UNIVERSIDAD DE LAS AMÉRICAS PUEBLA

Escuela de Artes y Humanidades

Departamento de Lenguas

Detección de agresividad, ofensividad y vulgaridad basada en rasgos lingüísticos y aplicada a un corpus de tweets

Paulina Alejandra Morán Méndez.

ID: 152820

Licenciatura en Idiomas

Dr. Antonio Rico Sulayes.

San Andrés Cholula, Puebla. Otoño 2019

Tesis que, para completar los requisitos del Programa de Honores presenta la estudiante **Paulina Alejandra Morán Méndez.**

Director de Tesis

Dr. Antonio Rico Sulayes

Presidente de Tesis

Dra. Myrna Iglesias Barrón

Secretario de Tesis

Dra. Brita Banitz

Detección de agresividad, ofensividad y vulgaridad basada en rasgos lingüísticos y aplicada a un corpus de tweets

Abstract
Social media is in constant growth and with it, its content as well. Since these are platforms that allow users to express their thoughts without much censorship, the publication of potentially damaging content is rapidly increasing. Thus, there is an urgent need to manage the content that could harm other users. However, because the great amount of content that is being produced, the detection of this harmful content has to be done automatically. This research uses as data the corpus produced in a project in which I collaborated. For that corpus, I and a colleague elaborated a diagram based on linguistic attributes to ensure an objective and concise tagging of offensive language. The corpus we annotated was created by the Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) for the MEX-A3T workshop. In the annotation project, three classes were proposed as qualities for the messages: offensive, aggressive and vulgar language. The results of the annotation were satisfactory, meaning a linguistic view presents improvements indeed. For this thesis, I use those classes and characteristics and apply them to an actual classification task. Therefore, the question to answer in this research is whether this linguistic properties used in the annotation, such as speech acts and linguistic variation, can serve as features for the classification of aggressiveness, offensiveness and vulgarity in a small sample of INAOE's corpus. I use Weka platform for the classification, applying two algorithms available with this software: Naïve Bayes and a decision tree. The results of the first experiment were high, a 73.33% of accuracy, and after a feature reduction, the accuracy improved to 86.66%.
Keywords: text classification, aggressiveness, vulgarity, offensiveness, speech acts.

Abstract
Los medios sociales están en constante crecimiento y, con ellos, también su contenido. Dado que se trata de plataformas que permiten a los usuarios expresar sus opiniones sin mucha censura, la publicación de contenidos potencialmente perjudiciales está aumentando rápidamente. Por lo tanto, existe una necesidad urgente de gestionar el contenido que podría perjudicar a otros usuarios. Sin embargo, debido a la gran cantidad de contenido que se está produciendo, la detección de este contenido nocivo tiene que hacerse automáticamente. Esta investigación utiliza como datos el corpus producido en un proyecto en el que colaboré. Para ese corpus, yo y una colega elaboramos un diagrama basado en atributos lingüísticos para asegurar un etiquetado objetivo y conciso del lenguaje ofensivo. El corpus que etiquetamos fue creado por el Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) para el taller MEX-A3T. En ese proyecto de etiquetado se propusieron tres clases como cualidades para los mensajes: lenguaje ofensivo, agresivo y vulgar. Los resultados del etiquetado fueron satisfactorios, lo que significa que una visión lingüística presenta mejoras. Para esta tesis utilizo esas clases y características y las aplico a una tarea de clasificación real. Por lo tanto, la pregunta a responder en esta investigación es si estas propiedades lingüísticas utilizadas en el etiquetado, como los actos de habla y la variación lingüística, pueden servir como características para la clasificación de la agresividad, ofensividad y vulgaridad en una pequeña muestra del corpus del INAOE. Para la clasificación, se utilizó la plataforma Weka, aplicando dos algoritmos disponibles con este software: Naïve Bayes y un árbol de decisión. Los resultados del primer experimento fueron altos, un 73,33% de precisión, y después de una reducción de los atributos, la precisión mejoró a 86,66%.
Palabras clave: clasificación del texto, agresividad, vulgaridad, ofensa, actos de habla.

Agradecimientos

Quiero agradecer a mi familia que siempre me han apoyado en cada etapa de mi vida. A mis padres porque no hay forma de que esté aquí sin ellos. Particularmente a mi mamá por ser una mujer tan increíble y que me ha hecho quien soy, que me impulsa siempre a trabajar arduamente y que sin ella no me hubiera sentado a empezar a escribir.

A mi hermana por su compañía, consejos, por los momentos de ocio y los necesarios breaks cuando los necesitaba.

A mis compañeras de carrera por todo su apoyo y por hacer de mis últimos semestres tan memorables. A Majo por todos los mensajes de apoyo cuando más los necesitaba y por aguantar mis constantes consultas con dudas de todo tipo.

A mis profesores por todo el conocimiento que he adquirido, tanto académico como personal. En específico a mi asesor el Dr. Antonio Rico, por su guía durante esta investigación y fuera de ella también.

A Víctor por todo su apoyo, su conocimiento y compañía. Gracias por todos los momentos juntos.

Índice

1. Introducción

Debido al aumento en el acceso a internet y a las herramientas allí disponibles, plataformas como las redes sociales se han vuelto la opción favorita de mucha gente para el intercambio de información. En 2019 el número aproximado de usuarios de redes sociales en todo el mundo es de 3,484 millones, un 9% más que el año anterior (Kemp, 2019). Esto implica que el uso de las redes sociales no sólo es enorme actualmente, sino que está en constante crecimiento. Sin embargo, este uso masivo implica la presencia de problemas. Turel & Qahri-Saremi (2016) mencionan que a pesar de los beneficios que el uso de las redes sociales han tenido para los usuarios, no todos sus efectos han sido positivos. El uso de estas redes también ha llevado a una serie de comportamientos problemáticos en forma de usos impulsivos, riesgosos y desventajosos. Este tipo de comportamientos se encuentran entre las consecuencias negativas que Gutiérrez-Esparza, Vallejo-Allende, & Hernández-Torruco (2019) consideran que afectan el bienestar de la sociedad. Es por esto que de acuerdo con el ex agente del FBI Douglass, nuevos delitos han surgido en paralelo con el nacimiento de nuevas tecnologías y métodos de comunicación en línea (Buoncompagni, 2018). De aquí surge la necesidad de detectar estos posibles delitos, cuyo origen se encuentra en los mensajes publicados. Dichos mensajes la mayor parte de las veces son textuales, por lo que detallar sus características lingüísticas para poder ubicarlos es primordial. En este sentido, Gutiérrez-Esparza et al. (2019) explican que donde más se encuentran agresiones es en las redes sociales (por ejemplo, Facebook y Twitter), los servicios de mensajes cortos, los foros, los sitios de encuestas con shit-posting (contenido agresivo de baja calidad), los blogs, los sitios web para compartir videos y las salas de chat, entre otros. Por esto, las nuevas formas de comunicación no sólo han producido efectos sociales positivos, sino que también se han

convertido en instrumentos de violencia, entendida ésta como una violación de las normas sociales (Buoncompagni, 2018). Algunas de las razones por las que estas plataformas fomentan la violencia es debido a sus características intrínsecas. Su accesibilidad y rápida adopción, hacen que sea imposible que los moderadores vigilen cada post realizado o que filtren su contenido. Buoncompagni (2018) hace un análisis acerca de la violencia en las redes sociales y explica que hay diversos estudios enfocados en el efecto de las pandillas en su uso de las redes sociales. El autor comenta que las redes sociales son un espacio perfecto para poder analizar el comportamiento de estas pandillas. En este tema, el autor compara cómo era la situación antes, donde los límites territoriales bastante estrictos limitaban los enfrentamientos entre pandilleros. En contraste, dentro de las plataformas digitales, esto ya no es posible y los lenguajes de odio, como comportamientos violentos, pueden circular mucho más libremente creando efectos a un nivel social más amplio.

Debido a los problemas mencionados, es muy importante analizar el contenido que se está subiendo a las redes sociales para llevar un mejor control de las plataformas y asegurar que los usuarios estén seguros. Sin embargo, debido a la cantidad masiva de contenido publicado el análisis de este se torna extremadamente difícil. Por lo tanto, existe una demanda apremiante de métodos para identificar automáticamente las publicaciones riesgosas (Ruppenhofer, Siegel, Wiegan, 2018). Es con esta motivación que surgen los workshops de detección de contenido riesgoso en redes sociales, especialmente en Twitter. Twitter es una red social que se caracteriza por la cantidad de usuarios que tiene, la forma en la que la plataforma les deja expresar ideas, la facilidad para el anonimato y la libertad para publicar contenido sin censura. Un ejemplo del uso negativo de Twitter es el caso de los jóvenes que

pertenecen a las bandas de Chicago. Ellos se han convertido en armas al usar una cuenta de Twitter y esta condición fomenta la violencia en el futuro (Buoncompagni, 2018).

Para poder contrarrestar las consecuencias negativas que viene con el mal uso de Twitter, diversos investigadores se han dado la tarea de diseñar sistemas de detección de lenguaje riesgoso. Estas investigaciones se han presentado en workshops donde a manera de competencia, personas de varias organizaciones se registran para ofrecer una solución a un problema y compiten para obtener la mejor solución. Normalmente a los equipos participantes se les otorga un conjunto de datos etiquetados manualmente siguiendo los lineamientos requeridos en la tarea. Estos datos son usados para entrenar un sistema que intenta solucionar el problema. Al final se les proporciona a los equipos un conjunto de datos sin anotación para la parte de prueba. Comparando los criterios de referencia, o la solución a los datos de prueba, se determina el mejor sistema de clasificación. Este tipo de competencias se enfocan en diversos temas con tareas de perfilamiento de autor, detección de *hate speech*, misoginia, ofensividad, entre otros. Para poder tener una mejor caracterización del lenguaje usado en tweets, los workshops se han hecho en varios idiomas. Un ejemplo de esto es el PAN, que es una serie de eventos científicos y tareas compartidas sobre texto digital forense y de estilometría. Las tareas propuestas para la versión de 2020 son: perfilamiento de celebridades: edición influencer, verificación de autoría entre dominios, perfiles de los difusores de noticias falsas en Twitter, y detección de cambios de estilo (PAN, 2019). También existe una versión de este tipo de competencia para el español (Markov, Gómez-Adorno, Jasso-Rosales, & Sidorov, 2018). Incluso existen competencias dedicadas exclusivamente al español de México, como el MEX-A3T "*Authorship and aggressiveness*

analysis in Twitter: case study in Mexican Spanish" (REDTTL - Red Temática en Tecnologías del Lenguaje, 2019).

El MEX-A3T presenta una tarea de detección automática de tweets agresivos u ofensivos en español de México basado en el entrenamiento de los sistemas de detección utilizando para ello un corpus de tweets etiquetados manualmente. Este workshop presenta un reto extra a la clasificación de la agresividad per se al involucrar también la variación lingüística. Esto, sin embargo, lo acerca a la realidad al mismo tiempo. Esta tarea es muy compleja y esto se vio reflejado en el bajo rendimiento reportado por la mayoría de los participantes en la primera edición del MEX-A3T (REDTTL - Red Temática en Tecnologías del Lenguaje, 2019). Uno de los problemas con los que se enfrentaron los competidores fue que el corpus de entrenamiento no se encontraba bien etiquetado. Este aspecto es muy importante debido a que para poder lograr una mejora en el rendimiento de los grupos participantes se les debe proporcionar un conjunto de datos de entrenamiento lo mejor etiquetado posible. Sin embargo, como ya se mencionó, el etiquetado manual de tweets fue inestable, inconsistente y por lo tanto ineficiente. Lo anterior está reflejado en el bajo acuerdo kappa, la medida de evaluación del etiquetado, el cual fue 0.5867 (Álvarez-Carmona et al., 2018).

Por consiguiente, el laboratorio de tecnologías del lenguaje del Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) realizó de enero a mayo de 2019 un proyecto de mejora del etiquetado, en el cual yo colaboré. En ese proyecto se llevó a cabo la elaboración de un diagrama de etiquetado de corpus de agresividad para poder lograr disminuir las inconsistencias del etiquetado manual y para procurar un etiquetado objetivo (OpenCor, 2019). Como resultado de dicha investigación, se encontró que la ofensividad era una clase

muy reducida y que no englobaba por completo al lenguaje de los tweets. Por eso, se propusieron tres clases mutuamente no excluyentes como descriptores del mensaje: agresividad, ofensividad y vulgaridad. Los resultados fueron muy satisfactorios, pues se obtuvo un coeficiente de kappa de 92%, el cual es mucho más alto que el inicial.

1.1 Preguntas de investigación

Con base en el proyecto explicado anteriormente, me surgió la idea de que para poder realizar la detección de contenido riesgoso en twitter de manera semi-automática se podrían aplicar las clases y los criterios del diagrama. Estos criterios se usarían no sólo en el etiquetado de corpus, sino también en un sistema de detección que utilice algoritmos de clasificación propios del aprendizaje automático. De allí se derivó la siguiente pregunta de investigación: ¿tomando en cuenta los elementos lingüísticos que se usaron para etiquetar un corpus, se podrán implementar éstos como atributos en un sistema de clasificación semi-automática basado en modelos de aprendizaje automático?

En el workshop del MEX-A3T sólo se usaron dos clases en la detección, agresivo vs. no agresivo, y los resultados obtenidos no fueron altos. Además, el enfoque general no fue lingüístico sino más bien computacional. Por lo tanto, se podrían derivar dos hipótesis de la pregunta antes expuesta.

(1) se pueden clasificar los tweets con mayor granularidad en tres clases (agresividad, ofensividad y vulgaridad) con una precisión significativa.

(2) los elementos lingüísticos propuestos en el diagrama pueden ser usados como atributos para la clasificación en tres categorías y tener igualmente resultados significativos.

Con lo anterior en mente, en esta investigación se usarán como base los criterios lingüísticos propuestos en el diagrama para la facilitación de etiquetado del corpus de tweets recopilado por el INAOE y se extenderá su uso para poder utilizarlos como atributos en un sistema de detección de las tres características del mensaje propuestas en el diagrama (agresividad, ofensividad y vulgaridad). En esta extensión del trabajo previo se tomará en cuenta la variación lingüística y los actos de habla –en particular el ilocutivo. Para la elaboración de este proyecto, se hizo una prueba más reducida en sólo una muestra del corpus del INAOE. Se usaron treinta tweets, un número pequeño que aseguró un análisis más profundo y veloz de cada texto. Debido al tamaño del conjunto de datos, la búsqueda de los atributos se realizó manualmente en acompañamiento del software AntConc, propio de la lingüística de corpus, y para la aplicación de los algoritmos de clasificación basados en aprendizaje automático se usó el software de Weka.

El resto de esta investigación se encuentra organizada de la siguiente manera. En la segunda sección se revisarán aspectos teóricos que permitirán comprender el trasfondo del problema y las bases de la solución planteada, así como se presentarán las tres clases propuestas: ofensividad, agresividad y vulgaridad. Posteriormente se hablará de la variación lingüística explicando el rol de ésta en las anteriores clases y la importancia de estudiarla en conjunto con dichas clases. A continuación, se presentarán los actos de habla para poder comprender lo que los enunciados hacen. Después, se hará un énfasis en el acto ilocutivo, que explicará cómo se usa este concepto para conocer el propósito del emisor al emitir un

enunciado. Seguido de esto se hablará de la lingüística de corpus y en esta parte se explicará la recolección del corpus a utilizar y su etiquetado. La última parte de esta segunda sección será sobre trabajos e investigaciones previas que se relacionen con la presente. De estos trabajos se hará un énfasis en el MEX-A3T, ya que fue la base para la elaboración de este trabajo. Después se presentará el proyecto de elaboración de un diagrama de etiquetado y finalmente el coeficiente de kappa.

En la tercera sección se mostrará la metodología usada para el análisis. Aquí se describirá el corpus usado para los experimentos realizados, el método de selección de datos, la selección de los atributos y las métricas para evaluar la clasificación: exactitud, precisión, recuerdo y valor F. Posteriormente, se expondrá la reducción de atributos y la técnica de aprendizaje supervisado. En esta última parte se incluirán también los n-gramas, que conforman algunos de los atributos utilizados, y los algoritmos de clasificación aplicados: Naïve Bayes y los árboles de decisiones.

En la cuarta sección se encuentran los resultados de los experimentos seguidos de un apartado de discusión donde estará el análisis. En este sentido es importante mencionar que en este trabajo se obtuvieron resultados altamente satisfactorios ya que, en un ejercicio pequeño, pero con datos obtenidos al azar, se logró obtener un resultado que rebasa el estado del arte de esta tarea en experimentos a gran escala. Posterior al análisis de estos resultados, en la quinta sección se ofrecen algunas conclusiones y líneas para posibles trabajos futuros.

