Detección de lenguaje ofensivo en Twitter basada en expansión automática de lexicones

por

Estefania Guzmán Falcón

Tesis sometida como requerimiento parcial para obtener el grado de

Maestra en Ciencias, en el area de Ciencias Computacionales

por el

Instituto Nacional de Astrofísica, Óptica y Electrónica

Diciembre, 2018

Tonantzintla, Puebla

Nombre de los asesores:

Dr. Luis Villaseñor Pineda

Dr. Manuel Montes y Gómez

Coordinación de Ciencias Computacionales

INAOE

Dr. Antonio Rico Sulayes

Departamento de Lenguas

UDLAP

En nuestro lenguaje diario hay un grupo de palabras prohibidas, secretas, sin contenido claro, y a cuya mágica ambigüedad confiamos la expresión de las más brutales o sutiles de nuestras emociones y reacciones. Palabras malditas, que sólo pronunciamos en voz alta cuando no somos dueños de nosotros mismos.

-Octavio Paz

Índice general

Índice de figuras

Índice de tablas

Agradecimientos

Agradezco el gran apoyo, tiempo y paciencia que ofrecieron mis asesores Dr. Luis Villaseñor Pineda, Dr. Manuel Montes y Gómez y Dr. Antonio Rico Sulayes para que se llevará a cabo este trabajo de tesis.

De igual forma, agradezco a mis sinodales: Dra. María del Pilar Gómez Gil, Dra. Claudia Feregrino Uribe y Dr. Aurelio López López por brindarme su tiempo, experiencia y orientación que permitieron la mejora del trabajo realizado.

También agradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico para la realización de mis estudios de maestría y de este trabajo por medio de la Beca No. 453970.

Agradezco al proyecto no. CB-2015-01-257383 por el apoyo para realizar una estancia de un mes en la universidad de Houston y al proyecto con no. FC-2016/2410 por la extensión de beca para terminar el trabajo de tesis.

Dedicatoria

Para mi familia,

por motivativarme cada día a ser mejor.

Para mi compañero de vida Carlos,

gracias por brindarme tu apoyo incondicional.

Resumen

Actualmente las redes sociales son el medio de comunicación más utilizado, en ellas las personas pueden interactuar con usuarios de diferentes lugares, compartir aspectos de su vida y expresar su opinión en diferentes temáticas. Los usuarios pueden manifestar libremente su criterio y ver los de otros, sin embargo, al ser un medio en el que todos tienen la total libertad de expresión, existen personas que aprovechan esto para promover o ejercer conductas como la discriminación, *bullying*, racismo, clasismo, sexismo y acoso. Este tipo de comportamientos son de gran preocupación debido a que suelen trascender las redes y perjudicar por completo la vida de la víctima. Plataformas como Facebook y Twitter han realizado campañas para incentivar la denuncia de esta clase de conflictos, sin embargo no todas las personas denuncian.

Debido a que los usuarios no denuncian, se han desarrollado diversos métodos para detectar discriminación y agresión en mensajes de redes sociales. La mayoría de las soluciones suelen requerir datos etiquetados manualmente para que los métodos aprendan a identificar los mensajes ofensivos. Estos suelen tener muy buenos resultados, sin embargo, los datos son escasos debido a la dificultad en la tarea de etiquetado. Por otra parte, los métodos que no requieren datos etiquetados manualmente tienen la ventaja de no depender de una tarea de etiquetado, pero se enfrentan al lenguaje en redes sociales que es informal y está en constante cambio.

Con base en lo anterior, los enfoques ya propuestos se encuentran limitados.

Por lo tanto en el presente trabajo se propone el desarrollo de un método que utiliza un diccionario de insultos expandido para realizar un etiquetado automático y un enfoque basado en aprendizaje el cual se encarga de identificar mensajes ofensivos en función de lo aprendido con los datos etiquetados automáticamente.

El enfoque propuesto está adaptado para el idioma inglés y español de México. Debido a que no existían datos para detectar lenguaje ofensivo en México, se desarrolló un conjunto de datos, el cual proporcionó un panorama amplio de la dificultad de la tarea de etiquetado.

Capítulo 1

Introducción

El número de personas que cuenta con acceso a internet ha incrementado drásticamente en los últimos años, este aumento se ve reflejado en las estadísticas obtenidas por la Unión Internacional de Telecomunicaciones[1] (ITU, por sus siglas en inglés). La ITU reporta que el número de internautas a nivel mundial en el año 2010 era de 1,991 millones, pero en la última estadística del 2016 se reportó 3,385 millones de usuarios.

En el caso de México, la asociación de internet[2] reporta que en 2010 el número de internautas era 30.6 millones y en el año 2016 pasaron a ser 70 millones [Asociación de Internet.mx, 2017]. Otro dato interesante que se reporta son las principales actividades de los usuarios en internet. En primer lugar destaca el uso de redes sociales, seguido del correo electrónico y en tercer lugar la búsqueda de información. Por lo tanto, los internautas mexicanos prefieren invertir más tiempo en redes sociales; las redes que más utilizan son Facebook[3], Whatsapp[4], Youtube[5] y

[1] www.itu.int

[2] www.asociaciondeinternet.mx

[3] www.facebook.com

[4] www.whatsapp.com

[5] www.youtube.com

Twitter[6].

Tomando en cuenta lo anterior, ¿por qué los usuarios invierten más su tiempo en las redes sociales?. La razón es porque los usuarios se pueden comunicar e incluso conocer nuevas personas, pero también estos medios suele emplearse como forma de propaganda para campañas politicas, manifestaciones sociales y publicidad en general [INMUJERES, 2016]. Los usuarios pueden expresar libremente su opinión respecto diferentes temáticas y ver lo que otros opinan, pero desafortunadamente antivalores sociales como la discriminación, racismo, clasismo y sexismo trascienden a estos medios. Quienes promueven estos antivalores en las redes sociales tienden a ser hostiles con aquellos que no compartan su pensamiento o simplemente disfrutan generar ambiente de tensión. Esta clase de usuarios suelen usar las redes sociales para cometer, promover o agravar una agresión hacia una persona o grupo [Derechos-Digitales, 2016].

Por lo tanto, algunas redes sociales alientan a los usuarios a combatir conductas dañinas, por ejemplo Facebook permite realizar denuncias[7] de comentarios, publicaciones, fotos o videos que pueden ser ofensivos o *spam*[8]. Además tiene opciones de filtros de groserías para páginas públicas, donde el administrador puede ocultar comentarios o publicaciones que las contengan[9]. En el caso de Twitter es diferente, en lugar de que algo sea removido de la plataforma, los usuarios pueden ocupar los filtros para no ver publicaciones o *hashtags*[10], y estos pueden permanecer a la vista de quienes si desean verlos. A pesar de que estos medios ya cuenten con mecanismos para fomentar un ambiente neutral de comunicación, la mayoría de los casos de agresión no son denunciados por los usuarios e incluso son socialmente aceptados a pesar de que se esté violentando los derechos de una persona [INMUJERES, 2016].

[6] twitter.com

[7] www.facebook.com/help/181495968648557

[8] Spam: publicidad basura. [RAE, 2017]

[9] www.facebook.com/help/131671940241729

[10] help.twitter.com/es/using-twitter/advanced-twitter-mute-options

Un ejemplo real que sucedió en México en el año 2016, fue el caso de una joven que se volvió un tema polémico en las redes, debido a la filtración de un video de su despedida de soltera. Este video provocó que la joven se volviera víctima de humillación y linchamiento público, e incluso en Twitter se volvió *trendingtopic*[11] bajo los *hashtags* #LadyCarolina y #Ladycuernos [Derechos-Digitales, 2016].

Por lo tanto es de gran interés el desarrollo de un sistema automático capaz de identificar mensajes ofensivos a personas y grupos vulnerables para una detección temprana de comportamientos abusivos o agresivos que se pueden viralizar en las redes. Esto ayudaría a que las víctimas no sufran de humillación pública o algún otro daño.

A lo largo de los años se han desarrollado diversas soluciones para atacar este problema, se ha partido desde la detección de groserías [Del Bosque and Garza, 2014], también empleando técnicas tradicionales de lenguaje natural y enfoques supervisados que usan datos etiquetados manualmente [Tulkens et al., 2016] los cuales han dado buenos resultados; además se ha recurrido a técnicas no supervisadas (enfoques basados en diccionarios) [Mubarak et al., 2017] y [Gitari et al., 2015], en las que no se requiere de datos etiquetados manualmente para su funcionamiento.

Sin embargo, los enfoques basados en diccionarios se encuentran con la limitante de que el lenguaje en redes sociales está en constante transformación. Las groserías poco a poco son más comunes en la comunicación cotidiana [Samghabadi et al., 2017] y para ofender los usuarios prefieren usar neologismos [12]. Sin embargo, también las técnicas supervisadas se ven afectadas por la escasez de datos, esto se debe a la dificultad en la tarea de etiquetado. El proceso de etiquetado es una tarea subjetiva en la que influyen aspectos sociales, culturales y creencias en el criterio del etiquetador [Rico-Sulayes, 2014] [Park and Fung, 2017], dificultando aún más el proceso para definir qué es ofensivo y qué no lo es.

[11] Trendingtopic: son las palabras clave más utilizadas en un plazo de tiempo concreto en Twitter.

[12] Vocablo, acepción o giro nuevo en una lengua. [RAE, 2017]

Con base en lo anterior, la detección de textos ofensivos se ve perjudicada por la dificultad en la tarea de etiquetado y la constante evolución del lenguaje en redes sociales, lo cual limita los sistemas ya propuestos. Por lo tanto en el presente trabajo se propone un método que involucra los puntos fuertes de los enfoques basados en diccionarios (no requiere datos etiquetados) y los supervisados (tienen buen rendimiento en la tarea); también se integra un proceso de enriquecimiento de diccionario.

El uso conjunto de enfoques basados en diccionarios y supervisados ya ha sido aplicado en [Lalji and Deshmukh, 2016], [Sabariah et al., 2015], [Zhang et al., 2011] y [Tan et al., 2008] pero adaptados a la tarea de análisis de sentimientos. Cabe destacar que la principal contribución del método propuesto es que está diseñado para la tarea de detección de textos ofensivos en Twitter, además cuenta con un método que se encarga de enriquecer el vocabulario, agregando jerga ofensiva usada en Twitter. Adicionalmente, se contribuye a que el método propuesto sea aplicable para los idiomas inglés y español de México.

Debido a que no se cuenta con datos de validación para el idioma español, se desarrolló un corpus que fue etiquetado manualmente para validar nuestro modelo.

1.1. Problemática

Diversas soluciones se han propuesto para la detección del textos ofensivos, las cuales se han enfocado más en *ciberbullying* [Chatzakou et al., 2017], racismo [Tulkens et al., 2016] y discurso de odio [Badjatiya et al., 2017] [Davidson et al., 2017].

Algunos de los anteriores trabajos emplean técnicas tradicionales de Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés) y métodos supervisados. Dichos trabajos suelen tener una buena aproximación para detectar ofensas explícitas, sin embargo no han podido tratar los siguientes problemas:

1. Identificar e ignorar mensajes que contienen groserías, lenguaje no formal y que no están ofendiendo a una persona o grupo. A continuación se muestran unos ejemplos de este caso:
   - Tweet 1: "¡Mis maestros son unos investigadores bien vergas y mis compañeros no se emocionan!, Weeee, no putas mames!"
   - Tweet 2: "#CosasQueMeMolestan que no me salgan las putas derivadas parciales"
2. Detectar textos ofensivos que usan palabras comunes con intención ofensiva y no cuentan con ninguna grosería. Además suelen contener expresiones elaboradas o referencias a un contexto sociocultural [Waseem et al., 2017b], por ejemplo:
   - Tweet 3: "Nos los vamos a enchufar. Les vamos a meter el chorizo hasta el fondo"
   - Tweet 4: "Ya María, no sabes cocinar nada! mejor vete a barrer o lavar tus calzones o tampoco sabes hacer eso #MasterChefMx"

Analizando lo anterior, para poder identificar esta clase de textos se requiere de un análisis semántico de la oración y de contexto. Los enfoques supervisados

no pueden detectar este tipo de tweets debido a que están limitados al contexto del corpus con el que fueron entrenados. La mayoría de los corpus reportados en el estado del arte, cuentan con pocas instancias en la clase ofensiva por lo tanto para que estos métodos funcionen en diferentes contextos se requiere de un gran volumen de textos ofensivos. Sin embargo, la dificultad del etiquetado reside en que no hay una clara definición de lenguaje ofensivo cuando se realiza el proceso de etiquetado [Park and Fung, 2017]. A continuación se enlistan un par de ejemplos que son difíciles de etiquetar.

- Tweet 5: "Gorda me la pones de tan solo mirarte"
- Tweet 6: "Muy bueno el botox Enrique, ni una arruga. Impecable."

1.2. Motivación

El lenguaje ofensivo suele ser la forma en la que se expresan los *bullies* y agresores, con la intención de atacar o intimidar. Por lo tanto, la detección de textos ofensivos es importante para la prevención y detección de antivalores en las redes sociales. La razón de trabajar en esta problemática es porque existe un gran interés en la comunidad por entender este comportamiento y combatirlo [Waseem et al., 2017a]. Sin embargo métodos propuestos cuentan con lexicones o ejemplos ofensivos limitados. Además se enfrentan al lenguaje variable de las redes sociales, el cual cuenta con diversas formas de escritura que dificultan el procesamiento de este tipo de textos [Chatzakou et al., 2017].

Considerando estas dificultades, en este trabajo se propone un método que integra un enfoque supervisado y uno no supervisado. La parte no supervisada hace frente en adquirir automáticamente instancias ofensivas y el enfoque supervisado es entrenado con lo datos etiquetados. Además este método es aplicado en el idioma inglés y español.

1.3. Objetivos

1.3.1. Objetivo general

Diseñar e implementar un método para la detección de textos ofensivos en Twitter que combina dos técnicas de aprendizaje, una basada en diccionarios y la otra supervisada.

1.3.2. Objetivos específicos

Los objetivos específicos se listan a continuación:

- Diseñar un método de expansión de lexicón que utiliza la similitud de contextos entre palabras.
- Diseñar un algoritmo basado en diccionarios para asignar una etiqueta a instancias no etiquetadas.
- Construir un conjunto de datos proveniente de Twitter en español de México.
- Evaluar y analizar el método propuesto para abordar la detección de textos ofensivos en Twitter.

1.4. Alcance y limitaciones

En este trabajo de tesis se abarca el diseño, implementación y evaluación del método propuesto para detectar lenguaje ofensivo en textos en inglés y español de México. El conjunto de datos utilizado para el idioma inglés proviene de foros de discusión sobre noticias y contiene diferentes tipos de inglés. En el caso de los datos en español, estos son de Twitter y fueron recuperados únicamente de la región centro del país en el periodo comprendido entre los meses de agosto a noviembre del 2017.

1.5. Organización de la tesis

La tesis está conformada por los siguientes capítulos.

- Capítulo 2: Marco teórico. En esta sección se describen definiciones y conceptos que son de gran relevancia para la comprensión de la solución propuesta.
- Capítulo 3: Trabajo relacionado. Se revisa el estado del arte y referentes en la detección de lenguaje ofensivo y el enfoque propuesto. El objetivo de este capítulo es conocer las técnicas que se han empleado para abordar este tema, además de analizar su alcance y limitaciones.
- Capítulo 4: Generación de corpus en español. Se presenta la metodología aplicada en la construcción del corpus de validación para el idioma español. Se describe cada paso: la extracción, los criterios de etiquetado y análisis de casos difíciles.
- Capítulo 5: Método propuesto. Se describe paso a paso el proceso de expansión de diccionario, el algoritmo basado en diccionarios, el enfoque supervisado aplicado y la evaluación del método.
- Capítulo 6: Configuración de experimentos y resultados. En esta sección se detallan los experimentos realizados, se reportan los resultados obtenidos y se realiza la comparativa con el estado del arte.
- Capítulo 7: Análisis y discusión. En esta sección se discute el rendimiento del método y las dificultades enfrentadas.
- Capítulo 8: Conclusiones y trabajo futuro. Por último en esta sección se describe la contribución del propuesto método y trabajo futuro.

Capítulo 2

Marco teórico

En este capítulo se presentan los conceptos y definiciones esenciales para comprender el método propuesto. En la primera sección se define qué es el lenguaje ofensivo y cuáles son sus características. En la siguiente sección se define qué es el aprendizaje supervisado, no supervisado y los enfoques que caen en esas categorías. Después, se detalla qué son los *word embeddings*, los diferentes algoritmos para su edificación y la variedad de herramientas para construirlos. Posteriormente, se procede con la explicación de las técnicas de representación de documentos. Y por último en las dos secciones restantes se presentan las medidas de similitud entre textos y las métricas aplicadas para evaluar nuestro conjunto de datos y método propuesto.

2.1. Lenguaje ofensivo

Diversos trabajos han nombrado al problema como detección de: lenguaje ofensivo [Chen et al., 2012], discurso de odio [Badjatiya et al., 2017] o lenguaje abusivo [Nobata et al., 2016]. Sin embargo no existe una definición concreta de qué es el lenguaje ofensivo. Debido al desacuerdo, algunos autores prefieren atacar el problema de forma particular, proponiendo soluciones a casos de *bullying* [Samghabadi et al., 2017],

racismo [Tulkens et al., 2016] y sexismo [Lee et al., 2010].

