeadas a alguno de los modelos emocionales mencionados en la fundamentación teórica.

Con el objetivo de construir un conjunto de datos de emoción básica cuyas etiquetas se asemejen a las de los modelos emocionales mencionados y que además el contenido fuera etiquetado por su mismo autor, se puso foco en la red social Facebook. Al momento de realizar la captura de los datos, dicha plataforma poseía 2.167 millones de usuarios activos lo que la convertía en la red social más importante del mundo (We_are_social, 2018). En esta red, distintos portales publican día a día diversas noticias de actualidad que los usuarios pueden comentar y a su vez reaccionar mediante las diversas etiquetas que provee la plataforma. Dicha fuente posee como ventaja los dos requisitos mencionados previamente. En primer lugar, el usuario al comentar y reaccionar a una noticia está de alguna manera etiquetando su propio comentario y, en segundo lugar, las reacciones presentes en la plataforma son asimilables varias de a las emociones básicas de Ekman, dado que las reacciones “LOVE” y “HAHA” pueden asociarse a alegría, “WOW” a sorpresa”, “SAD” a tristeza y “ANGRY” a enojo.

Para realizar la recopilación de datos, se recolectaron los títulos de noticias con su descripción junto con los comentarios y reacciones que los usuarios publicaron al interactuar con el contenido. Posteriormente se asocian los pares comentario – reacción publicados por el mismo usuario para una noticia determinada. La idea de tal asociación es que el comentario se

corresponde con la reacción y por ende con la emoción básica, ambas expresadas por un mismo usuario para una publicación específica.

Debido a la enorme cantidad de portales presentes en la plataforma, se decidió acotar la recopilación a un conjunto específico de los mismos. Por un lado, para obtener en su mayoría comentarios en el idioma de la región seleccionada y por el otro debido a restricciones en cuanto a tiempo y cantidad de solicitudes por día de la herramienta utilizada para la captura de datos. Los portales de noticias presentes en la red social seleccionados fueron Clarín; La Nación; Página12; El Cronista; Ámbito Financiero; Todo Noticias; Crónica; CNN en Español; C5N; Agencia Télam; Diario Deportivo Olé; Teleshow e Infobae. En los sitios seleccionados se publican diversos tipos de noticias. Algunos de ellos son especializados en economía, otros en programas de televisión, otros de interés general, etc. La selección se hizo por la variedad de temas que cubren y porque esos medios informativos se encuentran entre los más populares de Argentina (Becerra, 2018).

Los comentarios y las reacciones se recopilaron utilizando la herramienta *Facebook API Graph* (Facebook, 2020) que permite, mediante la creación de una aplicación, capturar comentarios y reacciones del usuario a publicaciones presentes en los diversos portales de noticias. Se recopilaron y almacenaron cerca de 1,8 millones comentarios etiquetados cada uno con su correspondiente reacción. La recopilación de datos se realizó durante cuatro años, entre enero de 2016 y diciembre de 2019 inclusive.

4.4 Preprocesamiento sobre el conjunto de datos recopilado

Una vez recopilado el conjunto de datos, la siguiente etapa es realizar un preprocesamiento de los textos del mismo. Para ello, primero se presentan los resultados experimentales de diversas investigaciones acerca de la efectividad de distintas técnicas de preprocesamiento aplicadas a conjuntos de datos de texto. Para luego, seleccionar algunas de ellas y aplicarlas al conjunto de datos recolectado, con el fin de verificar si y en qué medida producen mejoras en el desempeño de clasificadores basados en ML construidos con estos datos. Si bien, a priori, se podría pensar que el preprocesamiento produce siempre mejoras en el desempeño final, esto no siempre se verifica experimentalmente. Por tal motivo, a continuación, se presentan trabajos

de ambos casos, en primer lugar, de los que obtuvieron mejoras para luego detenerse en los que no.

Entre los trabajos que obtuvieron mejoras se encuentra (Baldwin et al., 2013). Según este, el texto proveniente de las redes sociales es generalmente más “ruidoso” que el texto tradicional. Sin embargo, este "ruido" se puede abordar mediante la aplicación de técnicas de preprocesamiento, como la normalización léxica, la identificación del idioma y/o el POS *tagging*. Una de las conclusiones más importantes obtenidas en dicho trabajo es que, una vez aplicadas, las técnicas de preprocesamiento reducen la variabilidad de los datos de tal forma que pueden ser analizados mediante técnicas tradicionales de minería de textos, al menos para los casos estudiados en dicho trabajo.

Otro trabajo que obtuvo mejoras (Krouska, Troussas, & Virvou, 2016) también aplicó diversas técnicas de preprocesamiento, como tokenización, vectorización de la entrada con ponderación TF-IDF, eliminación de “*stop words*” y lematización. El objetivo era comprobar si la aplicación de estas técnicas de preprocesamiento podía mejorar la precisión de una tarea de clasificación de textos. En esta investigación se concluyó que los clasificadores entrenados con texto preprocesado fueron generalmente más precisos que los entrenados con texto sin procesar.

En (Agrawal & Suri, 2019), la utilización de lematización, sinónimos de *WordNet* y normalización léxica (quitar signos de acentuación, números, palabras de paro, signos, etc.) produjo una merma en el desempeño de los clasificadores analizados en términos de F1. Con respecto a la representación de los textos de entrada, los autores destacan que los *embeddings* neuronales pre entrenados, pueden ser superados complementándolos con características tradicionales como n-gramas de palabras o caracteres, principalmente debido a que estos últimos son robustos a la hora de manejar errores gramaticales. En las pruebas realizadas mediante un análisis de ablación sobre el conjunto de datos estudiado, los autores reportan que los n-gramas de caracteres son la característica que más ganancia les reporta en términos de F1.

Sin embargo, en otros trabajos se evidencia que, no necesariamente, el pre procesamiento del conjunto de datos se traduce en una mejora en el *accuracy o F1* score del clasificador. Como ejemplo de esto, en el trabajo (Mosquera, Yoan, & Moreda, 2017) se identificaron

escenarios donde el pre procesamiento de los datos no solo no produce mejora, sino que empeora los resultados de tareas de Análisis de Sentimientos. En (Chandrasekar & Qian, 2016) el pre procesamiento de los datos no logra mejorar la exactitud de una de las clases analizadas.

El (Tellez et al., 2017) se realizó un análisis aplicando de manera sistemática diversas transformaciones sobre dos conjuntos de datos de redes sociales en español, TASS 2015 e INEGI, en este caso obtenidos de Twitter. Los autores llegaron a la conclusión de que no necesariamente una transformación de alto coste computacional sobre la entrada, como puede ser la lematización o el *stemming*, se traduce en una mejora en el desempeño de clasificadores basados en ML.

Por lo discutido previamente, puede verse que la efectividad de las técnicas de pre procesamiento más comunes dependen de la fuente de datos. Teniendo en cuenta esto, y los resultados de los trabajos mencionados en los párrafos anteriores, se decidió medir el impacto del pre procesamiento sobre el conjunto de datos recopilado. En esta sección se exponen y amplían los experimentos presentados en (Tessore et al., 2019) (Esnaola et al., 2019).

Para medir la efectividad de las distintas técnicas de preprocesamiento aplicadas se considera, en primer lugar, el porcentaje de reducción de OOV tokens que cada técnica produce. Y, en segundo lugar, la mejora en el desempeño de clasificadores entrenados con los textos preprocesados con respecto a los entrenados con el texto sin preprocesar. Con ambas métricas se busca establecer si para este conjunto de datos existe una relación entre la reducción de tokens OOV en los textos de entrada y la mejora en el desempeño de los clasificadores entrenados con estos datos.

El resto de la sección se estructura de la siguiente manera. En 4.4.1 se presentan las distintas técnicas de preprocesamiento aplicadas a los textos de entrada. En 4.4.2 se mide la efectividad de cada técnica en la reducción de tokens OOV. Finalmente, en 4.4.3 se verifica la variación en el desempeño de clasificadores basados en ML producidos por cada técnica de preprocesamiento.

4.4.1 Tareas de preprocesamiento aplicadas

Los textos de las redes sociales presentan características particulares, dada la presencia de enlaces, errores gramaticales, emoticones, lenguaje informal, entre otros. Este tipo de

construcciones, en general, no están presentes en textos más formales. Para construir un clasificador de este tipo de textos basado en ML, es necesario establecer qué tanto impactan dichas construcciones en el desempeño de los clasificadores. Por tanto, resulta oportuno aplicar técnicas de transformación de texto capaces de reducir la variabilidad de los tokens (Tellez et al., 2017). De esta manera, se identificaron las siguientes transformaciones: eliminación de etiquetas HTML, eliminación de nombres de usuario, procesamiento de siglas y abreviaturas, procesamiento de sílabas redundantes, eliminación de URLs, procesamiento de palabras incorrectas unidas por puntos, procesamiento de emoticonos, eliminación de tokens sin letras, eliminación de símbolos especiales, procesamiento de caracteres redundantes, procesamiento de hashtags, procesamiento de símbolos redundantes y conversión a mayúsculas, además de la tokenización del contenido.

La efectividad de cada transformación aplicada se mide a través del análisis de dos valores: el porcentaje de reducción de tokens OOV y el de tokens únicos. El proceso de corrección ortográfica requiere elegir una herramienta para tal fin. Entre las herramientas de revisión ortográfica analizadas, se seleccionó Hunspell en base a las recomendaciones de estudios previos (Clark & Araki, 2011), y se la personalizó utilizando un diccionario con la variante argentina del español (Bosio, 2014).

Estas tareas se ejecutaron por separado sobre el conjunto de datos recopilado para medir la eficacia individual. Aunque las tareas pueden combinarse, el orden en que se describen en el texto no implica ningún orden específico de aplicación, ya que se ejecutaron independientemente. En las siguientes subsecciones se explica cada una de ellas en detalle.

4.4.1.1 Procesamiento de acrónimos y abreviaturas

Los textos de redes sociales, suelen contener acrónimos y abreviaturas. Por tal motivo, se implementó una heurística para la detección de siglas y abreviaturas que utiliza una expresión regular para detectar la presencia de signos de puntuación con una densidad superior a la habitual dentro de un token. La frecuencia de aparición de tokens que contienen un punto también se considera para el proceso de selección. En el último paso se corrobora si el token es una palabra válida según Hunspell. La expresión regular puede verse en la siguiente fórmula:

$$\mathbf{\text{^(.)*([^.]){1}(\.){1}\$}}$$

Una vez aplicado este método, el resultado debe filtrarse manualmente para eliminar los falsos positivos. La Tabla 6 muestra ejemplos de los resultados obtenidos.

SR.	UDS.	ETC.	SRTA.
SRA.	MR.	A.M.	SERV.
DR.	UD.	CIA.	R.I.P.
Q.D.E.P.	U.S.A.	H.D.P.	PCIA.
UDES.	P.D.	AV.	DRA.
ARG.	GRAL.	JR.	LIC.
EJ.	APROX.	SRES.	T.V.
PROV.	SRAS.	ING.	SRS.
M.M.	DD.HH.	D.N.I.	T.N.
M.E.	S.O.S.	D.T.	D.E.P.
U.S.	U.S.	H.D.M.P.	DOC.
E.P.D.	Q.P.D.	Q.E.D.	PROF.
A.F.A.	Q.E.P.	P.R.	CAG.
EE.UU.	BS.AS.	GNO.	REP.
E.E.U.U.	PTE.	CAP.	F.M.I.
ATTE.	C.A.B.A.	E.U.	A.R.A.
Q.E.P.D.	C.F.K.	PD.	NTRA.

Tabla 6: Acrónimos y abreviaturas detectadas.

Los acrónimos y abreviaturas detectados se almacenan en una lista que se utiliza para filtrarlos de los comentarios, reemplazándolos por el índice del elemento de la lista. Este proceso conserva las posiciones de los acrónimos y abreviaturas en los comentarios y evita que Hunspell los marque como errores, porque todos los números son correctos según Hunspell.

A continuación, se presenta un ejemplo de la transformación del texto después de ejecutar esta tarea:

- Texto original: “Era una mujer muy joven con una voz muy particular Q.E.P.D.”
- Texto transformado: “Era una mujer muy joven con una voz muy particular 25”

4.4.1.2 Eliminación de tags HTML

Algunos de los comentarios incluyen etiquetas HTML dentro del texto. Este tipo de contenido no es particularmente útil para el objetivo final de este estudio, ya que es una pieza de metadatos que no transmite ningún significado emocional. Debido a esto se decidió eliminarlos con la ayuda de expresiones regulares.

A continuación, se presenta un ejemplo de la transformación del texto después de ejecutar esta tarea:

- Original: "¡BIENBENIDOS! ME ENCANTA VER ESTAS OBRAS <B> SI ESTAMOS EN TIEMPO DE CAMBIO <\ b> VA SER UN AÑO DIFERENTE".
- Transformado: "¡BIENBENIDOS! ME ENCANTA VER ESTAS OBRAS SI ESTAMOS EN TIEMPO DE CAMBIO VA SER UN AÑO DIFERENTE".

4.4.1.3 Eliminación de nombres de usuario

Los nombres de usuario generalmente no son reconocidos como tokens válidos por las herramientas de corrección ortográfica. Además, la connotación positiva o negativa de un nombre de usuario puede variar a lo largo del tiempo (Newell, Potharaju, Xiang, & Nita-Rotaru, 2014). Este aspecto puede causar problemas porque, por ejemplo, si aparece un nombre de usuario en muchos comentarios asociados a una emoción negativa, el clasificador puede vincular ese nombre de usuario con esa emoción negativa. En comentarios futuros, si el mismo nombre de usuario está acompañado por palabras ligeramente positivas, el clasificador puede indicar de todos modos que el comentario es negativo debido a la fuerte connotación negativa del nombre de usuario, esto produciría un error de clasificación.

A continuación, se presenta un ejemplo de la transformación del texto después de ejecutar esta tarea:

- Original: "$ 34.60 a la venta en el Banco Ciudad. Tenes info FALSA @infobae".
- Transformado: "$ 34.60 a la venta en el Banco Ciudad. Tenes info FALSA".

4.4.1.4 Corrección de palabras acentuadas

Muchas palabras en español llevan tilde, pero en el lenguaje informal presente en las redes sociales el uso del acento ortográfico generalmente se omite. Esto hace que muchas palabras queden en la categoría de OOV. Con el fin de corregir esto, se diseñó un procedimiento para que los errores de acentuación comunes/menores pudieran corregirse. En consecuencia, se consideró un error común/menor cuando una palabra clasificada como OOV tiene entre sus sugerencias de Hunspell sólo una opción, que coincide exactamente con la palabra incorrecta excepto por un carácter acentuado. Así, estas palabras fueron transformadas a su variante correcta.

A continuación, se presenta un ejemplo de la transformación del texto después de ejecutar esta tarea:

- Texto original: “que ruidito a Helicoptero”.
- Texto transformado: “que ruidito a Helicóptero”.

4.4.1.5 Eliminación de tokens sin letras

En esta tarea, se eliminan del comentario todos los tokens que no contienen al menos una letra, dado que muchos de estos símbolos no suelen asociarse con una emoción. Como ejemplo de esto se encuentran los números, las fechas o los símbolos están compuestos sólo por caracteres especiales.

A continuación, se presenta un ejemplo de la transformación del texto después de ejecutar esta tarea:

- Original: “Yo nací el 20/12/69 todos me dicen q tengo un año más xq nací en el 69... Pero a 10 días del 1970... A siq.. Pertenezco a la década del 70... Jjajajsa.. Y como es 20... A cuatro días de noche buena... Mucho no puedo festejar xq si lo dijo para el finde ya es navidad!!! Jajajjaaj pero dentro de dos años cae vienes justo para mis 50!!!! Ahí si... Joooodaaaa!!!! Jajajjaaj”.
- Transformado: “Yo nací el todos me dicen q tengo un año más xq nací en el Pero a días del A siq.. Pertenezco a la década del Jjajajsa.. Y como es A cuatro días de noche

buena... Mucho no puedo festejar xq si lo dijo para el finde ya es navidad!!! Jajajjaaj pero dentro de dos años cae vienes justo para mis Ahí si... Joooodaaaa!!!! Jajajjaaj".

4.4.1.5.1 Separación de palabras unidas incorrectamente

Muchas palabras de los comentarios están unidas erróneamente por un punto, generalmente debido a la escritura rápida. Esas palabras unidas provocan diversos OOV. En este sentido, se realizó un proceso para identificarlas, en primer lugar, y luego tratar de repararlas. Hunspell se usó para verificar palabras erróneamente unidas por un punto. Así, si al menos una de las palabras unidas por un punto es correcta de acuerdo con Hunspell, entonces el punto que las une es reemplazado por un espacio en blanco. A continuación, se presenta un ejemplo de la transformación del texto después de ejecutar esta tarea:

- Original: "Q.E.P.D.Los 5 amigos !!!!!".
- Transformado: "Q.E.P.D. Los 5 amigos !!!!!".

4.4.1.6 Procesamiento de emoticones

Los emoticones son elementos que también son frecuentes en textos tomados de las redes sociales. Pueden expresar una emoción, un estado de ánimo o hacer referencia a una situación u objeto. Muchos estudios los utilizan como una sugerencia para indicar la emoción del contenido analizado (Tian, Galery, Dulcinati, Molimpakis, & Sun, 2017). En esta tarea, se eliminan los emoticones con el objetivo de medir posteriormente el impacto de su presencia o ausencia en el desempeño de los clasificadores a construir. Para este propósito, se han utilizado dos bases de datos de emoticones: una para los de texto sin formato y la otra para los Unicode. Después de analizar todos los comentarios, las incidencias se marcaron con una etiqueta específica que indica al preprocesador que ignore el token.

A continuación, se presenta un ejemplo de la transformación del texto después de ejecutar esta tarea:

- Original: "LADRAN SANCHO"SEÑAL QUE CABALGAMOS ♥♥".
- Transformado: "LADRAN SANCHO"SEÑAL QUE CABALGAMOS ".