2. Marco Teórico

2.1 Ofensividad

Para poder comenzar a hablar acerca del tema que compete es necesario tener primeramente en claro cuáles son las tres características de los mensajes que se van a tratar a lo largo del texto. Estas son agresividad, ofensividad y vulgaridad. Acerca de la ofensividad, Kock, J., Delbecque, N., & Paepe (1998) comentan que este término tiene un significado de emotividad que se relaciona con el desafecto. Este significado, genérico, abarca un haz de sentidos que lo especifican: desprecio, ridiculización, burla, ironía, caricatura, exageración, sátira, descalificación, deformación, e injuria. En este sentido, algo muy importante de este significado y que se relaciona con su efecto emotivo, es la razón por la cual surge. Kock et al. explican que las causas que hacen surgir a la ofensividad pueden ser varias: fealdad, maldad, exceso, poquedad (p.1). Como se puede notar, la ofensividad es subjetiva y tiene diversos significados. Otra definición, un poco más concisa, la provee Jeshion (2013). Este autor menciona que las ofensas funcionan para menospreciar o deshumanizar, es decir, para señalar que sus objetivos no son dignos, de igual importancia o no merecen pleno respeto como personas y que, por tanto, son inferiores. Esta definición enfatiza el fin de herir emocionalmente al referente. El mismo autor propone que a la ofensividad la acompañan los siguientes cuatro atributos (p.232): 1) las ofensas funcionan para menospreciar de la misma manera, 2) las ofensas funcionan para menospreciar en el mismo grado, 3) la capacidad de las ofensas para menospreciar está estrechamente relacionada con la intención del orador de hacerlo, 4) la capacidad del insulto para menospreciar es convencional; es un rasgo de poder de las palabras mismas. Respecto a estos atributos, cabe mencionar que uno de los efectos del insulto es su capacidad de convertir un mensaje en una ofensa, y que esto depende completamente del propósito de quién la dice. El último atributo habla sobre el poder de las palabras. En este sentido, el lenguaje ofensivo más particularmente afecta a diferentes personas de diferentes maneras, ya que algunos miembros de la sociedad evitan activamente

el lenguaje soez y otros lo adoptan abiertamente (Wilson, 2017). La definición de ofensividad es complicada, el tratamiento del término lo muestra de manera consistente. Ávila-Cabrera (2015), comenta que hay una disparidad de acuñaciones a la hora de referirse a este tipo de lenguaje, ya que es considerado por algunos autores como "*dirty language* (Jay, 1980), *strong language* (Lung, 1998; Scandura, 2004), *bad language* (Azzaro, 2005; McEnery, 2006), *foul language* (Azzaro, 2005; Wajnryb, 2005), *rude language* (Hughes, 2006) y *emotionally charged language* (Díaz Cintas y Remael, 2007)" (p.17). Sin embargo, "ofensividad" es un término más amplio y permite englobar distintas características de los términos previamente mencionados.

Retomando la idea anteriormente expuesta acerca de que el efecto de la ofensa en la persona es subjetivo, no resulta buena idea enfocarse en este aspecto para determinar si un comentario es una ofensa o no. En este sentido, como ejemplo se podría tomar lo que comenta Penco (2017), que dice que el uso de términos despectivos es "apropiado" en grupos pequeños. Esto se debe a que las características de la ofensa dependen bastante del contexto en que se están diciendo, por lo que el autor también lo relaciona con el concepto de prejuicio. Este autor explica que una expresión es apropiada si sus presuposiciones son compartidas por los participantes en una conversación. Así, un término peyorativo es perfectamente aceptable en una conversación entre sujetos racistas, porque ellos ciertamente comparten los prejuicios vinculados al término peyorativo. Estos términos por lo tanto connotan los sentimientos con los que se encuentran cargados las ofensas.

Ávila-Cabrera (2015) está de acuerdo también con la contextualización del lenguaje ofensivo y comenta que este "existe en la mayoría de las culturas y la aceptación del mismo viene condicionada por diferentes aspectos tales como el tipo de sociedad, la cultura, las

creencias, etc." (p.17). El mismo autor realizó un estudio acerca del lenguaje ofensivo y el tabú relacionado con su traducción. Como parte de ese trabajo, elaboró una tabla donde expone los elementos que desde su perspectiva conforman al lenguaje ofensivo (p.18). Dicha tabla se encuentra a continuación (la traducción es mía, con excepción de los ejemplos que se han mantenido en la lengua original):

Tabla 1. Taxonomía del lenguaje ofensivo.

Categoría	**Subcategoría**	**Tipos**	**Ejemplos**
Ofensivo	Insultos abusivos.	Maldecir	Goddamn you!
		Tono despectivo	I'm sick of fucking hearing it.
		Insulto	Bunch of shithead.
		Juramento	I swear on my mother's grave.
	Palabrota	Palabrota o frase exclamatoria	Fuck a duck!
	Invectiva	Insulto sútil	Ich habe nicht mit ihnen gesprochen, Obersturmführer München. [I have not spoken to you Lieutenant Munich]

En la Tabla 1, se puede observar cómo la concepción del lenguaje ofensivo puede comúnmente incluir insultos, groserías, términos despectivos e incluso el tono del insulto. En esta tabla están expuestas las características del lenguaje ofensivo en la oralidad. Sin embargo, de forma escrita y en las redes sociales puede que se presente de manera diferente. El hecho de que en las plataformas en línea se tiene la ventaja de mantenerse anónimo y no tener contacto físico con el referente da forma a este tipo de lenguaje de una manera específica. El uso de lenguaje ofensivo es un problema común de comportamiento abusivo en las redes sociales en línea (Santos, Melnyk, & Padhi, 2018). Sin embargo, en estas plataformas se

presentan nuevos retos para detectar este tipo de lenguaje. Aunque hay evidencia de que motivos similares dan forma a comportamientos diferentes en contextos fuera de línea, no hay evidencia empírica comparable para los contextos en línea, y las normas en línea a veces difieren de las normas fuera de línea (Fichman & Sanfilippo, 2015). Fichman & Sanfilippo, explican que a estas cuestiones también se les suma la problemática de los atributos contextuales únicos de las comunidades en línea, tales como el anonimato y la desinhibición, el aumento de la conciencia de la comunidad en línea y la atención a la desviación (p.164).

Con base en las anteriores definiciones, la que se tomará en cuenta para este trabajo de manera más concisa es que el lenguaje ofensivo es aquel que busca insultar o humillar a un grupo o individuo, usualmente utilizando términos peyorativos o despectivos. Por eso, se seleccionaron los términos peyorativos como una forma de identificar el lenguaje ofensivo ya que es uno de los elementos de primordiales del mismo. Por ejemplo, Ávila-Cabrera, (2015) explica que los términos ofensivos hacen referencia a aquellas groserías, exclamaciones soeces, etc. y que se consideran peyorativas e insultantes. Así mismo, Kock, Delbecque, & Paepe (1998) también comentan que los peyorativos, junto con los aumentativos y diminutivos, forman el grupo de derivados llamados emotivos (también apreciativos, evaluativos), en oposición al resto, llamados referenciales (o también: apreciativos, significativos). Esto último significa que los peyorativos pertenecen a expresiones con una carga negativa, la misma de las ofensas. Un elemento más que se tomará en cuenta como parte del término ofensividad, son lo intensificadores. Estos cambian el sentido de una palabra que podría ser neutra a volverse ofensiva. Esto se puede ver en el siguiente tweet:

Chingas a tu madre pinche gordo Edwin Cardona... #DaleRiver 

La primera expresión “chingas a tu madre” es en sí ofensiva en el contexto mexicano. En este caso, el adjetivo gordo por sí solo funcionaría solamente como un descriptor. Sin embargo, acompañado del intensificador negativo “pinche” hace que se vuelva una ofensa. En este tema, Arce Castillo (1999) comenta que la intensificación supone, en general, un énfasis en la cuantificación de un término, conseguido gracias a la sustitución de los cuantificadores habituales mucho, muy, tan, tanto, etc., por fórmulas más expresivas (p.38).

2.2 Agresividad

Aunque la agresividad puede relacionarse con el concepto de ofensividad, una separación debe hacerse debido a que la primera conlleva un efecto diferente en el receptor. Por lo anterior, es importante tener en claro en qué consiste el término agresividad. En general, las formas de agresión son variadas y van desde expresiones de repugnancia y desprecio, hasta tratos, calumnias, insultos y odio. Ya que se está analizando este concepto en Twitter, se debe ver más específicamente en el ámbito específico de las redes sociales. Allí, si la agresión es aprobada por otros usuarios, puede escalar y provocar una "tormenta de fuego en línea", que se describe como una ola de comentarios negativos y furiosos en los medios sociales (Rösner, & Krämer, 2016). En este sentido, la agresividad tiene una connotación violenta que se realiza inicialmente con odio y termina creando más.

Igual que la ofensividad, la agresividad tiene lugar en las redes de manera muy común debido a la forma en la que estas plataformas están creadas. Rösner et al. (2016), explican que una de las razones principales de la agresión en línea se atribuye al anonimato en internet. Esto se remonta a la teoría de la desindividuación, que afirma que las personas pierden sus limitaciones internas y se sienten menos conscientes, inhibidas y responsables de su comportamiento cuando son anónimas (p.1). En este sentido, los mismos autores agregan

que los usuarios experimentan agresión verbal y comportamiento incivilizado incluso en plataformas web menos anónimas como los sitios de redes sociales (tales como lo son Facebook, Youtube y Twitter), donde la mayoría de las personas están registradas por su nombre real y comparten información personal. En estas plataformas lo autores explican que, de acuerdo con las teorías de influencia social, los individuos afectan las opiniones y conductas de los demás en el contexto social y tienden a conformarse con las normas sociales prevalentes de un grupo social común, sobre todo si se identifican con este grupo. De las características expuestas por los autores, los elementos que destacan en la caracterización de la agresividad son la presencia de odio y el objetivo de afectar la conducta de otros con los enunciados, ya que son características claras en estos mensajes.

Respecto a la plataforma que se analizará específicamente en este trabajo, un estudio elaborado por el Instituto de las Mujeres del Distrito Federal (2016), menciona que Twitter es la principal plataforma para los hashtags y campañas de odio (con hashtags que lo fomentan, como #Putipobreza, #Putiesposa, #Pobrezafilia). Twitter también es la red que cuenta con el mayor contenido pornográfico comparado con Facebook. Lo anterior se debe a las características de la plataforma. Twitter es un espacio abierto donde el intercambio es continuo y en tiempo real, mediante mensajes cortos de 140 caracteres. La información y mensajes se comparten de manera global, por lo que la mayoría de los perfiles son públicos. Al ser una red principalmente abierta, cuenta con políticas de privacidad menos restringidas que facilitan el anonimato de sus usuarios; además de que los perfiles en Twitter son más impersonales (p.25). Esto quiere decir que, en comparación con Facebook o Instagram, donde la información que se publica es sobre la persona que lo publica y muchas veces con sus datos personales, en Twitter muy pocas veces ese es el caso.

El tipo de contenido agresivo que se encuentra en Twitter, así como en otras redes sociales, se puede relacionar con el *hate speech*. Marciani Burgos (2013) explica que el llamado *hate speech* involucra expresiones ofensivas dirigidas contra ciertos grupos, principalmente identificados como minorías tradicionalmente excluidas por motivos de género, orientación sexual, raza, religión o alguna otra situación similar. De este tipo de agresividad, una problemática bastante grande y que dificulta su detección, es que las expresiones están dirigidas contra grupos y no contra individuos de forma particular. Por esto, no pueden subsumirse dentro de las figuras de la difamación, la calumnia o la injuria y como resultado, son más complicadas de detectar. Otra definición proporcionada por Davidson, Warmsley Macy, & Weber (2017) ayuda a ilustrar mejor este tipo de agresión. Los autores comentan que este término se refiere al lenguaje que se utiliza para expresar el odio hacia un grupo objetivo o que tiene la intención de ser despectivo, humillante o insultar a los miembros del grupo. En casos extremos, también puede tratarse de un lenguaje que amenace o incite a la violencia.

Davidson et al. (2017), mencionan que la diferencia entre *hate speech* y otro lenguaje ofensivo se basa a menudo en sutiles distinciones lingüísticas. Por ejemplo, los tweets que contienen la palabra *n*gger* tienen más probabilidades de ser etiquetados como incitación al odio que *n*gga*. Muchos pueden ser ambiguos, por ejemplo, la palabra "gay" puede usarse tanto peyorativamente como en otros contextos no relacionados con la incitación al odio. Pero ambos ejemplos se asocian directamente con la intención y el contexto. De acuerdo con los mismos autores, también se han aprovechado las características sintácticas para identificar mejor los objetivos y la intensidad de la incitación al odio. Por ejemplo, estos autores han utilizado frases en las que aparecen un sustantivo y un verbo relevantes (como

“matar” y “judíos”), el trigrama con categorías gramaticales combinadas "DT *jewish* NN", y la estructura sintáctica I <intensidad > <intención del usuario > <objetivo de odio >, (como en *I f*cking hate white people*). Por último, este estudio también menciona que los rasgos no lingüísticos como el género o la etnia del autor pueden ayudar a mejorar la clasificación de la incitación al odio, pero esta información a menudo no está disponible o no es fiable en los medios de comunicación social. Esta última información es bastante difícil de conseguir, debido a lo que ya se mencionó anteriormente acerca de la impersonalización de Twitter. Con estos ejemplos entonces se puede inferir que hay ciertos temas relacionados con discurso agresivo, tales como el racismo, la homofobia, la misoginia, entre otros.

Sobre estos temas, el racismo se define como “un fenómeno de exclusión, segregación e inferiorización” (Gall, O., 2004, p.235). Debido a que el racismo se enfoca en minorías, Munger (2017) comenta que las minorías y otras poblaciones vulnerables suelen ser objeto de acoso en línea en los sitios de redes sociales, a menudo en respuesta a la expresión de opiniones con las que los acosadores no están de acuerdo. El racismo, en muchos sentidos, es una extensión de los estereotipos y prejuicios, como la definición propuesta por Leone, donde el racismo es descrito como la creencia en la superioridad inherente de una raza en particular. Según este autor, el racismo niega la igualdad básica de la humanidad y correlaciona la capacidad con la composición física. Por lo tanto, asume que el éxito o el fracaso en cualquier esfuerzo social dependerá de la dotación genética y no del medio ambiente y el acceso a las oportunidades (Poliakov, 1996, citado en Samovar, Porter, & McDaniel, 2015). Así mismo, el lenguaje racista generalmente se identifica con relativa facilidad por el empleo extensivo de estereotipos denigrantes, o bien se apoya en mecanismos retóricos accesibles a otras formas de lenguaje discriminatorio como los contrarios:

nosotros/ellos que anteponen dos mundos adversos mediante el uso de dos conjuntos de palabras (Islas Asaïs, 2005).

Así como el racismo es un acto de violencia que se ayuda del lenguaje, el lenguaje sexista también tiene el mismo objetivo final sobre el referente. Islas Asaïs (2005) explica que el lenguaje sexista es el lenguaje que fomenta la discriminación de género contra las mujeres. El mismo autor también explica que el lenguaje sexista ha fomentado, "con el empleo de estereotipos insidiosos y asimetrías semánticas y sintácticas, una imagen de la mujer que desestima su contribución a la sociedad e incluso su presencia misma en ciertas áreas" (p.29). Esto último demuestra lo problemático de esta situación y por lo tanto su necesidad de caracterizarlo y así lograr evitarlo. Por ello, usando las definiciones anteriores de agresividad, la que se usará específicamente para su caracterización en este trabajo es la siguiente: el lenguaje agresivo es el que se usa con el fin de dañar o herir a un grupo o individuo aludiendo o incitando a la violencia. Dicha definición se puede ilustrar con los siguientes dos tweets provenientes del conjunto de prueba del corpus del INAOE.

(1) Apoco las putas sienten? Yo pensaba y tenía la terminación de q eran un trozo de carne 🍡 y q solo ve...

(2) Está Putita tenemos su Pack completo 30 Rt y la quemamos 💞💞💞😎

En estos ejemplos, se pueden ver términos peyorativos comunes para las mujeres como lo es la palabra "puta" y sus variantes. Así también como la temática de los tweets y la presencia de otra palabra relevante de notar: Pack. Este léxico presupone en sí una falta de respeto y un afán por incitar violencia contra el grupo aludido.

2.3 Vulgaridad

Una definición general de vulgaridad podría ser la siguiente: un vulgarismo es una palabra obscena, sucia, desagradable y/o indecente. Sin embargo, la aparición de una forma vulgar, lingüística y de un contenido obsceno puede estar relacionada con varias causas. También se relaciona con la cultura, la política, las analogías y la creatividad del lenguaje en términos de materia social (Kusumaningsih, Santosa, Subroto, & Djatmika, 2019). Sin embargo, este término resulta más complicado de describir ya que puede tener distintas connotaciones. Por ello, es importante conocer que este término tiene una raíz en la variación lingüística, en la forma en la que cada uno de los hablantes hacen uso del lenguaje con respecto a su contexto. Rico Sulayes (2014) comenta que el término vulgar corresponde a una variación lingüística de orden social, que pertenece a un nivel de lengua bajo, al margen de la norma estándar. "El término vulgar para Briz coincide con la variedad no estándar, pero a diferencia de ésta, representa usos incorrectos o fuera de la norma" (p.70). En este mismo tema, Rico Sulayes citando a Haensch (1982), explica que la parte del léxico que corresponde al léxico tabuizado o malsonante no constituye un nivel de lengua o una variedad lingüística en sí misma, sino más bien forma parte del léxico vulgar que se sitúa como una variación social. En esta variación, que será mejor explicada más adelante, se puede situar a la vulgaridad como perteneciente a las clases medias y bajas. Holmes (2013) explica que las personas que trabajan con salarios más bajos utilizan este vocabulario con más frecuencia, y claramente funciona como un marcador de solidaridad para este grupo. Sin embargo, en la actualidad es considerado vulgar por muchos de los hablantes de mayor edad, por lo que puede permanecer dentro de los niveles sociales más bajos. Alternativamente, puede extenderse al discurso informal de los jóvenes de los grupos sociales más altos, y así gradualmente se extiende hacia

arriba de esta manera. Con esto se puede observar en dónde es que sucede este tipo de lenguaje y esto permite saber sus características más fácilmente.