Entre las definiciones de los anteriores antivalores se superponen ciertas características: presencia de groserías[1], vocabulario discriminatorio[2], adjetivos despectivos[3] [Waseem et al., 2017b] y el uso de menciones (nombre, etiqueta de usuario o pronombre). Además otro elemento clave es la intención u objetivo por el cual fue escrito el mensaje. La mayoría de estos mensajes tienen como propósito hacer un daño físico, psicológico, social y/o económico para el receptor; vulnerando sus derechos humanos [Derechos-Digitales, 2016].

Tomando en cuenta las características anteriores, la definición de lenguaje ofensivo que utilizamos para este trabajo es la siguiente: *"El lenguaje ofensivo se refiere a las expresiones o palabras discriminatorias, despectivas y groseras; utilizadas con el propósito de hacer algún daño a una persona o colectividad"*.

Figura 2.1: Ejemplos de mensajes ofensivos en Twitter.

[1]Descortesía, falta grande de atención y respeto. [RAE, 2017]

[2]Discriminación: Dar trato desigual a una persona o colectividad por motivos raciales, religiosos, políticos, de sexo, etc. [RAE, 2017]

[3]Despectivo: Dicho de una palabra, que manifiesta idea de menosprecio. [RAE, 2017]

Por lo tanto, un mensaje es ofensivo si el propósito para que el fue escrito es para hacer algún daño al receptor, resaltando sus características de forma negativa empleando vocabulario discriminatorio, adjetivos despectivos y/o groserías. En la figura 2.1 se ilustran ejemplos de estos mensajes.

2.2. Aprendizaje supervisado

El objetivo del aprendizaje supervisado es aprender a mapear de $\mathcal{X}$ a $\mathcal{Y}$, dado un conjunto de entrenamiento hecho de pares (x_i, y_i). Donde, $y_i \in \mathcal{Y}$ son las etiquetas u objetivos de las instancias $x_i \in \mathcal{X}$. Un algoritmo supervisado es el encargado de realizar este mapeo y obtene reglas para la creación de una función de inferencia, la cual ayudará a predecir las etiquetas en el conjunto de prueba [Thomas, 2009]. Los datos de entrenamiento que se requieren son etiquetados manualmente por uno o varios humanos que se encargan de asignar una etiqueta a cada elemento [Kalita, 2015]. En la figura 2.2 se muestra el proceso y componentes del aprendizaje supervisado.

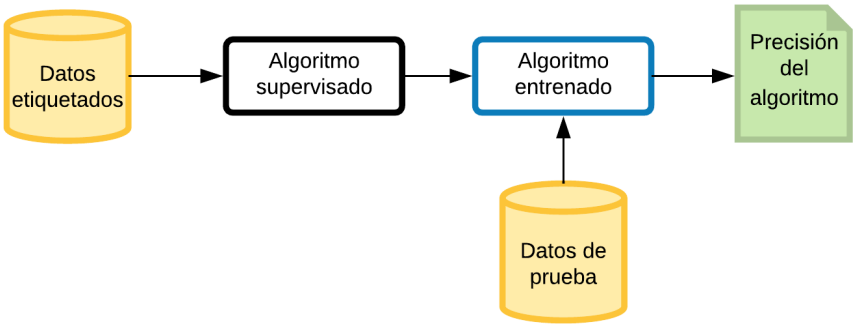

Figura 2.2: Pasos básicos del aprendizaje supervisada.

En la literatura se encuentran múltiples algoritmos de aprendizaje supervisado, como se comentó en el capítulo 1, estos han obtenido buenos resultados en la detección de textos ofensivos. A continuación, se introducen los más utilizados:

2.2.1. Máquinas de Soporte Vectorial (SVM)

Este modelo cuenta con algoritmos de aprendizaje asociados que analizan datos y reconoce patrones. Fue introducido por [Cortes and Vapnik, 1995], basándose en el principio de minimización de riesgos estructurales. Su diseño tiene como función resolver problemas de reconocimiento de patrones en dos clases, para hallar la superficie de decisión que separa los ejemplos de entrenamiento positivos y negativos de una categoría con margen máximo. Las máquinas de soporte vectorial encuentran el hiperplano h que separa los ejemplos de entrenamiento positivo y negativo con un margen máximo [Kalita, 2015], figura 2.3.

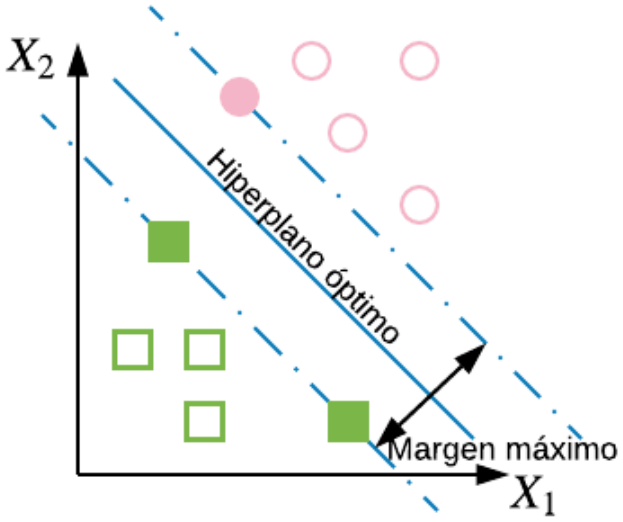

Figura 2.3: Hiperplano que separa las clases con margen máximo [Cortes and Vapnik, 1995].

2.2.2. Clasificador Naïve Bayes

Naïve Bayes es un clasificador probabilístico lineal que usa el teorema de Bayes y tiene como suposición principal que todos los atributos son independientes dado el valor de la variable de clase [Friedman et al., 1997]. El clasificador considera que algunos datos son inútiles (es decir, no afectan el resultado de la clasificación incluso eliminándolos) y otros tienen significados similares, por lo tanto se eliminan. De esta manera, el conjunto de datos puede ser más preciso. El clasificador Naïve Bayes es construido usando el conjunto de entrenamiento para estimar la probabilidad de cada clase dadas las características de una nueva instancia[Kalita, 2015]

El clasificador Naiïve Bayes es aplicado para aprender tareas donde cada instancia x es descrita por un conjunto de valores de atributo y donde la función objetivo $f(x)$ puede tomar cualquier valor de algún conjunto finito V. Un conjunto de ejemplos de entrenamiento de la función objetivo es proporcionado y una nueva instancia es presentada, descrita por la tupla de valores de atributo $(a_1, a_2, ..., a_n)$. El algoritmo entrenado es cuestionado para predecir el valor objetivo o la clasificación, para esta nueva instancia. El enfoque Bayesiano para clasificar una nueva instancia es asignar el valor objetivo más probable v_{MAP}, dado los valores de atributo $(a_1, ..., a_n)$ que describen a la instancia [Mitchell, 1997].

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid a_1, a_2, ..., a_n) \tag{2.1}$$

Se puede usar el teorema de Bayes para reescribir la expresión como:

$$v_{MAP} = \arg\max_{v_j \in V} \frac{P(a_1, a_2, ..., a_n \mid v_j) P(v_j)}{P(a_1, a_2, ..., a_n)} \tag{2.2}$$

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2, ..., a_n \mid v_j) P(v_j) \tag{2.3}$$

El clasificador Naiïve Bayes se basa en la suposición de que los valores de atributo son condicionalmente independientes dado el valor objetivo. Por lo tanto, la suposición es que dado el valor objetivo de la instancia, la probabilidad de observar la conjunción $a_1, a_2, .., a_n$ es solo el producto de las probabilidades para los atributos individuales: $P(a_1, a_2, .., a_n \mid v_j) = \prod_i P(a_i \mid v_j)$. Sustituyendo esto en la ecuación 2.3, se tiene el enfoque utilizado por el clasificador Naiïve Bayes en 2.4, donde v_{NB} denota la salida del valor objetivo por el clasificador.

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j) \tag{2.4}$$

2.3. Aprendizaje no supervisado

Los algoritmos de aprendizaje no supervisado, no requieren de datos etiquetados manualmente para su funcionamiento. Para asignarle una clase a cada elemento, se necesita de una serie de criterios que ayuden a la exploración de datos y descubrir patrones o estructuras que son de interés. De forma simple un algoritmo no supervisado tiene el objetivo de clasificar elementos sin conocimiento adicional, de modo que las instancias dentro de una clase sean más similares que las de otra clase [Kalita, 2015].

Diversos algoritmos de aprendizaje no supervisado se encuentran en la literatura, como vecinos más cercanos, enfoques basados en lexicón, etc. A continuación sólo se describe el enfoque aplicado en el presente trabajo.

2.3.1. Enfoque basado en lexicón

Un enfoque basado en lexicón no requiere datos de entrenamiento y su funcionamiento depende del diccionario aplicado [Sabariah et al., 2015]. El método realiza

la clasificación de un texto con base en la coincidencia de términos con los del diccionario y se calcula una puntuación para asignarle una clase, comúnmente este método clasifica la información en dos clases: Positiva o Negativa [Tan et al., 2008]. El método regularmente es aplicado para la tarea de detección de sentimientos, partiendo de un diccionario de opinión, que contiene palabras positivas y negativas. La clasificación se efectúa obteniendo la polaridad del texto por medio del conteo de palabras positivas y negativas [Tan et al., 2008]. En la figura 2.4 se ilustra cómo este enfoque es aplicado en análisis de sentimientos.

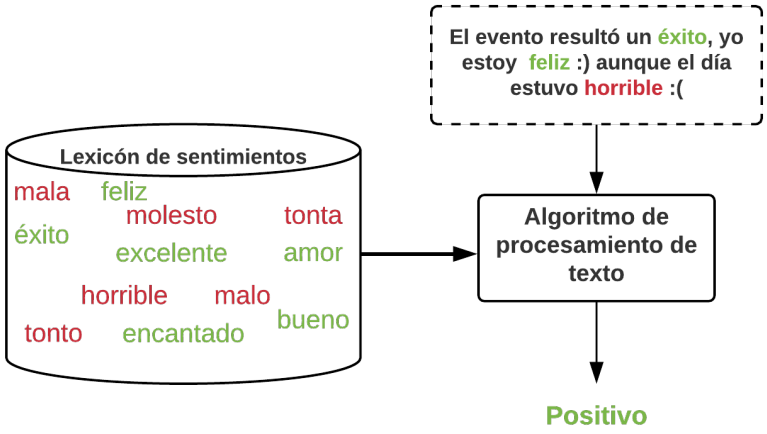

Figura 2.4: Enfoque basado en lexicón para la tarea de análisis de sentimientos.

2.4. *Word Embeddings*

Actualmente el uso de los *word embeddings* se ha vuelto popular en tareas de clasificación de textos. Pero antes de su popularidad los *word embeddings* eran conocidos como representaciones distribuidas de las palabras.

Una representación distribuida es un concepto fundamental para el conexionismo. En una red conexionista, una representación distribuida ocurre cuando algún

concepto o significado es representado por la red, pero ese significado está representado por un patrón de actividad a través de un número de unidades de proceso [Rumelhart et al., 1986a]. Por lo tanto, las representaciones distribuidas de las palabras en un espacio vectorial ayudan a los algoritmos de aprendizaje a lograr un mejor rendimiento en las tareas de procesamiento del lenguaje natural mediante la agrupación de palabras similares. Una de las primera aplicaciones de las representaciones distribuidas se remonta a 1986 por Rumelhart, Hinton y Williams [Rumelhart et al., 1986b], en el contexto del aprendizaje de representaciones distribuidas de símbolos.

Existen diversas herramientas para la generación de *word embeddings* como son *Weighted Textual Matrix Factorization (WTMF)* [Guo and Diab, 2012], GloVe [Pennington et al., 2014], Word2Vec [Mikolov et al., 2013a], etc. En el presente trabajo se utilizan *Word2Vec* y GloVe debido a los buenos resultados reportados en trabajos relacionados con detección de paráfrasis, sarcasmo [White et al., 2015] [Joshi et al., 2016] [Ghosh et al., 2015] y análisis de sentimientos.

2.4.1. *Word2Vec*

Word2Vec es una herramienta que fue desarrollada por el equipo Google Brain [Mikolov, 2014] y cuenta con dos arquitecturas para la generación de representaciones vectoriales de las palabras [Mikolov et al., 2013b]. Las arquitecturas toman como base la red neuronal[4] que planteó Bengio en 2003 [Bengio et al., 2003].

La red propuesta en [Bengio et al., 2003] es una *Feedforward* [5] *Neural Net Lan-*

[4] Las redes neuronales (artificiales) son sistemas de procesamiento de información, cuya estructura y principio de funcionamiento están inspirados en el sistema nervioso y el cerebro de animales y humanos [Kruse et al., 2016].

[5] Red neuronal que cuenta con estructura de red acíclica (no contiene bucles ni ciclos dirigidos) [Kruse et al., 2016].

guage Model (NNLM), esta red modela un vocabulario dadas N palabras y se produce una distribución de probabilidad sobre todas las palabras. Los elementos que componen a esta red son las capas de entrada, proyección, ocultas y salida, ver figura 2.5.

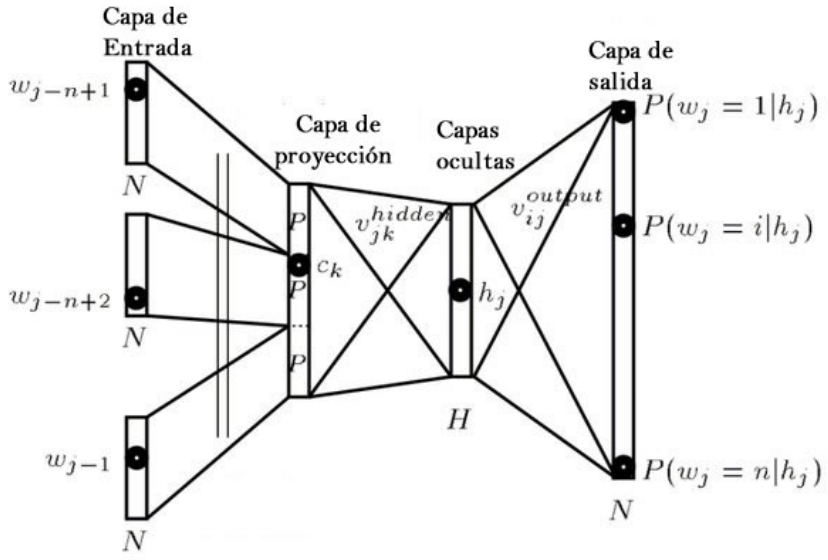

Figura 2.5: Representación de una NNLM. [Mikolov et al., 2010]

En la capa de entrada, las palabras N se codifican utilizando una codificación de 1 a V (V es el tamaño del vocabulario), luego se pasa a una capa de proyección P que tiene una dimensionalidad $N \cdot D$ (D es la dimensionalidad de los vectores de palabras), utilizando una matriz de representación compartida. Por consiguiente, cada vector 1 a V se reemplaza por un vector de palabra de la matriz compartida $N \cdot D$. Entre la capa de proyección y las capas ocultas hay una conexión densa y los resultados de las capas ocultas son utilizados por la capa de salida para calcular una distribución de probabilidad sobre todas las palabras en el vocabulario usando una función softmax[6].

[6]La función softmax transforma un vector dimensional k en otro vector dimensional k de valores reales, cada uno entre 0 y 1, sumando hasta 1[Michelucci, 2018].

El problema con la NNLM es su complejidad para el cálculo entre la capa de proyección y la capa oculta, porque los valores en la capa de proyección son densos. Otro elemento que afecta en la complejidad es la capa oculta, porque esta es usada para calcular la probabilidad de la distribución sobre todas las palabras en el vocabulario, resultando una capa de salida con dimensión del vocabulario [Mikolov et al., 2013b].

Tomando en cuenta los anterior, en [Mikolov et al., 2013a] se desarrollan dos modelos de redes neuronales poco profundas: *Continuous Bag-of-Words Model (CBOW)* y *Continuous Skip-gram Model*.

Modelo de Bolsa de palabras continua (CBOW)

Esta arquitectura es similar al modelo propuesto en [Bengio et al., 2003] pero se elimina la capa oculta no lineal y se comparte la capa de proyección para todas las palabras. Esta arquitectura fue llamada bolsa de palabras, ya que el orden de las palabras no influye en la proyección. La arquitectura CBOW predice la palabra actual basada en el contexto, ver figura 2.6.

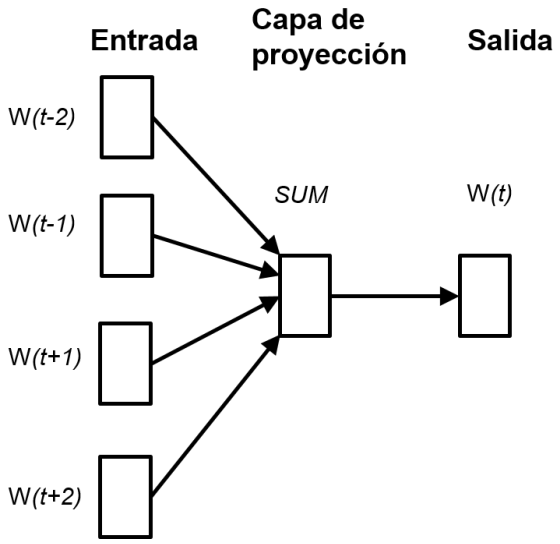

Figura 2.6: Arquitectura del CBOW.

Modelo *Continuous Skip-gram*

La arquitectura *Skip-gram* es similar a CBOW, la diferencia radica en su entrenamiento que se centra en encontrar representaciones de palabras que sean útiles para predecir las palabras circundantes en una oración o un documento. Dada una secuencia de palabras de entrenamiento $w_1, w_2, w_3, .., w_T$, el objetivo del modelo *Skip-gram* es maximizar la probabilidad promedio de *log*.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j} \mid w_t) \tag{2.5}$$

En 2.5, c es el tamaño del contexto de entrenamiento (que puede ser una función de la palabra central w_t). Una mayor c da como resultado más ejemplos de entrenamiento, por lo tanto, puede llevar a una mayor precisión, pero se ve afectado el tiempo de entrenamiento [Mikolov et al., 2013b]. En la figura 2.7 se muestra la arquitectura de *Skip-gram*.

