UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS
FACULTAD DE INGENIERÍA
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

INFLUENCIA DEL APRENDIZAJE COMPUTACIONAL BASADO EN TÉCNICAS DE MINERÍA DE TEXTOS EN LA CLASIFICACIÓN DE COMENTARIOS DE TEXTOS AGRESIVOS

Presentado por

ROSMERY JIMENEZ PALOMINO

TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

ANDAHUAYLAS – APURÍMAC – PERÚ
2018

UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS
FACULTAD DE INGENIERÍA
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

Presentado por

ROSMERY JIMENEZ PALOMINO

INFLUENCIA DEL APRENDIZAJE COMPUTACIONAL BASADO EN TÉCNICAS DE MINERÍA DE TEXTOS EN LA CLASIFICACIÓN DE COMENTARIOS DE TEXTOS AGRESIVOS

Asesor:

MSc. Herwin Alayn Huillcen Baca

ANDAHUAYLAS – APURÍMAC – PERÚ
2018

DEDICATORIA

Dedico este trabajo de investigación a Dios y a mis padres por haberme dado la vida y por estar siempre conmigo cuando más los necesito. Para ustedes papitos todo mi corazón y mi eterno amor.

Rosmery,

AGRADECIMIENTO

A Dios por permitirme estar aquí.

A mis padres Julián y Visitación por haberme dado la vida y por su apoyo incondicional para cumplir mis metas.

A mis hermanos Ronald y Beatriz por respaldar mis decisiones tomadas.

A mi asesor, el MSc. Herwin Alayn Huillcen Baca, por haberme guiado en la realización de este proyecto de investigación.

A mi mejor amiga, Pamela, por su apoyo moral.

A todos ellos, quiero expresarles mi más profundo agradecimiento y decirles que siempre los tendré en mi corazón.

“Porque la vida es vértigo y no una carrera. La vida es búsqueda y no una guarida”

TABLA DE CONTENIDO

LISTA DE IMÁGENES

Pág.

LISTA DE TABLAS

Pág.

RESUMEN

El ciberbullying es problema mundial y nuestro país no es ajeno a ello. Este va incrementado y las personas que son víctimas de ello son cada vez más. El propósito de esta investigación es ayudar a mitigar este problema, para lo cual se propone un modelo de clasificación de comentarios de textos agresivos.

La metodología que se empleó fue KDD (Knowledge Discovery in Databases), esta nos permitió determinar la influencia del corpus de conocimiento (Database) en la clasificación de comentarios de textos agresivos, basados en las técnicas de minería de textos.

Las etapas que se desarrollaron fueron: 1) Selección de datos, 2) Procesamiento, 3) Transformación, 4) Data Mining, 5) Interpretación y evaluación. En la primera etapa, se extrajo los comentarios de textos de las redes sociales. En la segunda etapa, se realizó la limpieza de los datos. En la tercera etapa, se convirtió los datos al formato ARFF de la herramienta weka. En la cuarta etapa, se hizo usó de las técnicas y algoritmos de clasificación de weka. En la quinta etapa, se evaluó e interpretó los resultados obtenidos de los algoritmos NaiveBayes, BayesNet, DecisionStump, J48, OneR, PART, SMO, IBK y KStart.

El menor Error Absoluto Medio fue de 0.09 adquirido por el algoritmo IBK con una data de 1561 comentarios de textos, teniendo una precisión de clasificación del 89.3%. Con su modelo de clasificación, se implementó un servicio web de clasificación de comentarios de textos agresivos al cual se denominó “duke classifier”.

El servicio web duke classifier, puede ser empleado por cualquier persona que desee integrarlo a su blog, microblogging, red social, etc. O estudiado por otros investigadores que estén interesados en el tema.

Palabras claves: Técnicas de minería de textos, weka, modelo de clasificación, corpus conocimiento, KDD, servicio web, análisis de sentimientos.

ABSTRACT

Cyberbullying is a global problem and our country is no stranger to it. This is growing and people who suffer from this are increasing. The purpose of this research is to help mitigate this problem. We propose a classification model of aggressive texts.

The methodology used was KDD (Knowledge Discovery in Databases), This allowed us to determine the influence of the corpus of knowledge (Database) in the classification of comments of aggressive texts, based on text mining techniques.

The stages that were developed were: 1) selection of data, 2) processing, 3) transformation, 4) data mining, 5) interpretation and evaluation. In the first stage, comments were extracted from texts of social networks. In the second stage, data cleaning was performed. In the third stage, the data was converted to the ARFF format of the weka tool. In the fourth stage, weka techniques and algorithms were used. In the fifth stage, the results obtained from the NaiveBayes, BayesNet, DecisionStump, J48, OneR, PART, SMO, IBK and KStart algorithms were evaluated and interpreted.

The lowest Mean Absolute Error was 0.09 acquired by the IBK algorithm with a data of 1561 textual comments, having a classification accuracy of 89.3%. With its classification model, a web service of classification of comments of aggressive texts was implemented to which it was denominated "duke classifier".

The web service duke classifier can be used by anyone who wants to integrate it into your blog, microblogging, social network, etc. Or studied by other researchers who are interested in the topic.

Key words: Mining techniques of texts, weka, classification model, knowledge corpus, KDD, web service, and sentiment analysis.

INTRODUCCIÓN

En la medida que el tiempo avanza, son cada vez más las personas que usan las redes sociales, blogs, microbloggings, foros, etc. que ofrece la red informática mundial WWW (World Wide Web), para que puedan emitir sus opiniones e ideas a través de comentarios, imágenes, videos, etc. En la actualidad, existen muchas personas que alentadas por el anonimato y/o distancia, vierten publicaciones, opiniones, etc. Con exagerada libertad, causando daño a sus víctimas, lo cual nos conlleva a hablar del gran problema de la WWW, que es el ciberbullying.

El ciberbullying es un problema muy complejo y difícil de abordar, se da de diversas maneras, ya sea a través de comentarios de textos, imágenes, videos, audios, etc. Actualmente existen varios estudios relacionados a la clasificación de opiniones, basados en el análisis de sentimientos, que combina técnicas de minería de texto y Procesamiento del Lenguaje Natural (PLN). Aunque haya varios trabajos relacionados a la temática, la mayoría usan textos en inglés, y realiza una clasificación en dos categorías que son negativa o positiva. Sin embargo, no te indican el grado de agresividad (ya se nominal o numérica) de los comentarios de textos. Es por ello que esta investigación tiene por objetivo general determinar la influencia del corpus de conocimiento en la clasificación de comentarios de textos agresivos, basados en las técnicas de minería de textos. El corpus de conocimiento está conformado por comentarios textos extraídos principalmente de las redes sociales de Facebook, twitter y youtube.

Combinado el corpus de conocimiento con las técnicas de minería de textos, se obtuvo varios modelos de clasificación. Con el modelo de clasificación que dio mejores resultados, se implementó un servicio web que clasifica comentarios de textos en las siguientes categorías: Muy agresivo, agresivo, poco agresivo o neutro.