Una vez comprendido lo arraigado que está este concepto a la cultura de los hablantes se entiende la naturaleza de este lenguaje. Debido a esta variación del lenguaje y su dependencia con el contexto social, Rico Sulayes también comenta que resulta un problema saber sus características específicas. Sin embargo, normalmente se relacionan las palabras mal sonantes o de connotación sexual con este término. En este sentido, el autor aclara que respecto al problema en la localización del léxico vulgar y el tabuizado, resulta útil la medida que toma Cheshire en su estudio del habla vernácula e informal de los adolescentes (1997). En este estudio se consideran las groserías como una categoría de estudio, pero las separa de las características de la lengua vernácula y las coloca en una categoría aparte. La separación funciona para lograr ser más detallado en estos distintos elementos. Sin embargo, se relacionan y por ello ambos elementos se explicarán junto con el concepto de vulgaridad por motivos de practicidad.

Debido a lo anterior, uno de los elementos que podría identificar este tipo de leguaje es el uso de palabras malsonantes. Estas palabras, de acuerdo con Khalaf & Rashid, (2019) representan un reflejo de la manipulación del lenguaje en diferentes esferas de la vida por diferentes personas para lograr ciertas funciones pragmáticas. Entre las diversas funciones de las groserías, se incluyen expresiones de enojo, frustración, molestia, sorpresa, poder, solidaridad, pertenencia a un grupo y felicidad, que van más allá del significado literal de tales palabras (Andersson y Trudgill 1990 citado en Khalaf & Rashid). Aunque se podría considerar que las palabras malsonantes son solamente vulgares, en realidad transmiten diversos sentimientos y transfieren emociones complejas. Hughes (1991) comenta que el uso

de groserías se basa en resonadores tan poderosos e incongruentes como la religión, el sexo, la locura y la nacionalidad. Estos abarcan una extraordinaria variedad de actitudes, incluyendo las violentas, las divertidas, las escandalosas, las absurdas, las casuales y las imposibles. Por esto, su uso representa muchas ideas y creencias del hablante que permiten conocer más a fondo las características de su lenguaje.

Hughes (1991) explica que en vista de que se tiene la común percepción de que el lenguaje se genera en una dispensación "patriarcal" o "falocrática", se ha desarrollado, especialmente en los insultos masculinos, una prevalencia de los términos de la anatomía femenina. Es en este sentido que surge la concepción de la vulgaridad desde el lado de las palabras con connotación sexual, que en la actualidad no se reservan solamente en partes femeninas expresadas por hombres, sino también de mujeres y haciendo referencia a partes del cuerpo masculinas. Por lo tanto, otro elemento más es el uso de palabras con referencia a situaciones sexuales, que se les podría llamar indecentes. En este sentido, la indecencia se puede definir como lenguaje que describe, en términos claramente ofensivos medidos por los estándares de la comunidad contemporánea para el medio de transmisión, actividades u órganos sexuales o excretorios (Kaye, & Sapolsky, 2009).

Finalmente, tomando en cuenta las concepciones generales de la vulgaridad, se llegó a una conclusión acerca de este término. Por lo tanto, es el que será usado para la investigación y es el siguiente: el lenguaje vulgar involucra expresiones soeces, con connotación sexual y en algunas ocasiones con doble sentido, pero que puede o no referirse a un individuo o colectividad. Para ilustrar esta definición, los siguientes ejemplos del corpus ayudan a comprenderla.

(3) Que gorda verga Papi me encanta y a mamar se ha dicho

(4) Qué rico me da mi marido Rt y Fv si les gusta mi culo yo digo que me veo muy gorda ustedes que opinan????

Con estos ejemplos se observa el uso de palabras vulgares, como “culo” y “verga” que hacen referencia a partes del cuerpo y a verbos que tratan con el acto de fornicación. Así, la connotación sexual queda mejor ilustrada.

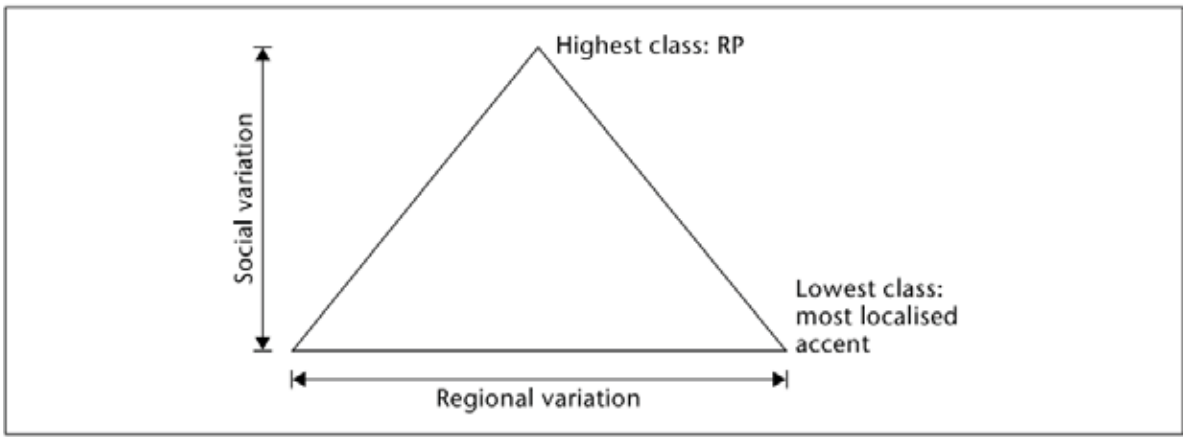
2.4 Variación lingüística

Como se puede observar, en los temas tratados anteriormente y tomando en cuenta las características del lenguaje empleado en los textos, este tipo de vocabulario no forma parte del “estándar”. Este tipo de vocabulario se asemeja más a un subestándar. Muchas expresiones forman parte de un grupo específico de la población de hablantes de español mexicano, lo que lo vuelve una variación del español. Debido a lo anterior, es relevante considerar estas variaciones para lograr una mejor comprensión del lenguaje usado y las posibles intensiones del autor y a esto se le llama variación lingüística.

La variación lingüística trata del estudio de los elementos lingüísticos que caracterizan los cambios en el uso de la lengua en general (Rico Sulayes, 2014). Estos tipos de cambios engloban distintos aspectos del lenguaje. Holmes (2013) explica que este vocabulario o la elección de palabras es un área de la variación lingüística. Pero la variación lingüística también ocurre en otros niveles del análisis lingüístico: sonidos, estructura de palabras (o morfología) y gramática (o sintaxis). Dentro de cada uno de estos niveles lingüísticos existe una variación que ofrece al hablante una variedad de formas de expresión.

Estos niveles nos proporcionan diferentes estilos lingüísticos para su uso en diferentes contextos sociales. Las opciones de variación pueden incluso involucrar diferentes dialectos de un idioma, o idiomas muy diferentes. Por esto es importante considerar esta variación cuando se trata del español en México, ya que el enfoque tiene que involucrar estos cambios. En estos cambios, también resulta de utilidad ubicar el contexto de los hablantes. Holmes comenta que la mayor parte de la variación lingüística se encuentra en el nivel socioeconómico más bajo, donde abundan las diferencias regionales. Más arriba en la escala social, la cantidad de variación observable se reduce hasta llegar a la cúspide de la variación del lenguaje. Esto se encuentra ejemplificado en la siguiente figura tomada de Holmes (p.139).

Figura 1. Variación de acento social y regional



En la Figura 1 se usa como ejemplo el inglés RP (*Really Posh*), que también se conoce como "*Queen's English*". Por lo que en la parte baja se encuentran las variaciones regionales acompañadas de las clases bajas y en la parte superior se encuentra el inglés estándar que también está acompañado de un incremento de la clase social. Por ello, el registro es un tema

que se debe analizar. Ávila-Cabrera (2015), explica que, al hablar de registro, nos referimos a "*a particular choice of diction or vocabulary regarded as appropriate for a certain topic or social situation*" (Hughes, 2006: 386 citado en Ávila-Cabrera). Es decir, el registro representa un ajuste de vocabulario dependiendo del contexto en el que se encuentre el hablante. Por ello, en las redes sociales, sobre todo en Twitter, no habría una gran necesidad de un ajuste que reprima el habla de los usuarios y estos se acercan más a su manera más natural de hablar. Sin embargo, se debe tener en cuenta que la situación del español en México es un poco diferente en comparación con el ejemplo expuesto en la figura anterior. En este país no existe en sí un lenguaje estándar respaldado por alguna institución mexicana equivalente a la realeza británica. No obstante, es claro que existe un modelo a seguir. Este lenguaje ideal es el que se usa en contextos académicos, se enseña en las escuelas y puede ser claramente observado, por dar un ejemplo, en la redacción de noticieros.

2.5 Actos del habla

Tomando en cuenta que el enfoque de esta investigación es analizar la intención del emisor y las características del lenguaje que le ayudan a comunicar esta, es importante saber cómo funcionan las intenciones de los enunciados y de qué manera englobarlas. Para esto, sirven como base los actos de habla. Grundy, (2008) menciona que los enunciados hacen mucho más de lo que transmiten las oraciones de manera literal y es relevante conocer lo que está más allá de las palabras, por ello el mismo autor comenta que conocer el significado literal de las oraciones no es suficiente para determinar qué es lo que se considera hacer, qué acto de habla se está realizando y cuándo se utiliza.

En este sentido, Austin (1962) propone que la producción de un enunciado realiza tres tipos de acciones: locutiva, ilocutiva y perlocutiva. La primera es la expresión lingüística por sí misma, su estructura sintáctica y su significado semántico literal. La ilocutiva es la fuerza o la acción del enunciado pretendido por el hablante. Por último, la perlocutiva es la reacción del destinatario al enunciado. El enfoque en el análisis pragmático es el segundo, el acto ilocutivo. Esto se debe a que este último se enfoca en lo que el emisor desea conseguir con el enunciado. Por lo tanto, tomar al acto ilocutivo como objeto de estudio en la detección de agresividad es una buena manera de conseguir resultados más precisos.

Sidorov (2013) comenta que, en investigaciones pertenecientes a la lingüística computacional, se ha empleado el uso de los actos de habla para la comunicación con los sistemas automáticos. En este sentido, el autor comenta que lo más común es que primero el robot reconoce la voz de la frase ingresada (realmente lo hace la computadora, es decir, el software de reconocimiento de voz). Después se realiza el procesamiento de la frase utilizando las técnicas tradicionales de la lingüística computacional (análisis morfológico, sintáctico, hasta llegar al análisis semántico), se analiza la frase en el contexto del diálogo, se determina el acto de habla de cada frase (su interpretación pragmática), se prepara una respuesta que corresponde a este acto de habla, se genera la frase de la respuesta y se pronuncia la respuesta en caso necesario. Finalmente, se realiza la acción en caso que la frase corresponda a un acto de habla que requiera alguna acción. Esto sucede en el caso de los sistemas de reconocimiento de voz más avanzados y que tienen mejor rendimiento. Por esto, valdría la pena intentar usar este enfoque, en especial el del acto ilocutivo, en la detección de ofensividad, agresividad y vulgaridad.

2.5.1 Actos ilocutivos

Como se mencionó en el anterior apartado, el acto de habla ilocutivo es de especial interés para el tipo de tareas como la que se tratan en esta investigación. Por ello, se describirá un poco más a fondo. Searle (1976) explica que en la realización de un acto ilocucionario el hablante intenta producir un cierto efecto, haciendo que el oyente reconozca su intención de producir ese efecto (acto perlocutivo).

A modo de poder catalogar las distintas maneras en las que se usa el lenguaje, Searle intentando resolver los problemas de las categorías básicas de Austin, propone su propia taxonomía. Esta consiste en los siguientes actos; asertivos, directivos, compromisorios, expresivos y declarativos. Dicha taxonomía será explicada a continuación.

a) Representativos: estos actos comprometen al orador a la verdad de una proposición expresada. Representan la creencia del orador de que algo puede ser evaluado como verdadero o falso.
b) Directivos: El punto ilocutorio de estos actos consiste en el hecho de que son intentos (de varios grados) por parte del orador para conseguir que el oyente haga algo. Por ejemplo, el orador puede hacer una pregunta, hacer una solicitud o emitir una invitación. En este sentido, esto se puede relacionar con la clase de agresividad, ya que obliga al referente a cambiar su comportamiento.
c) Compromisorios: aquí Searle comenta que la definición de Austin es suficiente y no requiere grandes cambios. El acto compromisorio es cuando el hablante se compromete a un curso de acción futuro. En la conversación, los actos más comunes son promesas y amenazas. Sin embargo, Searle expresa que los verbos que Austin enumera como comisivos no pertenecen en absoluto a esta clase. Dichos verbos son

"shall", "intend", "favor", entre otros. Estos más bien tienen como objetivo comprometer al orador (de nuevo en diferentes grados) a tomar alguna medida en el futuro. Aquí esto se puede relacionar con la parte violenta de la agresión, ya que habla de acciones futuras en relación con el objetivo del enunciado.

d) Expresivos: expresan el estado psicológico especificado en la condición de que hay sinceridad sobre un estado de cosas especificado en el contenido de la propuesta. Los paradigmas de los verbos expresivos son 'agradecer', 'felicitar', 'disculparse', 'condolerse', 'deplorar' y 'dar la bienvenida'. Conocer el estado emocional del emisor en su enunciado podría facilitar el diferenciar su mensaje entre ofensivo o no ofensivo.

e) Declarativos: estos actos son cambios que se producen a través de los enunciados. Produce la correspondencia entre el contenido proposicional y la realidad. El desempeño exitoso garantiza que el contenido proposicional corresponda al mundo. Como un jefe cuando nombra a un empleado gerente de un área tras decir las palabras “Desde hoy eres gerente”. Este cambio se ve reflejado en acciones del mundo real.

2.6 Lingüística de corpus

Debido a que este proyecto se basa en un corpus, es importante conocer su definición. Un corpus es una extensa colección de textos auténticos, a diferencia de los textos "ya elaborados"; normalmente están en formato electrónico, lo que permite enriquecerlos a

medida que se avanza, y responden a un conjunto específico de criterios en función de los objetivos de la investigación en cuestión (Bermúdez Bausela, 2016). El uso de corpus está más estrechamente relacionado con el campo y el método de la lingüística de corpus. Esta tiene un compromiso de utilizar datos lingüísticos empíricos "reales" para la investigación lingüística (Dobrić, 2016). McEnery & Hardie, (2011) comentan que la lingüística de corpus no trata directamente el estudio de algún aspecto particular del lenguaje. Más bien, es un área que se centra en un conjunto de procedimientos, o métodos, para el estudio de la lengua. Estos mismos autores definen a la lingüística de corpus como un conjunto de textos legibles por máquinas que se considera una base adecuada para estudiar un conjunto específico de cuestiones de investigación. Así, el conjunto de textos o corpus tratados es generalmente de un tamaño que desafía el análisis manual y visual, realizable dentro de un marco de tiempo razonable. En este sentido, los mismos autores comentan que surgen varias generalizaciones respecto a lo anterior. Una de ellas es que los corpus se explotan invariablemente utilizando herramientas que permiten a los usuarios buscar a través de ellos de forma rápida y fiable. Algunas de estas herramientas, como los concordantes, permiten a los usuarios ver las palabras en su contexto. La mayoría de estas herramientas también permiten la producción de datos de frecuencia de alguna descripción, por ejemplo, una lista de frecuencias de palabras, que enumera todas las palabras que aparecen en un corpus y especifica para cada palabra cuántas veces ocurre en ese corpus. Así, se permite caracterizar un texto por las palabras más usadas en él.

Por otro lado, McEnery & Hardie (2011) comentan que respecto al tipo de datos que se manejan en el corpus, típicamente en la lingüística de corpus los datos son textuales, de modo que cada archivo representa, por ejemplo, un artículo de periódico o una transcripción

ortográfica de alguna lengua hablada. Sin embargo, los archivos digitales dentro de un corpus no necesitan ser textuales, y hay ejemplos hoy en día de archivos de datos de video que se utilizan como textos de corpus. Los mismos autores remarcan las concepciones de la lingüística de corpus como un método o como disciplina. En esta diferencia, comentan que se relaciona con *Corpus-based* y *corpus-driven linguistics*. Los estudios *corpus-based* suelen utilizar los datos del corpus para explorar una teoría o hipótesis, normalmente la establecida en la literatura actual, con el fin de validarla, refutarla o perfeccionarla. La definición de la lingüística de corpus como método sustenta este enfoque del uso de los datos de corpus en la lingüística. La lingüística del *corpus-driven* rechaza la caracterización de la lingüística del corpus como método y se dice en cambio, que el propio corpus debería ser la única fuente de las hipótesis sobre la lengua. Así pues, se afirma que el corpus encarna su propia teoría del lenguaje. Una vez entendido el concepto general de la lingüística de corpus y sus dos concepciones respecto a sus aplicaciones en la investigación, surge entonces la necesidad de saber cómo se elabora un corpus que cumpla con las necesidades de los objetivos de la investigación o del investigador.