4.4.1.7 Procesamiento de caracteres redundantes

Para esta tarea, se consideraron los tokens OOV que tenían tres o más repeticiones de la misma letra consecutivamente. Las letras repetidas tres o más veces se simplificaron a una. En el caso de que un token OOV sólo posea dos repeticiones consecutivas de la misma letra, el reemplazo sólo se realiza si el token transformado resulta en una palabra correcta según Hunspell. Este proceso se implementa tal como se describe debido a que no hay muchas palabras correctas en español que contengan tres o más repeticiones de la misma letra, pero hay algunas que tienen dos repeticiones.

A continuación, se presenta un ejemplo de la transformación del texto después de ejecutar esta tarea:

- Original: “Peeero pará jajaajaaa estas si que las sacaste del fondo del río baagree bagreee ??????????????”.
- Transformado: “Pero pará jajaaja estas con el fondo del río bagre bagre ??????????????”.

4.4.1.8 Procesamiento de hashtags

Otra peculiaridad de los textos procedentes de las redes sociales es la inclusión de hashtags. Estos permiten agrupar diferentes comentarios al vincularlos con un tema específico. Suelen estar muy presentes en redes sociales como Twitter. Sin embargo, en Facebook, que es la red social utilizada en este estudio, su uso es marginal. Luego de analizar todo el *dataset*, se determinó que los hashtags representan menos del 0,8% del total de tokens. En consecuencia, se decidió eliminarlos de los comentarios. En los casos en los cuales la presencia de estos sea más representativa, lo cual suele ocurrir en los conjuntos de datos de Twitter, puede descomponerse el hashtag en las palabras que lo componen.

A continuación, se presenta un ejemplo de la transformación del texto después de ejecutar esta tarea:

- Original: “Creo que es con el único deportista Argentino con el cual todos estamos en un acuerdo que es el más grande .. a nivel nacional y latinoamericano. #ElPibeDe40”.

- Transformado: “Creo que es con el único deportista Argentino con el cual todos estamos en un acuerdo que es más grande a nivel nacional y latinoamericano”.

4.4.1.9 Procesamiento de símbolos redundantes

En este filtro se simplifican las repeticiones múltiples de los siguientes símbolos: ":", ";", "," y ".". En este caso, dos o más apariciones consecutivas de alguno de estos símbolos se simplifican en una.

A continuación, se presenta un ejemplo de la transformación del texto después de ejecutar esta tarea:

- Original: “Desde polonia. Un saludo .. genial el frío”.
- Transformado: “Desde polonia. Un saludo . genial el frío”.

4.4.1.10 Eliminación de URLs

Muchos de los comentarios incluyen URLs, estas no tienen significado por sí mismas sino más bien por el contenido al cual apuntan, ni tampoco pueden asociarse a una emoción específica (Tian et al., 2017). Por lo tanto, se decidió eliminarlas de los comentarios para facilitar la clasificación del texto. Esto se logró a través de expresiones regulares que permiten detectar si un texto es una URL.

A continuación, se presenta un ejemplo de la transformación del texto después de ejecutar esta tarea:

- Texto original: “¡Que golazo! https://media1.tenor.co/images/75024f582bd9bc28f896df80a68763f1/tenor.gif?itemid=10611559”
- Texto transformado: “¡Que golazo!”

4.4.1.11 Eliminación de sílabas redundantes

Los textos de redes sociales suelen utilizar lenguaje informal introducido adrede, por tal motivo, se realizó una clasificación de palabras mal escritas, que indica el número de ocurrencias en el conjunto de datos completo de comentarios y el porcentaje que representan sobre el total de palabras del *dataset*.

La Tabla 7 presenta un caso representativo donde las sílabas redundantes dan como resultado muchos tokens con el mismo significado, en este caso, diversas formas de interjección de la risa, con múltiples apariciones en el conjunto de datos. Este hecho refuerza la idea de que la eliminación de sílabas repetidas es una tarea necesaria.

Palabra	Ocurrencias	Porcentaje
jajaja	14429	7.33%
jajajaja	9354	4.75%
jaja	8092	4.11%
jajajajaja	4720	2.40%
jajajajajaja	2086	1.06%
jajajaj	1540	0.78%
jajajajajajaja	1515	0.77%
jajaj	1330	0.68%

Tabla 7: Diferentes formas para la interjección de la risa y sus apariciones en el conjunto de datos.

A continuación, se presenta un ejemplo de la transformación del texto después de ejecutar esta tarea:

- Original: “Obvio vamos a salir a festejar jajaja”.
- Transformado: “Obvio vamos a salir a festejar ja”.

4.4.1.12 Convertir el texto a mayúsculas

Con el fin de reducir el número de características a analizar por el clasificador a implementar, todas las palabras se convirtieron a mayúsculas. Esta conversión también presenta la ventaja de reducir los tokens OOV causados por los nombres de países, lugares o personas en minúsculas.

A continuación, se presenta un ejemplo de la transformación del texto después de ejecutar esta tarea:

- Texto original: “Vos sos arquero y armani es arquerazo papa ahi esta tu envidia”.
- Texto transformado: “VOS SOS ARQUERO YARMANI ES ARQUERAZO PAPA AHI ESTA TU ENVIDIA”.

4.4.2 Efectividad del preprocesamiento en la reducción de tokens OOV

Una vez aplicados, de manera independiente, todos los filtros de la sección anterior, es necesario medir la efectividad de los mismos en la reducción de tokens OOV. Para ello, se consideró una muestra aleatoria de 181.368 comentarios, que representa el 10% de todo el conjunto de datos.

Todos los procesos descritos en la sección anterior se aplicaron individualmente a dicha muestra. Como línea base de comparación se tomó la tokenización por espacios en blanco.

En la Tabla 8 se presentan los resultados de cada tarea. La efectividad se mide considerando el porcentaje de reducción de OOV sobre la línea de base, el porcentaje de reducción de tokens únicos sobre la línea de base y el tiempo consumido en segundos. Este último valor es una medida relativa para comparar la eficiencia entre las diferentes tareas.

4.4.3 Preprocesamiento y desempeño de clasificadores basados ML

Con el objetivo de transformar los textos de entrada para los algoritmos de ML, es decir todos los comentarios (en adelante referidos como “documentos”) a una representación numérica, se construyó una bolsa de palabras y se utilizaron tres enfoques de ponderación de términos: frecuencia de términos, pesado binario y frecuencia de término – frecuencia inversa en documentos.

Estos tres enfoques se seleccionaron con el objetivo de verificar si las distintas tareas de preprocesamiento aplicadas podrían tener mayor incidencia según la ponderación escogida. Asimismo, si bien hay estudios (Tellez et al., 2017) que demuestran que la combinación de q-gramas y n-gramas consigue un mejor desempeño que los n-gramas con n = 1, es decir, a nivel palabra, se escogió esta última representación dado que las tareas de preprocesamiento intentan reducir los tokens OOV, es decir, trabajan a nivel palabra y, como el objetivo es medir su efectividad, se intentó minimizar la incidencia de otros formatos, como q-gramas, que podrían conducir a mejoras por sí mismos en lugar de a causa de las tareas aplicadas.

Posteriormente, se entrenaron clasificadores basados en ML supervisado, donde cada comentario, posee una etiqueta que define la emoción asociada con el mismo. Las posibles emociones son: “me divierte”, “me encanta”, “me enoja”, “me sorprende” y “me entristece”,

estas se corresponden con las provistas por la red social Facebook. El *dataset* utilizado, que reúne el 10% del total de ejemplos muestrales, cuenta con 180.000 documentos etiquetados (36.000 documentos para cada emoción) escogidos aleatoriamente del total de datos disponibles, de manera tal que las diferentes clases están balanceadas.

Tarea de preprocesamiento	Reducción de OOV tokens (%)	Reducción de tokens únicos (%)	Segundos consumidos
Procesamiento de acrónimos y abreviaturas	0,2016	0,0419	1,09
Eliminación de tags HTML	0,0007	0,0008	0,21
Eliminación de nombres de usuario	0,0030	0,0050	0,22
Corrección de palabras acentuadas	7,9288	1,4125	15.810,14
Eliminación de tokens sin letras	40,9195	31,1682	2,73
Separación de palabras unidas incorrectamente	3,8822	16,3163	57,81
Procesamiento de emoticones	0,5204	0,6522	7,21
Procesamiento de caracteres redundantes	2,4183	2,8645	36,34
Procesamiento de hashtags	0,5995	0,6230	0,78
Procesamiento de símbolos redundantes	-1,4148	19,0824	2,00
Eliminación de URLs	0,0795	0,1242	1,83
Eliminación de sílabas redundantes	0,9030	0,1992	5,43
Convertir el texto a mayúsculas	0,1460	16,3428	0,16

Tabla 8: Resultados obtenidos para cada tarea de preprocesamiento.

Los algoritmos de ML seleccionados para realizar las pruebas fueron NB y SVM, la selección de estos algoritmos se debe a que se encuentran entre los más utilizados en las tareas

de clasificación de texto (Ravi & Ravi, 2015). Los scripts fueron implementados utilizando Python 3.6 junto con la biblioteca scikit-learn 0.20.0.

Para el caso de NB, se utilizaron los parámetros por defecto, mientras que para SVM, los parámetros utilizados fueron: *loss='modified_huber'*; *penalty='l2'*; *alpha=0.55e-4*; *random_state=42*; *max_iter=100*. La elección de parámetros por defecto responde a que el objetivo de este trabajo no es lograr la mejor exactitud en la clasificación, sino medir el impacto que cada tarea de preprocesamiento aplicada, ya sea de manera aislada o combinada con otras, produce sobre el desempeño del clasificador.

A continuación, se exhiben los resultados obtenidos luego de aplicar cada tarea de preprocesamiento primero de forma aislada y luego combinadas, previo al entrenamiento de los algoritmos NB y SVM. Para el caso de las tareas combinadas, se escogieron las cuatro tareas de preprocesamiento que mostraban un mejor desempeño y se las permutó según P_2^4, P_3^4 y P_4^4. La exactitud mostrada en cada caso surge de calcular el promedio de las diez ejecuciones realizadas para cada combinación probada, ya que se ejecutó *10-fold cross validation* para cada prueba con cada algoritmo. De todos modos, en las tablas siguientes, y por cuestiones de espacio físico en el documento, sólo se muestran los tres mejores resultados obtenidos (es decir, aquellos que alcanzaron una mayor exactitud), la línea base (es decir, con los documentos originales sin aplicar ninguna tarea de preprocesamiento, que se muestra resaltada en color gris) y los tres peores resultados (es decir, aquellos con la peor exactitud), todos ellos correspondientes tanto a la etapa de entrenamiento, como de pruebas (con datos previamente no observados por el algoritmo). Los resultados completos pueden encontrarse en (Esnaola & Tessore, 2019).

Cada una de las columnas de las tablas de resultados agrupan un conjunto de diez ejecuciones, donde la exactitud mostrada refleja el promedio de dichas corridas. Cada celda corresponde a la aplicación o no de la correspondiente tarea de preprocesamiento. En este sentido, el número 1 representa que la tarea se aplicó, en tanto que el número 0 indica lo contrario. Cuando en la columna se muestran números mayores que 1, debe interpretarse como que la tarea ha sido aplicada, pero además indica el orden en el que fue combinada para ese grupo de ejecuciones.

La Tabla 9, muestra el desempeño alcanzado, primero sobre los datos de entrenamiento y luego sobre los de prueba, utilizando los algoritmos de ML mencionados y frecuencia de términos como enfoque de ponderación.

Los resultados obtenidos demuestran que, en la etapa de entrenamiento ambos algoritmos alcanzan una mejor exactitud aplicando tareas de preprocesamiento individuales. Mientras que, evaluando la exactitud de los algoritmos con los datos de prueba, los mejores resultados se alcanzaron combinando cuatro tareas.

Exactitud obtenida en entrenamiento

	NB - tf							SVM - tf						
A	1	0	0	0	0	0	0	1	0	0	0	0	0	0
B	0	0	1	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	1	0	0	0	0
D	0	0	0	0	3	3	4	0	0	0	0	3	4	3
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	2	1	1	0	1	0	0	2	1	1
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	1	2	2	0	0	0	0	1	2	2
J	0	0	0	0	4	4	3	0	0	0	0	4	3	4
K	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Exactitud	0,5991	0,5979	0,5979	0,5979	0,5875	0,5874	0,5874	0,6953	0,6944	0,6943	0,6943	0,6785	0,6784	0,6784

Exactitud obtenida en pruebas

	NB - tf							SVM - tf						
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	2	3	3	0	0	0	0	2	4	4	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0	0	1
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	3	2	1	0	0	0	0	0	2	3	0	1	0	0
H	0	0	0	0	0	1	0	0	0	0	0	0	0	0
I	1	1	2	0	0	0	0	1	1	1	0	2	0	0
J	4	4	4	0	0	0	0	3	3	2	0	0	0	0
K	0	0	0	0	0	0	1	0	0	0	0	0	1	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Exactitud	0,4631	0,4631	0,4631	0,4599	0,4599	0,4598	0,4593	0,4547	0,4547	0,4546	0,4518	0,4517	0,4509	0,4508

A: Procesamiento de acronismos y abreviaturas
B: Eliminación de tags HTML
C: Eliminación de nombres de usuario
D: Corrección de palabras acentuadas
E: Eliminación de silabas redundantes
F: Eliminación de URL's
G: Separación de palabras unidas incorrectamente
H: Procesamiento de emoticones
I: Eliminación de palabras sin letras
J: Procesamiento de caracteres redundantes
K: Procesamiento de hastags
L: Procesamiento de símbolos redundantes
M: Convertir el texto a mayúsculas

Tabla 9: Desempeño de los clasificadores en entrenamiento y pruebas con ponderación frecuencia de término.

La Tabla 10 muestra el desempeño alcanzado, primero sobre los datos de entrenamiento y luego sobre los de prueba, utilizando los algoritmos de ML mencionados y pesado binario como enfoque de ponderación.

En esta variante se mantiene el hecho que las tareas individuales alcanzan una mayor exactitud en la etapa de entrenamiento, mientras que las permutaciones de tareas se desempeñan mejor cuando se evalúan los clasificadores con los datos de prueba. La exactitud, en general, es levemente superior con este formato de ponderación que con frecuencia de términos. Con este esquema de ponderación se alcanza la mayor exactitud para los datos de entrenamiento.

Exactitud obtenida en entrenamiento

	NB - Pesado binario							SVM - Pesado binario						
A	1	0	0	0	0	0	0	1	0	0	0	0	0	0
B	0	0	1	0	0	0	0	0	0	1	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	4	3	4	0	0	0	0	4	3	4
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	2	1	1	0	1	0	0	2	1	1
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	1	2	2	0	0	0	0	1	2	2
J	0	0	0	0	3	4	3	0	0	0	0	3	4	3
K	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Exactitud	0,6031	0,6019	0,6019	0,6019	0,5911	0,5910	0,5910	0,7021	0,7011	0,7011	0,7011	0,6858	0,6858	0,6857

Exactitud obtenida en pruebas

	NB - Pesado binario							SVM - Pesado binario						
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	2	2	2	0	1	1	0	4	4	3	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	3	0	4	0	2	3	0	3	2	4	0	2	0	0
H	0	0	0	0	0	0	0	0	0	0	0	0	1	0
I	1	1	1	0	3	2	0	1	1	1	0	1	0	0
J	4	3	3	0	0	0	0	2	3	2	0	0	0	0
K	0	0	0	0	0	0	1	0	0	0	0	0	0	1
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Exactitud	0,4636	0,4635	0,4635	0,4611	0,4611	0,4611	0,4607	0,4622	0,4620	0,4620	0,4593	0,4591	0,4591	0,4590

A: Procesamiento de acronismos y abreviaturas
B: Eliminación de tags HTML
C: Eliminación de nombres de usuario
D: Corrección de palabras acentuadas
E: Eliminación de silabas redundantes
F: Eliminación de URL's
G: Separación de palabras unidas incorrectamente
H: Procesamiento de emoticones
I: Eliminación de palabras sin letras
J: Procesamiento de caracteres redundantes
K: Procesamiento de hastags
L: Procesamiento de símbolos redundantes
M: Convertir el texto a mayúsculas

Tabla 10: Desempeño de los clasificadores en entrenamiento y pruebas con ponderación binaria.

La Tabla 11, muestra el desempeño alcanzado, primero sobre los datos de entrenamiento y luego sobre los de prueba, utilizando los algoritmos de ML mencionados y TF-IDF como enfoque de ponderación.

Se repite el hecho que las permutaciones de tareas tienen un mejor desempeño para los datos de prueba mientras que las tareas individuales alcanzan mejores resultados para el

entrenamiento. Con este esquema de ponderación se alcanzan la mayor exactitud para los datos de prueba.

Exactitud obtenida en entrenamiento														
	NB - tf x idf							SVM - tf x idf						
A	1	0	0	0	0	0	0	1	0	0	0	0	0	0
B	0	0	1	0	0	0	0	0	1	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	1	0	0	0	0
D	0	0	0	0	4	3	4	0	0	0	0	4	3	4
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	2	1	1	0	0	0	0	2	1	1
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	1	2	2	0	0	0	0	1	2	2
J	0	0	0	0	3	4	3	0	0	0	0	3	4	3
K	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Exactitud	0,6053	0,6041	0,6041	0,6040	0,5938	0,5938	0,5938	0,6240	0,6229	0,6229	0,6229	0,6128	0,6128	0,6128

Exactitud obtenida en pruebas														
	NB - tf x idf							SVM - tf x idf						
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	3	3	2	0	1	1	0	2	2	2	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	1	2	3	0	0	2	0	0	4	3	0	2	1	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	2	1	1	0	2	3	0	1	1	1	0	1	2	0
J	4	4	4	0	0	0	0	3	3	4	0	0	0	0
K	0	0	0	0	0	0	1	0	0	0	0	0	0	1
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Exactitud	0,4649	0,4649	0,4649	0,4621	0,4617	0,4617	0,4615	0,4736	0,4736	0,4736	0,4696	0,4694	0,4693	0,4688

A: Procesamiento de acronismos y abreviaturas
B: Eliminación de tags HTML
C: Eliminación de nombres de usuario
D: Corrección de palabras acentuadas
E: Eliminación de silabas redundantes
F: Eliminación de URL's
G: Separación de palabras unidas incorrectamente
H: Procesamiento de emoticones
I: Eliminación de palabras sin letras
J: Procesamiento de caracteres redundantes
K: Procesamiento de hastags
L: Procesamiento de símbolos redundantes
M: Convertir el texto a mayúsculas

Tabla 11: Desempeño de los clasificadores en entrenamiento y pruebas con ponderación frecuencia de término frecuencia inversa de documento.