CAPITULO I

1. PLANTEAMIENTO DEL PROBLEMA

1.1 Descripción del Problema

Con el inicio de la Web 2.0, muchos usuarios comenzaron a emplear las redes sociales para compartir sus ideas, opiniones acerca de algún producto, servicio e incluso de personas; varias de estas opiniones vertidas son hechas de buena manera, sin embargo, existen personas que alentadas por la distancia y a veces por el anonimato, utilizan los diferentes blogs, microblogging, redes sociales para emitir comentarios agresivos (insultos, vulgarismo, etc.).

Durante este tiempo se ha venido lidiando con un gran problema que es el ciberbullying. Según Smith (2006), el ciberbullying es un acto agresivo e intencionado llevado a cabo de manera repetida y constante, mediante el uso de formas de contacto electrónicas por parte de un grupo o de un individuo contra una víctima que no puede defenderse fácilmente (como se citó en Gonzáles, Márquez y López, 2015, p.213). Willard (2012), lo define como el envío y acción de colgar textos o imágenes dañinas o crueles por Internet u otros medios digitales de comunicación. Asanza, Flores y Berrones (2014), mencionan que las principales consecuencias que posee el ciberbullying son: Tensión, autoestima baja, pesimismo, tristeza, aversión al ambiente en que se encuentre, suicidio.

Un estudio general la dio a conocer el periódico Universia de España (2016), que revela datos increíbles sobre el avance del cyberbullying y sus consecuencias,1 de cada 4 casos de bullying es a través de las redes. Así mismo el estudio realizado por la fundación Mutua Madrileña y la fundación ANAR, muestra datos impresionantes sobre el acoso escolar desde las redes sociales; según la investigación, el 70% de las víctimas son mujeres y el 92% de los afectados presenta secuelas psicológicas; el 81% de los casos de cyberbullying se realizan desde WhatsApp, el 36% desde las redes sociales; ansiedad, depresión, soledad o baja autoestima son los problemas psicológicos más comunes entre las víctimas y la gran cantidad de niños sienten y experimentan pensamientos suicidas, creyendo que es la única forma de escapar (Universia, 2016).

Cataldi (2013) menciona que, tras el caso de Amanda Todd, la Asociación Americana de Pediatras publicó un estudio en su conferencia nacional que muestra que el ciberacoso no es la única razón que lleva al suicidio en la mayoría de los casos; sino también los problemas relacionados al ciberbullying; el estudio revela que hasta el 78% de los adolescentes que termina por suicidarse, fue acosado en la red y en la vida real.

Cabe indicar también que ESET (2013), desarrolló una encuesta en Latinoamérica en la que participaron cientos de jóvenes, de la cual se extrajo varias conclusiones; la principal es que el 30,7% aseguró haber sufrido de ciberacoso; otra conclusión que arrojó la encuesta fue que el ciberacoso se da a través de distintas formas: Burlas (52,6%), insultos (49,7%), apodos denigrantes (32,9%), amenazas (31,8%), difamación (27,2%); en cuanto a los medios en los que se dio la situación de maltrato, predominaron las redes sociales como Facebook, Twitter y Google+, con el 82,7% de los votos.

El ciberbullying es un problema mundial, es por ello que nuestro país no es ajeno a esta realidad. A continuación, se plasmará los problemas causados en el Perú.

RPP (2013), señala que Zoyla Uriarte Nuñez, ex directora de la UGEL de Chiclayo del departamento de Lambayeque, reconoció que se reportaron casos de “ciberbullying” en colegios secundarios de la ciudad; la funcionaria explicó que fueron los padres de familia quienes denunciaron ante esta instancia que sus hijos eran insultados y maltratados con fotos que publicaban sus amigos en las redes sociales para originar la burla colectiva del grupo y de otros estudiantes; mencionó la preocupación de que no haya un adecuado manejo de las redes en los alumnos y que ahora los maltratos ya no sean solo físicos, sino también a través de internet.

Así mismo en una publicación realizada por La República (2013), se muestra que al menos el 12.1% de los peruanos, entre los 10 y 18 años de edad, sufre del denominado “ciberbullying”; según lo que informó el psicólogo especialista en acoso escolar, Roberto Lerner, existen escolares que usan las tecnologías de la información (correo electrónico, teléfonos celulares y las redes sociales) para humillar, insultar o difamar a un tercero.

Perú21 (2014), otro periódico informativo del Perú dio a conocer que el ciberbullying se ha convertido en una amenaza que crece en forma alarmante en el Perú; los casos se presentaron principalmente en escolares a partir de los 12 años, según advirtió Cedro (Centro de información y educación para la prevención del abuso de drogas). Jorge Arnao, psicólogo de esa institución, señaló que el ciberbullying va en aumento en la medida en que cada vez hay más adolescentes que acceden a las redes sociales; lo preocupante es que este tipo de hostigamiento que se da a través de las redes sociales y es uno de los factores de suicidio de adolescentes; el especialista además explicó que este fenómeno se manifiesta de diversas formas, de acuerdo a la encuesta a escolares que se realizó entre 12 y 17 años de nueve ciudades del país, el 24% se da a través de la publicación de fotos retocadas e insultos, el 23% por memes, el 16% mediante comentarios ofensivos, 16% por discusiones a través de Facebook, un 7% por publicación de videos personales, entre otros; según explicó Arnao, también se determinó, que en el 77% de casos, los motivos de hostigamiento fueron por discriminación física y sexual, 55% son burlas por características físicas y 22% por el color de la piel.

En una noticia publicada por RPP (2015), se dio a conocer que los Psicólogos de la Comisión Nacional para el Desarrollo y Vida sin Drogas (Devida) explicaron que cada tres de 10 escolares con acceso a Internet manifestaron haber sido víctimas de “ciberbullying”, situación que los expone a desarrollar estrés y ansiedad, así como a consumir alcohol, tabaco y drogas ilegales como la marihuana; también mencionaron que existe asociación entre el bullying y el uso de drogas; ser un escolar víctima de ciberbullying está fuertemente relacionado con indicadores altos de consumo de sustancias: El 32.5% consume alcohol, el 21.3% fuma cigarrillos, el 5.2% tiene un uso reciente de marihuana, el 3.7% al éxtasis, el 2.9% a la cocaína y el 5.1% usa tranquilizantes.

Por todo lo indicado, se concluye que el ciberbullying es un gran problema mundial que va incrementando, así también va aumentando la cantidad de cibernautas que desde el anonimato vierten comentarios ofensivos, sin medir las consecuencias de éstos.

1.2 Formulación del Problema

1.2.1 Problema General

¿De qué manera influye el corpus de conocimiento en el proceso de la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico?

1.2.2 Problemas Específicos

- ¿De qué manera influye el corpus de conocimiento en la elección del mejor algoritmo de clasificación del grupo de algoritmos clasificadores de tal manera que tenga el menor Error Absoluto Medio en la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico?
- ¿De qué manera influye el corpus de conocimiento en la precisión de la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico?
- ¿De qué manera influye el corpus de conocimiento en el tiempo de generación de los Modelos de Clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico?
- ¿De qué manera influye el corpus de conocimiento en la cantidad de datos óptimos para la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico?