2.6.1 Recolección de un corpus.

Debido a lo explicado en la última sección, la construcción de corpus, y en particular la recopilación de datos surge como un tema crítico para la lingüística de corpus. En este sentido McEnery & Hardie (2011) explican que han surgido dos enfoques amplios para la elección de los datos que se deben recopilar según los límites establecidos en la investigación: el enfoque de corpus de monitoreo (monitor corpus), en el que el corpus se amplía continuamente para incluir más y más textos con el paso del tiempo; y el enfoque de corpus equilibrado o corpus de muestra (*balanced/sample corpus*), en el que se construye un corpus

de muestra cuidadoso, que refleja el lenguaje tal y como existe en un momento dado en el tiempo de acuerdo con un marco de muestreo específico.

Torruella & Listerri (1999) mencionan que lo primero que se debe de hacer es definir la finalidad concreta del corpus, ya que este punto va a condicionar todos los demás. El segundo paso es establecer los límites temporales, geográficos y/o lingüísticos que va a tener. Para esto lo autores proponen establecer fechas de inicio y término, definir las lenguas que se van a incluir y el área geográfica que se va a utilizar. El siguiente paso es determinar el tipo de corpus que se va a utilizar. Para lo anterior Torruella & Listerri proponen definir la cantidad de texto que se tome de cada documento, la codificación, anotaciones que se le hacen y la documentación que lo acompañe. El siguiente paso que los autores proponen es definir las proporciones de los diferentes grupos temáticos de los corpus. Esto se refiere, a por ejemplo en el caso del corpus elaborado por el INAOE, sobre las profesiones y procedencia de los autores de tweets. Aquí hay clases a las que pertenecen los autores, tales como administradores, artistas, estudiantes, entre otros. Torruella & Listerri explican que esto es de suma importancia, pero es complicado definir las proporciones. El siguiente paso es establecer la población y muestra. Se refiere a establecer de dónde se extraerán los datos, de qué personas y se recomienda entonces aplicar las formas de extracción de muestras, porque es muy complejo delimitar el total de una población y que sea significativo. El siguiente paso es delimitar el número y longitud de los textos de la muestra para evitar un corpus demasiado desequilibrado.

2.6.2 Etiquetado de corpus.

La anotación en el corpus es la práctica de añadir información lingüística interpretativa a un corpus. Un ejemplo común de anotación es añadir *tags*, o etiquetas, que indican la clase de palabra o características preseleccionadas y particulares de cada texto (Leech, 2005). En este sentido, el corpus del INAOE tiene este tipo de anotación ya que el conjunto de prueba viene con etiquetas binarias que especifican si hay o no agresividad en el tweet.

Sin embargo, también hay corpus sin etiquetas. Algunos investigadores afirman que el corpus no anotado es el corpus "puro". Esto, debido a que quieren investigar el corpus sin alteraciones con información que es sospechosa, posiblemente reflejando las predilecciones, o incluso los errores, del anotador. En este sentido, los posibles usos del corpus son bastante más amplios ya que no se limita al uso de las etiquetas establecidas por el anotador. Para otros, la anotación es un medio para hacer un corpus mucho más útil y representa un enriquecimiento del corpus original en bruto (Leech, 2005). Leech comenta que esto se debe a que un corpus que ha sido anotado de antemano ayudará en diversos tipos de procesamiento o análisis automáticos, ya que, a comparación de un ser humano, la computadora no puede hacer un análisis tan extenso al mismo nivel. Leech propone que para evitar complejidad innecesaria en el sistema, un etiquetado binario convierte la tarea en efectiva e ideal para los sistemas automáticos. El mismo autor comenta que cualquier tipo de anotación presupone una tipología -un sistema de clasificación- para los fenómenos que se representan. Pero la lingüística, como la mayoría de las disciplinas académicas, lamentablemente carece de acuerdo sobre las categorías que se utilizarán en dicha descripción. Por ello, abundan las diferentes terminologías. Es por esto que el proceso de etiquetado es tan complicado para quienes lo realizan y requiere de criterios específicos para facilitar este proceso, tanto para lingüistas, como para cualquier individuo que deba etiquetar, ya que no siempre se tiene la suerte de tener expertos en el lenguaje etiquetando corpus.

2.7 Trabajos anteriores.

En el tema de la detección de agresividad u ofensividad, existen diversos trabajos que han abordado esta problemática. La mayoría de ellos se realizó para el inglés, como es el caso de SemEval (International Workshop on Semantic Evaluation). Aquí se propusieron diversas tareas. Una de las tareas fue: *Identifying and Categorizing Offensive Language in Social Media* (OffensEval) (Zampieri et al., 2019). Como parte de esta tarea, se definieron tres subtareas, correspondientes a los tres niveles de su esquema de anotación que se explican a continuación:

a) Identificación de lenguaje ofensivo (104 equipos participantes)
b) Clasificación automática de los tipos de ofensas (71 equipos participantes)
c) Identificación del blanco de la ofensa (66 equipos participantes)

Para poder evaluar las tres subtareas se usó la puntuación F1 macro-promediada. En los resultados de la primera subtarea, el equipo con más alto rendimiento, NULI, utilizó como base BERT con parámetros predeterminados, pero con una longitud máxima de 64 frases y entrenado para 2 fases. Con este método obtuvieron una puntuación de 82,9% en la F1. En la subtarea (b), el mejor equipo, jhan014, utilizó un enfoque basado en reglas con un filtro de palabras clave basado en una lista de comportamiento en lenguaje de Twitter. Esta lista incluía cadenas como hashtags, signos, etc., logrando una puntuación F1 de 75,5%. Por último, en la subtarea (c) el mejor equipo, vradivchev anikolov, utilizó BERT después de probar muchos otros métodos de aprendizaje profundo. También utilizaron pre-procesamiento e incrustaciones de texto pre-entrenadas basadas en GloVe (Zampieri et al.,

2019). Como se puede observar, los resultados fueron bastante altos estaban acompañados de un diseño bastante complejo.

Un ejemplo más donde se aborda una tarea similar, sucedió en el GermEval 2018, que incluyó dos subtareas de identificación de lenguaje ofensivo. GermEval es una serie de campañas de evaluación de tareas compartidas que se centran en el Procesamiento del Lenguaje Natural para el idioma alemán. Esta tarea compartida consiste en iniciar y fomentar la investigación sobre la identificación de contenidos ofensivos en los microposts de lengua alemana. Los comentarios ofensivos deben ser detectados a partir de una serie de tweets alemanes. El taller en el que se discutió esta tarea se celebró conjuntamente con la Conferencia sobre Procesamiento del Lenguaje Natural KONVENS (Konvens, 2019). En esta parte de identificación de lenguaje ofensivo, hubo dos subtareas. La tarea 1 consistía en una clasificación binaria de poca granularidad, en donde se debía decidir si un tweet incluye alguna forma de lenguaje ofensivo o no. Los tweets tenían que ser clasificados en las dos clases: *OFFENSE* y *OTHER*. La tarea 2, de clasificación de 4 vías de alta granularidad, involucró cuatro categorías. Una clase OTRA no ofensiva y tres subcategorías de lo que es OFENSA en la Tarea 1. Estas subcategorías fueron: *insult*, *profanity* y *abuse*. En *profanity* se utilizan palabras profanas, sin embargo, el tweet no quiere insultar a nadie. En el de *insult* el tweet claramente quiere ofender a alguien. INSULT es la atribución de cualidades o deficiencias evaluadas negativamente o el etiquetado de las personas como indignas (en cierto sentido) o no valoradas. En la última clase de *abuse*, el tweet no sólo insulta a una persona, sino que representa la forma más fuerte de lenguaje abusivo. Por abuso se definió un tipo especial de degradación. Este tipo de degradación consiste en atribuir una identidad social a una persona que es juzgada negativamente por una mayoría (percibida) de la sociedad

(Ruppenhofer, Siegel, Wiegan, 2018). Esta es una clasificación bastante interesante del lenguaje ofensivo, ya que abarca las diferentes connotaciones que tiene y también permite una mejor caracterización que podría facilitar la clasificación automática de los equipos. Para poder evaluar este rendimiento se usaron las medidas comunes de evaluación: precisión, recuperación y valor F1. Los puntajes generales de desempeño alcanzados en la Tarea 2 son considerablemente más bajos que en la Tarea 1. Esto no es una sorpresa, ya que la Tarea 2 es considerablemente más difícil, al tener 4 en lugar de 2 clases. En la 1 el puntaje más alto fue de 76.77% y en promedio hubo un 66.35%. De la tarea dos el más alto fue 52.71% y en general de todos los equipos se obtuvo un promedio de 39.71%. Al ver los resultados se puede inferir que en realidad el incremento de clases no facilitó la clasificación. A pesar de que se pensaba que podría lograr un mejor entendimiento de los mensajes, no hubo buenos resultados.

Un trabajo realizado en español, similar al explicado anteriormente es el IberEval. Este tiene como objetivo fomentar y promover el desarrollo de las Tecnologías del Lenguaje Humano (HLT) para las lenguas ibéricas (castellano, portugués, catalán, vasco y gallego), mediante la creación de una serie de evaluaciones y un foro de discusión sobre los sistemas de Procesamiento del Lenguaje Natural. En el marco del IberEval 2018 (IberEval, 2018) se llevaron a cabo las siguientes tareas: 1) análisis de autoría y agresividad en Twitter: estudio de caso en español mexicano (MEXA3T), 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), y 6) detección de posición multimodal en

tweets sobre el referéndum catalán (MultiStanceCat) (Guzmán Falcón, 2018). De aquí, el trabajo que se relaciona intrínsecamente con esta tesis es la primera tarea, el MEX-A3T.

2.7.1 MEX-A3T

El objetivo del MEX-A3T es mejorar aún más la investigación en el análisis de autoría y agresividad, que son importantes tareas del Procesamiento del Lenguaje Natural, así como seguir impulsando el tratamiento computacional del español mexicano. El MEX-A3T@IberLEF2019 tiene las siguientes dos tareas:

1) Tarea de perfilamiento de autor: Consiste en determinar el sexo, ocupación y lugar de residencia de los usuarios a partir de sus tweets. El tema se centra en el análisis de los tweets generados por los usuarios mexicanos, lo que plantea retos adicionales relacionados con el tratamiento de una variedad de español con muchas particularidades culturales. En la nueva edición de 2019 (REDTTL - Red Temática en Tecnologías del Lenguaje, 2019), el tema considera el uso de texto e imágenes como fuentes de información. El propósito es explorar y estudiar la relevancia y complementariedad de la información multimodal para el perfil de los usuarios de los medios sociales.
2) Detección de agresividad: esta segunda tarea se enfoca en la detección de tweets agresivos en español mexicano.

Es de especial interés la tarea de agresividad para esta tesis ya que es de ahí donde parte este trabajo. En la página del workshop se especifica que el conjunto de datos para esta tarea se recopiló entre agosto y noviembre de 2017 de acuerdo con la siguiente metodología.

En primer lugar, se recogieron tweets basados en un vocabulario fijo extraído de un diccionario de "mexicanismos". Principalmente se consideró el subconjunto de palabras clasificadas como "vulgar" o "insultante" y se buscaron tweets que contenían al menos una de estas palabras. Luego, los tweets fueron etiquetados manualmente por dos personas como agresivos o no agresivos. Se proporcionó a los etiquetadores un manual de etiquetado basado en la premisa de que un mensaje ofensivo se caracteriza por menospreciar o humillar a una persona o a un grupo de personas. Por lo tanto, un mensaje ofensivo puede contener algunos de los siguientes elementos: apodos (asignados a la persona/personas a las que se dirige el mensaje, aludiendo a una discapacidad o defecto), bromas (siempre que tengan la intención de humillar o atacar), adjetivos despectivos (utilizados con la intención de humillar) y blasfemias (malas palabras o expresiones altisonantes utilizadas para atacar a una persona). Posteriormente se entregó este conjunto de prueba etiquetado manualmente para que los equipos participantes en la competencia entrenaran sus sistemas. En el etiquetado se consideró a la ofensa como un concepto bastante amplio que abarcaba diversos aspectos del lenguaje y que no necesariamente expresaban este término. Un ejemplo de esto se puede ver en el siguiente tweet procedente de este corpus.

> (5) @USUARIO Ojalá nunca te vayas, cabron, pero en cualquier equipo te irá excelente, te extrañaremos puto crack!

Aquí se hace uso de términos despectivos y malsonantes, como “cabrón” y “puto”. Sin embargo, el mensaje no es ofensivo, ya que sólo usa estas expresiones como intensificadores positivos. En el siguiente ejemplo, se puede ver un texto difícil de clasificar.

> (6) Todas estan feas menos la que me gusta.

Con las indicaciones que obtuvo el etiquetador, si tiene un término insultante (fea) debería ser ofensivo, sin embargo, aquí no se tiene un referente concreto. Es simplemente en general; todas. No hay un sujeto al que se dirija el mensaje por lo que no resulta ofensivo. Estos pequeños detalles no resultaban claros para los etiquetadores y por consiguiente para los sistemas de clasificación.

Posterior al entrenamiento de los sistemas con el conjunto de tweets etiquetados se les proporcionó un conjunto de prueba sin etiquetar para así probar sus sistemas de clasificación. En los resultados de los equipos participantes los sistemas no tuvieron puntajes altos. Se puede observar que el mejor rendimiento lo obtuvo el equipo de INGEOTEC, superando las otras aproximaciones con un valor F en la clase agresiva superior a 0.48 y una F-macro superior a 0.62. En general, el desempeño reportado muestra que la detección de agresividad en tweets es una tarea muy complicada. Sólo tres sistemas, INGEOTEC, CGP (fase 2) y Aragón López, superaron ambas líneas base (Álvarez-Carmona, Guzmán-Falcón, Montes-y-Gómez, Escalante, Villaseñor-Pineda, Reyes-Meza, y Rico-Sulayes, 2018). El equipo ganador usó como atributos n-gramas o secuencias de caracteres, n-gramas o secuencias de palabras, pero también incrustaciones de palabras y léxicos hechos a medida. Para la clasificación, aplicaron un conjunto de diferentes clasificadores, como las máquinas de soporte vectorial (SVM). Al analizar sus enfoques se puede notar la falta de un enfoque lingüístico efectivo.

2.7.2 Coeficiente de Kappa

Ya que en todos estos trabajos la clasificación se hace con base en el etiquetado manual de un corpus para el entrenamiento de los sistemas, éste debe ser lo más preciso posible. Para

poder saber qué tan bien los etiquetadores trabajaron el corpus se usa el coeficiente de kappa. Este coeficiente mide la concordancia entre dos examinadores en sus correspondientes clasificaciones de N elementos en C categorías mutuamente excluyentes, representado por la siguiente fórmula:

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

En esta fórmula Pr(a) representa el acuerdo observado real y Pr(e) un acuerdo de azar. Cohen, (1968) sugirió que el resultado Kappa se interprete de la siguiente manera: los valores ≤ 0 indican que no hay acuerdo, el rango de valores 0.01 – 0.20 indican un acuerdo que va de ninguno a leve, 0.21–0.40, un acuerdo justo, 0.41 – 0.60, uno moderado, 0.61–0.80, uno sustancial y 0.81 – 1.00 un acuerdo casi perfecto.