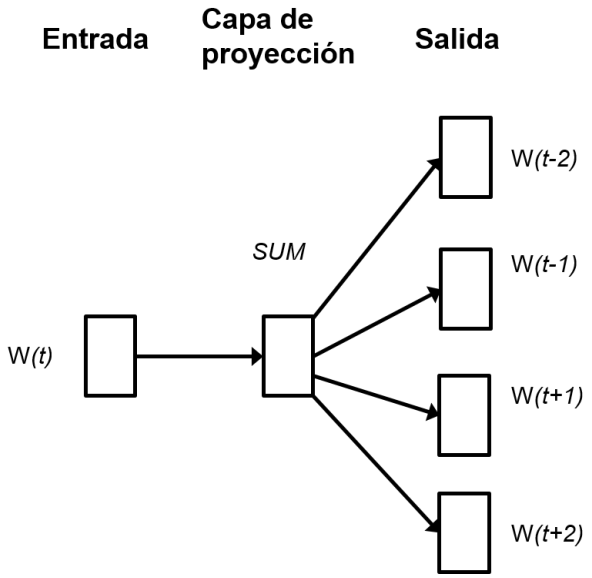

Figura 2.7: Arquitectura *Continuous Skip-gram.*

2.4.2. GloVe: Vectores Globales para la representación de las palabras

El modelo GloVe fue introducido por [Pennington et al., 2014] y está compuesto por dos enfoques: un método de factorización matricial y el otro basado en ventanas poco profundas. La idea principal de GloVe es aprender representaciones de las palabras con el objetivo de predecir palabras en un contexto local [Jurdzinski et al., 2017].

Antes de que GloVe aprenda vectores, primero se necesita crear una matriz de co-ocurrencia X, la cual contiene las co-ocurrencias de palabras en un conjunto de datos y se elige el tamaño de la ventana de contexto ws. Por lo tanto, una palabra j aparece en el contexto de una palabra i, si la distancia de j está dentro de la ventana ws de i. $X \in \mathbb{N}^{V \times V}$ (donde V es el tamaño del vocabulario en el conjunto de datos) es una matriz de co-ocurrencia palabra-palabra, lo que significa que X_{ij} es el número de apariciones de la palabra j en el contexto de la palabra i. Por lo tanto, j es una palabra de contexto. Los vectores de palabras se aprenden según la matriz X.

2.5. Representación de documentos

Uno de los pasos previos a la clasificación de documentos es la transformación de los textos a una representación, que ayuda al algoritmo a realizar la tarea de clasificado. Las representaciones aplicadas en el presente trabajo se explican a continuación.

2.5.1. Bolsa de Palabras (*BoW*)

Una de las técnicas más utilizadas para representar texto es la Bolsa de Palabras (*BoW: Bag of Words*). En esta técnica un texto es representado por un vector de características, cada palabra representa una rasgo. El espacio de atributos está conformado por todo el vocabulario de la colección de documentos, sin embargo, el espacio de rasgos no está ordenado por aparición debido a que la técnica ignora dicho orden.

Los vectores pueden tener pesado binario o de frecuencia. En el caso del pesado binario cada característica es marcada con 1 si al menos aparece una vez en el texto y 0 si no aparece. Para el pesado de frecuencia si el rasgo aparece n veces en el texto, se coloca n y en caso contrario se marca con 0. En la figura 2.8 se muestra BoW con pesado binario.

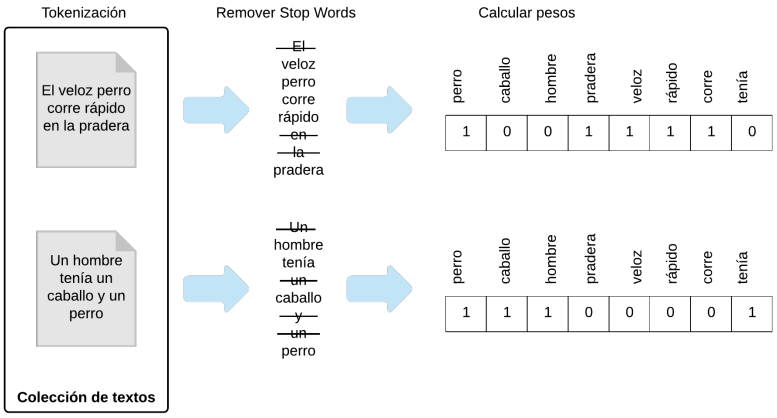

Figura 2.8: Proceso de representación de textos usando BOW [Grčar, 2012].

2.6. *N*-gramas

Los n-gramas son secuencias de elementos que aparecen en un conjunto de datos. Estos elementos pueden ser palabras, caracteres, fonemas, etc. La convención común es que n corresponde al número de elementos en una secuencia. Los n-gramas por lo regular son empleados para representar textos, considerando un número n de elementos y se obtene su probabilidad de ocurrencia en cada texto.

En la figura 2.9 se muestra un ejemplo de representación de textos utilizando secuencias de unigramas, bigramas y trigramas de palabras.

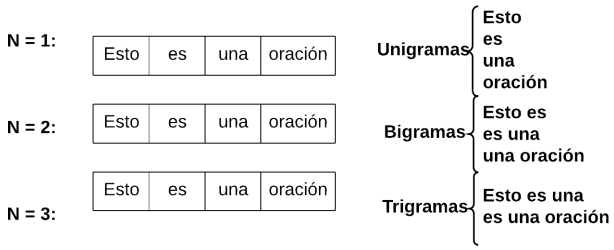

Figura 2.9: Proceso de representación de textos usando n-gramas de palabras.

2.7. Medida de similitud para textos

Para conocer la similitud que existe entre dos documentos, estos tienen que ser transformados a una representación vectorial y se mide el grado de similitud entre los dos documentos como la correlación entre los vectores. En la literatura existen diversas métricas para obtener la similitud y distancia entre documentos, pero a continuación solo se describe la métrica que fue aplicada en este trabajo.

2.7.1. Similitud coseno

Cuando los documentos se representan como vectores de términos, la similitud de dos documentos corresponde a la correlación entre los vectores. Esto se cuantifica como el coseno del ángulo entre vectores, es decir, la similitud del coseno. La figura 2.10 muestra el ángulo entre dos documentos en un espacio bidimensional.

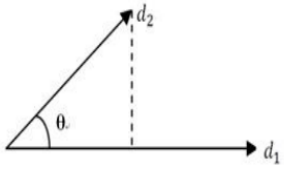

Figura 2.10: Ángulo entre dos documentos.

Dado dos documentos representados como $\vec{t_a}$ y $\vec{t_b}$, su similitud coseno es:

$$SIM_C(\vec{t_a}, \vec{t_b}) = \frac{\vec{t_a} \cdot \vec{t_b}}{|\vec{t_a}| \times |\vec{t_b}|} \tag{2.6}$$

Donde $\vec{t_a}$ y $\vec{t_b}$ son vectores m-dimensionales sobre el conjunto de términos $T = \{t_1, \ldots, t_m\}$. Cada dimensión representa un término con su peso en el documento, el cual no es negativo y está delimitado entre [0,1] [Huang, 2008].

2.8. Métricas de evaluación

En la tarea de etiquetado del conjunto de datos, se cuenta con diferentes etiquetadores que fungen como jueces para asignar una etiqueta a cada texto, dadas ciertas características. Al finalizar la tarea se requiere medir el acuerdo que tienen los jueces. Otro elemento que también debe ser medido es la exactitud del método propuesto después de realizar la clasificación. Por lo tanto, se requiere conocer cuales son las métricas que deben ser aplicadas al conjunto de datos y el método propuesto. A continuación se describen estas métricas.

2.8.1. Coeficiente Kappa

El kappa de Cohen [Cohen, 1968], simbolizado por la letra griega minúscula, κ es una medida estadística útil para pruebas de confiabilidad entre evaluadores [McHugh, 2012]. Los valores están entre el rango de -1 a +1, donde 0 representa la cantidad de acuerdo que se puede esperar de una posibilidad aleatoria, y 1 representa un acuerdo perfecto entre los evaluadores. En [Cohen, 1968] se sugirió que el resultado Kappa se interprete de la siguiente manera: los valores ≤ 0 indican que no hay acuerdo y 0.01 – 0.20 como ninguno a leve, 0.21–0.40 como justo, 0.41 – 0.60 como moderado, 0.61–0.80 como sustancial y 0.81 – 1.00 como un acuerdo casi perfecto. El cálculo de kappa de Cohen se puede realizar de acuerdo con la siguiente fórmula:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{2.7}$$

En 2.7 $Pr(a)$ representa el acuerdo observado real y $Pr(e)$ un acuerdo de azar. Se debe tomar en cuenta que el tamaño de la muestra consiste en el número de observaciones realizadas n, a través de las cuales se comparan los evaluadores. Cohen midió específicamente dos evaluadores en sus papeles, considerando dos clases a evaluar. El $Pr(e)$ se obtiene a través de la siguiente fórmula:

$$Pr(e) = \frac{(\frac{cm^1 \times rm^1}{n})(\frac{cm^2 \times rm^2}{n})}{n} \tag{2.8}$$

En 2.8, cm^1 representa la columna 1 marginal, cm^2 representa la columna 2 marginal, rm^1 representa la fila 1 marginal y rm^2 representa la fila 2 marginal. En la tabla 2.1 se muestra un ejemplo de como obtener los valores marginales entre 2 evaluadores y 2 clases.

		Evaluador 1		Filas	
		Clase 1	Clase 2	Marginales	
Evaluador 2	Clase 1	147	3	150	rm^1
	Clase 2	10	62	72	rm^2
Columnas marginales		157	65	222	n
		cm^1	cm^2		

Tabla 2.1: Datos de ejemplo para el cálculo kappa [McHugh, 2012].

2.8.2. Matriz de confusión

Una matriz de confusión de tamaño n x n, contiene información respecto a las clasificaciones reales y predichas realizadas por un sistema [Kohavi and Provost, 1998]. El número de clases se encuentra representado por n. En la tabla 2.2 se muestra la matriz de confusión para n=2, cada elemento de la matriz tiene el siguiente significado:

- a. Es el número de predicciones negativas correctamente clasificadas.
- b. Es el número de predicciones positivas incorrectamente clasificadas.
- c. Es el número de predicciones negativas incorrectamente clasificadas.
- d. Es el número de predicciones positivas correctamente clasificadas.

	Predicción negativos	Predicción positivos
Negativos actuales	a	b
Positivos actuales	c	d

Tabla 2.2: Matriz de confusión para el problema de clasificación de dos clases.

2.8.3. Medidas estándar

Dada la descripción de los elementos que componen a la matriz de confusión se procede a la definición de las siguientes medidas:

Exactitud

El valor se calcula como el número de elementos clasificados correctamente, sobre el número total de elementos clasificados. La fórmula para obtener la exactitud se observa en la expresión 2.9.

$$Exactitud = \frac{a+d}{a+b+c+d} \tag{2.9}$$

Precisión

El valor se calcula como el número de elementos positivos clasificados correctamente, entre el número total de elementos positivos clasificados. La fórmula para obtener la precisión se observa en la expresión 2.10.

$$Precision = \frac{d}{b+d} \tag{2.10}$$

Recuerdo

El valor se calcula como el número de elementos positivos clasificados correctamente, entre el número total de elementos positivos reales. La fórmula para obtener el recuerdo se observa en la expresión 2.11.

$$Recuerdo = \frac{d}{c+d} \tag{2.11}$$

Valor f

El valor f es una medida que combina la precisión y el recuerdo. Estas medidas ayudan a entender el desempeño del clasificador cuando abundan más elementos de una clase que en otra. La fórmula para obtener el valor f se observa en la expresión 2.12.

$$valor f = \frac{2 * Precision * Recuerdo}{Precision + Recuerdo} \tag{2.12}$$

Resumen del capítulo

En este capítulo se llevó a cabo un repaso de los conceptos y definiciones esenciales para la comprensión del método propuesto. La importancia de que el lector conociera estas definiciones ayuda a que se tenga una mejor comprensión de: la arquitectura y las métricas de evaluación del método que se describen en las secciones 5 y 6. Además, esta información facilita el entendimiento de los trabajos relacionados que se describen en el siguiente capítulo.

Capítulo 3

Trabajo relacionado

En este capítulo se expone el trabajo relacionado al enfoque propuesto. En la sección 3.1 se presentan las diferentes soluciones reportadas en la detección de lenguaje ofensivo, posteriormente en la sección 3.2 se muestran los enfoques que combinan técnicas basadas en el aprendizaje y en léxico, además una subsección en la que se describen enfoques basados en lexicón. Por último en la sección 3.3 se describe al ganador de la tarea de detección del lenguaje agresivo en IberEval 2018, donde el corpus del presente trabajo sirvió como base de evaluación.

3.1. Detección de lenguaje ofensivo

La mayoría de los trabajos en detección de lenguaje ofensivo son aquellos que están más relacionados con ciberacoso, racismo y sexismo. A continuación se describen estos trabajos y las dificultades a las que se enfrentaron.

En [Del Bosque and Garza, 2014], una escala fue diseñada para medir el nivel de agresividad en los tweets, utilizando como base algunas palabras relacionadas con el acoso escolar. En esta investigación se usaron diferentes metodologías para crear

esta escala, que va del 0 al 10 (donde diez es lo más agresivo). Se utilizaron enfoques basados en la frecuencia de insultos, en lexicones, supervisados, difusos, estadísticos y herramientas de análisis de sentimientos. Los autores observaron que la tarea de detección de textos agresivos está relacionada con el estudio de las emociones, debido a que la agresión es un sentimiento negativo. Sin embargo, el uso exclusivo del análisis de sentimientos para esta tarea no es suficiente a causa de que existen expresiones lingüísticas más elaboradas que no expresan alguna emoción. Los resultados que obtuvieron demostraron que la regresión lineal[1] resultó ser el mejor método para calificar los documentos y que el uso excesivo de lenguaje agresivo también parece ser una característica clave para la tarea.

En el caso de [Nand et al., 2016], la tarea es detectar si un mensaje contiene ciberacoso o no. Para realizar la tarea, los autores iniciaron con la creación de un conjunto de datos de validación que obtuvieron de Twitter. En la extracción de este conjunto utilizaron una serie de palabras frecuentes en los mensajes de intimidación: nerd, gay, perdedor, monstruo, emo, ballena, cerdo, gordo, aspirante, poser, puta, debe, morir, dormir, atrapado, chupar, puta, vivir, miedo, lucha, coño, coño, matar, polla, perra.

Los autores remarcan que algunas de las palabras anteriores son apodos o verbos que deben estar presentes con otras palabras para que se les considere insultos. Tres jueces participaron en el proceso de etiquetado y recibieron una serie de criterios para llevar a cabo la tarea. Uno de los requisitos era identificar si los tweets contienen insultos porque los agresores los usan para maldecir o humillar a sus víctimas. Para llevar a cabo la tarea de identificación usaron dos métodos; el primero es procesar los tweets en LIWC[2] y extraer sus características para aplicarlos a los clasificadores. El

[1]Una técnica clásica orientada a las estadísticas: también aprende un modelo que predice resultados basados en entradas con múltiples características; El modelo es una función lineal (de ahí el nombre) que mejor se ajusta a los datos.

[2]Consulta lingüística y recuento de palabras, LIWC calcula en cada texto el uso de diferentes categorías de palabras(emociones, estilos de pensamiento, preocupaciones sociales y partes del habla).

segundo fue usar las palabras como características. Como resultado, obtuvieron los siguientes valores f: 94.7 cuando usaron LIWC y cuando no lo usaban 84.7. Esto los lleva a concluir que los mensajes de acoso cibernético pueden no contener lenguaje vulgar y, en lugar de estos, se usan sinónimos o apodos.

Por otro lado, en [Mehdad and Tetreault, 2016] los autores intentan identificar si un comentario contiene lenguaje abusivo, analizando si un insulto fue escrito de forma diferente para evadir las listas negras. Por lo tanto, hicieron una primera aproximación con un análisis léxico y morfológico utilizando *word embeddings* y técnicas de parafraseo. Una de las dificultades que enfrentaron fue que las groserías estaban escritas con números y caracteres especiales. Además analizaron los comentarios a nivel de palabras y caracteres para identificar rasgos en la forma de escribir. Al realizar los experimentos, descubrieron que el análisis a nivel de caracteres arroja mejores resultados al igual que los *word embeddings*. Por lo tanto, concluyeron que si se llegara a realizar la fusión entre estas formas se podría obtener mejores resultados. Este trabajo utilizó el mismo conjunto de datos que [Nobata et al., 2016], pero obteniendo mejores resultados con un valor f=79 comparado con el valor f=78 de [Nobata et al., 2016].

Otro tipo de análisis se realizó en [Samghabadi et al., 2017], donde desarrollan un enfoque para detectar ciberacoso. El corpus usado fue compilado de ask.fm[3] , y cada publicación es una pregunta con varias respuestas. Para la identificación de mensajes negativos, aplicaron varios tipos de características clásicas (n-grams, char-n-grams, emoticones, SentiWordNet, LIWC) y nuevas características (Embeddings, LDA), y los combinaron. Los problemas que enfrentaron cuando hicieron la detección fueron que las palabras groseras se usaron en un contexto informal y no eran necesariamente ofensivas, además de que había mensajes tan cortos que no estaba claro si la intención era insultar. Al final, concluyen que las profanidades y vulga-

[3] Ask.fm es una red social construida en un formato de preguntas y respuestas. Donde se pueden realizar preguntas de forma abierta o anónima. https://ask.fm/

ridades abundan en textos escritos por adolescentes y en consecuencia el grado de negatividad de las blasfemias varía según el contexto.