1.3 Justificación

Debido a que son cada vez más las personas que son víctimas del ciberbullying en las diversas redes sociales, foros, blogs, etc. El presente proyecto propone un modelo de clasificación de comentarios de textos agresivos. Esta clasificación, se da en cuatro categorías nominales que indican el grado de agresividad de los comentarios de textos.

Para poder hacer posible la clasificación de los comentarios de textos, se generó varios modelos de clasificación, combinando las técnicas de minería de textos con el corpus de conocimiento construido, de los cuales se escogió el de

menor Error Absoluto Medio. Con el modelo más óptimo, se implementó un servicio web que servirá como una herramienta de filtración para las diversas redes sociales, foros, blogs, etc.

El ciberbullying es un problema mundial y por ende un tema complejo de abordar, se da de diversas maneras y una de ellas es a través de los comentarios de textos agresivos, pero estos también varían de acuerdo al lugar geográfico en el que se originan, es decir de país en país, es por ello que en esta investigación se analizó los comentarios de textos agresivos de habla hispana emitidos principalmente por los cibernautas peruanos en las diferentes redes sociales. De esta manera se ayudará a mitigar daños emocionales causados a personas de habla hispana que son víctimas del ciberbullying. Así mismo este proyecto de investigación y el aporte que se da a través del mismo, estará al alcance de cualquier investigador que quiera seguir realizando estudios relacionados a este tema.

1.4 Objetivos

1.4.1 Objetivo General

- Determinar la influencia del corpus de conocimiento en el proceso de la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico.

1.4.2 Objetivos Específicos

- Determinar la influencia del corpus de conocimiento en la elección del algoritmo de clasificación del grupo de algoritmos clasificadores de tal manera que tenga el menor Error Absoluto Medio en la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico.
- Determinar la influencia del corpus de conocimiento en la precisión de la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico.
- Determinar la influencia del corpus de conocimiento en el tiempo de generación de los Modelos de Clasificación de comentarios de textos

agresivos, basadas en las técnicas de minería de textos en un análisis léxico.

- Determinar la influencia del corpus de conocimiento en la cantidad de datos óptimos para la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico.

CAPITULO II

2. MARCO TEORICO

2.1 Antecedentes

2.1.1 Antecedentes a Nivel Nacional

De acuerdo con Samuel Vieyra, Consultor en Investigación y Desarrollo, en Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación (INFOTEC) existen varias aplicaciones para la Minería de textos, de las cuales destacan: Extracción de información, análisis de sentimientos o minería de opiniones, clasificación de documentos, elaboración de resúmenes, extracción de conocimiento (Lozada, 2016).

En la última década, el análisis de sentimientos, también conocido como minería de opiniones, ha despertado un creciente interés sobre todo para las marcas, que necesitan saber cuánta aceptación tienen sus productos en las diferentes personas que la consumen, empero, el análisis de opiniones es tan amplio que puede abarcar diferentes temas, a continuación, se muestra diferentes estudios realizados en este campo.

Modelo de Clasificación de Opiniones Subjetivas en Redes Sociales

La Revista Ciencia Tecnología e Innovación, muestra un artículo de Tapia, Ruiz y Chirinos (2014), en el cual se plasma el proyecto denominado “Modelo de clasificación de opiniones subjetivas en Redes Sociales”, que aplica las técnicas de Minería de datos para interpretar la opinión subjetiva (favorable o desfavorable) de los usuarios en las redes sociales, y analizar tendencias de un determinado producto, servicio o persona que sirva para la aplicación de marketing a fin de generar ventajas competitivas haciendo el uso de la analítica en la web 2.0. El Modelo usa técnicas de Minería de datos orientado al procesamiento de texto, específicamente utiliza técnicas de clasificación para generar descubrimiento a partir de predicciones que ejercen los atributos a una clase, el algoritmo de clasificación que se emplea es Redes Bayesianas. Los atributos a evaluar, que serán los predictores están conformadas por las palabras que se usaron para entrenar el modelo. Las palabras están contenidas en los comentarios que emiten los usuarios de las redes sociales,

estos comentarios se crearon a través de un sistema que se diseñó exclusivamente para esta investigación, teniendo como modelo una red social de Microblogging, específicamente Twitter. Con la utilización de un clasificador bayesiano de cálculo de probabilidad los resultados que se obtuvieron en la investigación fue la siguiente: Clasificación de 40% de opiniones favorables, 40% de opiniones desfavorables y un 20% de opiniones indeterminadas, con lo cual se puede ver que este clasificador es óptimo para este tipo de tareas.

Modelo computacional de minería de microblogs para el análisis del comportamiento del consumidor de telefonía celular

En la investigación titulada “Modelo computacional de minería de microblogs para el análisis del comportamiento del consumidor de telefonía celular” de Apaza (2016), menciona que los mensajes de Twitter están siendo cada vez más usados para determinar el sentimiento de los consumidores de servicios o productos. En este trabajo, se usó diversas técnicas computacionales, como son los modelos de aprendizaje de máquina. Para el uso adecuado de estas técnicas se realizaron diversos pasos desde la etapa del pre–procesamiento hasta la evaluación. Así mismo nos muestra el resultado del proceso de aplicación de diversas técnicas de Análisis de Sentimiento para poder asignar una polaridad positiva, negativa o neutral a los tweets de los usuarios de telefonía celular del Perú, teniendo como finalidad poder identificar el comportamiento que presentan los clientes de las empresas de telefonía celular a través de las opiniones vertidas en la red social Twitter. Para ello se siguió los siguientes procedimientos: a) Recolección de tweets de los seguidores de las empresas de telefonía celular (se extrajeron 26,917,539 tweets); b) Pre–procesamiento de la data obtenida para poder identificar elementos importantes de cada tweet; c) Filtrado de elementos poco relevantes, o ruido; y d) Clasificación de cada publicación basado en las características obtenidas en etapas previas. Además, se incluyó un diccionario de lexicones que es un conjunto de palabras que ya poseen una polaridad de sentimiento y además con ellas se pueden clasificar textos que las contengan. Los resultados obtenidos muestran que la introducción de un diccionario de lexicones incrementó el número de términos que pueden ser considerados para la clasificación. Así mismo, el uso de este diccionario al cual se le aumento nuevos términos permitió incrementar la tasa de clasificación en un 0,75%.

Finalmente, con ayuda de las técnicas de análisis de sentimiento, es posible explotar el contenido de redes sociales de manera que puedan servir a las corporaciones para la toma de decisiones, especialmente de servicio a sus usuarios.

2.1.2 Antecedentes a Nivel Internacional

Una aproximación supervisada para la minería de opiniones sobre tuits en español en base a conocimiento lingüístico.