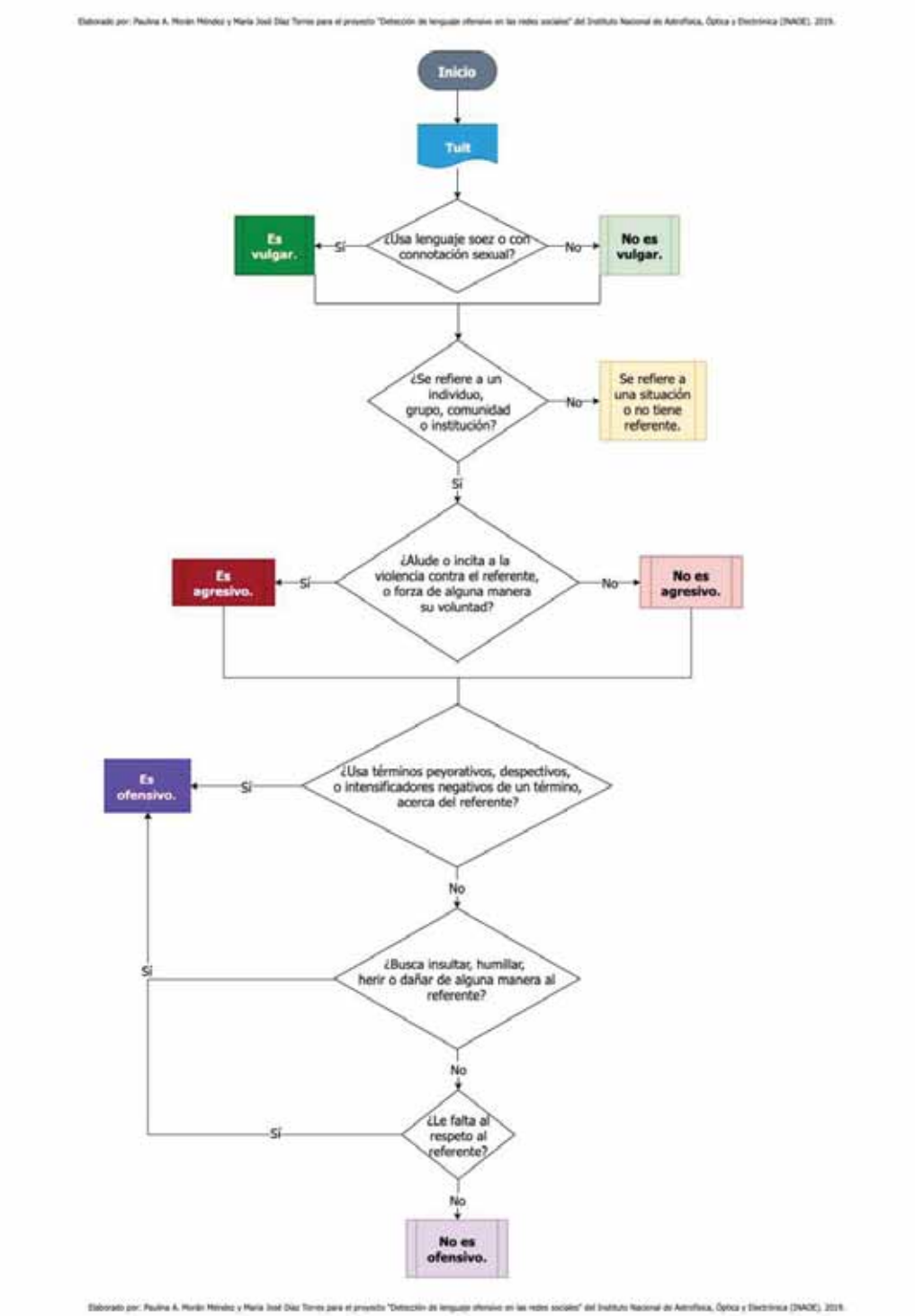
2.7.3 Diagrama de etiquetado manual de ofensividad.

Esta investigación se inserta en el contexto del workshop MEX-A3T. En esta competencia, se construyó un conjunto de datos orientado a la tarea. Como se explicó en la anterior sección, cada tweet fue anotado como ofensivo o no ofensivo por dos etiquetadores, siguiendo una serie de reglas definidas por los investigadores (Álvarez-Carmona, Guzmán-Falcón, Montes-y-Gómez, Escalante, Villaseñor-Pineda, Reyes-Meza, y Rico-Sulayes, 2018). No obstante, después de su revisión se corroboró que los datos no fueron anotados correctamente en muchos casos. Además de que los etiquetadores alcanzaron un acuerdo kappa de 0.5867 (Álvarez-Carmona et al., 2018), lo que significa que hay inconsistencias en el etiquetado elaborado por ambas personas. Por esto, surgió el interés de incrementar este acuerdo por

medio de un mejor etiquetado del corpus que resultará en un posible mejor resultado final de clasificación. Este trabajo lo realizamos en equipo colaborando con investigadores del laboratorio de tecnologías del lenguaje del INAOE y se presentó en la edición de este año del OpenCor, Latin American and Iberian Languages Open Corpora Workshop (OpenCor, 2019).

Con lo anterior en mente, el objetivo del proyecto de mejora fue definir los rasgos lingüísticos principales que caracterizan al lenguaje ofensivo, agresivo y vulgar manifestado a través de las redes sociales. Esto con el fin de establecer criterios que facilitaran su identificación y permitieran a personas no expertas en el tema etiquetar sistemáticamente un corpus para esta tarea. De aquí se produjo un diagrama con estos criterios y una mejora del acuerdo inter evaluador, obteniendo un 0.92 que fue considerablemente significativo. Dicho diagrama se encuentra a continuación en la Figura 2

Figura 2. Diagrama para mejora del etiquetado del corpus de MEX-A3T 2018



El diagrama en la Figura 2 comienza con la selección de un tweet a analizar. Para ilustrar el proceso se tomará el tweet "No es que estés gorda, lo gordo se quita. Es tu cara de caballo". Posteriormente se presenta la primera pregunta que cuestiona si el tweet "¿usa lenguaje soez o con connotación sexual?" Si la respuesta es positiva, entonces indica que es vulgar, de otra manera no lo es, como es el caso del ejemplo. La siguiente pregunta dice "¿se refiere a un individuo, grupo, comunidad o institución?" Esta pregunta busca hacer una temprana identificación de la agresividad, ya que si no tiene un blanco específico entonces ya no se seguirán con las siguientes preguntas. En el tweet ejemplo, el uso de la segunda persona da ya indicios claro de la existencia de un referente y debido a las expresiones utilizadas que hablan sobre las características de su rostro, se entiende que es humano. Por lo tanto, si es positiva la respuesta entonces la siguiente pregunta se asegura de saber si es agresivo al resaltar características como el incitar a la violencia o forzar la conducta del referente. En el ejemplo no se encuentran estos elementos de agresividad por lo que no es parte de las características de este mensaje en particular. Posteriormente, para determinar si es ofensivo se incita a observar si el tweet busca humillar o insultar. Si no se está seguro, se deberán tomar en cuenta la presencia de términos peyorativos, despectivos o intensificadores negativos y que busquen faltar al respeto usándolos. Con este último cuestionamiento termina el diagrama. En esta última pregunta, se observan en el mensaje ejemplo las expresiones "cara de caballo" y al resaltar la complexión de la persona, "gorda", es evidente que el objetivo es humillar al referente y faltarle al respeto. De esta manera podemos concluir que este mensaje no es vulgar o agresivo, pero sí ofensivo.

El resultado del coeficiente de Kappa significó en efecto que el uso de un diagrama basado en criterios lingüísticos facilita el etiquetado manual de agresividad ya que lo hace de

manera concisa. Así, se puede observar la importancia del punto de vista lingüístico cuando se está trabajando con lenguaje natural en áreas de la lingüística computacional y la tarea de anotación de corpus.

Capítulo 3. Metodología

Para poder analizar si una de las tres clases agresividad, vulgaridad y ofensividad son características de un mensaje, éstas tienen que ser reconocidas a través de atributos lingüísticos para ser detectadas de manera automática en mensajes de Twitter del corpus.

3.1 Descripción del Corpus.

Los datos a utilizar forman parte del corpus elaborado por el laboratorio de tecnologías del lenguaje del INAOE (LabTL). Este corpus fue recopilado con el propósito de ser usado para las tareas propuestas en el workshop titulado *Evaluation of Human Language Technologies for Iberian Languages* (IberEval) en el 2018. La tarea específica para la cual fue utilizado este corpus fue para la competencia de *Authorship and aggressiveness analysis in Twitter* (MEX-A3T). Como se mencionó anteriormente, el corpus completo fue producto del MEX-A3T. Este corpus consiste en 10,856 tweets. 7,700 tweets son parte del conjunto de entrenamiento. De estos, 4973 (un 65%) estaban etiquetados como no ofensivos, 2727 (35%) fueron clasificados como ofensivos. Por otro lado, el conjunto de test cuenta con 3156, de los cuales 2372 (un 75%) fueron etiquetados como no ofensivos y 784 (25%) como ofensivos.

Cada tweet fue etiquetado como ofensivo o no ofensivo por dos etiquetadores, siguiendo una serie de reglas definidas por los investigadores (Álvarez-Carmona et al., 2018). El corpus del MEX-A3T fue creado a partir de un conjunto de términos que sirvieron como

semillas para extraer los tweets. Estas fueron palabras que estuvieran clasificadas como vulgares no-coloquiales del Diccionario de Mexicanismos de la Academia Mexicana de la Lengua, así como palabras y hashtags identificados por el Instituto Nacional de las Mujeres como relacionadas con la violencia y acoso sexual contra la mujer en Twitter (Guzmán Falcón, 2018).

3.1.1 Selección del conjunto de datos para los experimentos

Usando el corpus descrito como punto de partida, se extrajo un conjunto de datos menor para poder hacer una clasificación que pusiera a prueba el modelo aquí esbozado. Para el conjunto de datos a utilizar en los experimentos, se decidió usar treinta tweets del conjunto de test. Diez por cada clase: agresividad, ofensividad y vulgaridad. Con el objetivo de evitar una predisposición a trabajar con tweets que cumplieran perfectamente con los atributos que se seleccionaron, se extrajeron los tweets de manera aleatoria usando el generador de números aleatorios de Google. En este, se estableció desde qué número empezar y hasta cuál terminar. En este caso el rango fue del 1 hasta 3156 del conjunto de test o prueba.

Una vez que se seleccionó la muestra de tweets que se usaron en los experimentos, se copió cada tweet en un archivo .txt. Para el primer experimento, se realizó una tabla donde se registraron las tres clases para cada uno de los tweets, así como los atributos seleccionados para identificar cada clase. Manualmente se revisó cada tweet anotando la presencia de los atributos en dicha tabla. Posteriormente se usó la plataforma Weka para la realizar la clasificación de los tweets con estos atributos por medio de varios algoritmos basados en aprendizaje automático.

Weka es una colección de algoritmos basados en aprendizaje automático para tareas de minería de datos. Contiene herramientas para la preparación de datos, clasificación, regresión, agrupación, minería de reglas de asociación y visualización (Eibe, Mark, & Witten 2016). Con este software se logró visualizar de manera más clara los resultados de los atributos seleccionados y el porcentaje de clasificación correcta por clase. Usando este software se convirtió el archivo de Excel en el que se produjo la tabla antes mencionada al formato nativo de Weka, con la extensión arff. Weka cuenta con una interfaz gráfica de usuario fácil de usar que aprovecha las capacidades de este software. Cada uno de los módulos de Weka está representado en la interface gráfica junto con una herramienta de Visualización (Eibe, et al.). En la parte de clasificación se seleccionó la opción de validación cruzada usando 10 iteraciones. La validación cruzada es un procedimiento estadístico que hace el experimento confiable y sus resultados generalizables a nuevas aplicaciones del clasificador. Esto permite trabajar con un mismo conjunto de datos y aún así tener un conjunto de entrenamiento y uno de prueba al dividir los mismos datos en varios conjuntos. Finalmente se usaron dos algoritmos de clasificación: Naïve Bayes y un árbol de decisión J48.

Para el segundo experimento, se seleccionaron las diez palabras plenas y las diez palabras vacías que mayor frecuencia tienen en el conjunto de textos. Así mismo, con el propósito de intentar obtener un mejor análisis de los textos y la estructura de su contenido, se extrajeron los cinco bigramas o secuencias de dos palabras más frecuentes. Se utilizó el software de AntConc para poder obtener los datos usados para este segundo experimento. AntConc es un kit gratuito de herramientas de análisis de corpus para concordancia y análisis de textos (Anthony, L., 2019). En este software se cargaron los treinta archivos con extensión

txt y en la parte de WordList se seleccionaron las palabras requeridas y se muestran en la Tabla 2 a continuación.

Tabla 2. 10 palabras más comunes y bigramas.

Función	**Contenido**	**Bigramas**
Que	Gorda	Dan asco
La	Mamar	De mamar
De	No	Están feas
Una	Pinche	Gata que
Y	Asco	Pinche gorda
A	Chichona	
Las	Feas	
Con	Gata	
Me	Prieta	
Le	usuario	

Después de obtener una lista de los bigramas, o secuencias de dos palabras adyacentes, y se seleccionaron los cinco más frecuentes, que también se pueden ver en la Tabla 2. Se repitió el mismo proceso con estos 25 atributos en Weka usando este nuevo archivo y se realizó la clasificación usando los mismos algoritmos explicados anteriormente.

Dado que en ambos experimentos fueron numerosos los atributos usados, se tomó la decisión de reducirlos. Para esto se siguió el proceso de selección de atributos con el que cuenta Weka para poder seleccionar aquellos que conllevan a mejores resultados. Se seleccionaron cuatro, dos del primer experimento y dos del segundo. Posterior a esto, se volvieron a registrar estos cuatro atributos y volvieron a ser procesados por Weka para obtener la clasificación. En vista de que los resultados no fueron tan altos, 63.33% y esperando obtener mejores, de decidió agregar más atributos. Después de esta revisión se agregaron a los cuatro atributos la inflexión verbal para que pueda ayudar al atributo de pronombres, y la de connotación sexual para que ayude a identificar mejor la clase de vulgaridad, que resultaba difícil de clasificar en los primeros experimentos.

3.2 Selección de atributos

Como se mencionó anteriormente, las clases en las que se separarán los textos son tres; agresivo, ofensivo y vulgar. Estas son no excluyentes, esto quiere decir que por ejemplo un mismo mensaje puede ser agresivo y vulgar, ofensivo y agresivo o sólo vulgar. En la categoría de lenguaje ofensivo, como se explicó en otra sección, su definición es: aquel que busca insultar o humillar a un grupo o individuo, usualmente utilizando términos peyorativos o despectivos. En este sentido, para poder detectar este tipo de lenguaje con más exactitud se usaron como términos peyorativos la lista que se encuentra en la Tabla 3.

Tabla 3. Palabras peyorativas

Afrancesado	Buey	Chupamedias/huevos
Analfabeto	Burro	Debilucho
Antipático	Canalla	Deforme
Apestoso	Cínico	Desagradable
Borracho	Cobarde	Gordo
Gringo	Idiota	Ignorante
Incompetente	Inepto	Infeliz
Insufrible	Inútil	machorra
Mantenido	Menso	Metido
Mísero	Mocoso	Roñoso
Vejestorio	Vividor	Feo
Loco	Chingar madre	Enano

En esta tabla se pueden encontrar palabras como *afrancesado*, que en un principio no es claro por qué se seleccionó. Sin embargo, esta palabra hace referencia a un individuo de manera despectiva ya que de cierta manera se burla del sujeto haciendo una comparación que no se podría relacionar directamente en otro contexto, donde se utilizaría un término más

respetuoso o neutro. Otro elemento a considerar fueron las palabras despectivas, que se pueden observar en la Tabla 4 que se muestra a continuación.

Tabla 4. Despectivos

Chacha	Naco	Ruco
Chusma	Ñero	Gentuza
chaka	prole	Tipejo
Mamarracho	debilucho	Baratilla
Culero	Hij* de tu _ madre	No mamar
Mentiroso(a)	Mamon(a)	

También se tomaron en cuenta los intensificadores negativos, que agregan un énfasis a un término que demuestra afectividad por parte del hablante, pero que en este caso será negativa, ya que esa es la emoción que se pretende transmitir por medio de las ofensas. Por lo tanto, la fórmula de un intensificador negativo que se usó para estas ofensas es el [intensificador] + adjetivo. Se pensó de esta manera porque el uso de un adjetivo solo sin acompañamiento de otra palabra en pocos casos es ofensivo, sin embargo, el efecto cambia cuando se acompaña del intensificador negativo. Por ejemplo, el adjetivo "gordo/a" por sí solo no tiene una carga negativa tan fuerte, puede simplemente representar una opinión y eso no afecta al referente. Sin embargo, si se dice [pinche] o [bien] gorda entonces la carga negativa del adjetivo se intensifica y es así cuando es mucho más probable que sea ofensivo.

En la categoría de lenguaje agresivo, se definió a éste como el que se usa con el fin de dañar o herir a un grupo o individuo aludiendo o incitando a la violencia. Este podría ser el más importante de identificar, debido a que cuando hay intercambios sociales en las redes sociales, estos se vuelven agresivos y puede que haya riesgos físicos. Por lo tanto, durante el

monitorio automático de los mensajes en línea es sumamente valioso poder predecir y prevenir dichas acciones violentas. Es así que para poder identificarlo se decidió usar los pronombres de segunda persona del singular y plural y los de tercera persona igualmente del plural y el singular. Esto, debido a que la agresividad cuenta como tal cuando es especialmente dirigida hacia alguien. En vista de que en español se puede evitar el uso explícito de los pronombres, también se tomaron en cuenta las inflexiones en las conjugaciones verbales con los pronombres anteriormente mencionados. Por ejemplo, se tomó en cuenta en un mensaje escrito la presencia del pronombre "tú" y cuando no estaba presente se observó el verbo, por ejemplo "muérete", que indica perfectamente que el mensaje está dirigido. Así mismo, a causa de la naturaleza de la plataforma de Twitter, se usarán los "@Usuario" que son menciones de usuarios. En un sentido parecido a la mención de usuarios, también se usarán los nombres propios para identificar que el mensaje esté dirigido. Específicamente, en la parte de violencia, se seleccionaron tres indicadores importantes de violencia: homofobia, racismo y misoginia. Dichos indicadores hacen alusión al *hate speech* que es una evidente muestra de violencia a través del lenguaje.

Para la detección de misoginia se usó un banco de palabras que se puede observar en la Tabla 5. Estas palabras provienen de un estudio realizado por el instituto de las mujeres del distrito federal en el 2016. En este estudio se analizó el impacto de las redes sociales con respecto a la violencia de género retratadas en el espacio digital. A parte de palabras por sí solas, en el estudio se usaron nubes semánticas, en las que detectaban las expresiones cercanas a las palabras más comúnmente usadas, por lo tanto, también hay expresiones que se usaron y que están expuestas en la Tabla 5 en anexos. Cuando las palabras de la Tabla 5 se encontraron en el tweet o también cuando una palabra estaba acompañada de los

pronombres y las inflexiones verbales anteriormente mencionadas, entonces se consideró como misógino el mensaje.