4.5 Selección y validación de las etiquetas del conjunto de datos

Luego de la recopilación y el preprocesamiento del conjunto de datos, resulta necesario determinar, qué etiquetas se asignan al contenido. Según lo discutido en el Capítulo 3, la DS permite recopilar grandes conjuntos de datos reduciendo la necesidad de clasificación manual. Esto facilita la creación de recursos para el Análisis de Sentimientos en lenguajes que no han tenido tanta atención como el inglés, como es el caso del español.

Si bien, como se vio en el estado del arte, DS se ha aplicado para la construcción de conjuntos de datos para el Análisis de Sentimientos en español, la mayoría de los trabajos se

enfocan en una clasificación polar (positiva-negativa) (Moctezuma et al., 2017) (Martín et al., 2020), en detrimento de los conjuntos de datos con etiquetas que asociadas a emociones. Esto ha llevado a que, recientemente, importantes congresos dedicados al Análisis de Sentimientos en español pongan su atención en la creación de recursos de emoción básica (García-Vega et al., 2020).

Por otro lado, se identificó también una carencia en la certificación de la calidad de conjuntos de datos construidos mediante DS, debido a que no existe un procedimiento estandarizado que indique los pasos a seguir para tal fin. Varios de los trabajos revisados no realizan una etapa de validación de manera de medir que tan certeras son las etiquetas asignadas al contenido y los que lo hacen construyen un proceso ad-hoc. Una de las principales dificultades en abordar este punto se da en cómo validar el conjunto de datos sin necesidad de recurrir a un etiquetado manual completo del mismo, lo cual anularía las ventajas de la DS. Otro inconveniente detectado es que existen ambigüedades acerca de qué métricas utilizar para la validación de etiquetas y finalmente cómo se relacionan esas métricas con los desempeños de clasificadores basados en ML entrenados a partir de los recursos recopilados.

Tanto la validación de las etiquetas como el interrogante de qué métricas utilizar son abordados en la presente sección. Mientras que la relación con el desempeño de clasificadores basados en ML construidos a partir de estos datos es tratada, entre otros temas, en el siguiente capítulo.

La validación de la calidad de los recursos construidos no resulta una excentricidad, sino que es un aspecto importante a tener en cuenta, ya que la utilidad de los conjuntos de datos emoción básica depende de la fiabilidad de las etiquetas asignadas al contenido. El objetivo final de los usuarios de este tipo de conjuntos de datos es predecir las emociones básicas y no etiquetas asignadas de manera arbitraria. En este sentido, es necesario establecer la fuerza del vínculo entre esas etiquetas y las emociones básicas.

4.5.1 Selección y filtrado de comentarios

El total de comentarios recopilados, antes de elegir solo aquellos que se consideran útiles para este trabajo, fue de 20.996.169. Sin embargo, de este total solo se seleccionaron los que poseen una reacción asociada para ser utilizada como etiqueta del mismo.

De las reacciones recopiladas LOVE, HAHA, ANGRY, SAD, LIKE y WOW, se consideraron para el conjunto de datos las primeras 4, se decidió excluir a la reacción WOW, asociada con sorpresa, del conjunto de datos. El motivo de esta decisión radica en que existe al día de hoy un debate abierto acerca de si considerar o no a la sorpresa como una emoción básica (Ortony, 2019) lo que ha llevado a prestigiosas investigaciones a excluirla de sus modelos (Susanto et al., 2020). También los comentarios asociados con la reacción LIKE se excluyeron deliberadamente del estudio, ya que su significado es más ambiguo dado que suele ocurrir que las personas lo utilizan para indicar que vieron una determinada publicación. Por otra parte, en caso de que el mismo usuario escribiera más de un comentario en respuesta a una publicación, se decidió seleccionar el primero de ellos, es decir el más antiguo, para asociarlo a la reacción. Esto es debido a que se considera que dicho comentario es el que refleja más fielmente el estado emocional del usuario al leer la noticia, los comentarios posteriores suelen ser interacciones con otros usuarios. Luego de estas consideraciones, el número de comentarios etiquetados se redujo a 1.716.413.

Debido a que no se reportó una mejora categórica en el desempeño de los clasificadores construidos en la sección 4.4, se decidió no utilizar las técnicas de preprocesamiento mencionadas previamente. No obstante, se aplicaron otras técnicas de preprocesamiento, las cuales se consideran necesarias para realizar el proceso de validación del conjunto de datos recopilado. Primero se procedió a una tokenización de los textos, para luego filtrar comentarios escritos en otro idioma, los menores a una longitud dada y los sospechosos de haber sido escritos por *trolls*. Los detalles de estos procesos se explican a continuación.

En primer lugar, se procedió a ***tokenizar*** los comentarios recopilados. Para ello se utilizó la clase *TweetTokenizer* de la biblioteca NLTK (Bird et al., 2009). Este proceso permitió, además, eliminar tokens que no se consideran útiles en el contexto de este trabajo como enlaces, signos, caracteres no imprimibles y las *stopwords* en español. Por ejemplo, en el siguiente comentario: “Señor Olé usted es diabolico.” *TweetTokenizer* obtiene seis tokens: “Señor”, “Olé”, “usted”,

“es”, “diabolico” y “.”, pero solo cuatro de esos tokens se consideran válidos porque el token "." es un signo de puntuación y el token "es" es una *stopword*. En consecuencia, el comentario en el ejemplo sólo tiene cuatro tokens válidos.

A continuación, se decidió ***filtrar de los comentarios con menos de tres tokens válidos***, debido a que por su longitud se dificultaba vincularlos a una emoción, por lo cual el número de comentarios se redujo a 1.261.783.

Luego del proceso anterior, se decidió ***validar el idioma*** de los comentarios. Aunque casi todos los usuarios que interactúan con los portales de noticias seleccionados comentan en español, existen algunos comentarios en otros idiomas. Por lo tanto, se aplicó otro filtro para eliminar los comentarios que no sean en español. Para este proceso, se utilizó la biblioteca “Python Bindings to CLD2” (Al-Rfou, 2020). Esto se aplicó a todos los comentarios restantes, dando un total de 1.035.045 comentarios que fueron escritos en español. Como consecuencia de que la detección del idioma es un proceso complejo que puede presentar falsos positivos, se realizó un paso extra de validación cruzada utilizando *Google Translate* (Han, 2015). Dado que la cantidad de solicitudes diarias de la API de *Google Translate* es limitada, se tomó una muestra relativamente pequeña de 1.400 comentarios, seleccionados al azar del conjunto anterior para realizar un proceso de validación cruzada. Todos los comentarios analizados con *Google Trans* fueron reconocidos en idioma español, lo que es un elemento adicional para confiar en los resultados obtenidos con la biblioteca CLD2

Finalmente, se decidió ***filtrar los comentarios provenientes de trolls***. Estos últimos presentan un comportamiento antisocial interpersonal prominente dentro de la cultura de Internet en todo el mundo, que incluye publicar comentarios agresivos, mensajes incendiarios y maliciosos en las secciones de comentarios para provocar, interrumpir y molestar deliberadamente a otros usuarios (Craker & March, 2016). Esos comentarios e interacciones no son deseables para este estudio, ya que no reflejan necesariamente una emoción o una reacción a un tema en particular. Estos usuarios escriben comentarios en varias publicaciones, frecuentemente con el mismo contenido, independientemente del tema de la publicación. Por lo tanto, el proceso de filtrado consiste en identificar todos los comentarios que potencialmente podrían haber sido publicados por estos usuarios y excluirlos del conjunto de datos. El proceso se realizó identificando primero todos los comentarios que aparecen más de una vez y luego

contando el número de apariciones. Este reveló que había 14.488 comentarios en el conjunto de datos que aparecieron al menos dos veces. Si se considera la totalidad de los comentarios recogidos inicialmente (20.996.169), este número asciende a 237.309 comentarios repetidos. Dichos comentarios, que representan alrededor del 1,399% del conjunto de datos, fueron excluidos. Después de aplicar este filtro, el número de comentarios útiles se ubicó en 1.020.557.

4.5.2 Descripción del conjunto de datos

Tras el proceso de selección y filtrado, el conjunto de datos, que se encuentra disponible en el material suplementario de (Tessore et al., 2021), tiene las siguientes características:

- Número de muestras: 1.020.557
- Número de atributos por muestra: cuatro más el atributo de clase
- Información de atributos:
    1. Id de muestra (tipo: numérico)
    2. Título de la publicación en Facebook (tipo: texto codificado en UTF8)
    3. Subtítulo de la publicación en Facebook (tipo: texto codificado en UTF8)
    4. Comentario del usuario a la publicación (tipo: texto codificado en UTF8)
    5. Clase, que es la reacción del usuario al post (HAHA, LOVE, ANGRY, SAD)
- Valores de atributos faltantes: 0
- Distribución de clases: HAHA: 338.835 (33,20%), LOVE: 159.830 (15,66%), ANGRY 436.357 (42,75%), SAD: 85.535 (8,38%).

La Tabla 12 muestra el número máximo, mínimo y promedio de tokens y caracteres en títulos, subtítulos y comentarios de publicaciones. Por otro lado, la Tabla 13 muestra la misma información, pero segmentada por reacción.

Nivel	**Título**			**Subtítulo**			**Comentario**		
	Max	**Min**	**Avg**	**Max**	**Min**	**Avg**	**Max**	**Min**	**Avg**
Token	41	0	12,65	388	0	22,57	1.218	3	19,36
Caracter	246	1	71,52	2.193	1	135,85	7.587	7	110,93

Tabla 12: Estadísticas a nivel de token y caracter para títulos, subtítulos y comentarios.

Otra información relevante sobre este conjunto de datos es cuánta superposición hay entre los tokens de las diferentes clases, es decir, las reacciones relacionadas con los comentarios. La Tabla 14 muestra el nivel de superposición considerando tokens únicos para cada reacción. Por ejemplo, esta tabla muestra que el 27% de tokens únicos de comentarios vinculados a la reacción HAHA están contenidos en los tokens únicos de comentarios relacionados con la reacción SAD y que, a la inversa, el 62% de los tokens únicos de comentarios que corresponden a la reacción SAD están presentes en los tokens únicos de comentarios relacionados con la reacción HAHA.

Reacción	Nivel	Título			Subtítulo			Comentario		
		Max	Min	Avg	Max	Min	Avg	Max	Min	Avg
ANGRY	Caracter	246	3	73,12	2.193	1	137,26	7505	7	119,03
	Token	41	1	12,92	388	0	22,84	1185	3	20,86
HAHA	Caracter	241	1	70,36	2.193	1	134,32	7587	10	100,39
	Token	35	0	12,38	388	0	22,14	1218	3	17,49
LOVE	Caracter	231	3	69,30	2.193	2	133,26	7347	9	103,01
	Token	41	1	12,23	388	0	22,12	1195	3	17,69
SAD	Caracter	241	3	72,11	2.193	4	139,57	5930	12	126,11
	Token	41	1	13,14	388	0	23,66	938	3	22,25

Tabla 13: Estadísticas de nivel de token y caracter para títulos, subtítulos y comentarios, segmentados por reacción.

También es importante considerar la longitud de los comentarios medidos en la cantidad de tokens y caracteres en ellas. Asumiendo que siguen aproximadamente una distribución normal, y utilizando la regla empírica, se consideraron todas las instancias alrededor del valor medio con un ancho de dos desviaciones estándar para producir histogramas más comprensibles.

	HAHA	SAD	LOVE	ANGRY
HAHA	1,00	0,27	0,35	0,48
SAD	0,62	1,00	0,54	0,66
LOVE	0,57	0,39	1,00	0,59
ANGRY	0,42	0,26	0,31	1,00

Tabla 14: Solapamiento de vocabulario entre las clases.

Las Figuras Figura 17, Figura 18 y Figura 19 muestran la frecuencia de instancias en términos de tokens para títulos, subtítulos y comentarios, respectivamente. De la Figura 20 a la Figura 22, muestran lo mismo, pero a nivel caracter.

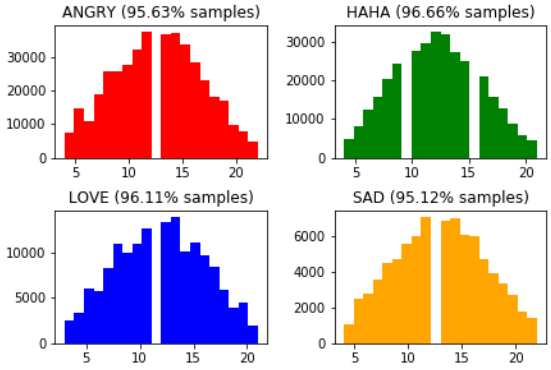


Figura 17: Histograma de títulos a nivel de token.

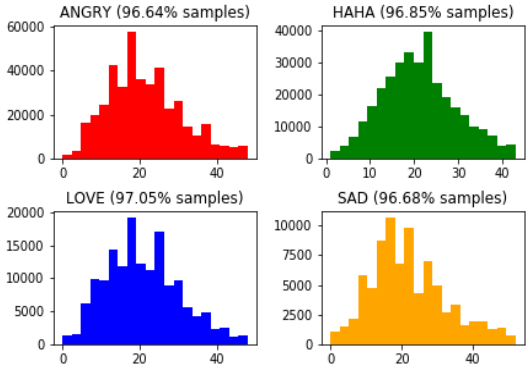


Figura 18: Histograma de subtítulos a nivel token.

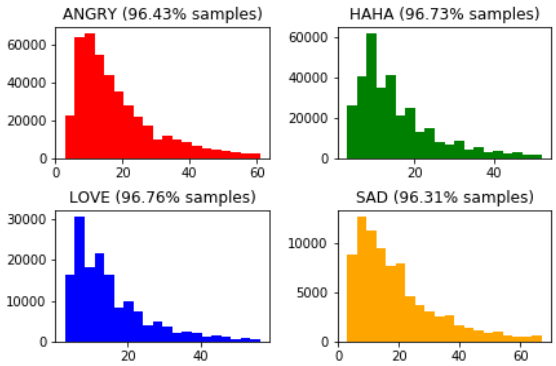


Figura 19: Histograma de comentarios a nivel token.

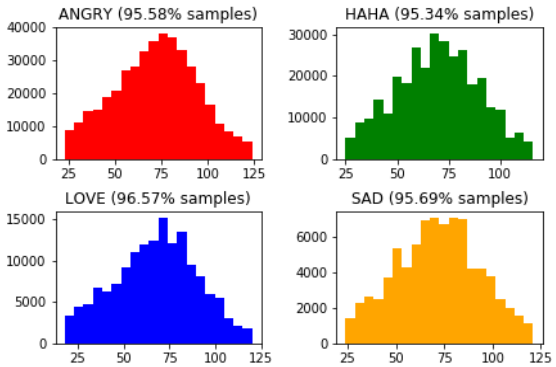


Figura 20: Histograma de títulos a nivel de caracter.

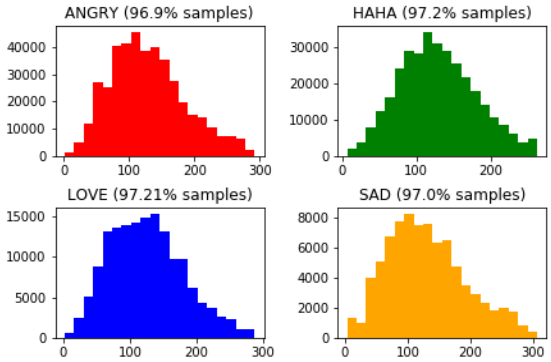


Figura 21: Histograma de subtítulos a nivel de caracter.

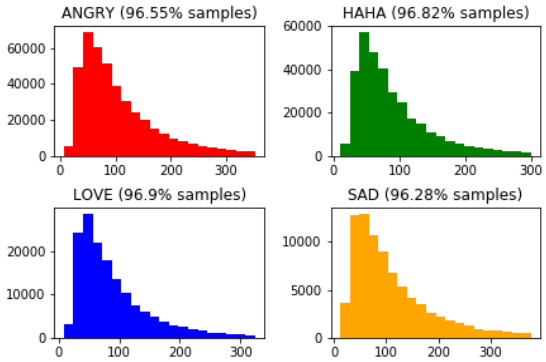


Figura 22: Histograma de comentarios a nivel de caracter.

Para la última métrica extraída del conjunto de datos, se identificó la frecuencia de cada término único en cada reacción, es decir, aquellos términos que están vinculados sólo a una reacción específica. Las siguientes figuras muestran las nubes de palabras para estos términos,

considerando cada reacción. La Figura 23 corresponde a la nube de palabras para los términos de comentario asociados con la reacción HAHA, la Figura 24 corresponde a la reacción ANGRY, la Figura 25 representa la reacción SAD y, finalmente, la Figura 26 muestra la reacción LOVE.

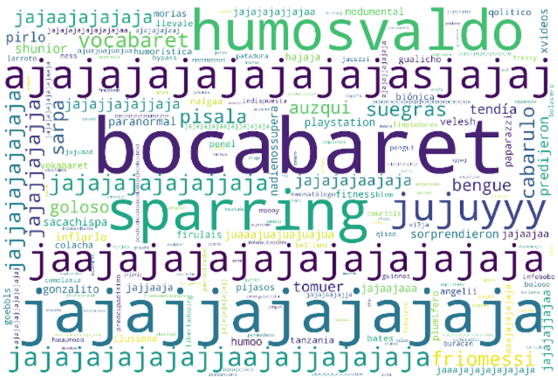


Figura 23: Nube de palabras para la reacción HAHA.