Vilares, Alonso y Gómez (2013), describen un sistema para la clasificación de la polaridad de twitters escritos en español (si la opinión es positiva, negativa o mixta). Se adopta una aproximación hibrida, que combina conocimiento lingüístico obtenido mediante PLN (procesamiento de lenguaje natural) con técnicas de aprendizaje automático. Como paso previo, se realiza una primera etapa de reprocesado para tratar ciertas características del uso del lenguaje en Twitter que consiste básicamente en la eliminación de ruidos de los comentarios de texto a través de los siguientes pasos: Tratamiento de emoticonos que se emplea para reflejar el estado de ánimo, para su identificación se utilizó cinco etiquetas muy positivo (emp), positivo (ep), neutro (eneu), negativo (en) y muy negativo (emp) recogido de Agarwal (2011); la normalización de URL's; las correcciones de abreviaturas más frecuentes, la normalización de risas, el tratamiento de elementos de Twitter (“@”y “#”); lo que da un aporte para poder tratar los comentarios de texto con ruidos. Se extrajo información morfológica, sintáctica y semántica, para utilizarla posteriormente como entrada a un clasificador supervisado. La propuesta se lleva a cabo sobre el corpus tass 2012, el principal aporte que da esta investigación para el proyecto a realizar es el panorama del tratamiento de twitters y los problemas con los que nos podemos encontrar al tratar con comentarios de textos en las redes sociales (ruidos).

Procesamiento de Lenguaje Natural en Sistemas de Análisis de sentimientos

Esta investigación, realizada por Dubiau (2013), muestra el análisis, evaluación y la comparación experimental de las técnicas de Procesamiento de

Lenguaje Natural para el análisis de sentimientos en textos no estructurados. Para validar esta investigación, se desarrolló una herramienta de análisis de sentimientos para textos en el idioma español, que tiene como objetivo clasificar documentos automáticamente según la polaridad de emociones (positivos o negativos) y provee las métricas necesarias para evaluar la performance de los distintos modelos de clasificación.

En la etapa de experimentación se utilizó dos sitios de críticas online como casos de estudio, siendo estás Guía Óleo (Críticas sobre Restaurantes) y Google Play (Críticas sobre Aplicaciones para Android), con los cuales se construyó un conjunto de datos principales con el que se realizó los experimentos de análisis de sentimientos y otro conjunto de datos que nos servirá para realizar pruebas a las que se les conoce como out-of-domain testing, siendo esta experiencia muy importante en modelos supervisados para analizar la efectividad de los clasificadores de forma objetiva.

Los modelos supervisados de Naïve Bayes, MaxEnt y SVM resultaron ser muy apropiados para clasificar información subjetiva en idioma español alcanzando hasta un 95% de accuracy (precisión) para los tamaños de corpus más grandes mientras que técnicas basadas en árboles de decisión no proveen buenos resultados. Además, se adaptó la técnica de clasificación no supervisada de Turney al idioma español con el que se obtuvo excelentes resultados considerando que no se requiere información previa ni un conjunto de datos de entrenamiento.

2.2 Bases Teórico Científicas

2.2.1 Base teórica (variable independiente)

X: Aprendizaje computacional.

El aprendizaje computacional se llevó a cabo combinando el corpus de conocimiento con las técnicas de minería de datos, para ello se definió qué es el corpus de conocimiento y las técnicas de minería de datos.

Corpus de Conocimiento: Es también conocido como conjunto de entrenamiento, son los datos utilizados en el entrenamiento de los algoritmos

de minería de datos. Estos datos ya tienen etiqueta, es decir son ejemplos ya clasificados (Álvarez, *et al,* 2007, p.112).

Técnicas de Minería de datos:

Como describe García y Molina (2012) la minería de datos utiliza varias técnicas, estas técnicas se clasifican en dos grandes categorías: Supervisada o predictiva y no supervisada o descriptiva. Para la clasificación supervisada se utiliza las siguientes técnicas: Tabla de Decisión, Árboles de Decisión, Inducción de Reglas, Bayesiana, Redes de Neuronas, Lógica Borrosa, Técnicas Genéticas, Clasificadores Funcionales (García et al., 2012). En este trabajo de investigación se trabajó con la clasificación supervisada y las técnicas que se usaron fueron: Naive Bayesina, Árboles de decisión, Reglas de Clasificación y Clasificadores Funcionales.

Las dimensiones de la variable independiente son:

X1: Precisión de la clasificación de comentarios de textos agresivos en categorías.

Para poder darle una definición a la variable independiente X1, primero se buscó el concepto de precisión.

Rojano (2014), menciona que la precisión es el intervalo admisible para el valor de una magnitud medida. Con lo mencionado por Rojano, diremos que la variable X1, es el valor que indicará cuan aceptable es la clasificación de un comentario de texto en una de las siguientes categorías: Agresivo, agresivo, poco agresivo o neutro.

X2: Tiempo de generación del Modelo de Clasificación de comentarios de textos agresivos.

Para poder explicar el concepto de la variable independiente X2, primero se indagó la definición de tiempo y luego de modelo.

Tiempo es una magnitud física que se utiliza para realizar la medición de lo que dura algo que es susceptible de cambio (Pérez y Merino, 2008).

> Un modelo de minería de datos, es un conjunto de datos, estadísticas y patrones que se pueden aplicar a los nuevos datos

> para generar predicciones y deducir relaciones; un modelo se genera aplicando algoritmos a un determinado conjunto de datos (Microsoft, 2016).

Con estas dos explicaciones, se definió la variable X2, como la duración que toma un algoritmo de minería de datos para de la generación de un modelo de clasificación, que posteriormente, servirá para la clasificación de comentarios de textos agresivos.

X3: Elección del algoritmo de clasificación con el menor MAE.

Para poder entender qué es el Error Absoluto Medio, primero se debe conocer la definición de Error Absoluto.

El Error Absoluto, es la diferencia entre el valor de la medida y el valor tomado como exacto (valor verdadero o valor probable), este error puede ser positivo o negativo, y dependerá de la medida del valor real (Arriagada y Peralta, 2008). Mientras que el Error Absoluto Medio, es el promedio de todos los Errores Absolutos (WordPress, 2017).

Se eligió al algoritmo con menor MAE, de tal manera que su promedio de Errores Absolutos de la clasificación de comentarios de textos agresivos sea mínimo (todo ello se observó en los experimentos realizados).

X4: Cantidad de datos óptimos para la clasificación de comentarios de textos agresivos.

Como menciona Han y Kamber (2006), el objetivo de seleccionar atributos (datos), es encontrar un conjunto mínimo de atributos de forma que la distribución resultante de probabilidad de las clases de datos sea lo más próxima de la distribución original.

Para N atributos hay 2N posibles subconjuntos. Por lo ende, una búsqueda exhaustiva del subconjunto óptimo es computacionalmente costoso si N es grande, es por ello que son necesarios métodos heurísticos que exploren el espacio de soluciones, estas técnicas se denominan attribute subset selection o feature subset selection (Han y Kamber, 2000).

2.2.2 Base teórica (variable dependiente)

La variable dependiente es:

Y: Clasificación de comentarios de textos agresivos.