Tabla 5. Palabras Misóginas

Putas	prostituta	culera	golfa	putita	perra
Zorra	vieja	Viejas + adjetivo	Intensificador negativo + viejas + adjetivo	moretones	Muestras de amor
putipobre	putinovia	putiesposa	putimorrita	putiprima	putirica
pack	urgida	feminazi	Hija de puta	Ojo morado	Putizorrra

De la tabla anterior, se muestran palabras como *moretones* y *ojo morado*. Estas se encontraron presentes en tweets donde se relata de manera explícita la violencia física hacia las mujeres. Por lo que se encontraron útiles para poder ubicar la misoginia.

Una característica más que se le suma en violencia es el racismo. Para esto, se tomó la presencia de palabras como "prieto/a" e "indio/a". Debido a que México es un país culturalmente diverso pero que tiene un sistema de clases bastante marcado. Para esta sección de racismo se consideró en específico lo que Gall (2004) considera como racismo de la desigualdad que "tiene su origen en la tradición comunitaria, afirman la diferencia, exaltan la pureza de las razas y separan a los grupos" (p.235). Por ello aquí se sitúa la diferencia entre las clases sociales, en especial el trato hacia las comunidades indígenas que desde hace años viven marginados de la sociedad. Como se puede observar, el racismo se vuelve una vía de transmisión y ejecución de violencia. Por lo tanto, es en extremo importante que se detecte con anticipación.

Por último, en la parte de agresividad se tomó en cuenta los actos de habla y en específico el ilocutivo, ya que representa la fuerza o la intención de la expresión hecha por el hablante. Se decidió verlo desde esta perspectiva porque el enfoque es tanto lo que se dice, como para qué se dice y qué se pretende lograr con el mensaje, y no lo que causa en el referente. Debido a que esto último no es controlable y resulta subjetivo. Esto es clave para poder identificar el tipo de texto ya que es en extremo común que se use un lenguaje indirecto o que general se juegue con él para poder transmitir ideas. En este sentido, tomando en cuenta la intención del mensaje se tomaron las palabras que propone Sidorov (2013) como parte de su clasificación de actos de habla basada en los verbos de habla. De dichos verbos se seleccionó la sección seis de motivación, los verbos de esta parte se pueden ver en el Anexo 1. De esta tabla se buscaron en el conjunto de tweets los siguientes verbos: Dictar, ordenar, requerir, decretar, mandar, disponer, comandar, requerimiento, impulsar, inducir, insinuar, desaprobar, prohibir, solicitar, exhorto, y pedir. Así mismo, la conjugación de todos los verbos en imperativo. Esto refiere a la definición de violencia que se explicó al principio; busca cambiar la conducta del referente. Por ello el imperativo da un buen indicio de cuándo se requiere que el referente modifique su conducta. Un ejemplo de esto se puede ver en el siguiente tweet tomado de la colección total:

(7) "¡Ya cállate alv maldito chairo, cierra el puto hocico pinche gorda femichaira!"

Este tweet se considera agresivo, por diversas características, en este caso hay dos imperativos: cállate y cierra, que sin duda obligan al referente a modificar sus acciones. Como un aspecto más se incluyeron palabras que difaman, como ratero, mentiroso, corrupto y traidor.

La última categoría de las características de un tweet que se propuso es vulgaridad. Para recordar, la definición que se usa de un tweet vulgar es aquel que involucra expresiones soeces, con connotación sexual y en algunas ocasiones con doble sentido, pero puede o no referirse a un individuo o colectividad. En este tipo de tweets el referente no es importante debido a que lo que lo caracteriza es que no pretende dirigir el mensaje hacia ningún sujeto en particular con afán de dañarlo. Rico Sulayes (2014) aporta un concepto general sobre el lenguaje vulgar. Él caracteriza el lenguaje vulgar como aquel que implica solidaridad, intimidad o familiaridad, es de un estilo espontáneo, informal, coloquial o casual. Como ya se explicó, este se considera también como un tabú, ya que contiene léxico malsonante o grosero, así como no estándar, de nivel de lengua bajo, popular o vernáculo. Es por esto, que en este caso la presencia de pronombres y la conjugación de los verbos no fueron consideradas. Los rasgos a considerar fueron la presencia de groserías y para esto se usó un diccionario de vulgaridades.

3.3 Métricas de evaluación

Con el propósito de poder medir los resultados y evaluar el desempeño del modelo se usarán cuatro métricas: exactitud, precisión, recuerdo y valor F. Dichas métricas serán explicadas a continuación.

3.3.1 Exactitud

La exactitud es la medida en la que más me basaré para poder evaluar los resultados de las clasificaciones. Es una medida muy comúnmente usada. La exactitud se calcula y expresa mediante una declaración del porcentaje del texto que ha sido correctamente clasificado cuando se compara con los datos de referencia o "verdad de terreno" (Aronoff, 1985). Por

esto, resulta útil especialmente en este tipo de experimentos. El valor se calcula como el número de elementos clasificados correctamente, sobre el número total de elementos clasificados (Guzmán Falcón, 2018).

3.3.2 Precisión

La precisión es un criterio frecuentemente usado en la Búsqueda y recuperación de información. Este se refiere a recuperar el mayor número posible de elementos relevantes (Raghvan & Gwang, 1989). El mismo autor explica que tanto el recuerdo, que se explicará a continuación, y la precisión se miden después de que el sistema determina un pedido de los documentos de su colección en respuesta a la consulta de un usuario. Este orden representa el juicio del sistema de cómo cada documento se relaciona con la necesidad del usuario.

3.3.3 Recuerdo

El recuerdo mide que se hayan recuperado tan pocos elementos no relevantes como sea posible en respuesta a una solicitud (Raghvan & Gwang, 1989). El recuerdo es particularmente útil en los sistemas de recuperación de información, como Google, y obedece a necesidades de otro tipo de clasificación.

3.3.4 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 (Guzmán Falcón, 2018). Por esto, el valor f es una medida en extremo importante y que es comúnmente usada al evaluar los workshops de clasificación.

3.4 Reducción de atributos

La selección de atributos es esencial para la detección de autoría, o de clases establecidas. Rico Sulayes (2017) explica que es común que se elaboren largas listas de atributos en colecciones de textos. El autor explica que esto en parte se debe a la proliferación de atributos que caracterizan la atribución de autoría, incluso cuando se dirige a pequeños corpus. Por esto, Rico Sulayes indica que la predeterminación de los atributos es el factor más importante para mejorar la precisión de la clasificación en esta tarea, incluso más que cualquier ajuste sutil del algoritmo de clasificación. Por ello, esto también se vuelve un proceso extremadamente complicado ya que determina el resultado final de la clasificación. Sin embargo, a veces puede ser un poco complicado trabajar con tantos atributos y podría dificultar el análisis de los resultados. Por esto, Rico Sulayes comenta una respuesta al problema de tener demasiadas características en las tareas de clasificación de textos en general ha sido la elaboración de listas de atributos reducidas en un intento de señalar y luego aplicar características altamente discriminatorias durante la clasificación. Entonces, el propósito es seleccionar aquellos atributos que se desempeñan mejor en la clasificación. Así, el mismo autor comenta que los pequeños conjuntos de atributos, especialmente discriminatorios, pueden mejorar la precisión de las tareas de clasificación ya que evitan el ruido introducido por las características redundantes y poco discriminatorias.

Un ejemplo de cómo se aplica la reducción de atributos se puede ver en el trabajo de Xu et al. (2015). En este estudio los autores intentan hacer una clasificación de múltiples etiquetas que no necesariamente son excluyentes entre sí y que, por tanto, dificultan la tarea. Decidieron realizar esta clasificación múltiple para poder lograr que las distintas clases sean incluyentes y se superpongan si es necesario. Para esto, construyeron los atributos específicos

de las etiquetas comprobando las distancias entre la muestra y todos los centros de agrupamiento. Sin embargo, la construcción de atributos específicos para la etiqueta de las clases puede verse acompañada del incremento en la colección de atributos y por lo tanto existe una gran cantidad de información redundante. Por esto, ellos decidieron aplicar un proceso de reducción de atributos en el que construyeron atributos específicos de la etiqueta comprobando las distancias entre la muestra y todos los centros de agrupación. Respecto a esto, los investigadores explican que como no hay semantema para los atributos particulares de la etiqueta construida, pueden considerarse como un conjunto de distancias. En base a su modelo y la selección de atributos, concluyeron que en efecto se mejora el rendimiento del aprendizaje multi-etiqueta al aplicar la reducción de las dimensiones de los atributos particulares a aquellos que son redundantes.

Rico Sulayes (2017) plantea que la solución posible detrás de la reducción de la lista de atributos también ha sido ampliamente resaltada en los estudios de atribución de autoría. Una solución alternativa para hacer frente al ruido de largas listas de atributos es el uso de algoritmos de clasificación de última generación especialmente resistentes a este ruido. Un claro ejemplo de este tipo de algoritmo son las máquinas de soporte vectorial (SVM). Estas han sido ampliamente utilizadas en la atribución de autoría.

En este trabajo, dado que se hizo uso de Weka para la clasificación, en la plataforma se utilizará la opción de "selección de atributos". La opción seleccionada fue CfsSubsetEval (*Correlation-based feature subset selection*). Este reductor evalúa el valor de un subconjunto de atributos considerando la capacidad de predicción individual de cada característica junto con el grado de redundancia entre ellos. Este reductor prefiere subconjuntos de características

que estén altamente correlacionadas con la clase y que tengan una baja intercorrelación (Hall, 1998).

3.5 Aprendizaje supervisado

Los métodos usados en este trabajo se relacionan con el de aprendizaje supervisado. En este se entrenan algoritmos utilizando datos precodificados y se valida su rendimiento en datos no codificados (Gibbons, Richards, Valderas, & Campbell, 2017). Esto es precisamente lo que se hace en los workshops explicados en las anteriores secciones al usar datos de entrenamiento manualmente etiquetados y posteriormente probar el sistema de clasificación con datos sin etiquetar. También esto es lo que se realizó en los experimentos de clasificación de este trabajo, pero en conjunto con la opción en Weka de validación cruzada ya que permite este mismo método separando el corpus que se está procesando.

Soto y Jiménez (2011) explican más a detalle que el aprendizaje supervisado es un método que posibilita una discriminación y clasificación difusa, pero requiere de un conjunto de ejemplares que hayan sido clasificados convencionalmente, en otras palabras, en categorías excluyentes. Este conjunto es llamado conjunto de entrenamiento o muestra de aprendizaje. Con base en los patrones que se encuentren, se determinan las categorías a las cuales pertenecen los nuevos ejemplares.

3.5.1 N-Gramas

Durante el proceso de extracción de atributos, con el fin de que estos lleven a una clasificación más precisa, una manera de obtener más información del mensaje es el uso de n-gramas o secuencias de elementos lingüísticos. Estos elementos pueden ser palabras o

caracteres. La extracción de atributos usando n-gramas ofrece un medio para retener parte del contexto en el que se usan las palabras. Un n-grama tokeniza secuencias (de longitud n) de palabras como atributos. Esto puede proporcionar mejor información que la simple estrategia de conteo de palabras utilizada en la matriz de términos y documentos (Gibbons et al., 2017). Es por esta información extra que proveen que son utilizados ya que funcionan un poco como pequeñas muestras de las concordancias.

Cuando sólo se tiene una simple lista de palabras es complicado saber qué utilidad tienen en el texto. Entonces para saber qué es lo que hace que los n-gramas sean más útiles que una simple lista de palabras tokenizadas, Koppel et al. (2011 citado en Sapkota, Bethard, Montes, & Solorio, 2015) comentan que los caracteres de n-gramas conllevan varias cosas: contenido léxico, contenido sintáctico e incluso estilo por medio de la puntuación y los espacios en blanco. Por esto aportan bastante información que puede seleccionarse conforme los objetivos del experimento que se esté elaborando. Sapkota et al. mencionan tres categorías en las que se pueden ordenar los n-gramas. Estas categorías están relacionadas con los tres aspectos lingüísticos hipotéticos para ser representados por n-gramas de caracteres: morfosintaxis (representados por n-gramas de afijos), contenido temático (representados por n-gramas como palabras) y estilo (representados por n-gramas basados en puntuación).

En este trabajo se usaron n-gramas de palabras cortos, sólo dos palabras adyacentes. En este sentido, Sapkota et al. (2015) explican que los n-gramas de caracteres son a menudo demasiado cortos para capturar palabras enteras, algunos tipos pueden capturar palabras parciales y otros tókenes relevantes para la palabra. Es decir que pueden ser tanto palabras como otros elementos de las oraciones de utilidad para el análisis de las oraciones. Estos autores consideran las siguientes características para este tipo de n-gramas: palabra entera;

un carácter n-grama que cubre todos los caracteres de una palabra que es exactamente n caracteres de largo, Mitad de Palabra; un carácter n-grama que cubre n caracteres de una palabra que es de al menos n + 2 caracteres de largo, y que no cubre ni el primer ni el último carácter de la palabra, Multi-palabra; n-gramas que abarcan múltiples palabras, identificada por la presencia de un espacio en el centro del n-grama. De estas características explicadas, la que se ocuparon fue la primera; palabra entera. Por ejemplo, si se tiene la oración "el perro corre", aquí habría tres tókenes. Por ello, si se decide usar bigramas con este ejemplo, hay dos bigramas posibles: el perro, perro corre. De esta manera estas dos podrían ya formar parte de los atributos. Para saber cuántos bigramas o cualquier cantidad de n-gramas hay, se usa la siguiente fórmula. En esta, x son la cantidad de palabras que hay en una oración y la N es la cantidad de gramas que se deseen obtener.

$$N\text{-gramas} = X - (N-1)$$

3.5.2 Clasificador Naïve Bayes

Para el proceso de clasificación en el modelo computacional se puede aplicar el algoritmo Naïve Bayes**.** Este algoritmo está basado en el teorema de Bayes (1763) y en la premisa de independencia de los atributos dada una clase. Asimismo, es uno de los métodos de aprendizaje supervisado más utilizados debido a que es posible adaptar para diversos experimentos como el análisis de emociones, perfilamiento y detección de autoría (Gutiérrez Esparza, Margain Fuentes, Ramírez del Real, & Canul Reich, 2017). El uso de este algoritmo presenta ventajas en la clasificación. Mosquera, Castrillón, & Parra, (2018) explican que a menudo proporciona una mejor precisión de clasificación en conjuntos de datos en tiempo real comparado con cualquier otro clasificador. Otra ventaja es que requiere una pequeña

cantidad de datos de entrenamiento. Así, el clasificador Naïve-Bayes aprende de los datos de entrenamiento y luego predice la clase de la instancia de prueba con la mayor probabilidad posterior.

Gutiérrez Esparza et al. (2017) incluyen en su trabajo una tabla acerca de cómo se evalúa el desempeño en la clasificación. Para esto, la matriz de confusión se considera el punto de partida para el cálculo de la medición del desempeño de un modelo predictivo. Esta matriz evalúa los resultados a partir de un conteo de los aciertos y errores de cada una de las clases en la clasificación. Dicha matriz se encuentra en la tabla a continuación.

Tabla 6. Matriz de confusión para clasificación binaria

	Valor actual	
Valor Predictivo	VP	FP
	FN	VN

En esta matriz VP se refiere a los verdaderos positivos, FP a los falsos positivos, FN a los falsos negativos y VN a los verdaderos negativos. VP mide la cantidad de mensajes que eran de una clase y que fueron clasificados como tal. FP es el Número de casos que la prueba declara positivos y que en realidad son negativos. VN es el número de casos que la prueba declara negativos y que son realmente negativos y FN es el número de casos que el sistema declara negativos y que en realidad son positivos. Los autores comentan que el análisis de esta matriz deriva de la comparación de los resultados de un modelo predictivo contra los valores reales. Por lo que este tipo de análisis resulta sencillo de comprender y es una buena fuente de análisis de los resultados.

3.5.3 Árbol de decisión

Otro algoritmo de clasificación común en este tipo de tareas son los árboles de decisión y también fueron usado en los experimentos. Los árboles de decisión son un método de aprendizaje supervisado no paramétrico utilizado para la clasificación y regresión (Han et al. 2006 citado en Fletcher & Islam, 2019). Los árboles de decisión no hacen suposiciones sobre la distribución de los datos subyacentes y son entrenados en datos etiquetados para clasificar correctamente los datos no vistos anteriormente (Fletcher & Islam). Es por esto que es útil en los experimentos de este trabajo. Los mismos autores comentan que usar este algoritmo también tiene varias ventajas sobre otros tipos de métodos de aprendizaje supervisado y que los hacen atractivos para los científicos de datos. Algunas de sus ventajas se encuentran a continuación: 1) alta interpretabilidad humana, 2) diseño no paramétrico, 3) habilidad para descubrir relaciones no lineales entre los atributos 4) resistencia a valores faltantes, 5) capacidad para manejar datos continuos y discretos, y 6) capacidad para manejar etiquetas no binarias.