Figura 24: Nube de palabras para la reacción ANGRY.

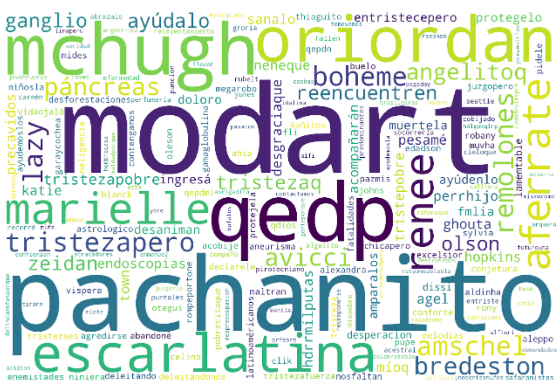


Figura 25: Nube de palabras para la reacción SAD.

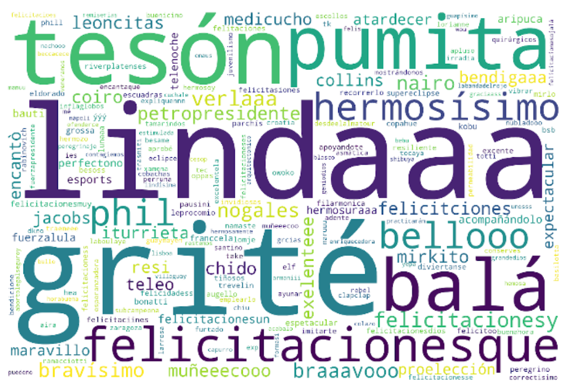


Figura 26: Nube de palabras para la reacción LOVE.

4.5.3 Etiquetado y medición del consenso sobre el conjunto de datos

Habiendo realizado la selección y el filtrado del contenido relevante del conjunto de datos, junto con el cálculo y representación de estadísticas relevantes, resulta pertinente pasar a la etapa de validación del mismo. Como se mencionó en el Capítulo 3, esta etapa crucial suele omitirse o realizarse de manera ad-hoc en los *datasets* recopilados con DS por lo que, en general, no se puede tener certeza acerca de las etiquetas asignadas al contenido. Este problema debe ser abordado ya que se pretende utilizar la DS para el armado de conjuntos de datos.

Para realizar tal validación en el presente capítulo, se utiliza la métrica Kappa de Fleiss mencionada previamente. Dicha métrica es muy frecuente para los conjuntos de datos construidos de manera manual. También, según lo comentado anteriormente en diversas investigaciones, se toma un nivel de consenso moderado, en la escala de Kappa de Fleiss, para considerar a los datos aptos para la construcción de clasificadores basados en ML.

Por otro lado, los datos deberían validarse sin necesidad de recurrir a un re etiquetado manual completo, debido a que se perderían las ventajas de la aplicación de DS. En tal sentido, en (Vosoughi et al., 2015) sobre un *dataset* de 18 millones de tweets clasificados de manera polar con DS, se tomó una muestra de apenas unos 3.000 para validar el mismo. Dicha muestra representa menos del 0,02% del total del conjunto de datos. También (W. Wang et al., 2012), utilizó una muestra de 400 tweets para validar un conjunto de datos de 2.488.982, representando igualmente menos del 0,02% del total.

Para la validación de los datos en el presente trabajo de tesis, se toma este último porcentaje como parámetro y se estiman los valores para extraer una muestra y realizar una validación estadística del conjunto de datos. Dicha muestra se re etiqueta por especialistas en psicología. Finalmente utilizando la etiqueta original y la proporcionada por los especialistas se calculan los valores de Kappa de Fleiss para establecer si los mismos cumplen con el requisito de un consenso moderado.

En consecuencia, sobre los 1.020.557 elementos resultantes del proceso de selección y filtrado, se selecciona una muestra aleatoria de cuádruplas (título, subtítulo, comentario, reacción), con el fin de estimar el valor del parámetro deseado, es decir la medida de acuerdo con Kappa de Fleiss para este conjunto de datos en particular. Para determinar el tamaño de la muestra se utiliza la fórmula de determinación de población finita. Esto se muestra en la siguiente ecuación:

$$n = \frac{N * Z_{\alpha}^{2} * p * q}{e^{2} * (N-1) * Z_{\alpha}^{2} * p * q}$$

Donde n es el tamaño de la muestra resultante, N es el tamaño de la población, Z es el parámetro estadístico en función del nivel de confianza, e es el margen de error, p es la probabilidad de éxito y $q = 1 - p$.

Como las proporciones de p y q son desconocidas para este conjunto de datos, sus valores se establecieron en 0,5 que es lo que se recomienda en tales casos. El nivel de confianza se fijó en el 95% y el error máximo permitido en el 5,23%. Estos dos últimos parámetros fueron los mejores posibles con los recursos disponibles, pero incluso así superan el porcentaje muestral utilizado por (Vosoughi et al., 2015) y (W. Wang et al., 2012) para realizar una validación similar. El tamaño de la muestra resultante fue de 247.

La muestra se dividió en diez conjuntos de 24 o 25 cuádruplas. Luego, cada conjunto se usó para construir un Formulario de Google con una tarea de clasificación por cuádrupla. El tamaño de los cuestionarios se determinó experimentalmente, ya que se observó que los conjuntos más grandes arrojaban peores resultados de concordancia; esto puede deberse a la pérdida de concentración del evaluador humano.

(Hsueh, Melville, & Sindhwani, 2009) sostiene que, para llevar a cabo la fase de etiquetado manual, es conveniente involucrar ***a expertos en la tarea***. Siguiendo esta sugerencia, se solicitó a psicólogos que realizaran la tarea de clasificación manual. A los participantes se les mostró un título de noticia, un subtítulo y un comentario, y luego se les pidió que seleccionaran qué emoción, entre cuatro opciones posibles (me enoja, me entristece, me divierte y me encanta) transmitía el comentario. A los participantes se les permitió seleccionar una segunda opción, pero esto no era obligatorio. Alrededor de 25 psicólogos participaron en esta tarea de clasificación. Cada comentario fue revisado por al menos tres personas, ya que la calidad de las anotaciones se puede mejorar a través de la validación cruzada y la verificación por parte de varios anotadores (Mercado, Villagra, & Errecalde, 2020). También se debe tener en cuenta que el análisis de las emociones en texto puede realizarse desde la perspectiva del escritor o del lector. El primero se refiere a las emociones que tuvo el autor cuando estaba escribiendo el mensaje, mientras que el segundo a la respuesta afectiva de un usuario al ser expuesto a un texto emocional (Yadollahi et al., 2017). En el caso de este trabajo de tesis, se instruyó a los anotadores que consideren la perspectiva del escritor del texto. Los resultados tabulados del proceso de validación del conjunto de datos pueden encontrarse en el material electrónico suplementario dos de (Tessore et al., 2021).

Luego del proceso de etiquetado manual sobre una muestra del conjunto de datos se calcularon varias métricas para establecer el nivel de consenso sobre las etiquetas. En primer

lugar, se calculó la Kappa de Fleiss global pero también se consideró cada reacción versus las demás, es decir, considerando una reacción como categoría y las tres restantes como otra categoría.

Otra medida para calcular la concordancia fue considerar como etiqueta válida la clase más votada para cada comentario, como se hizo en (Chatterjee et al., 2019). Se calculó la medida de acuerdo Kappa de Fleiss para la reacción original y la reacción que recibió la mayor cantidad de votos de los evaluadores por cada comentario. En el caso de un empate en la reacción más votada, se utilizó la respuesta secundaria opcional y la etiqueta original. El efecto logrado con esto fue calcular el consenso exclusivamente entre los etiquetadores y entre estos últimos y la etiqueta original del conjunto de datos.

Métrica	Acuerdo resultante
Fleiss Kappa global	0,49113
Fleiss Kappa ANGRY vs todos	0,4933
Fleiss Kappa HAHA vs todos	0,4989
Fleiss Kappa LOVE vs todos	0,5332
Fleiss Kappa SAD vs todos	0,4240

Tabla 15: Acuerdo entre etiquetadores humanos.

Finalmente, para obtener una mejor perspectiva de los casos desafiantes, los comentarios se analizaron utilizando BabelSenticNet (Vilares, Peng, Satapathy, & Cambria, 2018) para extraer los conceptos principales y la polaridad general de cada clase. Se realizó un análisis global utilizando nubes de conceptos; además, se analizaron varias muestras representativas de cada clase.

Los resultados generales se presentan en la Tabla 15, donde puede verse que el acuerdo es moderado. La puntuación global de Kappa de Fleiss es de 0,49 y, si se consideran las reacciones individuales, LOVE es la más alta y SAD la más baja. Como se puede ver en la Tabla 16, si la reacción original en el conjunto de datos se considera como otro revisor, la puntuación global de Kappa de Fleiss cae a 0,4426, pero aún dentro de la zona de acuerdo moderado, la reacción individual con el valor más alto de consenso sigue siendo LOVE, pero el valor más bajo ahora lo comparten ANGRY y SAD.

Métrica	Acuerdo resultante
Fleiss Kappa global	0,4426
Fleiss Kappa ANGRY vs todos	0,4071
Fleiss Kappa HAHA vs todos	0,4415
Fleiss Kappa LOVE vs todos	0,5452
Fleiss Kappa SAD vs todos	0,4081

Tabla 16: Acuerdo entre etiquetadores humanos y la etiqueta original.

En los resultados presentados en la Tabla 17, para dar más peso a la reacción original y filtrar posibles valores atípicos de clasificación manual, la reacción clasificada manualmente para cada comentario se decidió por votación. Luego se calculó Kappa de Fleiss entre la reacción más votada para cada muestra y la etiqueta original.

Métrica	Acuerdo resultante
Fleiss Kappa primera respuesta	0,4409
Fleiss Kappa primera y segunda respuestas	0,4036
Fleiss Kappa primera respuesta, empates como correctos	0,4701
Fleiss Kappa primera y segunda respuesta, empates correctos	0,4922

Tabla 17: Acuerdo entre la reacción más votada y la etiqueta original.

La segunda medida presentada en la Tabla 17 también considera la reacción secundaria (si es seleccionada) como un voto. En las dos últimas medidas de la misma tabla, si la votación está empatada y la reacción original está entre las más votadas, entonces el resultado de la votación se ajusta a la reacción original. Como se puede observar, todas las medidas también se encuentran dentro de la zona de acuerdo moderado.

Para visualizar dónde estaban los desacuerdos entre la etiqueta original y los evaluadores humanos, se construyó una matriz de confusión; el resultado se presenta en la Figura 27. Como se puede observar, ANGRY fue el más acertado en la predicción, pero también el que presentó más falsos positivos. Cada reacción se confundió con ANGRY, siendo HAHA el peor de los casos. Las tres reacciones restantes no presentaron problemas de clasificación entre sí.

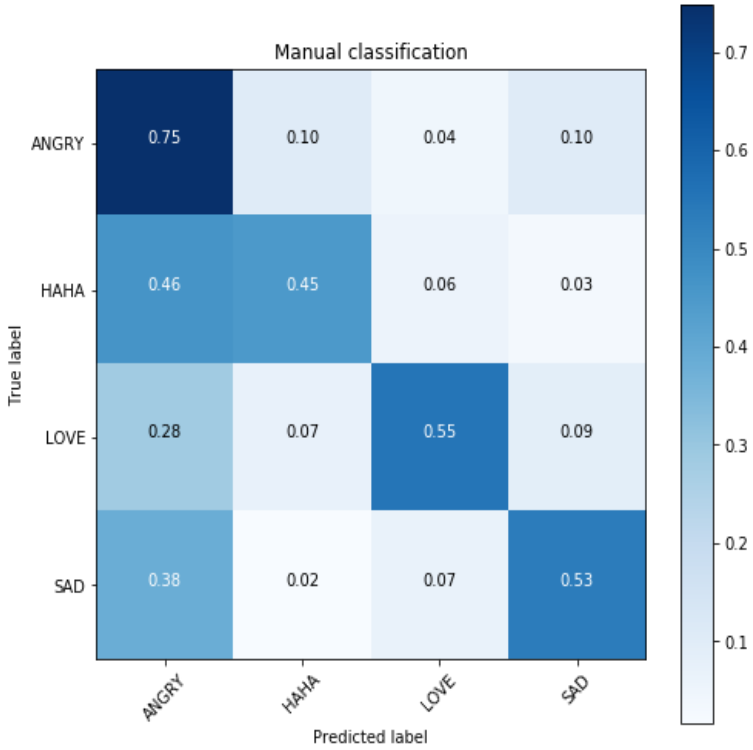


Figura 27: Clasificación manual versus etiqueta original.

4.5.4 Revisión de casos controvertidos

Como se puede ver en la matriz de confusión, la mayoría de los errores se dieron entre ANGRY y el resto de las reacciones. Para analizar la posible causa de estas clasificaciones erróneas, se utilizó la versión en español de BabelSenticNet (Vilares et al., 2018) para extraer y evaluar la polaridad de los conceptos mencionados en los comentarios.

Los resultados se presentan de la Figura 28 a la Figura 31, que muestran una nube de conceptos para cada reacción. El color de los conceptos en la nube indica su polaridad. Cuanto más intenso es el rojo, más negativo es el concepto; cuanto más intenso es el verde, más positivo es el concepto; los conceptos grises están cerca de neutral.

Además, para cada reacción, también se calculó la polaridad promedio de todos los conceptos. La reacción con la polaridad media más negativa es HAHA (-0,00725), seguida de ANGRY (0,02047), SAD (0,03534) y LOVE (0,0989). Esto podría explicar por qué los psicólogos clasificaron erróneamente muchos de los comentarios etiquetados con HAHA como

ANGRY. En cambio, la reacción con polaridad media más positiva, LOVE, fue la que menos se confundió con ANGRY.

Figura 28: Análisis de polaridad para la reacción HAHA.

Figura 29: Análisis de polaridad para la reacción ANGRY.

Figura 30: Análisis de polaridad para la reacción SAD.

Figura 31: Análisis de polaridad para la reacción LOVE.

La Tabla 18 presenta algunos casos representativos de las clasificaciones erróneas más comunes detectadas, que son HAHA, SAD y LOVE clasificadas como ANGRY. Las otras regiones de la matriz de confusión no presentaron un número relevante de errores de clasificación.

#	Reacción	Contenido		Errores
1	HAHA	Título	El incómodo momento que vivió Pedro Pablo Kuczynski al intentar besar el anillo del papa Francisco	2/3 como ANGRY
		Comentario	Rarísimo que un papa se guíe por cuestiones políticas. Rarísimo.	
2	LOVE	Título	Violador de nena atropellado por camión tenía herida de bala	2/3 como ANGRY
		Comentario	q se joda ojala q haya sufrido mucho antes de morir	
3	SAD	Título	A Maru Botana le echan en cara la muerte de su hijo	2/3 como ANGRY
		Comentario	Qué barbaro cuanto odio, estamos en democracia, por si no lo recuerdan, hay liberdad de expresión y Maru defendió su postura, más allá si se comparte o no, se merece todo el respeto. RESPETO que se esta perdiendo!	
4	HAHA	Título	Un hombre condenado por matar al bebé de su amante marchó contra el aborto en Río Grande	2/3 como ANGRY
		Comentario	**Jajajajajajajaja** la doble moral de los pro vida no acaba nunca. No se olviden que también marchan al lado de pedófilos y pederastas que en algún momento les van a meter las manos adentro del pantalón de sus propios hijos. Tienen el cerebro lavado. #YoNoMeMetoEnUteroAjeno	
5	SAD	Título	Se entregó el hombre acusado de violar a un nene en Chaco	2/3 como ANGRY
		Comentario	6 y 15 años de prisión? Nada más?!!!! Degenerado horrible y el nene arruinada su salud	

			física y mental por siempre, que horror, **estoy triste!!!!**	
6	HAHA	Título	Maju Lozano explotó contra Baby Etchecopar	3/3 como ANGRY
		Comentario	Esta gorda haciéndose la víctima por ser mujer **causa mucha mucha gracia**. Marmota, las bardeo por boludas, no por mujeres.	
7	LOVE	Título	Operativo contra los manteros en Liniers: desalojaron 475 puestos ilegales	2/3 como ANGRY
		Comentario	Si quieren trabajar qué paguen impuestos como todo comerciantes x eso ellos venden mas barato x que usan espacio publico y no pagan impuesto y de paso se traen la droga para vender en Argentina como nadie controla nada en este pais. Estamos fritos con esta gente	
8	HAHA	Título	Marcha contra ajuste de planes sociales	2/3 como ANGRY
		Comentario	El día que hagan un reclamo legítimo capaz que el pueblo los acompañe	
9	LOVE	Título	Comienza el juicio a Lázaro Báez por la ruta del dinero K	2/3 como ANGRY
		Comentario	Tiene que ser rápido las pruebas son contundentes hay pruebas de sobra no sé qué tanto tienen que estudiar	

Tabla 18: Ejemplos representativos de clasificaciones erróneas.

En algunos de los comentarios etiquetados como ANGRY, pero mal clasificados como HAHA, el autor del comentario probablemente considera que la persona mencionada en el tema tiene poca o ninguna credibilidad, y todo lo que esa persona hace o dice hace reír al comentarista. También estos comentarios contienen algunas palabras con una connotación muy negativa que pueden haber inducido al error a los revisores humanos. Este es el caso de los comentarios cuatro y ocho de la Tabla 18. Otros comentarios mal clasificados de esta manera fueron sarcásticos (como el comentario número uno).

En cuanto a los comentarios etiquetados como LOVE, pero clasificados como ANGRY, la persona que publicó el comentario está de acuerdo con el hecho reportado en las noticias, pero al hacerlo, también critica algo o a alguien. Se pueden ver ejemplos de este comportamiento en los comentarios dos, siete y nueve de la Tabla 18.

Los comentarios con etiqueta SAD y ANGRY son difíciles de distinguir. Un ejemplo de esto es el comentario tres en la Tabla 18. Una pista para distinguirlos puede ser que los comentarios SAD están escritos de una manera más respetuosa que los comentarios ANGRY.