Un comentario es una opinión, parecer, juicio o consideración que una persona hace acerca de otra persona o de algo; dicha mención puede desarrollarse por vía oral o por escrito; ahora con el uso de Internet, los lectores, oyentes o televidentes pueden realizar comentarios con gran facilidad y expresar sus opiniones en los sitios web y blogs (Pérez y Gardey, 2010). Al emitir mensajes agresivos, se pretende causar daño o menoscabar las cualidades de la persona interpelada (López y González, 2013). Así mismo El Gran Diccionario de la Lengua Española (2016), define a la clasificación como el ordenamiento de algo en clases o categorías.

Con este trabajo de investigación, lo que se quiere, es clasificar la variable dependiente en categorías, para ello se analizará las siguientes dimensiones:

Y1: Palabras y/o frases del comentario de texto que transmiten sentimientos negativos.

Y2: Palabras y/o frases que la cultura constituya como negativas.

Y3: Intensidad de las palabras y/o frases para graduar los sentimientos negativos trasmitidos en los comentarios de textos.

2.3 Definición Conceptual

En el presente proyecto, se han identificado varios conceptos que requieren ser detallados pues serán usados frecuentemente durante el desarrollo de este documento, por ello en esta sección se presentan conceptos teóricos relacionados al desarrollo de la solución del problema planteado.

2.3.1 Minería de datos

Existen muchas definiciones de Minería de Datos, pero la que más se aproxima a todas ellas es la que hace el Instituto SAS que define el concepto

de Data Mining como el proceso de seleccionar, explorar, modificar, modelizar y valorar grandes cantidades de datos con el objetivo de descubrir patrones desconocidos que puedan ser utilizados como ventaja comparativa respecto a los competidores (Pérez y Santín, 2006).

Asimismo, “La minería de datos puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos” (Pérez y Santín, 2007, p.1).

2.3.2 Técnicas de Minería de Datos

Como describe García y Molina (2012) la minería de datos utiliza varias técnicas y se clasifican en dos grandes categorías: Supervisada o predictiva y no supervisada o descriptiva. La que se utilizará en la presente tesis es la Técnica supervisada que la describiremos a continuación.

2.3.2.1 Técnicas Supervisadas

Advanced Tech Computing Group UTPL (2008) define que, a diferencia de la no supervisada, en las técnicas supervisadas se cuenta con un conocimiento a priori, lo que me permite la tarea de clasificar un objeto dentro de una categoría o clase, se cuenta con modelos ya clasificados (objetos agrupados que tienen características comunes). Dentro de esta técnica encontramos múltiples algoritmos de predicción y clasificación, pero las que se abarcaron fueron los algoritmos de clasificación.

2.3.3 Algoritmos de clasificación

Recordemos que los sistemas de clasificación supervisados son aquellos en los que, a partir de un conjunto de ejemplos clasificados (conjunto de entrenamiento), intentamos asignar una clasificación a un segundo conjunto de ejemplos (Sancho, 2016). Para la clasificación supervisada se utiliza las siguientes técnicas: Tabla de Decisión, Árboles de Decisión, Inducción de Reglas, Bayesiana, Redes de Neuronas, Lógica Borrosa, Técnicas Genéticas (García et al., 2012). En la presente tesis, se trabajó con algunas de ellas, con

la finalidad de poder medir y ver que algoritmo se aproxima más a los resultados que se desea obtener; las técnicas que se utilizaron fueron:

2.3.3.1 Técnicas Bayesianas

Son clasificadores estadísticos, Duda, Hart, Stork (2012) mencionan que pueden predecir tanto las probabilidades del número de miembros de clase, como la probabilidad de que una muestra dada pertenezca a una clase particular. Estos clasificadores se basan en el teorema de Bayes, cabe mencionar que han demostrado una alta exactitud y velocidad momento de trabajar con grandes bases de datos; un clasificador Bayesiano sencillo es el "Naive Bayesiano" (García et al., 2012). En la presente tesis, se usaron los siguientes clasificadores:

2.3.3.1.1 Naive Bayesiana

Naive Bayes es un método de clasificación probabilístico, es utilizado para la clasificación de nuevas instancias de un documento D dentro de un conjunto finito C de clases predeterminadas, esto quiere decir que, dada una clase C y un conjunto de palabras W del nuevo documento a clasificar, se calcula la probabilidad de que dicho documento se clasifique dentro de la categoría C, así se tiene:

$$P(C|w) = \frac{P(W|C)P(C)}{P(W)}$$

Dónde: P(C) es la probabilidad a priori de la clase, P(WC) es la probabilidad condicional de la palabra W dada la clase C (Bordignon, 2007).

Según Hernández, Ramírez y Ferri (2004) la topología de un clasificador de Naive Bayes es la siguiente:

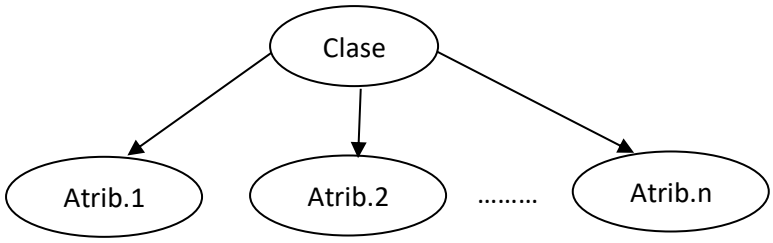

Imagen 1. Topología de un Clasificador Naive Bayes.
Fuente: (Hernández et al., 2004).

2.3.3.1.2 Redes Bayesianas

Una red bayesiana representa el conocimiento cualitativo mediante un grafo acíclico; este conocimiento se articula mediante la relación de independencia/ dependencia entre las variables que componen el modelo (Hernández, et al., 2004). Por ejemplo, una red bayesiana puede representar las relaciones probabilísticas entre enfermedades y síntomas. Dados los síntomas, la red puede ser usada para computar la probabilidad de la presencia de varias enfermedades (López, García, de la Fuente Sánchez y de la Fuente Solana, 2007).

Las Redes Bayesianas modelan un fenómeno mediante un conjunto de variables y las relaciones de dependencia entre ellas. Dado este modelo, se puede hacer inferencia bayesiana; es decir, estimar la probabilidad posterior de las variables no conocidas, en base a las variables conocidas. Estos modelos pueden tener diversas aplicaciones, para clasificación, predicción, diagnóstico, etc. Asimismo, pueden dar información interesante en cuanto a cómo se relacionan las variables del dominio, las cuales pueden ser interpretadas en ocasiones como relaciones de causa—efecto (wikidot, 2010).