De estas características, algunas son especialmente útiles en el experimento de este trabajo. El primer punto acerca de la alta interpretabilidad es de utilidad ya que, al tener un mayor entendimiento del sistema, el análisis de resultados resulta más claro. El tercer punto es útil ya que provee nuevas opciones para relacionar los atributos y encontrar diversas interpretaciones en los resultados. El último punto es el que resulta en extremo relevante, ya que en este caso se trabajan con tres clases, por lo que un algoritmo que trabaje con múltiples clases es necesario. Para poder hacer uso de este algoritmo Weka ofrece el algoritmo C4.5 también llamado J48. Saheed, Oladele, Akanni, & Ibrahim (2018) explican que este algoritmo utiliza la relación de ganancia como criterio de división para particionar el conjunto de datos. El algoritmo aplica una especie de normalización a medida que entra la información

utilizando un valor de "información dividida". El algoritmo árbol de decisión puede ser diseñado cuando se toman las muestras originales como la raíz del árbol de decisión.

Capítulo 4. Resultados y Discusión

Resultados

Para poder realizar un análisis de las tres clases de un mensaje (agresivo, ofensivo, vulgar), primero se tomaron las características de las tres clases extraídas detalladas en la sección de metodología. Dado que estas características se aplicaron a un conjunto pequeño de datos, se hizo la búsqueda de estos atributos manualmente y se registraron en una base de datos.

Al pasar estos datos por Weka usando Naïve Bayes se obtuvieron los resultados en las tablas 7 y 8. En estas tablas se puede observar un porcentaje de 73.33% y un valor F de 0.55. La clase de agresividad tuvo la precisión más baja con 0.66, después el ofensivo con 0.70 y el más alto fue el vulgar con 0.75.

Tabla 7. Resumen de resultados con 16 atributos Naïve Bayes

Instancias correctamente clasificadas	22	73.3333%
Instancias clasificadas incorrectamente	8	26.6667%
Número total de instancias	30	

Tabla 8. Precisión detallada por clase de 16 atributos NB.

	VP Índice	FP Índice	Precisión	Recuerdo	Valor F	
Clase	0.700	0.150	0.700	0.700	0.700	Agresivo
	0.900	0.250	0.643	0.900	0.750	Ofensivo
	0.600	0.000	1.000	0.600	0.750	Vulgar

Promedio ponderado	0.733	0.133	0.781	0.733	0.733	

Posteriormente, con un árbol de decisiones se obtuvieron los resultados detallados en la tabla 9 y 10. Con este clasificador se obtuvo la misma exactitud, 73.33%, por lo cual se presenta consistente el resultado incluso usando un clasificador diferente. La precisión es un poco menos siendo 0.741, el recuerdo fue el mismo con 0.733 y el valor F fue de 0.734. En esta ocasión las clases con el mejor resultado fueron dos: agresivo y ofensivo, ambas con un valor F de 0.737. Esta medida es menor a las dos mejores clases del anterior clasificador. Tomando en cuenta que ya fueron altos estos resultados, pero notando que fueron un total de 16 atributos, se decidió reducirlos. Por esto, se trató de encontrar aquellos que fueran más discriminatorios y que por consiguiente mejoraran los resultados con la clase de agresividad, que era la más baja.

Tabla 9. Resumen de resultados con 16 atributos usando J48.

Instancias correctamente clasificadas	22	73.3333%
Instancias clasificadas incorrectamente	8	26.6667%
Número total de instancias	30	

Tabla 10. Precisión detallada por clase de 16 atributos usando J48.

	VP Índice	FP Índice	Precisión	Recuerdo	Valor F	
Clase	0.700	0.100	0.778	0.700	0.737	Agresivo
	0.700	0.100	0.778	0.700	0.737	Ofensivo
	0.800	0.200	0.667	0.800	0.727	Vulgar
Promedio ponderado	0.733	0.133	0.741	0.733	0.734	

Para esto, se siguió la metodología explicada en secciones anteriores. Para un primer experimento de reducción, se seleccionaron las diez palabras y los 5 bigramas más frecuentes. Dichos atributos se muestran en la Tabla 11 a continuación.

Tabla 11. Atributos del segundo experimento.

Función	**Contenido**	**Bigramas**
Que	Gorda	Dan asco
La	Mamar	De mamar
De	No	Están feas
Una	Pinche	Gata que
Y	Asco	Pinche gorda
A	Chichona	
Las	Feas	
Con	Gata	
Me	Prieta	
Le	usuario	

En este segundo experimento los resultados se encuentran registrados en las tablas 12 y 13. Como se puede observar en las mismas, el porcentaje de exactitud obtenido con el clasificador Naïve Bayes fue 63.33%, la precisión promedio fue de 0.64 y el valor F fue 0.63. La menor precisión la tuvo la clase de agresión con 0.58, después la agresiva con 0.63 y la más alta fue vulgar con 0.71.

Tabla 12. Resumen de resultados del segundo experimento NB.

Instancias correctamente clasificadas	19	63.3333 %

Instancias clasificadas incorrectamente	11	36.6667 %
Número total de instancias	30	

Tabla 13. Precisión detallada por clase del segundo experimento con NB.

	VP Índice	FP Índice	Precisión	Recuerdo	Valor F	
Clase	0.700	0.200	0.636	0.700	0.667	Agresivo
	0.700	0.250	0.583	0.700	0.636	Ofensivo
	0.500	0.100	0.714	0.500	0.588	Vulgar
Promedio ponderado	0.633	0.183	0.645	0.633	0.630	

Al clasificar con un árbol de decisión J48 (tablas 14 y 15) resultó una exactitud bastante menor al primer experimento, siendo ésta de 30%. La clase con resultados más altos fue agresividad. Aun así, los resultados en general fueron mucho más bajos que con el otro clasificador e incluso usando el mismo en el primer experimento. La precisión fue 0.444, el recuerdo de 0.400 y el valor F de 0.421. En vista de que este segundo experimento no tuvo tan buenos resultados, se utilizó la reducción de atributos para seleccionar los que mejor se desempeñaron durante ambos experimentos.

Tabla 14. Resumen de resultados del segundo experimento J48.

Instancias correctamente clasificadas	9	30 %
Instancias clasificadas incorrectamente	21	70 %
Número total de instancias	30	

Tabla 15. Precisión detallada por clase del segundo experimento con J48.

	VP Índice	FP Índice	Precisión	Recuerdo	Valor F	
Clase	0.400	0.250	0.444	0.400	0.421	Agresivo
	0.400	0.500	0.286	0.400	0.333	Ofensivo
	0.100	0.300	0.143	0.100	0.118	Vulgar
Promedio ponderado	0.300	0.350	0.291	0.300	0.291	

Debido a los bajos resultados obtenidos al utilizar una selección de rasgos frecuentes, se usó la reducción disponible en Weka. Utilizando el selector de rasgos CfsSubsetEval (Correlation-based feature subset selection), las categorías de atributos que mejores resultados conseguían fueron: pronombres y términos relacionados con el racismo. Los resultados de la clasificación con los nuevos atributos se ven a continuación.

Tabla 16. Resumen de resultados del tercer experimento NB.

Instancias correctamente clasificadas	21	70 %
Instancias clasificadas incorrectamente	9	30%
Número total de instancias	30	

Tabla 17. Precisión detallada por clase del tercer experimento con NB.

	VP Índice	FP Índice	Precisión	Recuerdo	Valor F	
Clase	0.700	0.000	1.000	0.700	0.824	Agresivo
	1.000	0.300	0.625	1.000	0.769	Ofensivo
	0.400	0.150	0.571	0.400	0.471	Vulgar
Promedio ponderado	0.700	0.150	0.732	0.700	0.688	

Tomando estos atributos, usando el algoritmo de clasificación Naïve Bayes (tabla 16 y 17) el porcentaje de exactitud fue de 70%. No resultó un porcentaje bajo en general y sin

duda hubo un incremento en comparación con el segundo experimento. La precisión fue de 0.73, el recuerdo de 0.70 y el valor F de 0.68.

Usando el clasificador J48 (tablas 18 y 19) los resultados son considerablemente menores y poco significativos ya que solo la mitad fue clasificada correctamente con sólo un 50% de exactitud. La precisión es aún menor ya que sólo es 0.44, el recuerdo es de 0.50 y finalmente el valor F es de 0.43. La clase con menos valor F es la vulgar ya que obtuvo solamente 0.143, un recuerdo de 0.100 y precisión de 0.250. La clase con el valor F más alta fue la ofensiva, obteniendo 0.69. La ofensiva también obtuvo un recuerdo de 1.00 y una precisión de 0.52, aunque no fue la más alta. La más alta fue de la clase agresiva con 0.57.

Tabla 18. Resumen de resultados del tercer experimento J48.

Instancias correctamente clasificadas	15	50 %
Instancias clasificadas incorrectamente	15	50%
Número total de instancias	30	

Tabla 19. Precisión detallada por clase del tercer experimento con J48.

	VP Índice	FP Índice	Precisión	Recuerdo	Valor F	
Clase	0.400	0.150	0.571	0.400	0.471	Agresivo
	1.000	0.450	0.526	1.000	0.690	Ofensivo
	0.500	0.150	0.250	0.100	0.143	Vulgar
Promedio ponderado	0.700	0.250	0.449	0.500	0.434	

Cuando se establecieron las características del atributo de racismo, se incluyó la presencia de la palabra *prieta* y *prieto* como indicador de la clase de agresividad, por lo que tener ambas como atributos era repetitivo. Por esto, se decidió agregar más atributos. En este

sentido, se realizó un último experimento con los atributos: *de*, *prieta*, y los atributos de las categorías de pronombres, racismo, inflexión verbal y connotación sexual. Los resultados se encuentran a continuación en las tablas 20 y 21.

Tabla 20. Resumen de resultados del cuarto experimento NB.

Instancias correctamente clasificadas	26	86.6667 %
Instancias clasificadas incorrectamente	4	13.3333 %
Número total de instancias	30	

Tabla 21. Precisión detallada por clase del cuarto experimento con NB.

	VP Índice	FP Índice	Precisión	Recuerdo	Valor F	
Clase	0.800	0.00	1.000	0.800	0.889	Agresivo
	1.000	0.100	0.8333	1.000	0.909	Ofensivo
	0.800	0.100	0.800	0.800	0.800	Vulgar
Promedio ponderado	0.867	0.067	0.878	0.867	0.866	

Los resultados (tabla 20 y 21) fueron considerablemente mejores. La exactitud fue de 86.66%, siendo la más alta entre los anteriores porcentajes. En cuanto a las demás medidas, la precisión fue de 0.87, el recuerdo 0.86 y el valor F 0.86. La clase con el mayor valor F fue ofensividad, que obtuvo 0.90. Esta tiene una precisión de 0.83 y un recuerdo de 1.00. La clase con menor valor F fue la del vulgar, con 0.80. Esta misma clase tuvo un recuerdo de 0.80 y una precisión igual de 0.80.

Reproduciendo el ejercicio anterior con el clasificador J48, se obtuvo una exactitud de 80%. Esta igualmente es mayor a las ya obtenidas con este clasificador. Una precisión de 0.844, recuerdo 0.800 y valor F de 0.805. Estos resultados presentan cantidades altas y no hay una gran diferencia con Naïve Bayes, el otro clasificador usado. La clase con el valor F

más alto fue la vulgar, obteniendo 0.842. En esta clase también se obtuvo una precisión de 0.88, también la más alta de todas las clases y por último un recuerdo de 0.80, el segundo más alto en comparación con las otras dos clases restantes.

Tabla 22. Resumen de resultados del cuarto experimento J48.

Instancias correctamente clasificadas	24	80%
Instancias clasificadas incorrectamente	6	20 %
Número total de instancias	30	

Tabla 23. Precisión detallada por clase del cuarto experimento con J48.

	VP Índice	FP Índice	Precisión	Recuerdo	Valor F	
Clase	0.700	0.00	1.000	0.700	0.824	Agresivo
	1.900	0.250	0.643	0.900	0.750	Ofensivo
	0.800	0.050	0.889	0.800	0.842	Vulgar
Promedio ponderado	0.800	0.100	0.844	0.800	0.805	

En el último experimento cuyos resultados se muestran en las dos tablas anteriores, vale la pena observar la matriz de confusión que forma parte del resultado de clasificación. Ésta se presenta a continuación en la Tabla 24.

Tabla 24. Matriz de confusión del cuarto experimento.

Clasificado como:	**a**	**b**	**c**
Agresivo = a	8	0	2
Ofensivo = b	0	10	0
Vulgar = c	0	2	8

Se puede observar en esta matriz las cantidades de verdaderos positivos que sí son mayores. Los que tienen más de un error en la clasificación son el *a*, agresivo que en dos veces se clasificó como vulgar. Y también el *c*, vulgar que se clasificó la misma cantidad de veces como ofensivo.

4.1 Discusión

Viendo los resultados del primer experimento, se puede notar que en general con todos los atributos seleccionados el porcentaje de instancias correctamente clasificadas es alto, 73.33%. Este resultado es mayor que los resultados obtenidos en otras investigaciones de esta área por equipos mexicanos en la competencia de MEX-A3T. Por lo que podría resumirse que sí hay una mejoría en la clasificación al seleccionar atributos lingüísticos y esto parece ser consistente con ambos algoritmos, tanto Naïve Bayes como con J48. En el siguiente experimento, como ya se explicó anteriormente, hubo un enfoque en la frecuencia de palabras, método bastante común en este tipo de experimentos. De esta manera, se pretendió un enfoque más en el contenido y la estructura de los textos, ya que también se usaron bigramas como atributos. Esto para obtener un mejor porcentaje de instancias clasificadas correctamente. Sin embargo, el resultado no mejoró. Incluso hubo un resultado muy bajo de 30% con J48. Por lo tanto, la selección de los atributos lingüísticos resulta más útil que la frecuencia de palabras. Como sólo se seleccionaron los cinco bigramas más comunes de palabras, el contexto que estos proveen no resultó significativo. Otra opción que se podría tomar para poder obtener mejores resultados de ellos sería usar tókenes, caracteres o bigramas de POS. Ya que estos ofrecen más información del contexto y de la estructura de los enunciados, ayudando a comprender mejor este tipo de textos. Para el siguiente experimento usando la reducción de atributos, que como se explicó, buscaba facilitar el

manejo de atributos y eliminar aquellos que resultaban repetitivos, el porcentaje de exactitud incrementó. El incremento es en comparación con el segundo ejercicio que usaba atributos frecuentes. Este estuvo presente en ambos algoritmos, por lo que se puede observar que estos atributos sí fueron reducidos con éxito. Esto se debe a que se obtuvieron los que se desempeñaron mejor. Algo interesante aquí fue que en la reducción de atributos no se obtuvo ningún atributo significativo entre los bigramas. Especialmente porque en un principio se pensó que podrían ser los que mejor se desempeñarían ya que tienen la posibilidad de traer más contexto.

En el último experimento, los resultados fueron los mejores en comparación con los primeros. Los atributos: *de*, *prieta*, y los pertenecientes a las categorías de pronombres, racismo, inflexión verbal y connotación sexual fueron los mejores propuestos en todos estos experimentos. Esto puede deberse a que son determinantes particulares de cada clase. Por ejemplo, en una definición general, los atributos *prieta*. De las categorías de racismo, pronombres e inflexión verbal se relacionan inmediatamente con la clase agresiva. Así mismo, los atributos de connotación sexual con la clase de vulgaridad y en sí la ofensividad también con las categorías de los pronombres y la inflexión verbal. Aunque parecería que la clase de ofensividad es la que contaba con menos variedad de atributos en este último experimento, aparentemente fue suficiente. Y lo anterior incluso se ve reflejado en la matriz de confusión, ya que la clase que cuenta con más verdaderos positivos fue la de ofensividad. Por lo anterior, se puede decir que observar fenómenos lingüísticos como la presencia de pronombres y su efecto en la conjugación y, en otras palabras, o la presencia de un interlocutor del enunciado es clave de este tipo de lenguaje.

Capítulo 5. Conclusiones

En este trabajo se mostró el uso de criterios lingüísticos en la detección automática de agresividad. En este sentido, recordando la pregunta de investigación general planteada inicialmente: ¿tomando en cuenta los elementos lingüísticos que se usaron para etiquetar un corpus, se podrán usar como atributos en un sistema de clasificación automática? Para contestar esta pregunta se efectuó un análisis del lenguaje usado en los tweets. Por esto, basado en el proyecto de elaboración de un diagrama de etiquetado de ofensividad en un corpus de tweets elaborado por el LabTL del INAOE, se observaron los buenos resultados obtenidos del diagrama y se tomaron estas propuestas en una aplicación diferente. Cuando se analizó el corpus para diseñar el diagrama, se percibió que la clase de ofensividad era demasiado reducida y que no englobaba todas las características del lenguaje usado en los tweets. Por esto, se propusieron tres clases: ofensividad, agresividad y vulgaridad. Posteriormente, se describieron y se introdujo la posibilidad de usar el acto ilocutivo para una mejor caracterización de los mensajes. Así, se usaron todos estos elementos en el diagrama. Dicho proyecto sólo se realizó para etiquetar un corpus y se obtuvieron buenos resultados. Por lo tanto, en este trabajo se tomaron estas tres clases para clasificar los tweets y las características lingüísticas como atributos. Las características del diagrama eran generales, por lo que en este trabajo se realizó una revisión de la literatura analizando las definiciones de estas tres clases y las características que se les atribuía. Con base en esto se construyeron listas de palabras que respondan a estas características y de esta manera las clases se operacionalizaron, haciéndolas concisas y aplicables al proceso de clasificación. Para esto se elaboró una lista de palabras peyorativas, misóginas y despectivas, así como una mejor caracterización de las clases, que fueron suficientes para usarse como atributos. Posteriormente se realizó la clasificación de los tweets con estos atributos y en las tres clases. Por cada experimento se hizo un análisis de resultados y detección de posibles problemas

que se intentaron solucionar con un experimento posterior. Así, se llevaron a cabo un total de cuatro experimentos.