Para el caso de [Park and Fung, 2017], los autores trabajan con textos que contienen lenguaje racista y sexista. Ellos proponen un enfoque a dos pasos, el cual primero detecta el lenguaje abusivo en los textos y después los clasifica en tipos específicos. En este trabajo definen como objetivo principal comparar su enfoque en dos pasos contra uno de un solo paso y validar cual es mejor. La estructura de su enfoque incluye una red neuronal convolucional (CNN) para abordar la tarea de la detección de lenguaje abusivo, además cuentan con tres tipos de modelos de CNN que utilizan entradas a nivel de caracteres, nivel de palabra y uno híbrido que usa ambos; para la segunda fase solo aplican un clasificador binario.

Los modelos CNN tienen el objetivo de encontrar características relevantes que sean de utilidad para la segunda fase. Los componentes clave de estos modelos es el cálculo de diferentes tamaños de vectores de características. Los resultados que obtienen es un valor f = 82.7 cuando usa su enfoque híbrido en un solo paso y un valor f = 82.4 usando regresión logística en dos pasos. Como se observa tienen resultados similares, pero lo interesante es que la CNN híbrida ayuda encontrar esas características y se obtienen buenos resultados de clasificación por sí sola, sin embargo los autores no describen algún detalle de las características que encontró la CNN. Por último, concluyen que un elemento que puede complicar la detección de textos abusivos es la dificultad para definir qué es lenguaje abusivo, ya que el problema radica en que hay aspectos que son subjetivos y la carencia de contexto.

Tomando en cuenta los trabajos relacionados a la detección de lenguaje ofensivo, se pudo observar que existen dificultades con la existencia de variantes o nuevos sinónimos para las groserías y palabras que se utilizan para insultar, pero que no son groseros. Los trabajos anteriores utilizan el enfoque de datos etiquetados y obtienen resultados positivos, pero deben adaptarse a los nuevos dominios que se renuevan

continuamente en las redes sociales.

3.2. Combinación de métodos basados en lexicón y aprendizaje supervisado

Como se mencionó en el capítulo 1, enfoques que combinan métodos basados en lexicón y aprendizaje supervisado solo han sido aplicados para la tarea de análisis de sentimientos. Por lo tanto, en esta sección se expone la metodología que cada trabajo realizó para la integración de este tipo de métodos.

En [Tan et al., 2008] proponen un esquema para detección de sentimientos sin ejemplos anotados. El proceso de integración que desarrollan es el siguiente: primero utilizan el enfoque basado en lexicón para el etiquetado de una porción de ejemplos informativos. El lexicón usado para esta parte es un diccionario de opinión que está compuesto de palabras positivas y negativas. Las reglas que se aplican para el etiquetado son calculadas con base a los términos relativos positivos y negativos de cada texto con el corpus. Después realizan el entrenamiento de un clasificador centroide[4] con los datos etiquetados y por último se aplica el clasificador entrenado para evaluar su rendimiento. En el análisis del método, se reporta que el esquema propuesto tiene un mejor desempeño comparado con el enfoque basado en lexicón o con el de aprendizaje supervisado. Concluyen que su esquema puede tener aún mejoras, aunque funciona bien la fusión de estos métodos.

En [Sabariah et al., 2015] proponen un esquema para complementar la evaluación dada a un programa de televisión por medio del análisis de textos en Twitter. El proceso de integración que desarrollan es el siguiente: primero utilizan el enfoque

[4]El clasificador centroide es un algoritmo simple que asigna un elemento representativo de la clase como centroide y a su alrededor todos los elementos de la clase, cuando se va a clasificar una nueva instancia, esta debe estar cerca del centroide.

basado en lexicón para el etiquetado de textos extraídos de Twitter. El lexicón usado para esta parte es un diccionario de opinión que está compuesto de palabras positivas y negativas. Las reglas que se aplican para el etiquetado son primero asignarle una polaridad al texto, con base a la polaridad de cada palabra del texto; el paso siguiente es analizar si el tweet contiene negaciones y si hay presencia de estas la polaridad cambia a negativo. Después realizan el entrenamiento de un clasificador SVM con los tweets etiquetados y por último se aplica el clasificador entrenado para evaluar su rendimiento. En el análisis del método, se reporta que no se obtuvieron buenos resultados debido a que hay aspectos a mejorar de su metodología. Concluyen que el método basado en diccionario se debe mejorar, comentan que este enfoque requiere de un lexicón que contenga vocabulario acorde a la dificultad de la tarea.

3.2.1. Enfoques basados en lexicón

En esta sección se describirán trabajos que emplean enfoques basados en diccionarios, algunos están enfocados en la detección de lenguaje ofensivo y en otros en análisis de sentimientos.

El primer método que describiremos en esta sección es el que proponen en [Tang et al., 2014], este método consiste en construir un lexicón de sentimientos a gran escala empleando información de Twitter con un enfoque de representación de aprendizaje. Desarrollaron una arquitectura de red neuronal híbrida con una función con pérdida. La red que desarrollaron aprende de un conjunto de tweets de los cuales no están anotados, pero contienen emoticones positivos y negativos, lo cual toman como referencia de que la red cuenta con datos positivos y negativos. Además utilizaron palabras del *urban dictionary*[5] como semillas para alargar la lista pequeña y extraer datos de entrenamiento con el fin de construir un clasificador a

[5]El Urban Dictionary es un sitio web que contiene un diccionario de jerga en el idioma inglés. https://www.urbandictionary.com/

nivel frase.

Los autores identificaron que los lexicones pequeños no funcionan en Twitter debido a que los mensajes contienen lenguaje informal, argot y expresiones multi-palabra que no son cubiertas por lexicones tradicionales. Por lo tanto, los autores resaltan que es necesario un lexicón de gran escala, sin embargo, se debe analizar su precisión antes de aplicarlo a un método no supervisado

En [Ameur et al., 2017] se propone una nueva representación vectorial llamada emotional TF-IDF, la cual preserva los aspectos semánticos y sentimentales basándose en palabras como símbolos de emociones, con el fin de determinar la polaridad de un texto dado. Con las nuevas representaciones que se crean, también se realiza el enriquecimiento de los lexicones de sentimientos, por medio de una distancia que permita agregar una nueva palabra al lexicón. Examinan la palabra nueva contra todas las palabras existentes en el diccionario, agregan nuevas palabras con base en los comentarios que se van analizando, este método de expansión es basado en un corpus de Facebook, una parte del corpus fue empleada para expansión y la otra para validación. Punto importante de este trabajo, es que este contiene explicaciones de cómo la tarea se ha abordado con métodos de aprendizaje automático (estadísticos) y basado en lexicones (lingüísticos). La desventaja de este método es que se basa en los emoticones tanto para generar su representación vectorial inicial, como también para dar un grado de polaridad a los comentarios y realizar la expansión de los lexicones. Lo anterior se ve afectado porque los emoticones no siempre representan el sentimiento del texto, ya que hay casos donde el sentimiento es contrario al emoticón.

En [Tulkens et al., 2016] se presenta un enfoque basado en diccionarios para la detección de racismo en comentarios de redes sociales en Holandés. Elaboraron un corpus que emplearon para evaluar su enfoque, este corpus estaba etiquetado de comentarios racista y no racistas por 2 anotadores. Para el enfoque utilizaron 3 diccionarios, el primer diccionario fue creado recuperando términos posiblemente

racistas y más neutrales de los datos de entrenamiento. El segundo diccionario fue creado a través de la expansión automática utilizando un modelo de *Word2Vec*. El tercer diccionario fue creado filtrando manualmente expresiones incorrectas. Después de crear los diccionarios, entrenaron manualmente varias máquinas de soporte vectorial utilizando las distribuciones de palabras sobre las diferentes categorías en el diccionario como características. Como resultado se obtuvo que el mejor diccionario fue aquel que fue filtrado. El hecho de que la cobertura de expansión de diccionarios fuera incrementada indica que las palabras que fueron automáticamente añadidas solo ocurren en el corpus, pero no ayudan a dar un impacto significativo en el rendimiento.

Para el caso de [Gitari et al., 2015] se desarrolló un clasificador para detectar la presencia de discurso de odio en foros y blogs. En este trabajo los autores identifican que el problema del discurso de odio es abstracto dentro de tres principales áreas temáticas: raza, nacionalidad y religión. El modelo de clasificación que desarrollaron emplea técnicas de análisis de sentimientos y en particular detección de subjetividad para no solamente detectar si una oración es subjetiva sino también identificar su polaridad. Lo primero que realizaron en este trabajo fue la reducción del documento, removiendo oraciones objetivas usando características de subjetividad y semánticas relacionadas con lenguaje de odio. Después, crean un lexicón de expresiones de sentimientos que es empleado para construir el clasificador. En el último paso se entrena un clasificador que utiliza características creadas desde el lexicón y lo usa para probarlo en un documento. Concluyen que los mejores resultados fueron logrados cuando incluyen semántica, odio y atributos basados en tema. El trabajo es interesante, sin embargo, al momento en que se usan más rasgos de subjetividad, los resultados no se puede comparar con otros métodos debido a que no todos los enfoques que están basados en lexicón aplican subjetividad.

3.3. MEX-A3T en IberEval 2018: Análisis de autoría y agresividad en tweets de español mexicano.

IberEval[6] es un taller que surge en el año 2010 con el objetivo de fomentar y promover el desarrollo de las Tecnologías del Lenguaje Humano (por sus siglas en inglés, HLT) para las lenguas ibéricas (español, portugués, catalán, vasco y gallego). Este taller está compuesto de una serie de evaluaciones y un foro de discusión sobre los diferentes aspectos relacionados con la evaluación de los sistemas de procesamiento del lenguaje natural, enfatizando los principales problemas a los que se enfrentan estos sistemas con las lenguas ibéricas modernas.

En el marco de IberEval 2018 se llevaron a cabo las siguientes tareas: 1) Análisis de autoría y agresividad en Twitter: estudio de caso en español mexicano (MEX-A3T), 2) Identificación automática de misoginia (AMI), 3) Segunda tarea de reconocimiento y resolución de abreviaturas biomédicas (BARR2), 4) Discapacidad de anotación en documentos del dominio biomédico (DIANN), 5) Análisis de humor basado en anotaciones humanas (HAHA), 5)Detección de posición multimodal en tweets sobre el referéndum catalán (MultiStanceCat).

En MEX-A3T [Álvarez-Carmona et al., 2018] se llevaron a cabo dos tareas: análisis de autoría[7] y detección de agresividad para textos en español de México. El conjunto de datos que se desarrolló para este trabajo de tesis, sirvió como base para la evaluación de las soluciones que participaron para la tarea de detección de agresividad en MEX-A3T. Todos los trabajos que participaron, sus enfoques propuestos son supervisados porque utilizaron todo el conjunto de entrenamiento para obtener rasgos y entrenar. Los rasgos que más analizaron los participantes en esta tarea fueron

[6] http://gplsi.dlsi.ua.es/congresos/ibereval10/

[7] Análisis de autoría se describe como la extracción de información sobre una persona mediante el análisis de un documento o texto escrito por esa persona.

palabras, n-gramas de caracteres, rasgos sintácticos, palabras agresivas, *word embeddings*, características afectivas y etiquetas POS. Para la clasificación se emplearon redes neuronales [Aragón and López-Monroy, 2018], [Frenda and Banerjee, 2018], clasificadores SVM [Graff et al., 2018], [Ortega-Mendoza and López-Monroy, 2018], Naïve Bayes [Correa and Martin, 2018] y regresión logística [Gómez-Adorno et al., 2018].

El ganador de la tarea fue el enfoque propuesto por [Graff et al., 2018], ellos propusieron utilizar EvoMSA, que es una arquitectura de dos niveles para análisis de sentimientos que utiliza información de diferentes modelos sobre el texto actual analizado para obtener una predicción final mediante una vista de consenso, además integraron métodos basados en lexicones. Los modelos basados en lexicones son Up-Down y Bernoulli, el primero consiste en producir un recuento de palabras afectivas y el segundo predice la agresividad de un texto usando un léxico con palabras agresivas.

Resumen del capítulo

En este capítulo se describieron los trabajos relacionados a lenguaje ofensivo, enfoques basados en lexicones, métodos basados en lexicón y aprendizaje supervisado. De los trabajos relacionados a lenguaje ofensivo se analizaron las dificultades a las que se enfrentaron para crear un conjunto de datos para la tarea, así como el análisis de los casos que no pudieron cubrir sus métodos. Para los enfoques basados en lexicón se exploró la metodología para expandir los diccionarios y el proceso de clasificación de textos.Por último, en los métodos basados en lexicón y aprendizaje supervisado, se revisó la arquitectura que se diseñó y los casos que lograron cubrir estos métodos.

Con la revisión de estos trabajos se logró identificar los puntos fuertes de cada uno e integrarlos en este trabajo. Como son: La expansión de lexicones, el uso de representaciones basadas en diferentes diccionarios, el uso de análisis de sentimientos y el diseño de métodos que son basados en lexicón y aprendizaje supervisado.

Capítulo 4

Generación de corpus en español

En este capítulo se describe la metodología aplicada en la recuperación y etiquetado del corpus de validación en español. En la sección 4.1 se detalla la recopilación de tweets, a continuación en la sección 4.2 se explican las rúbricas para el proceso de etiquetado y por último en la sección 4.3 se describen las características del corpus.

El conjunto de datos construido para este trabajo de tesis sirvió para la evaluación de las soluciones propuestas en MEX-A3T de IberEval 2018.

4.1. Proceso de recopilación de tweets

Para llevar a cabo la extracción de los tweets, primero se obtuvo un conjunto de palabras que sirvieron como semillas para extraer los textos. Estas semillas fueron obtenidas del diccionario de mexicanismos [Academia Mexicana de la Lengua, 2010], del cual se consideraron solo aquellas palabras que estuvieran clasificadas como vulgares y que no tuvieran la clasificación de coloquiales. Además se incluyeron palabras y *hashtags* analizados en [INMUJERES, 2016], estas palabras están presentes en contextos de violencia y acoso sexual contra las mujer en Twitter. En total se

recolectaron 143 semillas, en la tabla 4.1 se muestran algunas de estas.

#QueAscoSerHomosexual
luchona
pendejo
pendeja
prieto
prieta
vergazos
golfas
puta
lameculos

Tabla 4.1: Muestra del vocabulario aplicado para la recuperación de tweets.

Después se realizó un script en Python montando la API TwitterSearch[1] para extraer los mensajes. Para garantizar que los tweets que se recuperaran fueran de México, se configuró el script para que realizara una búsqueda geolocalizada. Se consideró a la ciudad de México como punto central y se contemplaron todos los tweets que se encontraban dentro de un radio de 500 km. La aplicación se encargó de hacer la búsqueda por semilla y recuperar todos los tweets dentro de los rangos establecidos. La recuperación de los datos fue realizada durante el período comprendido entre los meses de agosto a noviembre del año 2017. En total se obtuvieron 15,905 tweets.

[1] https://github.com/ckoepp/TwitterSearch/

4.2. Proceso de etiquetado

Debido a que el procedimiento de etiquetado es una tarea subjetiva, se requirió dos personas que fungieron como jueces para marcar aquellos tweets que son ofensivos o no. Para que pudieran realizar la tarea se les proporcionó la definición descrita en la sección 2.1, una serie de rúbricas y ejemplos guía. Las rúbricas y ejemplos se describen a continuación.

4.2.1. Reglas de etiquetado

Regla 1. Para identificar un texto ofensivo, se debe analizar si el objetivo de este es para dañar a otra persona. Se considera un tweet ofensivo si contiene adjetivos peyorativos, lenguaje discriminatorio o groserías dirigidas a una persona o colectivo. Además, puede contener también alguno de los siguientes elementos:

- **Apodos:** Sobrenombre que se le asigna a la persona/personas a quien va dirigido el mensaje, aludiendo a una discapacidad o defecto
- **Bromas:** Las bromas son mensajes que serán considerados ofensivos siempre y cuando su intención sea causar daño

Ejemplos de tweets ofensivos:

Tweet 1: “Tu novia la gata esa que usa hashtag hasta para poner hola, tu novia la acapulqueña esa”

Tweet 2: “Deja de estar de calientagüevos, que te vas a ganar una madriza”

Tweet 3: “Es una tipa tan cagante que no tiene amigos”

En los mensajes anteriores, se usan sobrenombres como: gata; adjetivos despectivos: cagante, calientagüevos; o groserías: madriza. En todos los casos la ofensa

está dirigida a una persona, en el tweet 1 la intención es hacer menos a la novia de alguien, en el tweet 2 la intención es de atacar a una persona si no se comporta y en el tweet 3 se está resaltado el defecto de alguien. Por lo tanto, en los tres casos son ofensivos.

Regla 2. Para facilitar la identificación, se le invita al etiquetador ubicarse en los zapatos del receptor del mensaje.

Regla 3. Se debe considerar la existencia de adjetivos que no son groseros, pero que son ofensivos, ejemplos:

- Idiota
- Estúpido
- Baboso
- Tonto
- Tarado

Estas palabras se deben manejar con cuidado debido a que no son groserías y que están permitidas en un contexto formal.

Regla 4. Los casos especiales, son textos que no están dirigidos a una persona o colectivo. Estos textos deberán marcarse como no ofensivos.