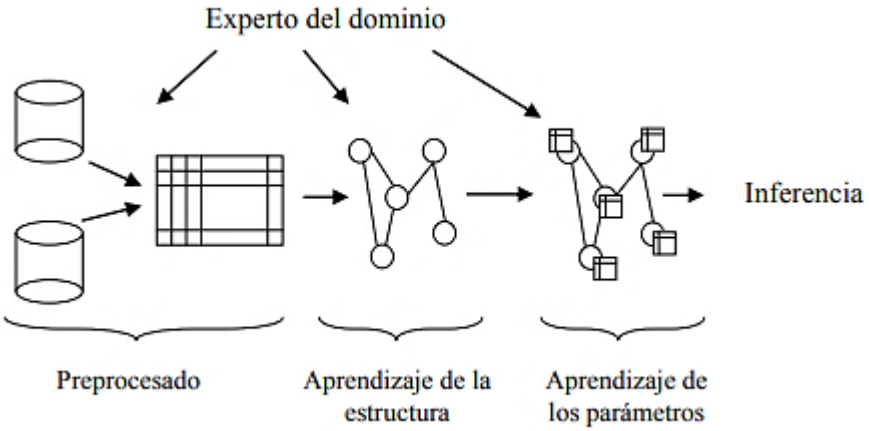

Imagen 2. Construcción de Redes Bayesianas partiendo de base de datos.
Fuente: (Rodríguez y Dolado, 2010).

2.3.3.2 Árboles de Decisión

2.3.3.2.1 Decision Stump (Árbol de un solo nivel)

García *et al*. (2012) mencionan que Decision Stump es un algoritmo sencillo que genera un árbol de decisión de un único nivel, utiliza un único atributo para construir el árbol de decisión; la elección del único atributo que formará parte del árbol se realizará basándose en la ganancia de información, a pesar de su simplicidad, en algunos problemas puede llegar a conseguir resultados interesantes; admite atributos numéricos como simbólicos y clases de ambos tipos también; el árbol de decisión tendrá tres ramas: Una de ellas será para el caso de que el atributo sea desconocido, y las otras dos serán para el caso de que el valor del atributo del ejemplo de test sea igual a un valor concreto del atributo o distinto a dicho valor.

2.3.3.2.2 J48

El algoritmo J48 como mencionan Boris y Cebrián (2012), forma parte de los algoritmos basados en árboles de decisión; la característica fundamental de este algoritmo es que incorpora una poda del árbol de la clasificación una vez que este haya sido construido, se podan aquellas ramas del árbol con menor capacidad predictiva; uno de los factores más importante para la poda es el **confidence level**, puesto que influye notoriamente en el tamaño y capacidad de predicción del árbol construido.

Una explicación simplificada de este parámetro de construcción del árbol es la siguiente: Para cada operación de poda, define la probabilidad de error

que se permite a la hipótesis de que el empeoramiento debido a esta operación es significativo; cuanto más baja se haga esa probabilidad, se exigirá que la diferencia en los errores de predicción antes y después de podar sea más significativa para no podar (García et al., 2012). El valor por defecto de este factor es del 25%, y conforme va bajando se permiten más operaciones de poda y por tanto llegar a árboles cada vez más pequeños; otra forma de variar el tamaño del árbol es a través de un parámetro que especifica el mínimo número de instancias por nodo, si bien es menos elegante puesto que depende del número absoluto de instancias en el conjunto de partida (García et al., 2012).

2.3.3.2 Reglas de Clasificación

Permiten la generación y contraste de árboles de decisión, o reglas y patrones a partir de los datos de entrada; la información de entrada será un conjunto de casos donde se ha asociado una clasificación o evaluación a un conjunto de variables o atributos; con esa información estas técnicas obtienen el árbol de decisión o conjunto de reglas que soportan la evaluación o clasificación; en los casos en que la información de entrada posea algún tipo de “ruido" estas técnicas pueden habilitar métodos estadísticos de tipo probabilístico para generar árboles de decisión recortados o podados, de igual manera esta técnica suele llevar asociada una alta interacción con el analista de forma que éste pueda intervenir en cada paso de la construcción de las reglas, bien para aceptarlas o modificarlas (García et al., 2012). A continuación, se muestra las técnicas de Reglas de clasificación que se utilizaran en la investigación:

2.3.3.2.1 Algoritmo 1R

Según García et al., (2012) es uno de los algoritmos más simples de reglas de clasificación, genera un árbol de decisión de un nivel expresado mediante reglas que consiste en seleccionar un atributo (nodo raíz) del cual nace una rama por cada valor, que va a parar a un nodo hoja con la clase más probable de los ejemplos de entrenamiento que se clasifican a través suyo; el Pseudocódigo del algoritmo se muestra a continuación:

```
1R (ejemplos) {
Para cada atributo (A)
Para cada valor del atributo (Ai)
Contar el número de apariciones de cada clase con Ai
Obtener la clase más frecuente (Cj|)
Crear una regla del tipo Ai -> Cj
Calcular el error de las reglas del atributo A
Escoger las reglas con menor error
}
```

Imagen 3. Pseudocódigo del algoritmo 1R.
Fuente: (García et al., 2012).

La clase debe ser simbólica, mientras los atributos pueden ser simbólicos o numéricos. En cuanto al error de las reglas de un atributo, consiste en la proporción entre los ejemplos que cumplen la regla y los ejemplos que cumplen la premisa de la regla; en el caso de los atributos numéricos, se generan una serie de puntos de ruptura a los cuales se le conoce como breakpoint, que discretizarán dicho atributo formando conjuntos; para ello, se ordenan los ejemplos por el atributo numérico y se recorren, de esta manera se van contando las apariciones de cada clase hasta un número m que indica el mínimo número de ejemplos que pueden pertenecer a un conjunto, para evitar conjuntos demasiado pequeños, por último, se unen a este conjunto ejemplos con la clase más frecuente y ejemplos con el mismo valor en el atributo (García et al., 2012).

La sencillez de este algoritmo es un poco insultante, tanto así que 1R no tiene ningún elemento de sofisticación y genera para cada atributo un árbol de profundidad cuya profundidad es 1, donde una rama está etiquetada por missing si es que aparecen valores desconocidos (missing values) en ese atributo en el conjunto de entrenamiento; el resto de las ramas tienen como etiqueta un intervalo construido de una manera muy simple, como se ha explicado antes, o un valor nominal, según el tipo de atributo del que se trate (García et al., 2012).

2.3.3.2.2 Algoritmo PART

PART (Generación de reglas por árboles de decisión parcial) fue desarrollado por el grupo neozelandés que construyó el entorno WEKA, este sistema se basa en las dos estrategias básicas para la inducción de reglas: El covering y la generación de reglas a partir de árboles de decisión; toma la estrategia del covering dado que genera una regla de decisión, elimina los

ejemplares que dicha regla cubre y continúa generando reglas hasta que no queden ejemplos por clasificar. Sin embargo, el proceso de generación de cada regla no es el usual; en este caso, para crear una regla, se genera un árbol de decisión podado, se obtiene la hoja que clasifique el mayor número de ejemplos, que se transforma en la regla, y posteriormente se elimina el árbol. Uniendo estas dos estrategias se consigue mayor flexibilidad y velocidad. Cabe mencionar que no se genera un árbol completo, sino un árbol parcial, un árbol parcial es un árbol de decisión que contiene brazos con subárboles no definidos. Para generar este árbol se integran los procesos de construcción y podado hasta que se encuentra un subárbol estable que no puede simplificarse más, en cuyo caso se para el proceso y se genera la regla a partir de dicho subárbol. A continuación, se muestra el Pseudocódigo de expansión de PART:

```
Expandir (ejemplos) {
  elegir el mejor atributo para dividir en subconjuntos
  Mientras (subconjuntos No expandidos)
             Y (todos los subconjuntos expandidos son HOJA)
    Expandir (subconjunto)
  Si (todos los subconjuntos expandidos son HOJA)
             Y (errorSubárbol >= errorNodo)
    deshacer la expansión del nodo y nodo es HOJA
```

Imagen 4. Pseudocódigo de expansión de PART. Fuente: (García et al., 2012).