Finalmente, los comentarios cuatro, cinco y seis en la Tabla 18 presentan sugerencias claras (resaltadas en negrita) sobre cuál es la reacción real. La clasificación errónea en esos casos podría deberse a la falta de concentración de los revisores humanos. Esto puede significar que los cuestionarios probablemente deberían ser más cortos o divididos en varias sesiones de etiquetado, de manera de aprovechar al máximo la capacidad de interpretación de los textos de los etiquetadores.

4.6 Conclusiones del capítulo

En el presente capítulo se realizó la recopilación, preprocesamiento y validación de un conjunto de datos de emoción básica en español. Los resultados del preprocesamiento y la validación presentaron ciertas particularidades que se explayan en las siguientes subsecciones.

4.6.1 Conclusiones acerca de preprocesamiento de conjunto de datos

El análisis de la aplicación de las diversas técnicas de preprocesamiento descritas en la sección 4.4.1 permite arribar a una serie de conclusiones que se presentan a continuación.

La aplicación de las tareas de preprocesamiento mejoró el porcentaje de tokens válidos según la herramienta de corrección de Hunspell. Las transformaciones aplicadas, además, no rompen estructuras complejas como siglas o abreviaturas.

Las tareas que permitieron obtener una mayor reducción en el porcentaje de tokens OOV fueron las de “eliminación de tokens sin de letras”, “procesamiento de palabras acentuadas”, “procesamiento de palabras incorrectas unidas por puntos” y “procesamiento de caracteres redundantes”. Las otras tareas no produjeron una mejora significativa. En la mayoría de los

casos, excepto en "procesamiento de símbolos redundantes" y "convertir el texto a mayúsculas", las tareas que redujeron la cantidad de tokens únicos también redujeron el porcentaje de tokens OOV. Se han obtenido numerosas siglas/abreviaturas aplicando el proceso descrito, pero se requiere una posterior limpieza y verificación manual, para eliminar posibles detecciones erróneas.

Sin embargo, según las pruebas realizadas, la reducción de tokens OOV no se traduce necesariamente en una mayor exactitud de clasificación cuando se las aplica en forma aislada. Aplicarlas en combinación, en general, conduce a obtener los mejores resultados, independientemente del algoritmo y del tipo de ponderación escogida. De todas formas, las mejoras obtenidas en todos los casos, respecto de la línea base, son mínimas, incluso cuando el porcentaje en la reducción de OOV tokens es alta, como por ejemplo en el caso de la tarea “eliminación de tokens sin letras” que consigue una reducción del 40,92% de OOV tokens y la exactitud del algoritmo SVM de hecho empeora respecto de la línea base.

Teniendo en cuenta lo anterior podemos concluir que las tareas de preprocesamiento aplicadas no producen una mejora categórica en el desempeño de los clasificadores basados en ML construidos al menos para este tipo de conjuntos de datos. Esto se condice con lo reportado en (Agrawal & Suri, 2019) y (Tellez et al., 2017) mencionado previamente. Más allá de esto, el paso realizado resulta necesario dado que no se puede conocer de antemano el resultado producido por las tareas de pre procesamiento aplicadas.

4.6.2 Conclusiones acerca de la selección y validación de etiquetas

En la sección 4.4, se presentó el proceso de validación de un conjunto de datos de noticias, comentarios y reacciones emocionales. El conjunto de datos consta de 1.020.557 comentarios, cada uno vinculado a un artículo de noticias (título y subtítulo) y una reacción específica (la clase de valor real). El número de entradas es significativamente mayor que otros conjuntos de datos de emoción básica etiquetados manualmente que se han creado para el idioma español (Díaz-Galiano et al., 2019), lo que se puede lograr fácilmente mediante el uso de etiquetas ruidosas para clasificar el contenido mediante DS. Ningún estudio, al menos para el idioma español, compara la confiabilidad de esas etiquetas contra el etiquetado manual, ni tampoco

vinculó las etiquetas directamente al comentario en lugar de vincularlas al artículo de la noticia original.

Como se vio en las secciones anteriores, el acuerdo medido entre los evaluadores humanos es muy similar al de los evaluadores humanos y la etiqueta original; todas las medidas presentadas se encuentran dentro de la zona de concordancia moderada, que otros autores (Gambino & Calvo, 2019) (Chatterjee et al., 2019) consideraron adecuada para el entrenamiento de clasificadores de emociones basados en ML.

Aunque la medida de concordancia es un poco más baja en comparación con los conjuntos de datos clasificados manualmente, se pueden crear *datasets* más grandes utilizando las pautas presentadas en este capítulo, ya que se requiere menor etiquetado manual, solo necesario para la etapa de validación.

La reacción ANGRY presentó un número significativo de falsos positivos; esto puede ser consecuencia de la actividad de *trolls* sin filtrar, por lo que refinar el proceso de filtrado de *trolls* puede ayudar a mejorar este problema. Esta confusión también podría deberse a un sarcasmo mal interpretado en los comentarios.

5. Construcción de clasificadores y utilización de información contextual

Contenido

Resumen

Después de completar la recopilación, preprocesamiento y validación de etiquetas, este capítulo se centra en las etapas finales del proceso diseñado. Se aborda la selección de un formato de representación y, posteriormente, la elección de un algoritmo de ML para la clasificación. El objetivo principal de este capítulo es doble: por un lado, medir el rendimiento que los clasificadores basados en ML pueden alcanzar al utilizar un conjunto de datos recopilado con DS; y, por otro lado, evaluar los efectos del uso de IC en dichos clasificadores. Se busca obtener una mayor comprensión del potencial y los beneficios de incorporar contexto en los clasificadores, para determinar si contribuye a mejorar su rendimiento general. El capítulo se estructura de la siguiente manera, en la sección 5.1 se realiza una introducción a los temas a tratar. Posteriormente en la sección 5.2 se detallan los formatos de representación y algoritmos de clasificación adoptados. En la sección 5.3 se configuran los clasificadores y se presentan los resultados experimentales. Finalmente, en la sección 5.4 se presentan las conclusiones del capítulo.

5.1 Introducción

En el presente capítulo se desarrollan las dos últimas etapas del proceso que se muestra en la Figura 32. Estas son en primer lugar la selección de un formato de representación para el texto y, en segundo lugar, la selección de un algoritmo de ML. En dichas etapas se privilegió la selección de formatos de representación y algoritmos actuales, pero también que permitieran establecer alguna comparación en cuanto al desempeño que puede lograrse mediante conjuntos de datos recopilados utilizando DS con respecto a los que recurrieron al etiquetado manual.

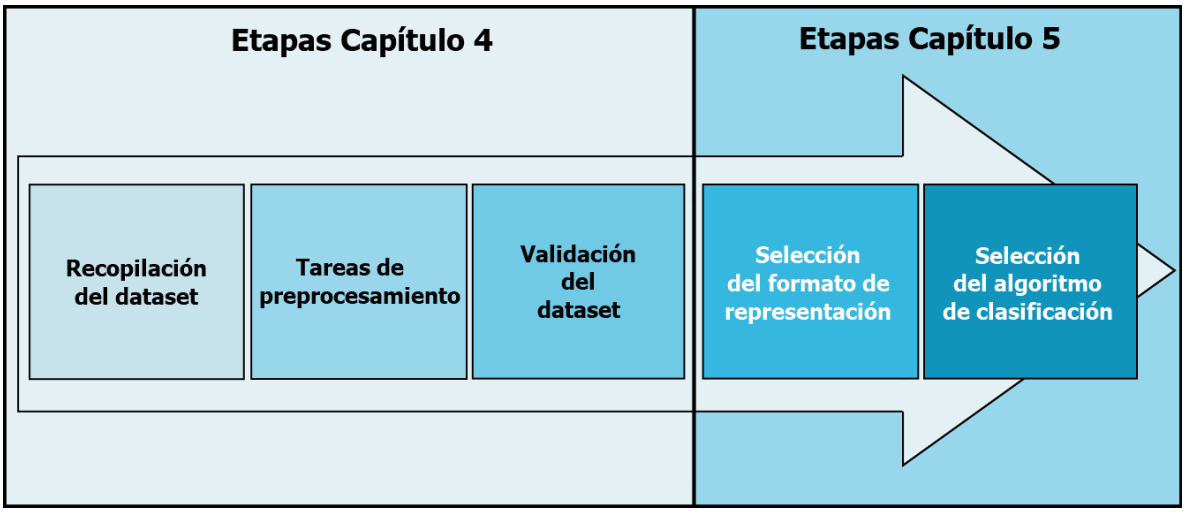


Figura 32: Etapas de proceso discriminadas por capítulo.

Como se mencionó anteriormente, la clasificación de emoción básica es una de las tareas del Análisis de Sentimientos que se suele realizar mediante el uso de varias técnicas de ML. Uno de los principales problemas del Análisis de Sentimientos es la escasa disponibilidad de recursos etiquetados para entrenar adecuadamente los algoritmos de clasificación supervisados, esto es particularmente preocupante en idiomas distintos del inglés, como el español, donde la escasez de estos recursos es la norma. Además, la mayoría de los conjuntos de datos de emoción básica disponibles en español son bastante pequeños, y contienen unos pocos cientos (o miles) de muestras. Por lo general, las muestras sólo contienen un texto breve (con frecuencia un comentario) y una etiqueta (una emoción básica), omitiendo IC crucial que puede ayudar a mejorar los resultados de la tarea de clasificación.

En el conjunto de datos recopilado, además de cada comentario etiquetado, se extrajo el título y el subtítulo de la noticia que lo generó, como IC del contenido. En este capítulo, se

construyen clasificadores con la arquitectura propuesta de la línea base de la competencia SemEval 2019. Esta se basa en la utilización de *embeddings* neurales (GloVe) para la representación de la entrada y un tipo particular de RNN llamadas LSTM. El conjunto de datos utilizado allí, etiquetado de manera manual, está compuesto por diálogos en inglés, cada diálogo tiene tres intervenciones, siendo las dos primeras contexto y la última el texto que se quiere clasificar.

Al utilizar la arquitectura mencionada se busca, por un lado, poder establecer si un clasificador entrenado utilizando un conjunto de datos recopilado mediante DS puede lograr un desempeño similar a uno entrenado mediante datos de clasificación manual, pero además medir el impacto de la IC en la performance del clasificador basado en ML. En el evento mencionado se presentaron diversas configuraciones para los clasificadores que obtuvieron rendimientos superiores a la línea base, sin embargo, en esta tesis se decidió comparar contra esta última debido a que el objetivo final es contrastar el desempeño de clasificadores entrenados con datos obtenidos mediante DS versus datos etiquetados manualmente, en lugar de comparar algoritmos o arquitecturas de clasificación entre sí.

Para la comparación mencionada se realizan dos experimentos. En el primero de ellos, se utiliza el máximo posible de muestras para el entrenamiento, de manera de verificar la ganancia en desempeño que produce la IC en clasificadores basados en ML. Mientras que en el segundo se efectúa un sub muestreo del conjunto de datos recopilado en esta tesis, de manera de igualar el tamaño de los datos de entrenamiento a utilizar al estudio de SemEval 2019 para, por un lado, verificar el desempeño que puede alcanzarse con un conjunto de datos recopilado mediante DS versus etiquetado manual, y por el otro medir el impacto que tiene el uso de IC con ambos conjuntos de datos.

Parte de los resultados expuestos en este capítulo, fueron publicados previamente en (Tessore, Esnaola, Ramón, Lanzarini, & Baldassarri, 2022).

5.2 Selección del formato de representación y el algoritmo de clasificación

Con respecto al formato de representación, se ha establecido que los enfoques tradicionales que involucran vectores dispersos como bolsas de palabras o caracteres no logran capturar información relacionada con el significado de la palabra y tienen limitaciones para mensajes de texto cortos como tweets (Mukherjee, Sahana, & K. Mahanti, 2017). Esto también se aplica al bigrama/trigrama de palabras/n-caracteres, etc. Probablemente por este motivo, muchos estudios (Agrawal & Suri, 2019) (Bae et al., 2019) (Basile et al., 2019) (Huang et al., 2019) (Liang, Ma, & Xu, 2019) (Winata et al., 2019) (Xiao, 2019) utilizan como formato de representación las *embeddings* neuronales, de los cuales los más comúnmente adoptados son Word2Vec (Mikolov et al., 2013), FastText ("FastText by fastText Team," 2017) y GloVe (Pennington et al., 2014). Estas últimas representaciones tienen la ventaja de que codifican el contexto de una determinada palabra en el mismo *embedding*. Las publicaciones originales de *embeddings* neuronales mencionadas presentaron recursos para el idioma inglés. Sin embargo, a lo largo de los años, se han creado recursos para el idioma español (Cañete, 2019) (Cardellino, 2016).

Por otro lado, entre los algoritmos de clasificación más comunes en la literatura se encuentran las RNN que son un tipo específico de ANN. Las RNN permite obtener y procesar información a partir de datos secuenciales, esto las hace particularmente útiles para el NLP porque permiten capturar las dependencias secuenciales y temporales de los datos de entrada. Es importante señalar que la literatura más reciente se inclina hacia variantes de RNN, siendo LSTM la más común (Chatterjee et al., 2019), estas últimas a diferencia de las RNN clásicas no presentan el inconveniente del desvanecimiento del gradiente explicado en el Capítulo 2.

Por otro lado, si bien el conjunto de datos de SemEval-2019-Task 3 consiste en diálogos textuales y el utilizado en el presente documento contiene comentarios en respuesta a publicaciones en las redes sociales, se considera que esto no representa un problema importante ya que estas respuestas son una forma de comunicación o diálogo.

La elección de *embeddings* neuronales, en particular GloVe, y de redes LSTM para los experimentos de este capítulo, responde a su actualidad, pero también a la necesidad de

comparar el desempeño del clasificador construido contra una línea base similar. La arquitectura mencionada fue utilizada en SemEval-2019-Task 3 (Chatterjee et al., 2019), que, además, entre todos los artículos relevados, es el más similar en cuanto a características del conjunto de datos y objetivos. El trabajo mencionado también utiliza IC para la clasificación de emociones básicas; sin embargo, no compara el comportamiento del sistema con y sin IC, además el conjunto de datos allí utilizado se etiquetó manualmente, lo que permite de alguna manera comparar el desempeño que se puede lograr con un clasificador entrenado con dicho tipo de conjunto de datos contra uno recopilado mediante DS, como es el caso del presente trabajo de tesis.

Al adoptar el mencionado trabajo como línea base es necesario remarcar que el objetivo que se persigue es medir cuánto la IC mejora el desempeño del clasificador en lugar de construir el mejor clasificador posible.

5.3 Configuración de los clasificadores

En esta etapa se llevaron a cabo dos experimentos, el primero para medir el impacto de la IC en una tarea de clasificación de emoción básica y el segundo para establecer cómo se comporta un conjunto de datos recopilado mediante DS en comparación con una contraparte clasificada manualmente. El formato de *embedding* utilizado es GloVe (Pennington et al., 2014) y el algoritmo de clasificación seleccionado fue LSTM. Esto se implementó utilizando el *framework Keras* (“Keras,” 2021).

Para la representación de los datos en idioma español, se seleccionaron los *embeddings* en GloVe proporcionados por *Spanish Billion Words Corpus*. Estos se calcularon a partir de un conjunto de datos de 1.400 millones de palabras en español, que contenían 855.380 vectores de 300 dimensiones cada uno. Estos vectores se usaron para construir una capa de *embeddings* no entrenable. El tamaño del vocabulario se limitó a 20.000 palabras.

Luego, la capa anterior se conectó con una capa LSTM de 128 dimensiones con un *dropout* de 0,2 para evitar el *overfitting*. Por último, se añadió una capa densa con la función sigmoide como activación. Para evaluar el desempeño del sistema se seleccionó la función entropía cruzada categórica. Se puede ver un diagrama de la arquitectura utilizada en la Figura 33, donde "*None*" debe reemplazarse con el número de muestras en el conjunto de datos.

La extensión máxima de la entrada se estableció en 100 palabras, suficiente para el 95% de las muestras compuestas por título de noticia, descripción y comentario. Se tomó esta decisión porque si se quisiera abarcar el contenido más extenso, la longitud máxima se habría incrementado significativamente y los beneficios de esta acción no habrían superado las desventajas tanto en la complejidad del sistema como en el tiempo de procesamiento.

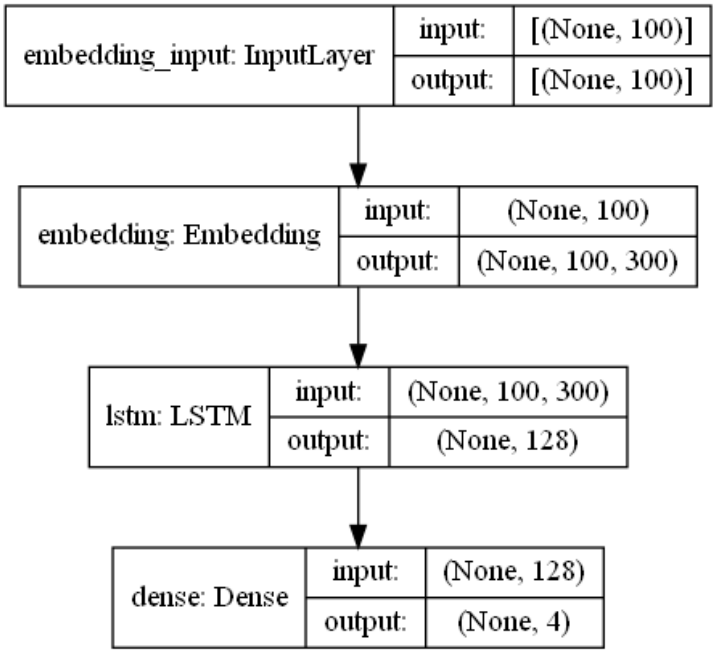


Figura 33: Arquitectura de los clasificadores construidos.