Los casos especiales se en listan a continuación:

- **Ofensa dirigida a un objeto:** Mensaje que contiene una ofensa que va dirigida al clima, las circunstancias de la vida u objetos.
- **Ofensa dentro de un diálogo:** Mensaje que describe un diálogo y dentro del diálogo se publica una ofensa.
- **Ofensa en una historia o citas:** Mensaje que narre lo que una persona dijo en X situación, haciendo un recuento de una ofensa o citando una ofensa.
- **Auto-Ofensa:** Mensaje que contenga una ofensa dirigida a la misma persona que escribió el mensaje.

- **Ofensa en anuncios:** Mensajes que contengan promoción de un video o persona. (Anuncios de prostitutas o videos pornográficos, utilizan groserías o adjetivos despectivos)

En seguida se muestran ejemplos de estos casos.

Tweet 4: "Aquí me juego la vida, o leo el libro o leo las diapos, porque nuestro capítulo es de mil putas hojas. *literal*" (Ofensa dirigida a un objeto)

Tweet 5: "-¡fui yo! ¡yo putos, yo! -pero ISIS, fue un sismo -QUE FUI YO LES DIGO, HIJOS DE SUS PUTAS MADRES" (Ofensa dentro de un dialogo)

Tweet 6: "Soy una enamoradiza sin remedio". -La emperatriz de todas las putas." (Ofensa en cita)

Tweet 7: "Atendiendote apartir de las 5 pm zona centro #SQUIRT #MILF #CULOS #NALGONA #HOTWIFE #SCORT #PUTAS" (Ofensa anuncio)

Regla adicional: Suelen haber excepciones en los casos donde hay ofensas dentro de un diálogo, historia o cita, donde después de éstas suele haber una opinión del autor del mensaje, es importante analizar esa opinión debido a que puede contener una ofensa a una persona /grupo.

En resumen, el tweet debe ser marcado con 1 si el tweet cumple los criterios de ofensivo y no cae en los casos especiales. En caso de que no cumpla los criterios o caiga en los casos especiales será marcado con 0.

4.2.2. Medición de acuerdo entre etiquetadores

Después de la definición de reglas, se les pidió a los etiquetadores analizarlas y posteriormente se llevó a cabo una prueba piloto de etiquetado en la que se verificó el acuerdo entre los jueces y la claridad de las reglas, ambos jueces etiquetaron los mismos mensajes. Con esta prueba se resolvieron dudas, además se realizaron ajustes

a las reglas existentes y se agregó la regla 2. La tabla 4.2 muestra el acuerdo en la prueba piloto y en etiquetado posterior.

Sesiones	Kappa
Prueba piloto	0.4240
Después de la prueba	0.5867

Tabla 4.2: Acuerdo antes y después del entrenamiento.

4.3. Descripción del corpus

Una vez terminado el proceso de etiquetado, se contaba con un total de 15,905 tweets. Posteriormente se consideraró solo aquellos tweets en los que ambos jueces habían tenido un acuerdo y los elementos seleccionados fueron normalizados, reemplazando el nombre de usuario con la etiqueta @USUARIO y las urls con <URL>. Por último, se realizó de forma aleatoria la selección de instancias que componen al conjunto de entrenamiento y prueba. En la Tabla 4.3 se muestra la distribución del corpus de lenguaje ofensivo en español. Como se observa, la distribución en la partición de entrenamiento es similar a la partición de prueba. La clase mayoritaria es la no ofensiva y la clase minoritaria es la ofensiva.

Clase	Entrenamiento (%)	Prueba (%)
No ofensivos	4,973 (65)	2,372 (75)
Ofensivos	2,727 (35)	784(25)
Σ	7700	3156

Tabla 4.3: Distribución del corpus.

Resumen del capítulo

En este capítulo se revisó el procedimiento para recopilar los datos, el número de etiquetadores, las reglas de etiquetado y la descripción del conjunto de datos resultante. Una vez revisada la metodología del corpus, se procede en el siguiente capítulo a revisar el método que se diseñó para detectar lenguaje ofensivo en textos.

Capítulo 5

Método propuesto

En este capítulo se describe cómo se realizó el diseño de nuestro enfoque basado en lexicones con un método supervisado para la detección de textos ofensivos. En cada sección se describe paso a paso cada recurso, métodos y reglas que componen al método. El orden de descripción de cada componente es organizado con la intención de que se conozca la secuencia en que fue requerida cada pieza del método.

5.1. Lexicón de insultos

En trabajos como [Nand et al., 2016] se comenta que el uso de las groserías destaca en mensajes ofensivos, debido a que los agresores las utilizan para atacar a sus víctimas. Por esta razón se decidió recolectar palabras usadas para insultar o maldecir. Para contar con un diccionario adecuado, se tomó en cuenta que en [Samghabadi et al., 2017] reportan que el uso de las groserías en la redes sociales se ha banalizado con el paso del tiempo, esto hace que estas palabras ya no sean exclusivas para ofender sino también forman parte del vocabulario informal. Por lo tanto el diccionario que empleamos contiene pocos insultos que son empleados en un contexto informal sin carga ofensiva.

El lexicón utilizado para el idioma inglés fue recuperado de *noswearing.com*. El sitio cuenta con 349 groserías y palabras para maldecir. El interés por este diccionario se debe a que la mayor parte del vocabulario está compuesto de insultos que no son tan comunes en el contexto informal. Para la tarea se usaron la mayoría de los elementos del diccionario, sin embargo se excluyeron expresiones de dos o más palabras debido a que el método propuesto sólo analiza a nivel palabra y no con expresiones, por ejemplo, *camel toe, nut sack, etc.* La longitud final del diccionario es de 336 palabras, parte del vocabulario se muestran en la Tabla 5.1. El diccionario completo se puede consultar en el apéndice A.

Palabras				
anus	bitch	clitfuck	dickface	mothafucka
arse	bitchass	cock	dike	motherfucker
arsehole	bitchtits	cockbite	dipshit	peckerhead
ass	bollocks	cumbubble	dumshit	pussies

Tabla 5.1: Extracto del diccionario de maldiciones.

Para el caso del idioma español, el lexicón fue extraído del diccionario de insultos *"Para insultar con propiedad"* [Montes-de Oca-Sicilia, 2016], el cual contiene más de 2,000 insultos, sin embargo solo se consideraron 83 palabras debido a que algunas se encuentran en desuso y la gran mayoría son aplicadas como parte del lenguaje coloquial. En la tabla 5.2 se muestran algunos de los insultos considerados. El diccionario completo se puede consultar en el apéndice B.

Palabras				
depravado	gentuza	imbécil	marrano	nefasto
depravada	hipócrita	inepto	marrana	ratero
escoria	ilusa	inepta	inútil	mierda
fantoche	ilusa	mandilon	nefasta	despreciable

Tabla 5.2: Extracto del diccionario de insultos.

5.2. Métodos de expansión de vocabulario

Los enfoques basados en diccionarios han dado buenos resultados, sin embargo, los diccionarios son recursos estáticos y no se adaptan a la evolución constante del lenguaje en las redes sociales [Tulkens et al., 2016]. Al detectar esta limitante, se consideró enriquecer el diccionario.

Para llevar a cabo esta tarea se implementaron dos métodos diferentes de expansión, dado un lexicón de insultos y un conjunto pre-entrenado de *word embeddings*. Ambos métodos comparten la idea de identificar nuevos insultos considerando la similitud contextual con un vocabulario conocido por medio de los word embeddings. Sin embargo, los métodos difieren en la forma en que calculan esta similitud y en el criterio aplicado para realizar la inserción de una nueva palabra al lexicón. Las siguientes subsecciones describen a detalle cada método. Para su descripción se asume que $\mathcal{L} = \{l_1, \ldots, l_n\}$ es el lexicón inicial de n insultos y $\mathcal{W} = \{(w_1, e(w_1)), \ldots, (w_n, e(w_m))\}$ es el conjunto pre-entrenado de embeddings, donde cada par representa una palabra y su vector correspondiente.

5.2.1. Método de expansión local

Este método expande el lexicón con las k palabras más similares de cada elemento inicial. El propósito es añadir diferentes variaciones léxicas y semánticas de los insultos iniciales, en la figura 5.1 se muestra este proceso. El método cuenta con dos pasos principales:

- **Expansión de palabra.**
  Por cada palabra del lexicón original $l_i \in \mathcal{L}$:
  (i) Se extrae su embedding de $\mathcal{W}$, denotado como $e(l_i)$.
  (ii) Se utiliza la similitud de coseno, se compara $e(l_i)$ contra el embedding $e(w_i)$ de cada $w_i \in \mathcal{W}$.
  (iii) Se extraen las k palabras más similares a l_i, definiendo el conjunto $E_i = (w_1, \ldots, w_k)$.
- **Expansión del diccionario.** Una vez determinado el conjunto de las k palabras más similares para cada insulto, los conjuntos E_1 a E_n, se unen para crear el lexicón expandido $\mathcal{L}_\mathcal{E} = \mathcal{L} \cup E_1 \ldots \cup E_m$

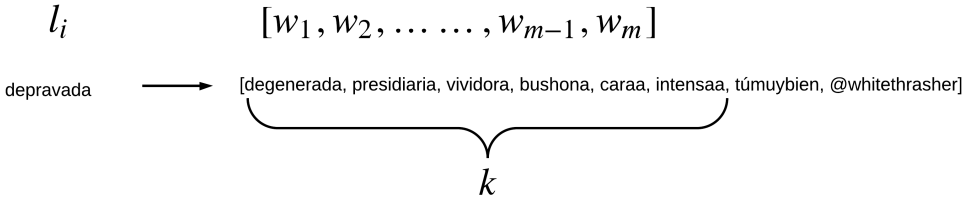

Figura 5.1: Expansión por palabra del método local.

5.2.2. Método de expansión global

Este método tiene como objetivo enriquecer el lexicon con palabras fuertemente relacionadas al argot ofensivo, sin necesidad de hacer asociaciones a una palabra en particular del vocabulario inicial. El propósito es encontrar palabras que tengan

contextos similares a todo el diccionario, en la figura 5.2 se muestra este proceso. Los pasos que componen a este método son los siguientes:

- **Calcular vector promedio del lexicón original.**

  (i) Se extrae el embedding $e(l_i)$ para cada palabra $l_i \in \mathcal{L}$.

  (ii) Se calcula el promedio de estos embeddings para obtener el vector que describe al lexicón entero $e(\mathcal{L})$, llamado el embedding de contexto.

- **Expansión del diccionario.**

  (i) Usando la similitud de coseno se compara $e(\mathcal{L})$ contra el embedding $e(w_i)$ para cada $w_i \in \mathcal{W}$.

  (ii) Se extrae las k palabras más similares a $e(\mathcal{L})$, definiendo el conjunto $E_K = (w_1, \ldots, w_k)$.

  (iii) Se insertan las palabras extraídas en el lexicón original para construir el nuevo lexicon, $\mathcal{L}_\mathcal{E} = \mathcal{L} \cup E_K$.

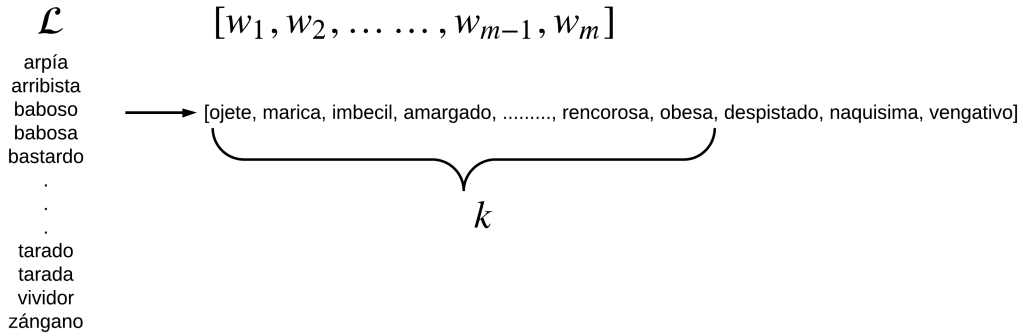

Figura 5.2: Expansión por palabra del método global.

5.3. Lexicón de sentimientos

En [Del Bosque and Garza, 2014] y [Waseem et al., 2017b] se observó que empleando únicamente insultos en el proceso de detección no era suficiente, debido a que hay palabras a las que se les da una carga negativa para ofender y no son in-

sultos. Por lo tanto, la detección de lenguaje ofensivo podría estar relacionada con análisis de sentimientos. Por consiguiente, en el método se consideró agregar un diccionario adicional para afinar el proceso de etiquetado automáico. El diccionario que se integra es uno enfocado a análisis de opiniones que está compuesto de palabras positivas y negativas. En este diccionario no se llevó a cabo el proceso de expansión debido a que el enriquecimiento de este tipo de diccionario requiere de condiciones y características que no son parte del alcance del método propuesto.

El lexicón de opinión utilizado para el idioma inglés es el que se reporta en [Hu and Liu, 2004], se consideró este lexicón debido a que cuenta con un gran volumen de palabras negativas. En total el diccionario cuenta con 6,800 palabras de las cuales 4,783 son negativas y 2,007 son positivas. Debido a que no se pudo hallar un lexicón con la misma longitud para el idioma español se decidió realizar la traducción automática de este, utilizando la API del traductor de Google.

5.4. Extracción de rasgos ofensivos

En los trabajos relacionados a lenguaje ofensivo se han identificado una serie de atributos que distinguen a los textos ofensivos de los que no lo son. Uno de los atributos que se considera en esta metodología se observa en [Del Bosque and Garza, 2014], dónde el número de insultos es un indicador, debido a que entre más groserías tenga un texto resulta que existe mayor probabilidad de sea ofensivo a comparación de los que tienen un menor número.

Otros atributos se describen en la tipología propuesta en [Waseem et al., 2017b], aquí se identifica que una ofensa es directa si se hace uso de menciones, pronombres personales e identidades nombradas. Además menciona que las ofensas dirigidas a grupos de personas emplean un vocabulario despectivo relacionado a esos grupos. También es reportado otro atributo en [Gitari et al., 2015], donde la polaridad ne-

gativa de un texto es un indicador del cual se puede tomar ventaja para ofensas explícitas. Explica que las ofensas explícitas no son ambiguas y que tienen como objetivo potencial ofender. Otro elemento que se comenta en [Waseem et al., 2017b], es que los trabajos en lenguaje ofensivo que emplean word embeddings parecen prometedores en cuestión de capturar términos relacionados al lenguaje ofensivo. Tomando en cuenta lo anterior se consideraron los siguientes atributos para cada texto:

- Número de insultos
- Número de palabras positivas
- Número de palabras negativas
- Número de veces que es empleado el pronombre personal en 2da persona
- Número de veces que es empleado el pronombre personal en 3era persona
- Distancia coseno con el diccionario de insultos
- Distancia coseno con el diccionario de palabras negativas
- Distancia coseno con el diccionario de palabras positivas

Además se integró el número de palabras para analizar si hay una relación entre la longitud de un texto y el lenguaje ofensivo.

5.5. Etiquetado automático

Una de las partes primordiales del método propuesto es el enfoque basado en diccionarios, ya que este se encarga de asignarle una etiqueta a los datos que cumplen con ciertos criterios para declararlos ofensivos o no ofensivos, después estos son usados para entrenar al clasificador supervisado.

Para la descripción del enfoque se asume que $\mathcal{T} = \{t_1, \ldots, t_m\}$ representa el conjunto de textos no etiquetados, $\mathcal{W}$ es el conjunto pre-entrenado de embeddings, m es el número de insultos que debe contener un texto para ser considerado ofensivo, $\mathcal{T}' = \{(t'_1, c(t'_1)), \ldots, (t'_m, c(t'_m))\}$ representa los datos de entrenamiento selecciona-

dos, donde cada par representa el texto y su clase asignada automáticamente.

$nInsultos_i$, $nNegativas_i$ y $nPositivas_i$ corresponde al número de insultos, palabras negativas y palabras positivas que contiene t_i. $distIns_i$, $distPos_i$, $distNeg_i$ son las distancias de coseno que hay entre t_i con el lexicón de insultos ($\mathcal{L}$), las palabras positivas ($\mathcal{P}$) y las palabras negativas ($\mathcal{N}$) del lexicón de opinión .

Para realizar el etiquetado de las instancias, se consideraron 2 reglas, las cuales ayudaron a elegir aquellos elementos que son más probables de ser ofensivos y no ofensivos. Las reglas son las siguientes: 1) Un texto se considera ofensivo si el número de insultos es mayor al umbral m y el número de palabras negativas es mayor al de positivas. 2) Un texto es no ofensivo si no tiene groserías, ni palabras negativas y el número de palabras positivas es mayor a 0. La etiqueta que se le asigna a los textos ofensivos es "1" y para los no ofensivos "0". A continuación en el algoritmo 1 se ilustra el método basado en diccionarios aplicando los atributos de la sección 5.4 y reglas para el etiquetado automático.

Algorithm 1 Etiquetado automático

procedure LEXICONMETHOD

Input: Textos no etiquetados, embeddings, lexicón de insultos $\mathcal{L}$, lexicón de palabras positivas $\mathcal{P}$, lexicón de palabras negativas $\mathcal{N}$ y umbral de groserías .