Se tiene un ejemplo de generación de árbol parcial con PART

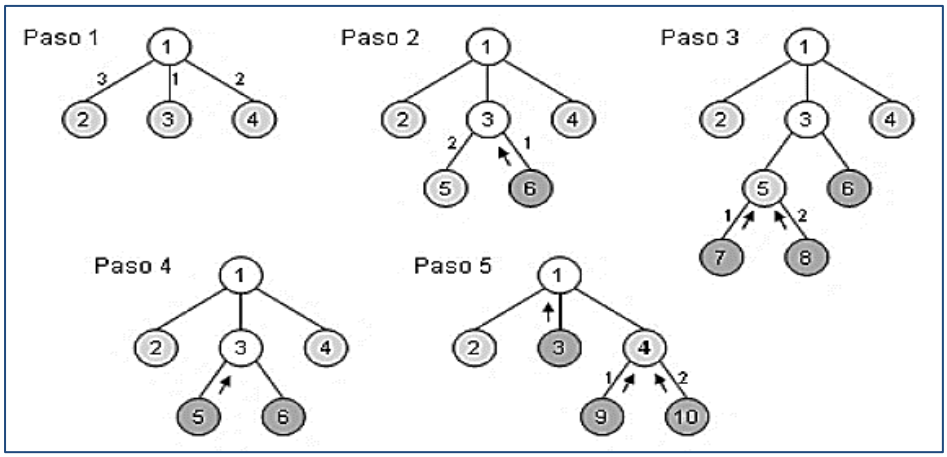

Imagen 5. Ejemplo de generación de árbol parcial con PART. Fuente: (García et al., 2012).

En la figura 4 se presenta un ejemplo de generación de un árbol parcial donde, junto a cada brazo de un nodo, se muestra el orden de exploración (orden ascendente según el valor de la entropía). Los nodos con relleno gris claro son los que aún no se han explorado y los nodos con relleno gris oscuro los nodos hoja. Las flechas ascendentes representan el proceso de backtracking. Por último, en el paso 5, cuando el nodo 4 es explorado y los nodos 9 y 10 pasan a ser hoja, el nodo padre intenta realizar el proceso de podado, pero no se realiza el reemplazo (representado con el 4 en negrita), con lo que el proceso, al volver al nodo 1, finaliza sin explorar el nodo 2.

Una vez generado el árbol parcial se extrae una regla del mismo. Cada hoja se corresponde con una posible regla, y lo que se busca es la mejor hoja; en el algoritmo PART se considera mejor hoja aquella que cubre un mayor número de ejemplos. Por último, PART permite que haya atributos con valores desconocidos tanto en el proceso de aprendizaje como en el de validación y atributos numéricos, tratándolos exactamente como el sistema C4.5.

2.3.3.3 Clasificadores Funcionales

2.3.3.3.1 Algoritmo SMO

Pérez y Cordoso (2010) mencionan que el aprendizaje de máquinas de vectores de soporte es un método supervisado que ha demostrado buenas propiedades para la categorización de documentos; uno de los algoritmos utilizados para entrenar máquinas de vectores de soporte usando un kernel polinomial es SMO (sequential minimal optimization).

El algoritmo SMO como dice Gala (2013) es un método de descomposición en el que la idea es elegir subproblemas del problema inicial para que sea menos costoso computacionalmente; en el caso del problema de optimización cuadrática de las SVM implica al menos dos multiplicadores de Lagrange, ya que deben cumplir la restricción $\sum_i^l \alpha_i y_i = 0$, es decir, los subproblemas que se elijan deben tener al menos tamaño dos. Podemos observar que la simplificación de SMO es que permite optimizar analíticamente, por tanto, el algoritmo se compone de dos pasos; el primero es una técnica heurística para elegir los dos mejores multiplicadores; el segundo es una técnica analítica para optimizarlos (Gala, 2013). Las

máquinas de Vectores Soporte resuelven problemas tanto de clasificación como de regresión, en la mayoría de los trabajos, los algoritmos de máquinas de Vectores Soporte se emplean para la clasificación (Gala, 2013).

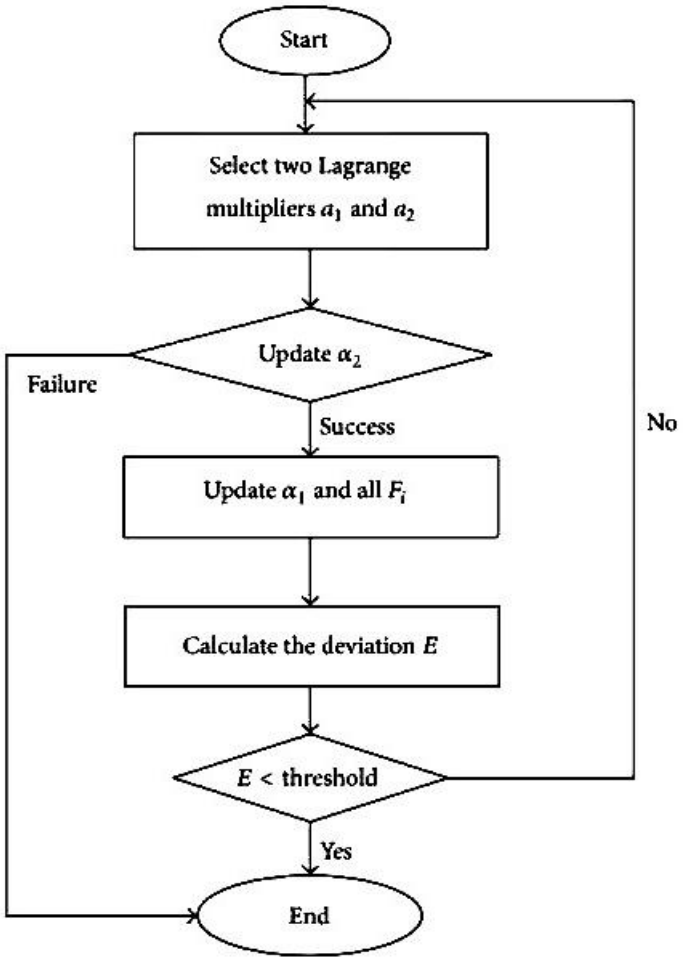

Imagen 6. Procedimientos del entrenamiento del algoritmo SMO.
Fuente: (Fenghua H. y Luming Y, 2014).