Para ambos experimentos, el sistema fue entrenado en cinco *folds*. En cada uno de ellos, el conjunto de entrenamiento se dividió a su vez en un 80% de entrenamiento y un 20% de validación. Luego de este proceso de validación cruzada, el clasificador fue posteriormente entrenado con todos estos datos (entrenamiento y validación). En todos los casos se estableció un tamaño de *batch* de 200 muestras.

5.3.1 Efecto de considerar la información contextual

El objetivo de este experimento es medir el impacto de la IC en el desempeño de un clasificador de emoción básica para el idioma español, utilizando el conjunto de datos recopilado a lo largo de este trabajo de tesis. Para este experimento, el clasificador descrito en la sección anterior fue entrenado dos veces. Primero, las muestras estaban compuestas de

comentarios, cada uno asociado con una emoción básica particular, es decir, sin IC. Luego, se agregó a las muestras un título de noticia y una descripción como contexto.

Las clases se balancearon realizando un sub muestreo. Los datos resultantes se dividieron aleatoriamente en un conjunto de entrenamiento de 273.712 muestras y un conjunto de pruebas de 68.428.

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Clase ANGRY	**Precision**	0,453	0,450	0,458	0,441	0,456
	Recall	0,490	0,471	0,477	0,508	0,456
	F1	0,471	0,460	0,468	0,472	0,456
Clase SAD	**Precision**	0,604	0,589	0,595	0,592	0,602
	Recall	0,555	0,576	0,558	0,565	0,563
	F1	0,579	0,583	0,576	0,578	0,582
Clase HAHA	**Precision**	0,518	0,506	0,526	0,527	0,503
	Recall	0,529	0,557	0,522	0,517	0,595
	F1	0,523	0,530	0,524	0,522	0,545
Clase LOVE	**Precision**	0,591	0,619	0,580	0,609	0,622
	Recall	0,577	0,542	0,597	0,557	0,547
	F1	0,584	0,578	0,588	0,582	0,582
Todas las clases	**Accuracy**	0,5377	0,5364	0,5385	0,5369	0,5504
	Macro Precision	0,5413	0,5411	0,5399	0,5423	0,5458
	Macro Recall	0,5378	0,5364	0,5384	0,5368	0,5403
	Macro F1	0,5396	0,5388	0,5392	0,5395	0,5431

Tabla 19: Resultados de validación sin IC.

La Tabla 19 presenta los resultados, de cada uno de los *folds*, sobre los datos de validación para el clasificador construido, entrenado sin contexto (es decir, solo comentarios). La Tabla 20 muestra lo mismo, pero en este caso el clasificador fue entrenado con IC.

		Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**
Clase ANGRY	**Precision**	0,565	0,552	0,540	0,567	0,530
	Recall	0,505	0,529	0,578	0,484	0,562
	F1	0,534	0,540	0,558	0,522	0,545
Clase SAD	**Precision**	0,713	0,733	0,736	0,729	0,733
	Recall	0,706	0,692	0,677	0,689	0,687
	F1	0,709	0,712	0,706	0,708	0,709
Clase HAHA	**Precision**	0,623	0,608	0,623	0,589	0,617
	Recall	0,613	0,648	0,627	0,669	0,633
	F1	0,618	0,627	0,625	0,627	0,625
Clase LOVE	**Precision**	0,663	0,720	0,705	0,687	0,713
	Recall	0,751	0,723	0,708	0,732	0,695
	F1	0,704	0,711	0,707	0,709	0,674
Todas las clases	**Accuracy**	0,6435	0,6481	0,6474	0,6433	0,6444
	Micro Precision	0,6409	0,6482	0,6512	0,6430	0,6480
	Micro Recall	0,6437	0,6478	0,6475	0,6436	0,6443
	Micro F1	0,6423	0,6480	0,6493	0,6433	0,6462

Tabla 20: Resultados de validación con IC.

Los resultados para los datos de prueba del clasificador, entrenado con y sin contexto, se encuentran en la Tabla 21 y la Figura 34. Como se puede ver en las figuras y tablas de esta sección, el uso de IC mejora todas las métricas de rendimiento del clasificador, es decir *accuracy*, *precision*, *recall* y F1.

		ANGRY	HAHA	SAD	LOVE	TODAS
Precision	**Sin IC**	0,456	0,503	0,608	0,637	0,5509
	Con IC	0,562	0,601	0,77	0,705	0,6593
Recall	**Sin IC**	0,467	0,584	0,581	0,548	0,5449
	Con IC	0,55	0,679	0,66	0,727	0,6543
F1	**Sin IC**	0,461	0,54	0,594	0,589	0,5479
	Con IC	0,556	0,637	0,771	0,716	0,6568
Accuracy	**Sin IC**	0,4671	0,5836	0,5807	0,5484	0,5449
	Con IC	0,5503	0,6791	0,6604	0,7275	0,6544

Tabla 21: Resultados de pruebas con y sin IC.

En la presente sección se compara la influencia de la IC en el proceso de construcción de un clasificador de emociones utilizando una red LSTM. Las Tabla 19 y la Tabla 20 muestran resultados consistentes de *accuracy*, *precision*, *recall* y F1 para cada *fold* y clase. Además, las medidas de rendimiento que se muestran para el clasificador entrenado con IC superan al entrenado sin ella. Este también es el caso de los resultados de las pruebas que se presentan en la Tabla 21, que muestran una ganancia de rendimiento de alrededor del 10% para cada métrica.

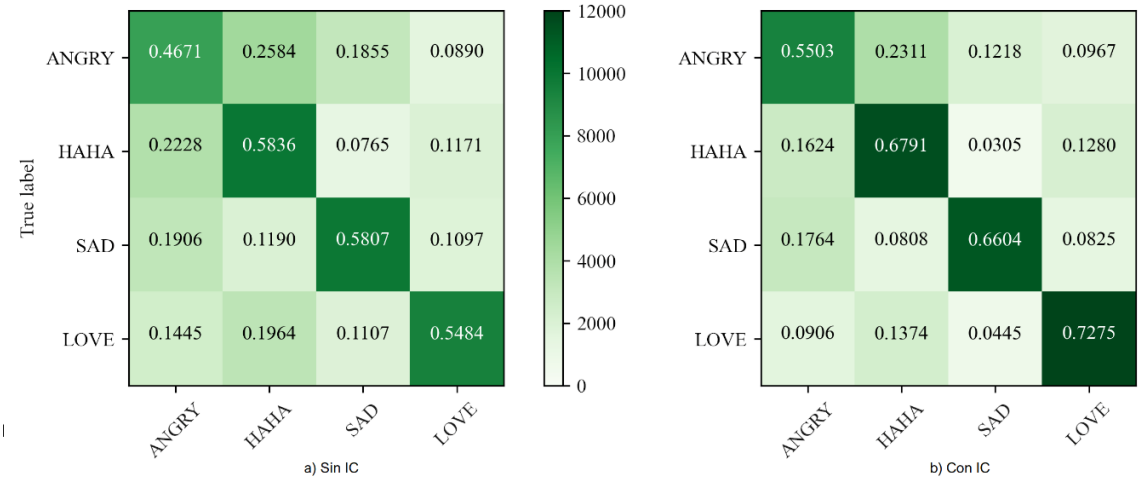


Figura 34: Matrices de confusión para los resultados de pruebas.

5.3.2 Comparación con los resultados obtenidos en otros estudios similares

Para medir la mejora del desempeño del clasificador a través de la incorporación de la IC y realizar comparaciones, se construyó una línea de base utilizando los conjuntos de datos con y sin IC para SemEval-2019-Task 3 y este estudio.

Cabe señalar que (Chatterjee et al., 2019) no informa el rendimiento del clasificador sin contexto. Sin embargo, dado que el conjunto de datos utilizado está disponible, las partes que representan contexto son reconocibles y se describe la arquitectura del modelo, se decidió recrear esos resultados, tanto en presencia o ausencia de IC. Estos resultados se denominan SemEval-2019-Task 3*, para señalar que no se corresponden fielmente con los reportados en el trabajo original de la competencia.

La comparación entre ambos estudios se realiza midiendo, y posteriormente comparando, el porcentaje de mejora obtenido en cada métrica de rendimiento evaluada al incorporar la IC. Por otro lado, si bien no es posible realizar una comparación estricta, ya que (Chatterjee et al., 2019) utiliza un conjunto de datos diferente, datos en inglés y otra fuente de *embeddings*, se busca poder establecer si un clasificador entrenado utilizando un conjunto de datos recopilado mediante DS puede lograr un desempeño aceptable con respecto a uno entrenado mediante datos de clasificación manual.

	SemEval-2019-Task 3*			**Este trabajo**		
	Sin Contexto	**Con Contexto**	**Mejora (%)**	**Sin Contexto**	**Con Contexto**	**Mejora (%)**
Accuracy	0,8403	0,8448	0,5355	0,4742	0,5444	14,8039
Precision	0,4755	0,4719	-0,7571	0,4783	0,5570	16,4541
Recall	0,6689	0,7215	7,8637	0,4740	0,5437	14,7046
F1	0,5559	0,5706	2,6444	0,4762	0,5503	15,5607

Tabla 22: Resultados de la prueba para el clasificador entrenado con el conjunto de datos sub muestreado.

Además, para poder trabajar con conjuntos de datos comparables en tamaño y distribución de clases, aunque sean diferentes, el conjunto de datos recopilado en esta tesis se sub muestreó para que coincidiera con el tamaño del conjunto de datos utilizado en la Task-3 de SemEval-

2019. Los resultados de la prueba, con y sin IC, se presentan en la Tabla 22 tanto para la Task-3* de SemEval-2019 como para el clasificador construido en este estudio. En esta tabla se muestra el porcentaje de mejora alcanzado mediante el uso de IC con el fin de establecer una base de comparación entre los estudios. La Figura 35 da más detalles con la matriz de confusión.

Estos resultados muestran que, si bien el clasificador SemEval-2019 Task 3* obtuvo poca o ninguna ganancia mediante el uso IC en casi todas las métricas, su uso tuvo un gran impacto para el conjunto de datos recopilado en este trabajo de tesis, ya que se logró una mejora significativa.

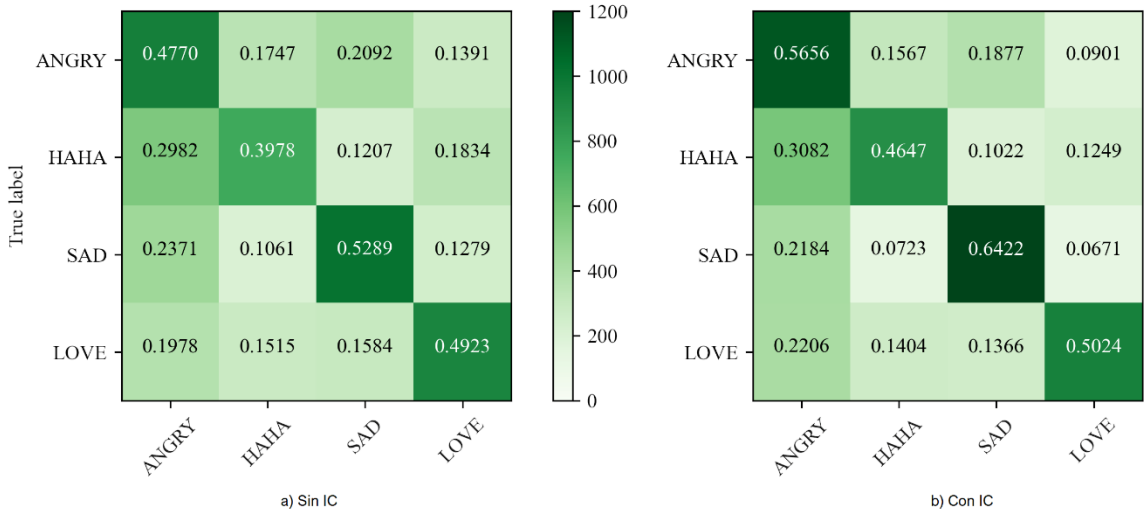


Figura 35: Matriz de confusión para los datos de prueba (conjunto de datos sub muestreado).

5.4 Conclusiones del capítulo

Los resultados muestran que la IC, como los titulares de noticias o resúmenes, ayuda a mejorar el desempeño de clasificadores basados en ML sobre un conjunto de datos de emoción básica obtenido mediante DS.

Al efectuar un sub muestreo, para igualar el tamaño del conjunto de datos elaborado en esta tesis con respecto al que se compara, se determinó que el desempeño obtenido por el clasificador construido es muy similar al presentado en dicha competencia. Por otro lado, al utilizar la totalidad de las muestras recopiladas se obtiene una mejora considerable, lo que

demuestra las ventajas de la utilización de la DS para el Análisis de Sentimientos en la tarea de clasificación de emoción básica.

Los experimentos realizados mostraron que el uso de IC produjo una mejora en el desempeño de los clasificadores construidos tanto individualmente para cada clase como globalmente. Como los clasificadores basados en ML pueden beneficiarse del uso de IC, los investigadores que planean crear conjuntos de datos de emociones básicas deberían considerar capturar estos datos que generalmente están disponibles pero que suelen ignorarse.

El Experimento 1 reveló el impacto del uso de IC para un conjunto de datos de emoción básica recopilado mediante DS, esto se evidencia en la mejora en el desempeño de los clasificadores construidos con IC (Tabla 21 y la Figura 34 b) contra los que no utilizaron IC (Tabla 20 y Figura 34 a)

El Experimento 2 mostró que la ganancia de rendimiento de IC en un conjunto de datos de títulos, subtítulos y comentarios de redes sociales resultó ser mayor que la obtenida para un conjunto de datos de diálogos textuales, al menos para los conjuntos de datos analizados en este estudio. Además, como muestra la Tabla 22, las puntuaciones F1 para ambos clasificadores entrenados con IC son similares, y es importante señalar que el conjunto de datos utilizado en SemEval2019-Task 3 se anotó manualmente, mientras que el utilizado en este estudio se construyó de forma semiautomática utilizando DS.

También es importante señalar que la tarea de clasificación descrita en este capítulo fue más difícil que la tarea en la competencia SemEval-2019-Task 3, ya que requirió clasificar 4 clases en lugar de 3. Además, el conjunto de datos SemEval-2019-Task 3 incluyó la categoría “otros”, lo que hace que generalmente el contenido más desafiante caiga en dicha categoría.

Por otro lado, la mayor cantidad de muestras disponibles en el conjunto de datos utilizado permitió que el clasificador entrenado con IC superara el puntaje F1 obtenido en SemEval-2019-Task 3*. Dado que este conjunto de datos se compiló de forma semiautomática con DS, el proceso de recopilación puede servir como guía para los investigadores que trabajan en el área, no solo para construir conjuntos de datos más grandes, sino también para construir clasificadores de emociones básicas que logren mejores desempeños.

6. Conclusiones y trabajos futuros

Contenido

Resumen

Este capítulo resume las contribuciones de esta Tesis Doctoral en lo que respecta a la metodología para la creación de recursos para el Análisis de Sentimientos presentada. La Sección 6.1 expone las principales conclusiones que se pueden extraer de este trabajo, mientras que la Sección 6.2 sugiere algunas líneas de investigación para trabajos futuros.

6.1 Contribuciones de la tesis

Las principales contribuciones del presente trabajo de tesis son el desarrollo de una metodología para el Análisis de Sentimientos para la creación de conjuntos de datos de emoción básica y su posterior uso para el entrenamiento de clasificadores basados en ML. A partir de la ejecución de dicha metodología en el presente trabajo, se recopiló un conjunto de datos de emoción básica para el idioma español con un número de muestras mucho mayor a los que se encuentran actualmente en la literatura.

A su vez, la ejecución de la metodología permitió obtener conclusiones específicas de cada etapa de la misma.

La etapa en la cual se ejecutan las tareas de limpieza y preprocesamiento del conjunto de datos, permitió demostrar que las mismas no necesariamente se traducen en una mejora categórica en el desempeño de clasificadores basados en ML, con lo cual es necesario realizar una comparación del desempeño del clasificador construido con los datos originales contra los datos luego de ejecutar el preprocesamiento. No obstante, algunas de las tareas aplicadas resultan imprescindibles para el proceso de validación del conjunto de datos, los ejemplos más claros de esto es el filtrado de comentarios provenientes de *trolls* y de comentarios de escasa longitud, ya que ambos se vuelven extremadamente difíciles de clasificar por los etiquetadores humanos.

La etapa de validación del conjunto de datos involucró un muestreo del conjunto de datos y el re etiquetado de dicha muestra por expertos en psicología para posteriormente calcular varias métricas basadas en la Kappa de Fleiss. Dicha métrica no solía ser tenida en cuenta para la validación de conjuntos de datos construidos con DS, posiblemente porque se presuponía necesario el re etiquetado de todo el conjunto de datos. La ejecución de la etapa de validación, permitió concluir que es posible construir conjuntos de datos de emoción básica con nivel de consenso moderado, medido por Kappa de Fleiss. Dicho nivel de consenso es considerado aceptable por distintos autores para la construcción de clasificadores basados en ML.

Por último, la construcción de los clasificadores permitió verificar su conocida sensibilidad al aumento del tamaño del conjunto de datos utilizado para el entrenamiento. Esto puede verse claramente al comparar los resultados obtenidos al entrenar el clasificador con todos los datos

disponibles contra el entrenado con el conjunto de datos sub muestreado. Por otro lado, la utilización de IC produjo una mejora categórica en todas las métricas de los clasificadores construidos, en consecuencia, es recomendable que investigadores que sigan la metodología propuesta en este trabajo de tesis presten atención también a la captura de la misma, ya que generalmente se encuentra accesible y suele ser ignorada. También puede concluirse que los clasificadores entrenados con conjuntos de datos construidos mediante DS pueden lograr rendimientos comparables a los entrenados con conjuntos de datos con etiquetado manual. Esto se evidencia de las métricas calculadas en el Experimento 2 del Capítulo 5.