Output: Datos de entrenamiento seleccionados y etiquetados

for all $t_i \in \mathcal{T}$ **do**

$nInsultos_i \leftarrow intCount(t_i, \mathcal{L})$

$nNegativas_i \leftarrow intCount(t_i, \mathcal{N})$

$nPositivas_i \leftarrow intCount(t_i, \mathcal{P})$

$distPos_i \leftarrow getDistance(W, t_i, \mathcal{P})$

$distNeg_i \leftarrow getDistance(W, t_i, \mathcal{N})$

$distIns_i \leftarrow getDistance(W, t_i, \mathcal{L})$

if $nInsultos_i >= m$ **then**

if $nNegativas_i > nPositivas_i$ **then**

$\mathcal{T}' \leftarrow add(t_i, 1)$

end if

if $nInsultos_i == 0$ **then**

if $distPos_i > 1$ AND $distNeg_i == 0$ **then**

$\mathcal{T}' \leftarrow add(t_i, 0)$

end if

end if

end if

end for

end procedure

5.6. Proceso del método propuesto

El flujo de trabajo de nuestro método inicia con un lexicón que contiene insultos, este cuenta con un número limitado de vocablos. Por lo tanto, el primer procedimiento que realiza nuestro método es la expansión de este lexicón aplicando una de las técnicas descritas en la sección 5.2. Después, se obtiene un conjunto de datos sin etiquetar, estos datos son analizados por el método propuesto en la sección 5.5, que requiere adicionalmente un lexicón de insultos expandido y otro de sentimientos. En seguida se obtienen un subconjunto de los datos originales etiquetados automáticamente.

Posteriormente se entrena el clasificador supervisado con los datos obtenidos, para este método se utiliza el clasificador SVM. Por último se evalúa el método con el cálculo de métricas (Precisión, Recuerdo, F-measure, etc). En la figura 5.3 se ilustra la arquitectura descrita.

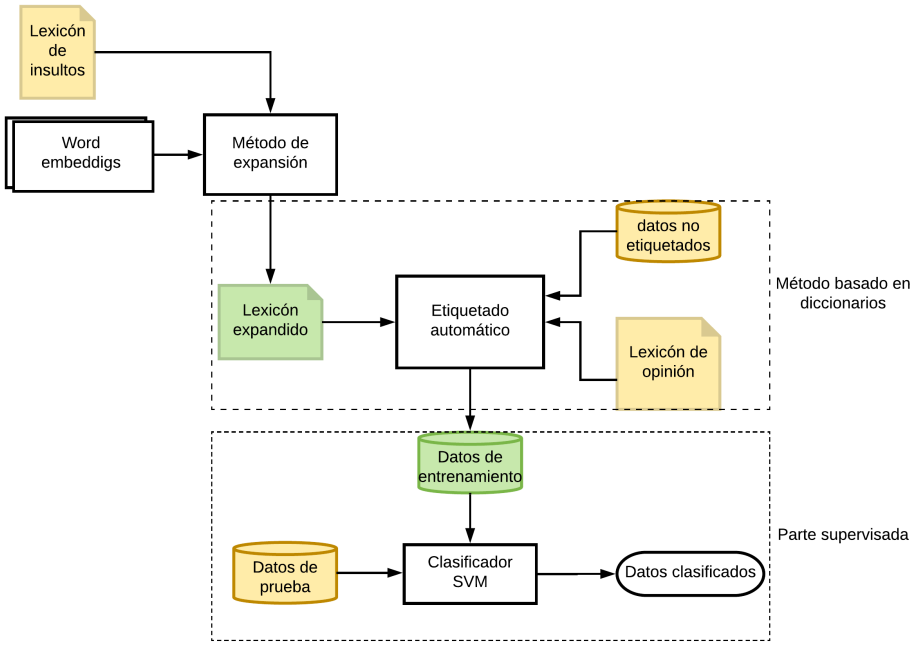

Figura 5.3: Arquitectura del método propuesto.

Resumen del capítulo

En este capítulo se revisó la arquitectura del método propuesto, desde los métodos de expansión de diccionarios hasta los rasgos que deben identificar para su funcionamiento. Por lo tanto, una vez que se dio conocer cada parte de este método, se procede a validar que tan eficiente es este para detectar lenguaje ofensivo en textos en español e inglés. En el siguiente capítulo se describen los resultados obtenidos.

Capítulo 6

Experimentos y resultados

En este capítulo se presenta la configuración experimental con la que se lleva a cabo la evaluación del método propuesto y los resultados de los experimentos. En las secciones 6.1 y 6.2 se exponen los recursos utilizados en el idioma inglés y español. Posteriormente, en la sección 6.3 se describen las configuraciones aplicadas a cada módulo de nuestro método y por último en la sección 6.4 se expone a detalle la evaluación del método y la comparación con el estado del arte.

6.1. Conjunto de datos

Kaggle: Detectando insultos en comentarios sociales[1]

El corpus de Kaggle es uno de los pocos conjuntos de datos en inglés que está enfocado a la tarea de forma general, está compuesto de comentarios que son parte de conversaciones en blogs o foros. Los datos han sido utilizados en investigaciones sobre la detección de textos racistas y ciberacoso [Samghabadi et al., 2017]. El conjunto cuenta con dos categorías asociadas: el comentario es insulto o neutral.

[1] https://www.kaggle.com/c/detecting-insults-in-social-commentary

La tabla 6.1 muestra como se distribuye cada clase en el conjunto de entrenamiento y prueba.

Clase	Entrenamiento (%)	Prueba (%)
No ofensivos	2898 (73.5)	1954 (73.8)
Ofensivos	1049 (26.5)	693(26.2)
Σ	3947	2647

Tabla 6.1: Información del conjunto de datos.

Para el idioma español el conjunto de datos ocupado es el que se describe en el capítulo 4, ahí se detalla la metodología aplicada, el proceso de etiquetado y su distribución.

6.2. *Word embeddings*

En el presente trabajo, los modelos de *Word embeddings* que se usaron para el proceso en inglés son [Pennington et al., 2014] y [Godin et al., 2015].

En [Pennington et al., 2014], los embeddings se generaron a partir de 2,000 millones de tweets utilizando *GloVe*, por otro lado, en [Godin et al., 2015] se construyeron con *Word2Vec* y se entrenaron con 400 millones de tweets.

Los embeddings utilizados para el idioma español fueron generados por el Laboratorio de Tecnologías del Lenguaje del INAOE. La herramienta que se empleó para su generación fue gensim[2] la cual cuenta con una implementación de *Word2Vec* en Python y el entrenamiento fue realizado con tweets de 10,000 usuarios, de cada usuario se extrajeron en promedio 1,500 mensajes.

[2]https://radimrehurek.com/gensim/

6.3. Configuración de módulos

6.3.1. Configuración del método local de expansión

Para llevar a cabo la expansión con este método, se requiere de tres elementos: un lexicón a expandir, un umbral k y word embeddings. Para identificar el umbral ideal, los experimentos se realizaron con diferentes umbrales para obtener el adecuado. Las pruebas se realizaron para los dos idiomas con los siguientes valores de umbral: 5, 10, 25, 50, 100, 500 y 1000.

6.3.2. Configuración del método global de expansión

En este método al igual que el anterior, requiere de 3 elementos de entrada. Pero el valor del umbral k requiere ser más grande comparado con el método local, por lo tanto las pruebas se realizaron con los siguientes valores de umbral: 1,000, 1,500, 2,000, 4,000 y 8,000.

6.3.3. Configuración del etiquetado automático

Para realizar una selección fina de elementos ofensivos y no ofensivos, se consideró un umbral de m igual 2 y 3. Se tomaron estos umbrales debido a que quedaban menos de 3 elementos ofensivos para el entrenamiento si se aumentaba el valor de m para cuando se usa el diccionario antes de expandir. Estos umbrales se aplicaron para ambos idiomas.

6.3.4. Representaciones de los textos

Antes de que los textos seleccionados pasen al proceso de entrenamiento del clasificador, se requiere representar los datos. Las representaciones utilizadas en estos experimentos fueron bolsa de palabras (BOW), 3-gramas de caracteres y la otra representación fue utilizar los rasgos ofensivos descritos en 5.4. Para BOW y 3-gramas de caracteres su pesado se realizó dada la frecuencia bruta de las palabras.

6.3.5. Parámetros de SVM

Los experimentos fueron llevados a cabo en Python utilizando la herramienta *Scikit-learn* que cuenta con múltiples algoritmos de aprendizaje entre ellos *SVM*. La configuración del clasificador es simple, se utilizó un kernel lineal.

6.4. Resultados

El primer paso del enfoque es la expansión del diccionario de insultos, por lo tanto se evaluó cual de los métodos se adapta mejor al proceso. La evaluación se realizó sobre los datos de entrenamiento de cada conjunto de datos.

Para el idioma inglés y español se obtuvieron mejores resultados con la expansión global que con la expansión local, pero se tomó el mejor de cada expansión para realizar los experimentos en el etiquetado automático. Para el caso del inglés la expansión se realizó con dos tipos de embeddings, la diferencia que radica entre ellos es que en *Word2Vec* se observaron variaciones léxicas de los elementos expandidos y en cambio con *GloVe* se obtenían palabras relacionadas en contexto. En la tabla 6.2 se muestra esta diferencia.

Embedding	Palabra	Expansión
Glove	ass	bitch, nigga, dick, shit, dumb, nasty, damn, bitches, ugly, fuck
Word2Vec	ass	azz, asss, axx, a$$, assss, asx, asssss, asz, a*s, asssss
Glove	bitch	ass, nigga, fuck, shit, bitches, hoe, hell, dick, damn, lmao
Word2Vec	bitch	bxtch, bish, b*tch, bitxh, btch, bitchh, bitchhh, bitchhhh, b***h
Glove	cock	dick, tits, pussy, sucking, cocks, dildo, arse, milf, balls, anal, lick
Word2Vec	cock	Cock, dick, c*ck, c**k, cocks, penis, d*ck, clit, schlong, cock-

Tabla 6.2: Ejemplos de expansión utilizando *GloVe* y*Word2Vec*.

Tomando en cuenta lo anterior, los lexicones empleados para el etiquetado en el idioma español fueron el expandido local 100 y el global 8,000. Para inglés se decidió tomar los embeddings de *GloVe* porque con ellos se podrían encontrar palabras relacionadas al contexto ofensivo. Los diccionarios que se aplicaron para los experimentos fueron el local expandido 50 y el global expandido 4,000.

En la segunda parte del método se realizó el etiquetado automático, se seleccionaron aquellas instancias que conformarían el conjunto de entrenamiento para el clasificador. Los corpus utilizados cuentan con etiquetas manuales para el entrenamiento, pero estas no fueron tomadas en cuenta. En la tabla 6.3 se muestra los datos iniciales, en las tablas 6.4 y 6.5 se muestra cuantos elementos fueron etiquetados automáticamente y el porcentaje que representan de los datos iniciales.

Idioma	Datos sin etiqueta
Inglés	3,947
Español	7,700

Tabla 6.3: Información de datos sin etiqueta.

Diccionario	Valor *m*	Etiquetados ofensivos	Etiquetados no ofensivos	Total(%)
Original	2	197	214	411(10.4 %)
	3	87	214	301(7.6 %)
Local (50)	2	801	115	916(23.2 %)
	3	508	115	623(15.7 %)
Global (4,000)	2	644	173	817(20.7 %)
	3	644	173	817(20.7 %)

Tabla 6.4: Resultados del etiquetado automático para el idioma inglés.

Diccionario	Valor *m*	Etiquetados ofensivos	Etiquetados no ofensivos	Total(%)
Original	2	45	113	158(2.05 %)
	3	6	113	119(1.5 %)
Local (100)	2	1060	30	1090(14.1 %)
	3	412	30	442(5.7 %)
Global (8,000)	2	2358	8	2366(30.7 %)
	3	1418	8	1426(18.5 %)

Tabla 6.5: Resultados del etiquetado automático para el idioma español.

Como se puede observar, la mayoría de los subconjuntos representan menos del 25 % de los datos originales. A pesar de que sean pocos datos, las instancias que ahora predominan son las ofensivas, con estos datos de entrenamiento veremos que impacto tiene utilizar un conjunto reducido, no etiquetado manualmente y que cuenta con más instancias ofensivas.

El último paso del método fue entrenar el clasificador con los documentos etiquetados y validar, en las tablas 6.6 y 6.7 se muestran los resultados.

Diccionario	Representación	Valor *m*	Valor f no ofensivo	Valor f ofensivo	Exactitud
Original	BOW	2	0.82	0.37	0.72
		3	0.85	0.23	0.75
	3-gram char	2	0.82	0.41	0.72
		3	0.84	0.28	0.74
	Rasgos	2	0.73	0.51	0.65
		3	0.82	0.37	0.72
Local (50)	BOW	2	0.27	0.43	0.36
		3	0.61	0.48	0.56
	3-gram char	2	0.37	0.45	0.41
		3	0.61	0.47	0.55
	Rasgos	2	0.58	0.50	0.54
		3	0.71	0.51	0.63
Global (4,000)	BOW	2	0.57	0.49	0.54
		3	0.57	0.49	0.54
	3-gram char	2	0.63	0.51	0.58
		3	0.63	0.51	0.58
	Rasgos	2	0.68	**0.53**	0.62
		3	0.68	**0.53**	0.62

Tabla 6.6: Resultados de la clasificación para el idioma inglés.

Con los valores obtenidos se puede notar que al utilizar el método con el diccionario de insultos sin expandir, se tiene un bajo valor f en la clase de interés, pero al realizar la expansión se puede visualizar que el incremento del vocabulario hace una notable mejora. Revisando cada representación aplicada, resulta tener mejores resultados el uso de los rasgos calculados como atributos.

Diccionario	Representación	Valor *m*	Valor f no ofensivo	Valor f ofensivo	Exactitud
Original	BOW	2	0.85	0.13	0.75
		3	0.85	0.002	0.75
	3-gram char	2	0.85	0.15	0.75
		3	0.85	0	0.75
	Rasgos	2	0.78	0.31	0.67
		3	0.85	0.08	0.75
Local (100)	BOW	2	0.01	0.39	0.25
		3	0.02	0.39	0.25
	3-gram char	2	0.02	0.39	0.25
		3	0.07	0.40	0.27
	Rasgos	2	0.53	0.40	0.47
		3	0.77	0.39	0.66
Global (8,000)	BOW	2	0.002	0.39	0.24
		3	0.003	0.39	0.24
	3-gram char	2	0	0.39	0.24
		3	0	0.39	0.24
	Rasgos	2	0.24	0.40	0.33
		3	0.53	**0.41**	0.47

Tabla 6.7: Resultados de la clasificación para el idioma español.

A pesar de que los conjuntos de datos son distintos no solamente en el idioma sino también en tamaño, temática y complejidad, se obtiene un valor f alto al aplicarle la misma configuración de parámetros para ambos. Los resultados coincidieron en que para ambos, el método funciona bien con el método global de expansión, con un valor m=3 para el proceso de etiquetado y el uso de los rasgos calculados como atributos.

Después de haber obtenido los resultados es importante realizar una comparativa de nuestros resultados contra el estado del arte y métodos tradicionales supervisados. A continuación en la figura 6.1 se muestra la comparativa para el idioma inglés y en la figura 6.2 para el español. Los valores reportados en las figuras es el valor f de la clase ofensiva.

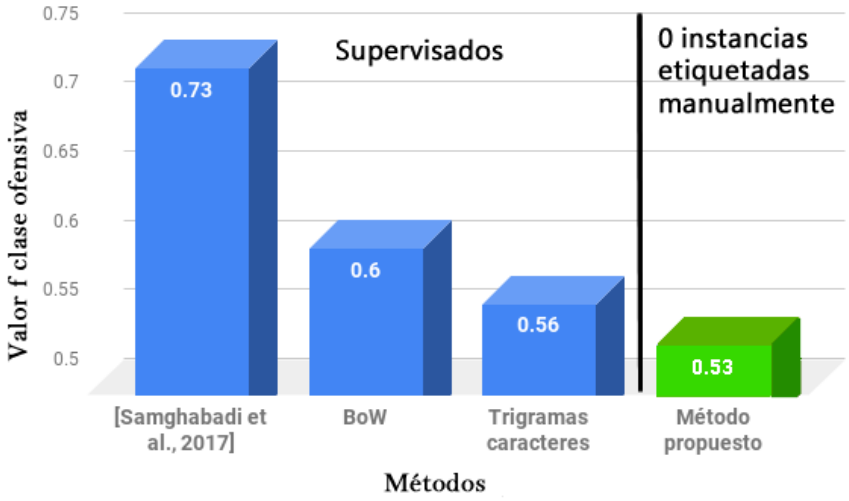

Figura 6.1: Comparativa de métodos para el idioma inglés.

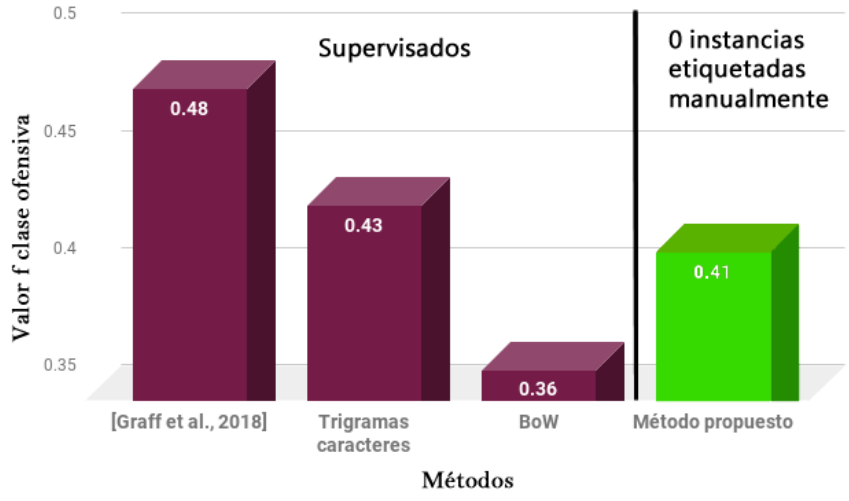

Figura 6.2: Comparativa de métodos para el idioma español.