Recordando las dos hipótesis derivadas de la pregunta de investigación inicial, la primera buscaba comprobar si se pueden clasificar los tweets en tres clases con una precisión significativa. La segunda proponía probar si los elementos lingüísticos propuestos en el diagrama pueden ser usados como atributos para la clasificación y tener igualmente resultados significativos. Acerca de la primera hipótesis, aquí se logró usar las tres clases para la clasificación y esto permitió obtener resultados significativos. Para poder apreciar los resultados aquí obtenidos se puede recordar lo mencionado sobre el workshop de GermEval, donde se plantearon una serie de tareas similares. Allí, la tarea que estableció múltiples clases obtuvo el resultado más bajo con un 52% de valor F. Por otro lado, en esta investigación se obtuvo un valor F más alto siendo 86% con el último experimento, mayor que en el GermEval. Por consiguiente, se puede concluir que se tuvo más éxito en clasificar textos con tres clases en este trabajo. Respecto a la segunda hipótesis, aquí se logró hacer una clasificación con buenos resultados usando atributos lingüísticos; obteniendo un valor F de 0.73 tan sólo en el primer experimento usando todos los atributos. Este resultado se puede comparar con el valor F de 0.48, obtenido por el mejor equipo participante en el MEX-A3T que realizó la misma tarea, pero con un enfoque diferente. Analizando los resultados obtenidos al contestar estas dos hipótesis, se puede afirmar que un enfoque lingüístico es más eficaz incluso cuando la clasificación se realiza en varias clases, tres en este caso. Sobre todo, es importante enfatizar que, en un ejercicio de menor tamaño presentado en esta tesis, pero con datos obtenidos al azar, se logró obtener un resultado que rebasa el estado del arte de esta tarea en experimentos

a gran escala. Con base en estos resultados, en el futuro se podría entonces partir de este trabajo y aplicarlo a datos más numerosos.

Finalmente, se debe tomar en cuenta que las condiciones en las cuales se elaboró este trabajo son diferentes a las que se usaron en las investigaciones con cuyos resultados se han comparado los de esta investigación. Por esto, se deberían buscar maneras en las que los experimentos puedan ser expandidos para corroborar los resultados obtenidos.

5.1 Trabajo Futuro

Debido a que se obtuvieron resultados significativos en la clasificación de tweets, se propone que más adelante se aplique el modelo aquí presentado a un conjunto de datos más numeroso para confirmar estos resultados. Sin embargo, estos resultados son ya altamente significativos. Esto porque en el ejercicio similar del MEX-A3T no hubo algún equipo con resultados parecidos o con un enfoque lingüístico que responda específicamente a las necesidades del lenguaje a analizar. Por ello, ampliar la investigación es sumamente relevante. En vista de que la muestra utilizada consistía de 30 tweets, se recomendaría aplicarlo en el conjunto de prueba del corpus del INAOE completo, es decir 3156 tweets. De esta manera, se tendrán resultados más generalizables debido a que se estaría trabajando con más opciones y ejemplos del lenguaje que se pretende analizar. Para esto, se sugiere que se diseñe un sistema de clasificación automática en vez de la búsqueda manual de atributos en los tweets. Esto es con el propósito de hacer más eficiente el análisis de los datos. Así mismo, también habría una necesidad de hacer los atributos elementos totalmente específicos. En esta investigación se empezó este proceso con la construcción de listas concretas de palabras para ubicarlas en los

textos dependiendo de cada clase. Por esto, se sugiere que estas listas se usen como base y se extiendan. Lo anterior beneficiaría la identificación de este tipo de mensajes.

En esta investigación se tomó un enfoque más dirigido a la parte semántica de los textos. Por ejemplo, con el uso de bigramas de palabras. Por lo anterior, en un futuro se recomienda extenderlo al análisis sintáctico, usando n-gramas de categorías gramaticales (POS, por sus siglas en inglés). Esto permitiría un mejor análisis de la composición de los textos dependiendo de su clase. Esto también permitiría una experimentación más grande de los n-gramas, como por ejemplo de signos de puntuación y una mezcla de caracteres con palabras, como @usuario + adjetivos.

Los resultados presentados en este trabajo sirven ya como incentivo para realizar las mejoras mencionadas y como prueba de que un enfoque lingüístico en las tareas de clasificación de textos sin dudas presenta ventajas para el análisis de tweets del español de México.

Referencias

Álvarez-Carmona, M. Á., Guzmán-Falcón, E., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Reyes-Meza, V., & Rico-Sulayes, A. (2018). Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. En *Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), (Vol. 6)*. Sevilla: España.

Anthony, L. (2019). AntConc (Version 3.5.8) [Software Computadora]. Tokyo, Japan: Waseda University. Disponible en https://www.laurenceanthony.net/software

Arce Castillo, Á. (1999). Intensificadores en español coloquial. *Anuario de Estudios Filológicos*, *22*(1), 37–48.

Aronoff, S. (1985). The minimum accuracy value as an index of classification accuracy. Photogrammetric Engineering and Remote Sensing, 51(1), 99-111.

Ávila-Cabrera, J. J. (2015). Propuesta de modelo de análisis del lenguaje ofensivo y tabú en la subtitulación. VERBEIA. *Revista de Estudios Filológicos. Journal of English and Spanish Studies*, 0, 8-27.

Bermúdez Bausela, M. (2016). The Importance of Corpora in Translation Studies: A Practical Case. Research-publishing.net. Research-publishing.net.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin, 70*(2), 213–220.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. *In Eleventh international aaai conference on web and social media*.

Dobrić, N. (2016). Corpora in Applied Linguistics : Current Approaches. Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

Eibe Frank, Mark A. Hall, and Witten I. H. (2016). The WEKA Workbench. Online Appendix for "*Data Mining: Practical Machine Learning Tools and Techniques*", Morgan Kaufmann, Fourth Edition, 2016.

Esparza, G. G., Fuentes, M. D. L. M., del Real, T. A. R., & Reich, J. C. (2017). Un modelo basado en el Clasificador Naïve Bayes para la evaluación del desempeño docente. *Revista Iberoamericana de Educación a Distancia*, *20*(2), 293–313.

Fichman, P., & Sanfilippo, M. R. (2015). The Bad Boys and Girls of Cyberspace: How Gender and Context Impact Perception of and Reaction to Trolling. *Social Science Computer Review*, *33*(2), 163–180. https://doi.org/10.1177/0894439314533169

Fletcher, S., & Islam, Md. Z. (2019). Decision Tree Classification with Differential Privacy: A Survey. *ACM Computing Surveys*, *52*(4), 1–33. https://doi.org/10.1145/3337064

Gall, O. (2004). Identidad, exclusión y racismo: reflexiones teóricas y sobre México. *Revista mexicana de sociología,* 66(2), 221-259.

Gibbons, C., Richards, S., Valderas, S., & Campbell, J. (2017). Supervised Machine Learning Algorithms Can Classify Open-Text Feedback of Doctor Performance With Human-Level Accuracy. *Journal Of Medical Internet Research*, *19*(3), 1438–8871.

Grundy, P. (2008). *Doing pragmatics* (3rd ed). London: Hodder Education.

Gutiérrez Esparza, G., Margain Fuentes, M. de L., Ramírez del Real, T. A., & Canul Reich, J. (2017). Un modelo basado en el Clasificador Naïve Bayes para la evaluación del

desempeño docente. *Revista Iberoamericana de Educación a Distancia*, *20*(2), 293–313.

Guzmán Falcón, E. (2018) *Detección de lenguaje ofensivo en Twitter basada en expansión automática de lexicones* [Tesis de Maestría]. Instituto Nacional de Astrofísica, Óptica y Electrónica. http://inaoe.repositorioinstitucional.mx/jspui/handle/1009/1722

Hall, M. A. (1998). Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand.

Holmes, J. (2013). *An introduction to sociolinguistics* (4. ed). London: Routledge.

Hughes, G. (1991). *Swearing: A social history of foul language, oaths, and profanity in English*. Oxford, UK ; Cambridge, Mass., USA: Blackwell.

IberEval. (2018). Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018). Recuperado el 15 de noviembre de 2019, de Proceedings website: http://ceur-ws.org/Vol-2150/

Instituto de las Mujeres del Distrito. (2016). *CDMX Ciudad Segura y Amigable para la Mujeres y las Niñas*. Recuperado de Programa Anual PAIMEF.

Islas Asaïs, H. (2005). *Lenguaje y discriminación* (1a ed.). Recuperado de http://repositorio.dpe.gob.ec/handle/39000/944

Jeshion, R. (2013). Expressivism and the offensiveness of slurs. *Philosophical Perspectives*, *27*, 231–259.

Kaye, B. K., & Sapolsky, B. S. (2009). Taboo or Not Taboo - That is the Question: Offensive Language on Prime-Time Broadcast and Cable Programming. *Journal of Broadcasting and Electronic Media*, (Issue 1), 22.

Kemp, S. (2019). *The Global State of Digital in 2019 Report* (Núm. 4; p. 221). Londres: Hootsuite.

Khalaf, A. S., & Rashid, S. (2019). Pragmatic Functions of Swearwords in the Amateur Subtitling of American Crime Drama Movies into Arabic. *International Journal of Asia Pacific Studies*, *15*(1), 97–131. https://doi.org/10.21315/ijaps2019.15.1.4

Kock, J. de, Delbecque, N., & Paepe, C. de (Eds.). (1998). *Estudios en honor del profesor Josse de Kock*. Leuven: Leuven University Press.

Konvens. (2019). Konvens 2019. Recuperado el 2 de septiembre de 2019, de GermEval website: https://2019.konvens.org/germeval

Kusumaningsih, D., Santosa, R., Subroto, E., & Djatmika. (2019). Pedagogical Values in Indonesian Lyrics of Dangdut Songs: Evidences of Language Vulgarism and Gender Exploitation. *Journal of Social Studies Education Research*, *10*(3), 311–331.

Leech, G. (2005). "Adding Linguistic Annotation" in *Developing Linguistic Corpora: a Guide to Good Practice*, ed. M. Wynne. Oxford: Oxbow Books: 17-29. Available online from http://ota.ox.ac.uk/documents/creating/dlc/

McEnery, T., & Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Marciani Burgos, B. (2013). El lenguaje sexista y el hate speech: un pretexto para discutir sobre los límites de la libertad de expresión y de la tolerancia liberal. *Revista Derecho del Estado* [fecha de Consulta 30 de Agosto de 2019]. ISSN: 0122-9893

Mosquera, R., Castrillón, O. D., & Parra, L. (2018). Máquinas de Soporte Vectorial, Clasificador Naïve Bayes y Algoritmos Genéticos para la Predicción de Riesgos Psicosociales en Docentes de Colegios Públicos Colombianos. *Información Tecnológica*, *29*(6), 153–162. https://doi.org/10.4067/S0718-07642018000600153

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. Political Behavior, 39(3), 629-649.

Mosquera, R., Castrillón, O. D., & Parra, L. (2018). Máquinas de Soporte Vectorial, Clasificador Naïve Bayes y Algoritmos Genéticos para la Predicción de Riesgos Psicosociales en Docentes de Colegios Públicos Colombianos. *Información Tecnológica*, *29*(6), 153–162. https://doi.org/10.4067/S0718-07642018000600153

OpenCor. (2019). Latin American and Iberian Languages Open Corpora Forum. Recuperado el 11 de octubre de 2019, de Latin American and Iberian Languages Open Corpora Forum website: http://opencor.gitlab.io/es/

PAN. (2019). PAN @ CLEF 2020. Recuperado de https://pan.webis.de/clef20/pan20-web/index.html#index-tasks

Penco C., (2017). Prejudice and Presupposition in Offensive Language. *Nordicum-Mediterraneum*, (3), A2.

Poliakov, Leon, (1996). *The Aryan Myth: A History of Racist and Nationalistic Ideas In Europe* (Barnes & Noble Books ISBN 0-7607-0034-6.

Raghvan, V., & Gwang, J. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS*, *7*(3), 205–229.

REDTTL - Red Temática en Tecnologías del Lenguaje. (2019). MEX-A3T. Recuperado el 1 de septiembre de 2019, de MEX-A3T: Authorship and aggressiveness analysis in Twitter website: https://sites.google.com/view/mex-a3t

Rico Sulayes, A. (2014). *De vulgaridades, insultos y malsonancias: El diccionario del subestándar mexicano* (1a ed.). Mexicali: Universidad Autónoma de Baja California.

Rico Sulayes, A. (2017). Reducing Vector Space Dimensionality in Automatic Classification for Authorship Attribution. *Revista Ingeniería Electrónica, Automática y Comunicaciones*, *38*(3), 26–35.

RuG at GermEval: Detecting Offensive Speech in German Social Media. (2018). *Proceedings of the GermEval 2018 Workshop*, 63–70.

Ruppenhofer J., Siegel M., Wiegan M., (2018). Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. *Proceedings of the GermEval 2018 Workshop* (pp.1-10). Austrian Academy of Sciences.

Rösner, L., & Krämer, N. C. (2016). Verbal Venting in the Social Web: Effects of Anonymity and Group Norms on Aggressive Language Use in Online Comments. *Social Media + Society*, 2(3), 205630511666422. https://doi.org/10.1177/2056305116664220

Saheed, Y. K. 1., Oladele, T. O, Akanni, A. O. 1., & Ibrahim, W. M. (2018). Student Performance Prediction Based on Data Mining Classification Techniques. *Nigerian Journal of Technology*, 37(4), 1087–1091.

Samovar, L. A., Porter, R. E., & McDaniel, E. R. (2015). *Communication between cultures* (9. ed). Boston, Mass: Wadsworth Cengage Learning.

Santos, C. N. D., Melnyk, I., & Padhi, I. (2018). *Fighting offensive language on social media with unsupervised text style transfer*. arXiv preprint arXiv:1805.07685.

Searle, J. R. (1976). A classification of illocutionary acts. *Language in society*, *5*(1), 1–23.

Sebastián Canós, J. (2018). Misogyny identification through SVM at IberEval 2018. *IberEval*, *2150*, 229–233. España.

Sidorov, G. (2013). Clasificación de actos de habla en diálogos basada en los verbos de habla. *Research in Computing Science*, *68*, 137–153.

Soto, C., & Jiménez, C. (2011). Aprendizaje supervisado para la discriminación y clasificación difusa. *Dyna*, 78(169), 26-33.

Teresa Turell, M. (2011). The Language of Defamation Cases. *International journal of speech language and the law*, 18(1), 169–173. https://doi-org.udlap.idm.oclc.org/10.1558/ijsll.v18i1.169

Torruella, J., & Llisterri, J. (1999). Diseño de corpus textuales y orales. Filología e informática. *Nuevas tecnologías en los estudios filológicos*, 45-77.

Wilson, K. (2017). Offensive Language and Serious Harm: Application of the Films, Videos, and Publications Classification Act 1993. Victoria University of Wellington Law Review, (Issue 1), 163.

Xu, S., Yang, X., Yu, H., Yu, D.-J., Jingyu, Y., & Tsang, E. C. (2015). Multi-label learning with label-specific feature reduction. *Knowledge-Based Systems*, *104*, 52–61.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). *arXiv:1903.08983 [cs]*. Recuperado de http://arxiv.org/abs/1903.08983

Anexos

Anexo 1. Verbos de motivación (Sidorov, G.,2013, p.146-147).

6	Motivación	1 Ordenar 2 Encargar 3 Incitar 4 Pedir 5 Invitar 6 Llamar	1 dictar, ordenar, requerir, decretar, mandar, disponer, comandar, // mandato, orden, requerimiento, consigna, mandato, disposición 2 cometer, confiar, dejar, delegar, encomendar, recomendar // cargo, comisión, encargo, encomienda, orden, tarea, cometido, comisión,
		7 Mandar 8 Inclinar 9 Permitir 10 Prohibir	delegación, embajada, mandato, requisitoria, requisitorio 3 impulsar, inducir, insinuar, incitación, aludir, // alusión, insinuación, reticencia 4 demandar, impetrar, pedir, peticionar, postular, pretender, recurrir, solicitar, suplicar, // alocución, alzada, dirección, manejo, pedida, recurso, ruego, solicitud, suplicación, trata, exhorto, instancia, pedida, pedido, pedimento, petición, postulación, requerimiento, ruego, solicitud, súplica 5 invitar, // invitación 6 llamarse, nombrar, // llamado 7 dictar, ordenar, requerir // mandato, orden, requerimiento, 8 disponer, colocar, decidir, determinar, inclinar, poner en orden, // inclinación 9 permitir, dejar, autorizar, admitir, conceder, aceptar 10 defender, desaprobar, interdecir, prohibir, proscribir, vedar