6.2 Trabajos futuros

Los siguientes pasos de esta investigación deberían estar enfocados, en primer lugar, en probar la metodología para el Análisis de Sentimientos descrita para distintos conjuntos de datos, tipos de datos de entrada, idiomas, procesos de validación y algoritmos de clasificación. En segundo lugar, utilizar distintos tipos de etiquetas para clasificar automáticamente el contenido. Por último, otro aspecto a considerar es utilizar distintos tipos de IC y medir su influencia en el desempeño de los clasificadores basados en ML.

Probar la metodología para distintas fuentes de datos permitiría dar certezas acerca de la utilidad de las tareas de preprocesamiento, verificando si la escasa mejora mostrada para el *dataset* recopilado en este trabajo es algo habitual o un hecho aislado. Por otra parte, ejecutar el proceso de validación de etiquetas sobre otros conjuntos de datos (particularmente los extraídos de redes sociales), podría arrojar más pistas acerca de qué fuentes de datos suelen cumplir con los requisitos de consenso moderado que la mayoría de los investigadores consideran aceptable.

El idioma de los datos a recopilar debería ser tenido en cuenta a la hora de aplicar la metodología descrita. En primer lugar, adaptando las tareas de preprocesamiento a realizar, ya que determinadas tareas, por ejemplo, la corrección de acentos, pueden tener utilidad en algunos idiomas y no en otros. Otro aspecto a considerar es que la efectividad de las etiquetas recopiladas mediante DS puede variar de un idioma al otro, por lo que resulta necesario repetir los experimentos para cada uno de los distintos idiomas. Particularmente sí que persigue el objetivo de realizar Análisis de Sentimientos multilenguaje.

Hace ya un tiempo, se ha popularizado el Análisis de Sentimientos multimodal, en consecuencia, otro aspecto interesante a considerar sería incorporar otros formatos de entrada, como pueden ser audio, video, imágenes, o incluso combinaciones de los anteriores. Esto naturalmente requeriría la adaptación de toda la metodología desde el preprocesamiento, validación y la posterior construcción de clasificadores.

Con respecto al proceso de validación de las etiquetas, sería interesante realizar diversas variaciones en los cuestionarios para medir el impacto en los resultados del consenso. En primer lugar, variar la longitud de los cuestionarios de validación, ya que, se presume que la extensión de los mismos puede impactar en la concentración de los etiquetadores. En segundo lugar, estudiar los efectos si el re etiquetado es presencial o a distancia.

Por otro lado, si se desea optimizar el desempeño del clasificador a construir, podrían adoptarse otras configuraciones, por ejemplo, las de los sistemas presentados en la competencia SemEval-2019-Task 3 como sugerencias (Agrawal & Suri, 2019) (Bae et al., 2019) (Basile et al., 2019) (Huang et al., 2019) (Liang et al., 2019) (Winata et al., 2019) (Xiao, 2019) y verificar si se comporta de manera similar, probando otros tipos de IC (datos de clima, geolocalización, historial del usuario, estado de ánimo del usuario, etc.) para mejorar el proceso de clasificación como se muestra en los artículos discutidos en el estado de la cuestión y, por último, verificar si el comportamiento mostrado también se da en otros tipos de conjuntos de datos u otros algoritmos de ML. También se debe explorar el uso de información semántica, ya que puede ayudar a mejorar el desempeño de la clasificación (K. Li, 2021). Otro posible aspecto a considerar es la experimentación con otros formatos de representación novedosos como es el caso de BERT (Devlin, Chang, Lee, & Toutanova, 2019).

El presente trabajo de tesis, sin embargo, tiene algunas limitaciones. Primero, en relación con el uso de IC en conjuntos de datos de DS, dado que las etiquetas ya están presentes en los datos, los compiladores no pueden elegir cómo categorizar el contenido. Además, se deben realizar más pruebas para demostrar la consistencia de este tipo de mejora utilizando IC con otros tipos de conjuntos de datos y de las que puedan surgir nuevas recomendaciones de manera de poder optimizar la metodología planteada.

Referencias

Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, *2*(7), 1–24. https://doi.org/10.1002/eng2.12189

Agarwal, B., Mittal, N., Bansal, P., & Garg, S. (2015). Sentiment analysis using common-sense and context information. *Computational Intelligence and Neuroscience*, *2015*. https://doi.org/10.1155/2015/715730

Aggarwal, C. C. (2018). Machine Learning for Text: An Introduction. In *Machine Learning for Text* (pp. 1–16). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-73531-3_1

Aggarwal, C. C., & Zhai, C. X. (2012). *Mining Text Data*. (C. C. Aggarwal & C. Zhai, Eds.), *Mining Text Data* (Vol. 9781461432). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4614-3223-4

Agrawal, P., & Suri, A. (2019). NELEC at SemEval-2019 Task 3: Think Twice Before Going Deep. In S. M. M. Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki (Ed.), *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval-2019)* (ACL, pp. 266–271). Minneapolis, Minnesota, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S19-2.pdf

Ahmad Refaee, E. A. (2016). *Sentiment Analysis for Micro-blogging Platforms in Arabic*. Heriot-Watt University.

Al-Rfou, R. (2020). PYCLD2 - Python Bindings to CLD2. Retrieved from https://pypi.org/project/pycld2/

Alegria, I., Aranberri, N., Comas, P. R., Fresno, V., Gamallo, P., Padró, L., … Zubiaga, A. (2015). TweetNorm: a benchmark for lexical normalization of Spanish tweets. *Language*

Resources and Evaluation, *49*(4), 883–905. https://doi.org/10.1007/s10579-015-9315-6

Alm, C. O. (2008). *Affect in Text and Speech*. https://doi.org/10.1.1.172.9934

Alm, C. O., & Sproat, R. (2005). Emotional Sequencing and Development in Fairy Tales. In J. Tao, T. Tan, & R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction. ACII 2005. Lecture Notes in Computer Science, vol 3784.* (pp. 668–674). Springer Berlin Heidelberg. https://doi.org/10.1007/11573548_86

Alswaidan, N., & Menai, M. E. B. (2020). *A survey of state-of-the-art approaches for emotion recognition in text*. *Knowledge and Information Systems*. Springer London. https://doi.org/10.1007/s10115-020-01449-0

Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4629 LNAI*(June), 196–205. https://doi.org/10.1007/978-3-540-74628-7_27

Arnold, M. (1960). *Emotion and personality*. Columbia University Press.

Baccianella, S., Esuli, A., & Sebastiani, F. (2008). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 7th Conference on Language Resources and Evaluation LREC10* (pp. 417–422). Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

Bae, S., Choi, J., & Lee, S. (2019). SNU IDS at SemEval-2019 Task 3: Addressing Training-Test Class Distribution Mismatch in Conversational Classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 312–317). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2054

Baeza-Yates, R., & Riberio-Neto, B. (2011). *Modern Information Retrieval 2nd edition*. Addison-Wesley Longman Publishing Co., Inc. https://doi.org/10.5555/553876

Bakker, I., van der Voordt, T., Vink, P., & de Boon, J. (2014). Pleasure, Arousal, Dominance: Mehrabian and Russell revisited. *Current Psychology*, *33*(3), 405–421.

https://doi.org/10.1007/s12144-014-9219-4

Balahur, A., Turchi, M., & Steinberger, R. (2011). Resource Creation and Evaluation for Multilingual Sentiment Analysis in Social Media Texts. *Proceedings of the Language Resources and Evaluation Conference*, 4265–4269. Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/965_Paper.pdf

Baldwin, T., Cook, P., Lui, M., Mackinlay, A., & Wang, L. (2013). How Noisy Social Media Text , How Diffrnt Social Media Sources ? *Proc. IJCNLP 2013*, (October), 356–364. Retrieved from http://rankings.big-boards.com

Banko, M., & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01* (Vol. 56, pp. 26–33). Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/1073012.1073017

Basile, A., Franco-Salvador, M., Pawar, N., Štajner, S., Chinea Rios, M., & Benajiba, Y. (2019). SymantoResearch at SemEval-2019 Task 3: Combined Neural Models for Emotion Classification in Human-Chatbot Conversations. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 330–334). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2057

Batista, R. (2015). Propuesta de Red cubana Aurora para la colaboración médica a través de Infomed utilizando un enfoque de redes sociales. In *Conference: Cuba Salud 2015. Convención Internacional en Salud Pública*.

Becerra, M. (2018). Medios digitales en Argentina: la película y la foto. *Letra P*. Retrieved from https://www.letrap.com.ar/nota/2018-9-20-16-3-0-medios-digitales-en-argentina-la-pelicula-y-la-foto

Bengio, Y., Goodfellow, I., & Courville, A. (2015). *Deep Learning*.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc. Retrieved from http://www.nltk.org/book/

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *ACL 2007 - Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 440–447.

Bollen, J., & Mao, H. (2011). Twitter Mood as a Stock Market Predictor. *Computer*, *44*(10), 91–94. https://doi.org/10.1109/MC.2011.323

Bosio, S. (2014). Diccionario para corrección ortográfica en español de OpenOffice.org. Versión localizada para Argentina. Retrieved from https://github.com/elastic/hunspell/tree/master/dicts/es_AR

Brooke, J., Tofiloski, M., & Taboada, M. (2009). Cross-linguistic sentiment analysis: From English to Spanish. *International Conference Recent Advances in Natural Language Processing, RANLP*, 50–54.

Buechel, S., & Hahn, U. (2017). EMOBANK: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, *2*, 578–585. https://doi.org/10.18653/v1/e17-2092

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., … Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project, 1–15. Retrieved from http://arxiv.org/abs/1309.0238

Cambria, E. (2016). Affective Computing and Sentiment Analysis. *IEEE Intelligent Systems*, *31*(2), 102–107. https://doi.org/10.1109/MIS.2016.31

Cambria, E., Havasi, C., & Hussain, A. (2012). SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, FLAIRS-25* (pp. 202–207).

Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment Analysis Is a Big Suitcase. *IEEE Intelligent Systems*, *32*(6), 74–80. https://doi.org/10.1109/MIS.2017.4531228

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, *28*(2), 15–21. https://doi.org/10.1109/MIS.2013.30

Cañete, J. (2019). FastText at Bot Center repo. Retrieved from https://github.com/BotCenter/spanishWordEmbeddings

Cardellino, C. (2016). Spanish Billion Words Corpus and Embeddings. Retrieved from https://crscardellino.github.io/SBWCE/

Carletta, J. (1996). Squibs and Discussions: Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, *22*(2), 248–254.

Chandrasekar, P., & Qian, K. (2016). The Impact of Data Preprocessing on the Performance of a Naive Bayes Classifier. In *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)* (pp. 618–619). Atlanta, Alabama, USA: IEEE. https://doi.org/10.1109/COMPSAC.2016.205

Chatterjee, A., Narahari, K. N., Joshi, M., & Agrawal, P. (2019). SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 39–48). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2005

Chaturvedi, I., Cambria, E., & Vilares, D. (2016). Lyapunov filtering of objectivity for Spanish Sentiment Model. In *2016 International Joint Conference on Neural Networks (IJCNN)* (pp. 4474–4481). Vancouver, British Columbia, Canada: IEEE. https://doi.org/10.1109/IJCNN.2016.7727785

Chen, S. Y., Hsu, C. C., Kuo, C. C., Huang, T. H. K., & Ku, L. W. (2019). Emotionlines: An emotion corpus of multi-party conversations. In *LREC 2018 - 11th International Conference on Language Resources and Evaluation* (pp. 1597–1601).

Chiny, M., Chihab, M., Bencharef, O., & Chihab, Y. (2022). Netflix Recommendation System

based on TF-IDF and Cosine Similarity Algorithms, (Bml 2021), 15–20. https://doi.org/10.5220/0010727500003101

Clark, E., & Araki, K. (2011). Text normalization in social media: Progress, problems and applications for a pre-processing system of casual English. *Procedia - Social and Behavioral Sciences*, *27*(Pacling), 2–11. https://doi.org/10.1016/j.sbspro.2011.10.577

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Craker, N., & March, E. (2016). The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences*, *102*, 79–84. https://doi.org/10.1016/j.paid.2016.06.043

Dabrowski, J. J., & De Villiers, J. P. (2015). A unified model for context-based behavioural modelling and classification. *Expert Systems with Applications*, *42*(19), 6738–6757. https://doi.org/10.1016/j.eswa.2015.04.061

De Albornoz, J. C., Plaza, L., & Gervas, P. (2012). SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, 3562–3567.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Díaz-Galiano, M. C., García-Vega, M., Edgar, C., Chiruzzo, L., Garcia-Cumbreras, M. A., Eugenio, M. C., … Miranda, S. (2019). Overview of TASS 2019 : One More Further for the Global Spanish Sentiment Analysis Corpus. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) Co-Located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)*, *2421*(1), 550–560. Retrieved

from http://ceur-ws.org/Vol-2421/TASS_overview.pdf

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124–129. https://doi.org/10.1037/h0030377

Elnagar, A., Khalifa, Y. S., & Einea, A. (2018). Hotel arabic-reviews dataset construction for sentiment analysis applications. *Studies in Computational Intelligence*, *740*, 35–52. https://doi.org/10.1007/978-3-319-67056-0_3

Esnaola, L., & Tessore, J. P. (2019). Pruebas realizadas para medición de efectividad de tareas de preprocesamiento sobre texto extraído de Facebook. Retrieved from https://tinyurl.com/y3v7nzwv

Esnaola, L., Tessore, J. P., Ramon, H., & Russo, C. (2019). Effectiveness of preprocessing techniques over social media texts for the improvement of machine learning based classifiers. In *2019 XLV Latin American Computing Conference (CLEI)* (pp. 1–10). IEEE. https://doi.org/10.1109/CLEI47609.2019.235076

Facebook. (2020). Facebook API Graph. Retrieved from http://developers.facebook.com

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34. https://doi.org/10.1145/240455.240464

Feldman, R., & Sanger, J. (2006). *The Text Mining Handbook*. *The Text Mining Handbook*. Cambridge University Press. https://doi.org/10.1017/CBO9780511546914

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/https://doi.org/10.1037/h0031619

Gambino, O. J., & Calvo, H. (2019). Predicting emotional reactions to news articles in social networks. *Computer Speech & Language*, *58*, 280–303. https://doi.org/10.1016/j.csl.2019.03.004

García-Vega, M., Díaz-Galiano, M. C., García-Cumbreras, M., Del Arco, F. M. P., Montejo-

Ráez, A., Jiménez-Zafra, S. M., … Moctezuma, D. (2020). Overview of TASS 2020: Introducing Emotion Detection. In *CEUR Workshop Proceedings* (Vol. 2664, pp. 163–170).

Ghazi, D., Inkpen, D., & Szpakowicz, S. (2015). Detecting Emotion Stimuli in Emotion-Bearing Sentences. In A. Gelbukh (Ed.), *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9042, pp. 152–165). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-18117-2_12

Gholipour Shahraki, A., & Zaïane, O. R. (2017). Lexical and Learning-based Emotion Mining from Text. *University of Alberta, Canada*, 24–55. Retrieved from https://pdfs.semanticscholar.org/da07/08bf99942992bcbfe1919d6755bc8168d46e.pdf

Ghosal, D., Akhtar, M. S., Chauhan, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2018). Contextual Inter-modal Attention for Multi-modal Sentiment Analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3454–3466). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1382

Ghosh, S. (2020). Machine Learning Word Embeddings An Application of Neural Networks.

Go, A., Bhayani, R., & Huang, L. (2009). *Twitter Sentiment Classification using Distant Supervision*. *CS224N project report, Stanford* (Vol. 1 (12)).

Hearst, M. A. (1997). TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, *23*(1), 33–64.

Hekkert, P., & Desmet, P. (2002). The Basis of Product Emotions. https://doi.org/10.1201/9780203302279.ch4

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *Journal for Language Technology and Computational Linguistics*, *20*(1), 19–62. https://doi.org/10.21248/jlcl.20.2005.68

Hsueh, P., Melville, P., & Sindhwani, V. (2009). Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria. In E. Ringger, R. Haertel, & K. Tomanek (Eds.), *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing - HLT '09* (p. 27). Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/1564131.1564137

Hu, X., & Liu, H. (2012). Text analytics in social media. *Mining Text Data*, *9781461432*, 385–414. https://doi.org/10.1007/978-1-4614-3223-4_12

Huang, C., Trabelsi, A., & Zaïane, O. (2019). ANA at SemEval-2019 Task 3: Contextual Emotion detection in Conversations through hierarchical LSTMs and BERT. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 49–53). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2006

Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing of Contents, 1–621.

Justo, R., Alcaide, J. M., Torres, M. I., & Walker, M. (2018). Detection of Sarcasm and Nastiness: New Resources for Spanish Language. *Cognitive Computation*, *10*(6), 1135–1151. https://doi.org/10.1007/s12559-018-9578-5

Kao, A., & Poteet, S. R. (2007). *Natural Language Processing and Text Mining*. (A. Kao & S. R. Poteet, Eds.). London: Springer London. https://doi.org/10.1007/978-1-84628-754-1

Kaur, W., Balakrishnan, V., Rana, O., & Sinniah, A. (2019). Liking, sharing, commenting and reacting on Facebook: User behaviors' impact on sentiment intensity. *Telematics and Informatics*, *39*(June), 25–36. https://doi.org/10.1016/j.tele.2018.12.005

Keras. (2021). Retrieved from https://keras.io/

Kilgarriff, A., & Fellbaum, C. (1998). *WordNet: an electronic lexical database*. (C. Fellbaum, Ed.), *Language* (Vol. 76). The MIT Press. https://doi.org/10.7551/mitpress/7287.001.0001

Krouska, A., Troussas, C., & Virvou, M. (2016). The effect of preprocessing techniques on

Twitter Sentiment Analysis, (July). https://doi.org/10.1109/IISA.2016.7785373

Kumawat, D., & Jain, V. (2015). POS Tagging Approaches: A Comparison. *International Journal of Computer Applications*, *118*(6), 32–38. https://doi.org/10.5120/20752-3148

Layton, R., Watters, P., & Dazeley, R. (2012). Recentred local profiles for authorship attribution. *Natural Language Engineering*, *18*(3), 293–312. https://doi.org/10.1017/S1351324911000180

Lazarus, R. S. (1991). Progress on a cognitive-motivational-relational theory of emotion. *American Psychologist*, *46*(8), 819–834. https://doi.org/10.1037/0003-066X.46.8.819

Li, K. (2021). Haha at emoevales 2021: Sentiment analysis in spanish tweets with cross-lingual model. *CEUR Workshop Proceedings*, *2943*(September), 49–58.