Como se puede observar en los resultados para inglés y español, el método no iguala ni supera al estado del arte y tampoco a los métodos tradicionales, con excepción del idioma español que supera a BOW. Pero la comparación anterior no se puede realizar directamente debido a que los métodos anteriores tienen la ventaja de haber utilizado todo el conjunto de entrenamiento y sus etiquetas. En el caso del enfoque propuesto, para el idioma inglés se usó apenas el 20.7 % de los datos y para el español el 18.5 %, los cuales fueron etiquetados automáticamente.

Aunque no podemos probar los métodos propuestos por [Samghabadi et al., 2017] y [Graff et al., 2018] con un conjunto de entrenamiento reducido, se puede realizar la comparación con los métodos tradicionales. En las figuras 6.3 y 6.4 se muestra una reducción de datos para los enfoques tradicionales, los número reportados son el f-score de la clase de interés.

Las reducciones fueron la siguientes: para el idioma inglés se redujeron los datos de entrenamiento al 21 % y para español al 20 %. Los elementos fueron elegidos al azar y la elección de estos se realizó diez veces y se obtuvo la media.

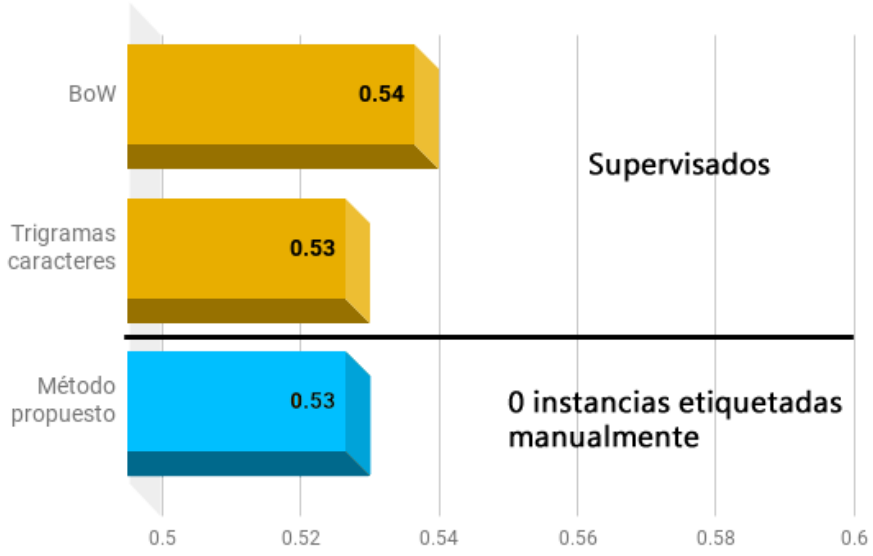

Figura 6.3: Comparativa de métodos con reducción en inglés, valor f.

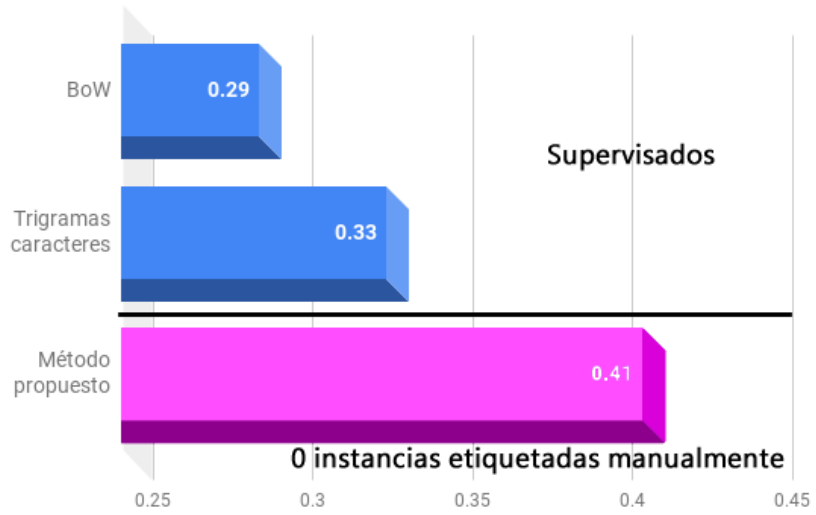

Figura 6.4: Comparativa de métodos con reducción en español, valor f.

Realizando la reducción se puede observar que a nuestro enfoque en el idioma inglés, funciona casi igual que si se utilizaran datos manualmente etiquetados y en español nuestro método es mejor.

Analizando los experimentos realizados, queda claro que al usar solo un subconjunto del entrenamiento provocó que no obtengamos resultados iguales o mayores al estado del arte. Pero los resultados obtenidos son cercanos a lo reportado y el enfoque propuesto no requirió datos etiquetados manualmente.

Capítulo 7

Análisis y discusión

En este capítulo se revisarán los elementos que impactaron en el rendimiento del enfoque propuesto. Se examinan elementos como el etiquetado automático, la calidad de los lexicones de insultos iniciales y la naturaleza de los corpus utilizados para entrenar y validar.

7.1. Análisis del etiquetado automático

Para determinar qué tan exacto fue el etiquetado automático, se tomaron las etiquetas reales de los datos de entrenamiento y se compararon contra las etiquetas automáticas. Primero comenzamos con la parte del etiquetado automático y los lexicones de insultos ya que ambos son usados en conjunto. Para el idioma inglés en la tabla 7.1 se expone qué tan exacto fue el algoritmo de etiquetado al identificar instancias ofensivas como no ofensivas. Para el caso de reconocer instancias ofensivas, el algoritmo obtiene mejores resultados antes de expandir el diccionario de insultos debido a que aproximadamente más del 70 % de los documentos identificados ofensivos si lo son, pero si se realiza la expansión del diccionario, decae a un 57 %. Esto refleja que algunas de las palabras que se integraron al diccionario pueden ser usadas

en contextos no ofensivos y está seleccionando instancias como ofensivas cuando no lo son. Para el caso del etiquetado de textos no ofensivo se observa que no decrementa sino al contrario mejora de 93 % a 95 %. Esto quiere decir que las reglas para identificar textos no ofensivos se ven beneficiadas con la expansión del lexicón de insultos.

Diccionario	Valor *m*	Etiquetados ofensivos	Ofensivos correctos(%)	Etiquetados no ofensivos	No ofensivos correctos(%)
Original	2	197	140(71)	214	200(93.4)
	3	87	68(78.1)	214	200(93.4)
Local (50)	2	801	405(50.5)	115	110(95.6)
	3	508	280(55.1)	115	110(95.6)
Global (4,000)	2	644	369(57.2)	173	165(95.3)
	3	644	369(57.2)	173	165(95.3)

Tabla 7.1: Comparación etiqueta automática vs etiqueta real inglés.

Textos
"fucking weird game man"
"thats a pathetic flop fuck you jvg"
"Exactly. Like, I don't really give a shit how fat a person may or may not feel, but it really is shitty when your clothes don't fit. And not in a I'm so gross way but in a I can't very well leave the house naked!! way."
"Eh, Billy Joel's not a bad guy. Music's fucking awful, of course."
"No, but you can't just treat someone like a piece-of-shit just because you think your shit don't stink. No one is above the law, so by law you have the right to ask a person to leave your property if they are disturbing the peace or damaging property, or stealling."

Tabla 7.2: Ejemplos en inglés que se etiquetaron incorrectamente como ofensivos.

Algunos ejemplos de instancias que se están clasificando como ofensivas se ilustran en la tabla 7.2. Como se puede observar algunos de los textos mal clasificados utilizan más de una grosería y palabras negativas, pero no están agrediendo a una persona, sino a un objeto o situación. Estos textos al contener una longitud mayor a la de un tweet, contienen más elementos que provocaron que fueran mal clasificados.

Para el caso de los textos que si eran ofensivos, pero se clasificaron no ofensivos se muestran en la tabla 7.3, se recurre a ofensas relacionadas al ámbito político como el caso de *"libTURD"*[1] también recurren al sarcasmo, empleando palabras no ofensivas y cambiando el sentido con otras en conjunto.

Textos
"why don't your wife blow all off the liberals for a quarter each and donate the money,,,????"
"You sir are not intelligent"
"Go away libTURD, your kind just can't stand that the man told the truth, something you would never understand! "

Tabla 7.3: Ejemplos ofensivos en inglés que se etiquetaron incorrectamente.

En el idioma español el cálculo de elementos correctamente etiquetados se muestra en la tabla 7.4. Antes de la expansión se tiene un 80 % de textos ofensivos correctamente identificados y con la expansión cae a menos del 45 %. Pero en los textos no ofensivos mejora la selección pasando de un 70 % correctamente etiquetados a un 100 %. Es importante notar que para la clase no ofensiva se reduce considera-

[1]Combinación de liberal y mierda. Un Libturd generalmente se refiere a la inclinación política de izquierda hacia la extrema izquierda que cree que sus puntos de vista liberales son los únicos puntos de vista válidos y que cualquiera que esté en desacuerdo con ellos pertenece a la derecha religiosa. Definición de www.urbandictionary.com

blemente el número de instancias para entrenar.

Diccionario	Valor *m*	Etiquetados ofensivos	Ofensivos correctos(%)	Etiquetados no ofensivos	No ofensivos correctos(%)
Original	2	45	36(80)	113	80(70.7)
	3	6	6(100)	113	80(70.7)
Local (100)	2	1060	532(50.19)	30	26(86.6)
	3	412	254(61.6)	30	26(86.6)
Global (8,000)	2	2358	967(41.01)	8	8(100)
	3	1418	657(46.3)	8	8(100)

Tabla 7.4: Comparación etiqueta automática vs real español.

En los experimentos para el idioma español se identificó que el mayor problema fue etiquetar elementos no ofensivos como ofensivos. Pasó igual que en los comentarios en inglés, los textos cuentan con elementos ofensivos que son empleados en agresiones no dirigidas a personas, lo cual indica que el método de etiquetado automático no ha sido capaz de diferenciar cuando una ofensa es para una persona, objeto o situación. Cabe destacar que a diferencia del inglés, en español se puede notar que en un texto corto aplican de forma rebuscada las groserías para describir más detalladamente lo que se desea expresar. En la tabla 7.5 se muestran algunos tweets incorrectamente clasificados.

Tweets
“puta madre quiero suicidarme por la hdp de la gastritis”
“Que poca madre que en pleno siglo XXI exista esa pobre mentalidad de discriminacion”
“@USUARIO JAJAJAJAJAJAJAJ la puta madre te amo”
“No putas mames. ¿Que vergas acabo de ver?! #estoyContenida ”
“Vale verga no se ni como putas se lee esa cantidad de dinero”

Tabla 7.5: Tweets no ofensivos incorrectamente etiquetados.

Analizando lo anterior, esto no solamente recae en el método sino también en el lexicón de insultos, ya que solo está limitado a un conjunto de insultos empleados de forma general, no incluye elementos ofensivos de regiones o elementos socioculturales ofensivos presentes en el dataset. Por lo tanto, al expandir el lexicón no se pueden capturar los elementos ofensivos socioculturales y argot relacionado, debido a que los embeddings se entrenan con textos obtenidos de otras regiones o corresponden a temáticas opuestas a los conjuntos de datos. Por otra parte el análisis individual de palabras impacta en el rendimiento del método, debido a que se pueden identificar ciertas expresiones que están compuestas de dos o más insultos, pueden ser expresiones usadas en el lenguaje informal y si se tratan solas no se obtiene el significado real. En la siguiente sección se muestran ejemplos de estas expresiones presentes en los corpus.

7.2. Análisis de los conjuntos de datos

Los corpus empleados para validar el enfoque propuesto no solo difieren en el idioma, sino también el las temáticas que tratan cada uno, la longitudes de los textos y la fuente de la que provienen. En la sección anterior se identificó que los tópicos

que trata el corpus en inglés son referentes a política y contexto social de los Estados Unidos. Los textos además cuentan con una longitud superior a los 280 caracteres, debido a que estos no pertenecen a Twitter sino son comentarios extraídos de blogs de noticias. Parte de esto se puede constatar en la tabla 7.6, 7.7 y 7.8 donde se muestra el vocabulario y expresiones para el dataset en inglés.

Palabra	# apariciones
people	286
obama	122
country	69
president	91
government	51
american	51

Tabla 7.6: Palabras frecuentes en el conjunto de entrenamiento en inglés.

Bigrama	# apariciones
gay marriage	19
united states	14
ass nigga	10
president obama	9
black people	9
piece shit	9
white house	8
tax payers	7

Tabla 7.7: Bigramas frecuentes en el conjunto de entrenamiento en inglés.

Trigrama	# apariciones
people born gay	4
barack hussein obama	3
xa president obama	3
single idiotic comment	3
stupid pro amd	3

Tabla 7.8: Trigramas frecuentes en el conjunto de entrenamiento en inglés.

Para el caso del español, el contexto social es de México, los mensajes no excedan los 280 caracteres porque su origen es de Twitter y las temáticas que abarca son política, racismo, clasismo, homofobia, sexismo, entretenimiento, fútbol y cultura. A diferencia del corpus en inglés, los insultos abundan por igual en textos ofensivos y no ofensivos, lo cual marca una dificultad mayor para discernir. En 7.9, 7.10 y 7.11 se muestra el vocabulario que abunda para cada caso.

Después de haber analizado estos conjuntos de datos se llega a la conclusión de que ambos son de naturaleza distinta ya que el conjunto de datos de *Kaggle* el lenguaje es más formal y los textos son de mayor longitud porque provienen de blogs relacionados a noticias. Por otra parte el dataset en español, los mensajes son de Twitter, donde el lenguaje suele ser coloquial y más complejo para analizar por la falta de contexto en los mensajes.

Tipo de texto	Palabra	# apariciones
Ofensivo	putos	649
	madre	516
	putas	477
	verga	382
	hdp	301
No ofensivo	verga	1102
	madre	952
	putas	750
	loca	742
	putos	529

Tabla 7.9: Comparación vocabulario entre clases.

Tipo de texto	Bigrama	# apariciones
Ofensivo	puta madre	117
	chinguen madre	43
	mil putas	41
	chingar madre	41
	chinga madre	41
No ofensivo	vale verga	102
	puta madre	87
	vale madre	53
	valer verga	45
	poca madre	32

Tabla 7.10: Comparación bigramas entre clases.

Tipo de texto	Trigrama	# apariciones
Ofensivo	hijo puta madre	24
	hijos puta madre	22
	hijo mil putas	16
	hijos putas madres	15
	hija puta madre	13
No ofensivo	madre teresa calcuta	20
	voy volver loca	9
	bien pinche loca	9
	litros agua loca	8
	vale verga vida	6

Tabla 7.11: Comparación trigramas entre clases.

Capítulo 8

Conclusiones y trabajo a futuro

8.1. Conclusiones

En este trabajo de tesis se contribuye con la creación de un método que integra un enfoque basado en diccionarios con un clasificador supervisado para la detección de mensajes ofensivos sin requerir datos etiquetados manualmente.

La primera parte que se desarrolló de nuestro enfoque, fueron los dos métodos de enriquecimiento de vocabulario que tienen como objetivo capturar palabras relacionadas a contextos ofensivos. El método que mejor resultado dio para el inglés y español fue el de expansión global, el cual considera todo el contexto del diccionario de insultos para obtener palabras relacionadas al lenguaje ofensivo. La expansión del diccionario de insultos ayudó a enriquecer el método basado en lexicones que cumplió con la función de realizar el etiquetado automático de las instancias no etiquetadas y con estos datos entrenar un clasificador supervisado. A pesar de que el algoritmo de etiquetado automático solo etiquetó una porción de los datos (20.7 % para inglés y un 18.5 % para español), estos fueron suficientes para entrenar al clasificador y obtener resultados cercanos a lo reportado. Los textos que nuestro enfoque detecta

son aquellos que tienen ofensas directas, donde el receptor, el vocabulario y la intensión son claros. Sin embargo nuestro método no pudo cubrir aquellas instancias que cuentan con elementos lingüísticos más elaborados como el sarcasmo, palabras y expresiones que tienen carga negativa con base en un contexto sociocultural.

Nuestro método se vio afectado para el caso del idioma inglés porque los textos analizados provenían de una fuente relacionada a foros y al realizar la expansión del vocabulario los embeddings utilizados provenian de Twitter. El corpus contaba con un lenguaje más formal comparado con el que se observa en Twitter. En el caso del español la dificultad que encontramos recae en el origen del corpus, la gran mayoría de insultos estaban presentes tanto en tweets ofensivos como no ofensivos, esto nos remarcó la banalización del uso de insultos en mensajes de redes sociales. Se observaron tweets que contienen lenguaje relacionado a política, programas de televisión y fútbol para ofender a otra persona o grupo, por lo tanto la carga negativa recae a un nombre propio o palabra común para volverla ofensiva.

Un elemento importante de nuestro trabajo fue la creación del corpus de lenguaje ofensivo para el idioma español de México, el cual sirvió para la evaluación de las soluciones propuestas en IberEval 2018. Durante el proceso de creación del corpus se analizó la dificultad de proponer una definición de lenguaje ofensivo y entender el porqué es tan subjetiva la tarea de etiquetado. A pesar de las dificultades en la tarea de etiquetado, se pudo observar como los mexicanos expresan sus emociones e insultan, la mayoría de los tweets reflejaban enojo o tristeza hacia un objeto o situación, cargadas de varios insultos. Esta clase de textos requieren un análisis adicional para identificar si es a una persona u objeto al que se ofende y agregar ese rasgo al proceso de clasificación, debido a que en la gran mayoría de los textos parecía que se estaba ofendiendo a alguien cuando en realidad solo era al clima o al transporte público. Otro aspecto identificado fue la cantidad de anuncios relacionados a prostitución, que usan groserías que denigran a una mujer para llamar la atención de los caballeros, la mayoría de estos anuncios utilizan los *hashtags* para poner los

insultos. La identificación de este tipos de tweets podría ayudar a combatir la trata de blancas, debido a que la mayoría de las personas ofertadas en esos tweets eran menores de edad.