Li, Y., & Su, H. (2017). DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset - ACL Anthology, 986–995. Retrieved from https://aclanthology.org/I17-1099/

Liang, X., Ma, Y., & Xu, M. (2019). THU-HCSI at SemEval-2019 Task 3: Hierarchical Ensemble Classification of Contextual Emotion in Conversation. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 345–349). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2060

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. *Synthesis Lectures on Human Language Technologies* (Vol. 5). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-02145-9

Liu, C., Osama, M., & de Andrade, A. (2019). DenS: A dataset for multi-class emotion analysis. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 6293–6298. https://doi.org/10.18653/v1/d19-1656

Liu, V., Banea, C., & Mihalcea, R. (2017). Grounded emotions. *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, *2018-Janua*,

477–483. https://doi.org/10.1109/ACII.2017.8273642

Lövheim, H. (2012). A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical Hypotheses*, *78*(2), 341–348. https://doi.org/10.1016/j.mehy.2011.11.016

Lovins, J. B. (1968). Development of a Stemming Algorithm*. *Mechanical Translation and Computational Linguistic*, *11*(1 and 2), 22–31.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, *1*, 142–150.

Majumder, N., Poria, S., Gelbukh, A., & Cambria, E. (2017). Deep Learning-Based Document Modeling for Personality Detection from Text. *IEEE Intelligent Systems*, *32*(2), 74–79. https://doi.org/10.1109/MIS.2017.23

Majumder, N., Poria, S., Peng, H., Chhaya, N., Cambria, E., Gelbukh, A., & Cambria, E. (2019). Sentiment and Sarcasm Classification With Multitask Learning. *IEEE Intelligent Systems*, *34*(3), 38–43. https://doi.org/10.1109/MIS.2019.2904691

Manjrekar, O. N., & Dudukovic, M. P. (2019). Identification of flow regime in a bubble column reactor with a combination of optical probe data and machine learning technique. *Chemical Engineering Science: X*, *2*, 100023. https://doi.org/10.1016/j.cesx.2019.100023

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press. https://doi.org/10.1108/00242530410565256

Martín, C., Aguilar, R. M., Torres, J. M., & Díaz, S. (2020). Supervisión remota en el entrenamiento de un clasificador de sentimientos en comentarios turísticos. In *XXXIX Jornadas de Automática* (pp. 644–650). https://doi.org/10.17979/spudc.9788497497565.0644

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and

applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Melville, P., Gryc, W., & Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (Vol. 35, pp. 1275–1284). New York, NY, USA: ACM. https://doi.org/10.1145/1557019.1557156

Mercado, V., Villagra, A., & Errecalde, M. (2020). Political Alignment Identification : a Study with Documents of Argentinian Journalists. *Journal of Computer Science & Technology*, *20*(1), 43–52. https://doi.org/https://doi.org/10.24215/16666038.20.e05

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems*, *2*. Retrieved from http://arxiv.org/abs/1310.4546

Miner, G., Elder, J., Hill, T., & Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. *Academic Press, Waltham, MA*. Elsevier. https://doi.org/10.1016/C2010-0-66188-8

Mizgajski, J., & Morzy, M. (2019). Affective recommender systems in online news industry: how emotions influence reading choices. *User Modeling and User-Adapted Interaction*, *29*(2), 345–379. https://doi.org/10.1007/s11257-018-9213-x

Moctezuma, D., Graff, M., Miranda-Jiménez, S., Tellez, E. S., Coronado, A., Sánchez, C. N., & Ortiz-Bejar, J. (2017). A Genetic Programming Approach to Sentiment Analysis for Twitter: TASS'17. *CEUR Workshop Proceedings*, *1896*, 23–28. Retrieved from http://ceur-ws.org/Vol-1896/p1_ingeotec_tass2017.pdf

Mohammad, S. M. (2012). Emotional tweets. **SEM 2012 - 1st Joint Conference on Lexical and Computational Semantics*, *1*, 246–255.

Mohammad, S. M., & Bravo-Marquez, F. (2017a). Emotion intensities in tweets. **SEM 2017*

- *6th Joint Conference on Lexical and Computational Semantics, Proceedings*, 65–77. https://doi.org/10.18653/v1/s17-1007

Mohammad, S. M., & Bravo-Marquez, F. (2017b). WASSA-2017 shared task on emotion intensity. *EMNLP 2017 - 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA 2017 - Proceedings of the Workshop*, 34–49. https://doi.org/10.18653/v1/w17-5205

Mosquera, A., Yoan, G., & Moreda, P. (2017). On Evaluating the Contribution of Text Normalisation Techniques to Sentiment Analysis on. *Procesamiento Del Lenguaje Natural*, (58), 29–36. Retrieved from http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5409/3173

Muhammad, A., Wiratunga, N., & Lothian, R. (2016). Contextual sentiment analysis for social media genres. *Knowledge-Based Systems*, *108*, 92–101. https://doi.org/10.1016/j.knosys.2016.05.032

Mukherjee, I., Sahana, S., & K. Mahanti, P. (2017). An Improved Information Retrieval Approach to Short Text Classification. *International Journal of Information Engineering and Electronic Business*, *9*(4), 31–37. https://doi.org/10.5815/ijieeb.2017.04.05

Munezero, M., Montero, C. S., Sutinen, E., & Pajunen, J. (2014). Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text. *IEEE Transactions on Affective Computing*, *5*(2), 101–111. https://doi.org/10.1109/TAFFC.2014.2317187

Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., & Wilson, T. (2013). SemEval-2013 task 2: Sentiment analysis in Twitter. In *SEMEVAL 2013 - 2nd Joint Conference on Lexical and Computational Semantics* (Vol. 2, pp. 312–320).

Nasukawa, T., & Yi, J. (2003). Sentiment analysis. In *Proceedings of the 2nd international conference on Knowledge capture* (Vol. 31, pp. 70–77). New York, NY, USA: ACM. https://doi.org/10.1145/945645.945658

Newell, A., Potharaju, R., Xiang, L., & Nita-Rotaru, C. (2014). On the Practicality of Integrity

Attacks on Document-Level Sentiment Analysis. In *Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop* (Vol. 2014-Novem, pp. 83–93). New York, NY, USA: ACM. https://doi.org/10.1145/2666652.2666661

Nielek, R., Ciastek, M., & Kopeć, W. (2017). Emotions make cities live. In *Proceedings of the International Conference on Web Intelligence* (pp. 1076–1079). New York, NY, USA: ACM. https://doi.org/10.1145/3106426.3109041

Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*, *31*(1), 527–541. https://doi.org/10.1016/j.chb.2013.05.024

Ortony, A. (2019). Some "basic" emotions are not emotions at all: Surprise! In *Language and Emotion: An International Handbook*.

Osorio Angel, S., Peña Pérez Negrón, A., & Espinoza-Valdez, A. (2021). Systematic literature review of sentiment analysis in the Spanish language. *Data Technologies and Applications*, *55*(4), 461–479. https://doi.org/10.1108/DTA-09-2020-0200

Paice, C. D. (1990). Another stemmer. *ACM SIGIR Forum*, *24*(3), 56–61. https://doi.org/10.1145/101306.101310

Pang, B., & Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Retrieved from http://arxiv.org/abs/cs/0409058

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, *3*(1), 115–124.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Vol. 31, pp. 1532–1543). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Picard, R. (1997). *Affective Computing*. MIT Press.

Plaza-Del-Arco, F. M., Jiménez-Zafra, S. M., Montejo-Ráez, A., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021. *Procesamiento de Lenguaje Natural*, *67*, 155–161. https://doi.org/10.26342/2021-67-13

Plaza-del-arco, F. M., Molina-González, M. D., Jiménez-Zafra, S. M., & Martín-Valdivia, M. T. (2018). Lexicon Adaptation for Spanish Emotion Mining. *Procesamiento Del Lenguaje Natural*, *61*, 117–124. https://doi.org/10.26342/2018-61-13

Plaza-Del-Arco, F. M., Strapparava, C., Alfonso Ureña-López, L., & Teresa Martín-Valdivia, M. (2020). EmoEvent: A multilingual emotion corpus based on different events. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, (May), 1492–1498.

Plutchik, R. (1980). A General Psychoevolutionary Theory of Emotion. In *Theories of Emotion* (Vol. 1, pp. 3–33). ACADEMIC PRESS, INC. https://doi.org/10.1016/b978-0-12-558701-3.50007-7

Plutchik, R. (2001). Nature of emotions.

Pool, C., & Nissim, M. (2016). Distant supervision for emotion detection using Facebook reactions. Retrieved from http://arxiv.org/abs/1611.02988

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

Poria, S., Cambria, E., Hazarika, D., Mazumder, N., Zadeh, A., & Morency, L. P. (2017). Multi-level multiple attentions for contextual multimodal sentiment analysis. *Proceedings - IEEE International Conference on Data Mining, ICDM*, *2017-Novem*, 1033–1038. https://doi.org/10.1109/ICDM.2017.134

Poria, S., Cambria, E., Winterstein, G., & Huang, G. Bin. (2014). Sentic patterns: Dependency-

based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, *69*(1), 45–63. https://doi.org/10.1016/j.knosys.2014.05.005

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2020). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* (pp. 527–536). https://doi.org/10.18653/v1/p19-1050

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Porter, M. F. (2001). Snowball: A language for stemming algorithms, 1–13. Retrieved from http://snowball.tartarus.org/texts/introduction.html

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J. C., Kern, M., Ungar, L., & Shulman, E. P. (2016). Modelling valence and arousal in facebook posts. *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA 2016 at the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 9–15. https://doi.org/10.18653/v1/w16-0404

Ravi, K., & Ravi, V. (2015). *A survey on opinion mining and sentiment analysis: Tasks, approaches and applications*. *Knowledge-Based Systems* (Vol. 89). https://doi.org/10.1016/j.knosys.2015.06.015

Rosá, A., & Chiruzzo, L. (2021). Emotion classification in Spanish: Exploring the hard classes. *Information (Switzerland)*, *12*(11). https://doi.org/10.3390/info12110438

Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 502–518). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S17-2088

Roth, B., Barth, T., Wiegand, M., & Klakow, D. (2013). A survey of noise reduction methods

for distant supervision. In *AKBC 2013 - Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, Co-located with CIKM 2013* (pp. 73–77). San Francisco, California: Association for Computing Machinery. https://doi.org/10.1145/2509558.2509571

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Saif, H., Fernandez, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014* (pp. 810–817).

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management*, *52*(1), 5–19. https://doi.org/10.1016/j.ipm.2015.01.005

Sailunaz, K., Dhaliwal, M., Rokne, J., & Alhajj, R. (2018). Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, *8*(1). https://doi.org/10.1007/s13278-018-0505-2

Sandoval-Almazan, R., & Valle-Cruz, D. (2018). Facebook impact and sentiment analysis on political campaigns. In *Proceedings of the 19th Annual International Conference on Digital Government Research Governance in the Data Age - dgo '18* (pp. 1–7). New York, New York, USA: ACM Press. https://doi.org/10.1145/3209281.3209328

Scandar, M. G. (2019). *Emoción y Comprensión Lectora : Relación entre Niveles de Valencia , Activación y Dominancia y la Comprensión de Textos Expositivos y Argumentativos*. https://doi.org/10.13140/RG.2.2.34321.38246

Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, *66*(2), 310–328. https://doi.org/10.1037/0022-3514.66.2.310

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment

analysis: A review and comparative analysis of web services. *Information Sciences*, *311*, 18–38. https://doi.org/10.1016/j.ins.2015.03.040

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, *52*(6), 1061–1086. https://doi.org/10.1037/0022-3514.52.6.1061

Sintaha, M., Bin Satter, S., Zawad, N., Swarnaker, C., & Hassan, A. (2016). *Cyberbullying Detection Using Sentiment Analysis in Social*.

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. *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1631–1642.

Stone, P. J., & Hunt, E. B. (1963). A computer approach to content analysis: Studies using the general inquirer system. In *AFIPS Conference Proceedings - 1963 Spring Joint Computer Conference, AFIPS 1963* (pp. 241–256). https://doi.org/10.1145/1461551.1461583

Strapparava, C., & Mihalcea, R. (2007). SemEval-2007 task 14: Affective text. *ACL 2007 - SemEval 2007 - Proceedings of the 4th International Workshop on Semantic Evaluations*, (June), 70–74.

Susanto, Y., Livingstone, A. G., Ng, B. C., & Cambria, E. (2020). The Hourglass Model Revisited. *IEEE Intelligent Systems*, *35*(5), 96–102. https://doi.org/10.1109/MIS.2020.2992799

Suttles, J., & Ide, N. (2013). Distant Supervision for Emotion Classification with Discrete Binary Values. In *International Conference on Intelligent Text Processing and Computational Linguistics* (Vol. 2, pp. 121–136). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-37256-8

Taller de Análisis de sentimientos en Español (TASS). (n.d.). Retrieved from http://tass.sepln.org

Tellez, E. S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Siordia, O. S., & Villaseñor, E. A. (2017). A case study of Spanish text transformations for twitter sentiment analysis. *Expert Systems with Applications*, *81*, 457–471. https://doi.org/10.1016/j.eswa.2017.03.071

Tessore, J. P., Esnaola, L. M., Lanzarini, L., & Baldassarri, S. (2021). Distant Supervised Construction and Evaluation of a Novel Dataset of Emotion-Tagged Social Media Comments in Spanish. *Cognitive Computation*. https://doi.org/10.1007/s12559-020-09800-x

Tessore, J. P., Esnaola, L. M., Ramón, H. D., Lanzarini, L., & Baldassarri, S. (2022). Contextual information usage for the enhancement of basic emotion classification in a weakly labelled social network dataset in Spanish. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-022-13750-x

Tessore, J. P., Esnaola, L. M., Russo, C. C., & Baldassarri, S. (2019). Comparative analysis of preprocessing tasks over social media texts in Spanish. In *Proceedings of the XX International Conference on Human Computer Interaction* (pp. 1–8). New York, NY, USA: ACM. https://doi.org/10.1145/3335595.3335632

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, *63*(1), 163–173. https://doi.org/10.1002/asi.21662

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558. https://doi.org/10.1002/asi.21416

Tian, Y., Galery, T., Dulcinati, G., Molimpakis, E., & Sun, C. (2017). Facebook sentiment: Reactions and Emojis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 11–16). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1102

Vanzo, A., Croce, D., & Basili, R. (2014). A context-based model for sentiment analysis in

twitter. *COLING 2014 - 25th International Conference on Computational Linguistics, Proceedings of COLING 2014: Technical Papers*, 2345–2354.

Vapnik, V. (2006). *Estimation of Dependences Based on Empirical Data*. *The Statistician* (Vol. 33). New York, NY: Springer New York. https://doi.org/10.1007/0-387-34239-7

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999. https://doi.org/10.1109/72.788640

Vaughan, B. (2011). *Naturalistic Emotional Speech Corpora with Large Scale Emotional Dimension Ratings.* https://doi.org/10.21427/D7GK59

Vilares, D., Peng, H., Satapathy, R., & Cambria, E. (2018). BabelSenticNet: A Commonsense Reasoning Framework for Multilingual Sentiment Analysis. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1292–1298). Bangalore, India: IEEE. https://doi.org/10.1109/SSCI.2018.8628718

Villa Monte, A. (2019). *Generación automática inteligente de resúmenes de textos con técnicas de Soft Computing*.

Voleti, V. (2017). Intuition behind LSTM.

Vosoughi, S., Zhou, H., & Roy, D. (2015). Enhanced Twitter Sentiment Classification Using Contextual Information. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 16–24). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-2904

Wang, B., Liakata, M., Zubiaga, A., Procter, R., & Jensen, E. (2016). SMILE: Twitter emotion classification using domain adaptation. *CEUR Workshop Proceedings*, *1619*(Saaip), 15–21.

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012). Harnessing twitter "big data" for automatic emotion identification. *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International*

Conference on Social Computing, SocialCom/PASSAT 2012, (June 2014), 587–592. https://doi.org/10.1109/SocialCom-PASSAT.2012.119

Wang, Z., Ho, S.-B., & Cambria, E. (2020). A review of emotion sensing: categorization models and algorithms. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-019-08328-z

We_are_social. (2018). Global Digital Report 2018: World's Internet users pass the 4 billon mark.

Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. J. (2005). Information Retrieval and Text Mining. *Text Mining*, 85–102. https://doi.org/10.1007/978-0-387-34555-0_4

Wilson, T., Wiebe, J., & Hoffmann, P. (2010). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *International Journal of Computer Applications*, *7*(5), 12–21. https://doi.org/10.5120/1160-1453

Winata, G. I., Madotto, A., Lin, Z., Shin, J., Xu, Y., Xu, P., & Fung, P. (2019). CAiRE_HKUST at SemEval-2019 Task 3: Hierarchical Attention for Dialogue Emotion Classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 142–147). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2021

Xiao, J. (2019). Figure Eight at SemEval-2019 Task 3: Ensemble of Transfer Learning Methods for Contextual Emotion Detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 220–224). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2036

Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, *53*(6), 4335–4385. https://doi.org/10.1007/s10462-019-09794-5

Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys*, *50*(2).

https://doi.org/10.1145/3057270

Yelp. (2014). Yelp Open Dataset. Retrieved from https://www.yelp.com/dataset

Yusof, N. N., Mohamed, A., & Abdul-Rahman, S. (2018). A review of contextual information for context-based approach in sentiment analysis. *International Journal of Machine Learning and Computing*, *8*(4), 399–403. https://doi.org/10.18178/ijmlc.2018.8.4.719

Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, *31*(6), 82–88. https://doi.org/10.1109/MIS.2016.94

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2014). A Survey of Affect Recognition Methods : Audio , Visual and Spontaneous Expressions, (February 2009). https://doi.org/10.1109/TPAMI.2008.52