8.2. Trabajo futuro

Como trabajo a futuro planeamos mejorar los métodos de expansión siendo más estrictos con las palabras que se agregan. Esto significa agregar criterios que permitan discernir entre palabras coloquiales y palabras ofensivas. Otro aspecto a mejorar sería recabar más datos para generar modelos de *word embeddings* con textos específicos por zona ya que el contexto sociocultural influye considerablemente para la interpretación de los insultos. Por último, otra mejoría en nuestro enfoque es el algoritmo de etiquetado automático para que se evalúen más elementos del contexto, agregando un lexicón de expresiones fijas y *embeddings* en los criterios.

Apéndice A

Lexicón de palabras ofensivas inglés

El lexicón recuperado de ***noswearing.com***.

1. anus
2. arse
3. arsehole
4. ass
5. assjabber
6. assbag
7. assbandit
8. assbanger
9. assbite
10. assclown
11. asscock
12. asscracker
13. asses
14. assface
15. assfuck
16. assfucker
17. assgoblin
18. asshat
19. asshead
20. asshole
21. asshopper
22. assjacker
23. asslick
24. asslicker
25. assmonkey
26. assmunch
27. assmuncher
28. assnigger
29. asspirate
30. assshit
31. assshole
32. asssucker
33. asswad
34. asswipe
35. axwound
36. bampot
37. bastard
38. beaner
39. bitch
40. bitchass
41. bitches
42. bitchtits
43. bitchy
44. blowjob
45. bollocks
46. bollox
47. boner
48. brotherfucker
49. bullshit
50. bumblefuck
51. buttplug
52. butt
53. buttfucka
54. buttfucker
55. cameltoe
56. carpetmuncher
57. chesticle
58. chinc
59. chink
60. choad
61. chode
62. clit
63. clitface
64. clitfuck

65. clusterfuck
66. cock
67. cockass
68. cockbite
69. cockburger
70. cockface
71. cockfucker
72. cockhead
73. cockjockey
74. cockknoker
75. cockmaster
76. cockmongler
77. cockmongruel
78. cockmonkey
79. cockmuncher
80. cocknose
81. cocknugget
82. cockshit
83. cocksmith
84. cocksmoke
85. cocksmoker
86. cocksniffer
87. cocksucker
88. cockwaffle
89. coochie
90. coochy
91. coon
92. cooter
93. cracker
94. cum
95. cumbubble
96. cumdumpster
97. cumguzzler
98. cumjockey
99. cumslut
100. cumtart
101. cunnie
102. cunnilingus
103. cunt
104. cuntass
105. cuntface
106. cunthole
107. cuntlicker
108. cuntrag
109. cuntslut
110. dago
111. damn
112. deggo
113. dick
114. dickbag
115. dickbeaters
116. dickface
117. dickfuck
118. dickfucker
119. dickhead
120. dickhole
121. dickjuice
122. dickmilk
123. dickmonger
124. dicks
125. dickslap
126. dicksucker
127. dicksucking
128. dicktickler
129. dickwad
130. dickweasel
131. dickweed
132. dickwod
133. dike
134. dildo
135. dipshit
136. doochbag
137. dookie
138. douche
139. douchebag
140. douchewaffle
141. dumass
142. dumbass
143. dumbfuck
144. dumbshit
145. dumshit
146. dyke
147. fag
148. fagbag
149. fagfucker
150. faggit
151. faggot
152. faggotcock
153. fagtard
154. fatass
155. fellatio
156. feltch
157. flamer
158. fuck
159. fuckass
160. fuckbag
161. fuckboy
162. fuckbrain
163. fuckbutt
164. fuckbutter
165. fucked
166. fucker
167. fuckersucker
168. fuckface
169. fuckhead
170. fuckhole
171. fuckin
172. fucking
173. fucknut
174. fucknutt
175. fuckoff
176. fucks
177. fuckstick
178. fucktard
179. fucktart
180. fuckup

181. fuckwad
182. fuckwit
183. fuckwitt
184. fudgepacker
185. gay
186. gayass
187. gaybob
188. gaydo
189. gayfuck
190. gayfuckist
191. gaylord
192. gaytard
193. gaywad
194. goddamn
195. goddamnit
196. gooch
197. gook
198. gringo
199. guido
200. handjob
201. heeb
202. hell
203. ho
204. hoe
205. homo
206. homodumbshit
207. honkey
208. humping
209. jackass
210. jagoff
211. jap
212. jerk
213. jerkass
214. jigaboo
215. jizz
216. junglebunny
217. kike
218. kooch
219. kootch
220. kraut
221. kunt
222. kyke
223. lameass
224. lardass
225. lesbian
226. lesbo
227. lezzie
228. mcfagget
229. mick
230. minge
231. mothafucka
232. mothafuckin
233. motherfucker
234. motherfucking
235. muff
236. muffdiver
237. munging
238. negro
239. nigaboo
240. nigga
241. nigger
242. niggers
243. niglet
244. nutsack
245. paki
246. panooch
247. pecker
248. peckerhead
249. penis
250. penisbanger
251. penisfucker
252. penispuffer
253. piss
254. pissed
255. pissflaps
256. polesmoker
257. pollock
258. poon
259. poonani
260. poonany
261. poontang
262. porchmonkey
263. prick
264. punanny
265. punta
266. pussies
267. pussy
268. pussylicking
269. puto
270. queef
271. queer
272. queerbait
273. queerhole
274. renob
275. rimjob
276. ruski
277. sandnigger
278. schlong
279. scrote
280. shit
281. shitass
282. shitbag
283. shitbagger
284. shitbrains
285. shitbreath
286. shitcanned
287. shitcunt
288. shitdick
289. shitface
290. shitfaced
291. shithead
292. shithole
293. shithouse
294. shitspitter
295. shitstain
296. shitter

297. shittiest
298. shitting
299. shitty
300. shiz
301. shiznit
302. skank
303. skeet
304. skullfuck
305. slut
306. slutbag
307. smeg
308. snatch
309. spic
310. spick
311. splooge
312. spook
313. suckass
314. tard
315. testicle
316. thundercunt
317. tit
318. titfuck
319. tits
320. tittyfuck
321. twat
322. twatlips
323. twats
324. twatwaffle
325. unclefucker
326. vag
327. vagina
328. vajayjay
329. vjayjay
330. wank
331. wankjob
332. wetback
333. whore
334. whorebag
335. whoreface
336. wop

Apéndice B

Lexicón de palabras ofensivas español

El lexicón extraído del diccionario de insultos ***"Para insultar con propiedad"*** [Montes-de Oca-Sicilia, 2016].

1. arpía
2. arribista
3. baboso
4. babosa
5. bastardo
6. bobo
7. boba
8. bocafloja
9. bocona
10. buey
11. burgués
12. cabezahueca
13. caca
14. canijo
15. canija
16. chismoso
17. chismosa
18. cínico
19. cobarde
20. cochino
21. cochina
22. dañada
23. depravado
24. depravada
25. desgraciado
26. desgraciada
27. déspota
28. despreciable
29. engendro
30. engreído
31. escoria
32. fantoche
33. gentuza
34. guarro
35. guarra
36. guey
37. hipócrita
38. hocicón
39. hocicona
40. huevón
41. huevona
42. idiota
43. iluso
44. ilusa
45. imbécil
46. inepto
47. inepta
48. infeliz
49. ingrato
50. ingrata
51. insolente
52. inútil
53. jodido
54. jodida
55. lambiscón
56. lamehuevos
57. lelo
58. lela
59. mamón
60. mamona

61. mandilon
62. marrano
63. marrana
64. menso
65. mensa
66. mierda
67. nefasto
68. nefasta
69. odioso
70. odiosa
71. orate
72. patán
73. pedorro
74. pedorra
75. pinche
76. ratero
77. ratera
78. ruco
79. ruca
80. tarado
81. tarada
82. vividor
83. zángano

Bibliografía

[Academia Mexicana de la Lengua, 2010] Academia Mexicana de la Lengua (2010). *Diccionario de mexicanismos*. Siglo XXI Editores, México.

[Álvarez-Carmona et al., 2018] Álvarez-Carmona, M. Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H. J., Villasenor-Pineda, L., Reyes-Meza, V., and Rico-Sulayes, A. (2018). Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, pages 74–96.

[Ameur et al., 2017] Ameur, H., Jamoussi, S., and Hamadou, A. B. (2017). Sentiment lexicon enrichment using emotional vector representation. In *Computer Systems and Applications (AICCSA), 2017 IEEE/ACS 14th International Conference on*, pages 951–958. IEEE.

[Aragón and López-Monroy, 2018] Aragón, M. E. and López-Monroy, A. P. (2018). Author profiling and aggressiveness detection in spanish tweets: MEX-A3T 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, pages 134–139.

[Asociación de Internet.mx, 2017] Asociación de Internet.mx (2017). 13 Estudio sobre los hábitos de los usuarios de internet en méxico 2017. https://www.asociaciondeinternet.mx/es/component/remository/Habitos-de-Internet/13-Estudio-sobre-los-Habitos-de-los-Usuarios-de-Internet-en-Mexico-2017/lang,es-es/?Itemid=.

[Badjatiya et al., 2017] Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, pages 759–760.

[Bengio et al., 2003] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

[Chatzakou et al., 2017] Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., and Vakali, A. (2017). Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 13–22. ACM.

[Chen et al., 2012] Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, pages 71–80. IEEE.

[Cohen, 1968] Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213–220.

[Correa and Martin, 2018] Correa, S. and Martin, A. (2018). Linguistic generalization of slang used in mexican tweets, applied in aggressiveness detection. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the*

Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018., pages 119–127.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[Davidson et al., 2017] Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

[Del Bosque and Garza, 2014] Del Bosque, L. P. and Garza, S. E. (2014). Aggressive text detection for cyberbullying. In *Mexican International Conference on Artificial Intelligence*, pages 221–232. Springer.

[Derechos-Digitales, 2016] Derechos-Digitales (2016). *Internet en México: Derechos Humanos en el entorno digital*. Derechos Digitales, México.

[Frenda and Banerjee, 2018] Frenda, S. and Banerjee, S. (2018). Deep analysis in aggressive mexican tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, pages 108–113.

[Friedman et al., 1997] Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, 29(2-3):131–163.

[Ghosh et al., 2015] Ghosh, D., Guo, W., and Muresan, S. (2015). Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1003–1012.

[Gitari et al., 2015] Gitari, N. D., Zuping, Z., Damien, H., and Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

[Godin et al., 2015] Godin, F., Vandersmissen, B., De Neve, W., and Van de Walle, R. (2015). Multimedia lab @ acl wnut ner shared task: Named entity recognition for twitter microposts using distributed word representations. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 146–153.

[Gómez-Adorno et al., 2018] Gómez-Adorno, H., Bel-Enguixa, G., Sierra, G., Sánchez, O., and Quezada, D. (2018). A machine learning approach for detecting aggressive tweets in spanish. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, pages 102–107.

[Graff et al., 2018] Graff, M., Miranda-Jiménez, S., Tellez, E. S., Moctezuma, D., Salgado, V., Ortiz-Bejar, J., and Sánchez, C. N. (2018). INGEOTEC at MEX-A3T: author profiling and aggressiveness analysis in twitter using μtc and evomsa. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, pages 128–133.

[Grčar, 2012] Grčar, M. (2012). Text mining and text stream mining tutorial.

[Guo and Diab, 2012] Guo, W. and Diab, M. (2012). Learning the latent semantics of a concept from its definition. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 140–144. Association for Computational Linguistics.

[Hu and Liu, 2004] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 168–177.

[Huang, 2008] Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56.

[INMUJERES, 2016] INMUJERES (2016). "CDMX Ciudad Segura y Amigable para la Mujeres y las Niñas". https://inmujeres.cdmx.gob.mx/storage/app/media/Estudios_Diagnosticos/PlanAccionesPublicas$_R$*edesSociales.pdf*.

[Joshi et al., 2016] Joshi, A., Tripathi, V., Patel, K., Bhattacharyya, P., and Carman, M. (2016). Are word embedding-based features useful for sarcasm detection? *arXiv preprint arXiv:1610.00883*.

[Jurdzinski et al., 2017] Jurdzinski, G. et al. (2017). Word embeddings for morphologically complex languages. *Schedae Informaticae*, 2016(Volume 25):127138.

[Kalita, 2015] Kalita, D. (2015). Supervised and unsupervised document classification-a survey. *International Journal of Computer Science and Information Technologies*, 6(2):1971–1974.

[Kohavi and Provost, 1998] Kohavi, R. and Provost, F. (1998). Confusion matrix. *Machine learning*, 30(2-3):271–274.

[Kruse et al., 2016] Kruse, R., Borgelt, C., Braune, C., Mostaghim, S., and Steinbrecher, M. (2016). *Computational intelligence: a methodological introduction*. Springer.

[Lalji and Deshmukh, 2016] Lalji, T. and Deshmukh, S. (2016). Twitter sentiment analysis using hybrid approach. *International Research Journal of Engineering and Technology*, 3(6):2887–2890.

[Lee et al., 2010] Lee, T. L., Fiske, S. T., and Glick, P. (2010). Next gen ambivalent sexism: Converging correlates, causality in context, and converse causality, an introduction to the special issue. *Sex Roles*, 62(7-8):395–404.

[McHugh, 2012] McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

[Mehdad and Tetreault, 2016] Mehdad, Y. and Tetreault, J. (2016). Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303.

[Michelucci, 2018] Michelucci, U. (2018). *Feedforward Neural Networks*, pages 83–136. Apress, Berkeley, CA.

[Mikolov, 2014] Mikolov, T. (2014). *Facebook Research. Consultado en Febrero, 2018.* https://research.fb.com/people/mikolov-tomas/.

[Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

[Mikolov et al., 2010] Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association.*

[Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine learning.* McGraw Hill series in computer science. McGraw-Hill.

[Montes-de Oca-Sicilia, 2016] Montes-de Oca-Sicilia, M. d. P. (2016). *Para insultar con propiedad, Diccionario de insultos.* Editorial Otras Inquisiciones, México.

[Mubarak et al., 2017] Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56.

[Nand et al., 2016] Nand, P., Perera, R., and Kasture, A. (2016). "How bullying is this message?": A psychometric thermometer for bullying. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 695–706.

[Nobata et al., 2016] Nobata, C., Tetreault, J. R., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 145–153.

[Ortega-Mendoza and López-Monroy, 2018] Ortega-Mendoza, R. M. and López-Monroy, A. P. (2018). The winning approach for author profiling of mexican users in twitter at mex.a3t@ibereval-2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, pages 140–148.

[Park and Fung, 2017] Park, J. H. and Fung, P. (2017). One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

[Rico-Sulayes, 2014] Rico-Sulayes, A. (2014). *De vulgaridades, insultos y malsonancias: el diccionario del subestándar mexicano.* Selección Anual para el Libro Universitario. Universidad Autónoma de Baja California.

[Rumelhart et al., 1986a] Rumelhart, D. E., Hinton, G. E., and McClelland, J. L. (1986a). A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(26):45–76.

[Rumelhart et al., 1986b] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986b). Learning representations by back-propagating errors. *Nature*, 323:533–536.

[Sabariah et al., 2015] Sabariah, M. K., Effendy, V., et al. (2015). Sentiment analysis on twitter using the combination of lexicon-based and support vector machine for assessing the performance of a television program. In *Information and Communication Technology (ICoICT), 2015 3rd International Conference on*, pages 386–390. IEEE.

[Samghabadi et al., 2017] Samghabadi, N. S., Maharjan, S., Sprague, A., Diaz-Sprague, R., and Solorio, T. (2017). Detecting nastiness in social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 63–72.

[Tan et al., 2008] Tan, S., Wang, Y., and Cheng, X. (2008). Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*, pages 743–744.

[Tang et al., 2014] Tang, D., Wei, F., Qin, B., Zhou, M., and Liu, T. (2014). Building large-scale twitter-specific sentiment lexicon: A representation learning approach. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 172–182.

[Thomas, 2009] Thomas, P. (2009). Semi-supervised learning by olivier chapelle, bernhard schölkopf, and alexander zien (review). *IEEE Trans. Neural Networks*, 20(3):542.

[Tulkens et al., 2016] Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., and Daelemans, W. (2016). The automated detection of racist discourse in dutch social media. *Computational Linguistics in the Netherlands Journal*, 6(1):3–20.

[Waseem et al., 2017a] Waseem, Z., Chung, W. H. K., Hovy, D., and Tetreault, J. (2017a). Proceedings of the first workshop on abusive language online. In *Proceedings of the First Workshop on Abusive Language Online*.

[Waseem et al., 2017b] Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017b). Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

[White et al., 2015] White, L., Togneri, R., Liu, W., and Bennamoun, M. (2015). How well sentence embeddings capture meaning. In *Proceedings of the 20th Australasian Document Computing Symposium, ADCS 2015, Parramatta, NSW, Australia, December 8-9, 2015*, pages 9:1–9:8.

[Zhang et al., 2011] Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., and Liu, B. (2011). Combining lexicon-based and learning-based methods for twitter sentiment analysis. In *HP Laboratories Technical Report*. HPL-2011-89.