2.3.3.3.2 Perceptron Multicapa

Un perceptron de múltiples capas (MLP) es un modelo de red neuronal artificial de feedforward que asigna conjuntos de datos de entrada a un conjunto de salidas apropiadas; un MLP consiste en múltiples capas de nodos en un gráfico dirigido , con cada capa totalmente conectada al siguiente; excepto para los nodos de entrada, cada nodo es una neurona (o elemento de procesamiento) con una función de activación no lineal; MLP utiliza una técnica de aprendizaje supervisado llamado backpropagation

para entrenar la red; MLP es una modificación del perceptron lineal estándar y puede distinguir datos que no son linealmente separables (Morales, 2016).

El perceptron multicapa según Morales (2016), consiste en tres o más capas (una capa de entrada y una capa de salida con una o más capas ocultas) de nodos activadores no lineales y, por lo tanto, se considera una red neural profunda; dado que un MLP es una red totalmente conectada, cada nodo en una capa se conecta con un cierto peso wij. A cada nodo de la siguiente capa; algunas personas no incluyen la capa de entrada al contar el número de capas y hay desacuerdo acerca de si wij. Debe ser interpretado como el peso de i a j o al revés.

El aprendizaje ocurre en el perceptron cambiando los pesos de la conexión después de que cada pedazo de datos se procesa, basado en la cantidad de error en la salida comparada al resultado esperado (Morales, 2016).

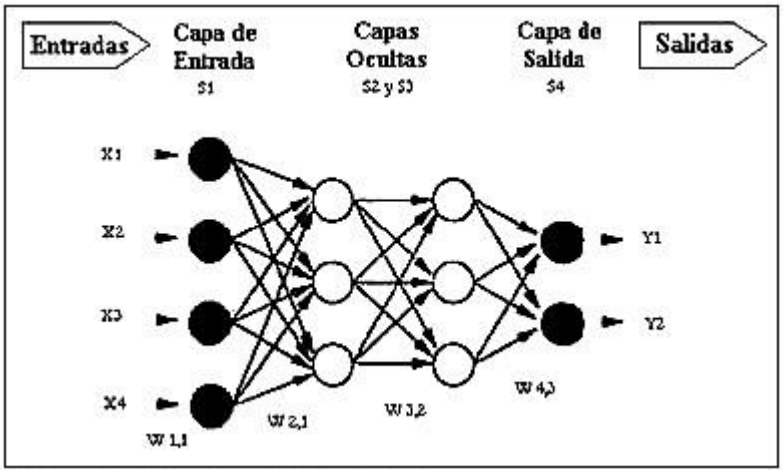

Imagen 7. Red neuronal multicapa.
Fuente:(Villamil y Delgado, 2007).

2.3.3.4 Clasificadores Lazy

2.3.3.4.1 Algoritmo IBK

Este algoritmo está basado en instancias, por ello consiste únicamente en almacenar los datos presentados. Cuando una nueva instancia es encontrada, un conjunto de instancias similares relacionadas es devuelto desde la memoria y usado para clasificar la instancia consultada. Se trata, por tanto, de un algoritmo del método

lazy learning. Este método de aprendizaje se basa en que los módulos de clasificación mantienen en memoria una selección de ejemplos sin crear ningún tipo de abstracción en forma de reglas o de árboles de decisión (de ahí su nombre, lazy, perezosos). Cada vez que una nueva instancia es encontrada, se calcula su relación con los ejemplos previamente guardados con el propósito de asignar un valor de la función objetivo para la nueva instancia. La idea básica sobre la que se fundamenta este algoritmo es que un nuevo caso se va a clasificar en la clase más frecuente a la que pertenecen sus K vecinos más cercanos. De ahí que es conocido como método K-NN: K Nearest Neighbours (Boris y Cebrián, 2012).

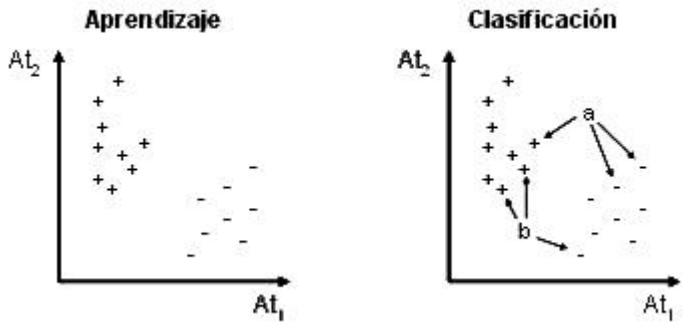

Imagen 8. Aprendizaje y clasificación con KNN.
Fuente: (Molina y García, 2011).

2.3.4 Análisis de sentimientos o minería de opiniones

Cuando hablamos de Análisis de Sentimientos, también llamada Minería de Opiniones, nos estamos refiriendo a una serie de aplicaciones de técnicas del procesamiento del lenguaje natural y minería de textos, que tienen como objetivo la extracción de información subjetiva a partir de contenidos generados por los usuarios, como puedan ser comentarios en blogs, redes sociales, etc. (Cortizo, 2011).

Anteriormente se dio conceptos de los términos, algoritmos y herramientas que se emplearán en el presente proyecto; se utilizarán los algoritmos de minería de datos enfocados a la minería de opiniones.

Para llevar a cabo la tarea de la clasificación de comentarios, se utilizará un framework de minería de datos, es un entorno de trabajo que ya tiene implementado los diferentes algoritmos de clasificación.

2.3.5 Framework de minería de texto

Un Framework es un entorno o ambiente de trabajo para desarrollo; es una suite de software para minería de datos, el framework de minería de datos que se utilizará es:

2.3.5.1 Weka

Weka es un conjunto de librerías Java para la extracción de conocimientos desde bases de datos; contiene las herramientas necesarias para realizar transformaciones sobre datos, tareas de clasificación, regresión, clustering, asociación y visualización. Una de las propiedades más interesantes de este software es su factibilidad modificar métodos y para añadir extensiones (Ferri, 2012).

2.3.6 Metodología KDD

KDD (Knowledge Discovery in Databases), es el proceso de descubrimiento de conocimiento en bases de datos más conocido, esta se compone de las siguientes fases:

- **Recopilación de datos:** El primer paso, consiste en tomar los datos que se desea analizar para integrarlos en un solo repositorio. Al unir las fuentes de datos en un mismo almacén, es necesario contar con un proceso que lea las diferentes fuentes, los limpie y adecue a la estructura que tiene la data warehouse para su almacenamiento. Este proceso es llevado cabo mediante el sistema ETL (Extraction-Transformation-Load).
- **Selección, limpieza y transformación de datos:** El objetivo de esta fase del proceso de KDD es tomar los datos almacenados, sobre los cuales se aplican diferentes tipos de técnicas de selección, limpieza y transformación de datos. En principio, no hay un orden establecido para ellas, sino que, se planifica en función del formato que se desea dar a los datos que conformarán la vista minable.

- **Data mining:** Una vez obtenida una vista minable de los datos, el siguiente paso es aplicar las técnicas de data mining para obtener modelos que representan a dichos datos.
- **Interpretación y evaluación de modelos:** Tras la obtención de los modelos de data mining, el último paso es evaluar la calidad de dichos modelos y realizar una interpretación de los mismos para obtener el conocimiento buscado (Lara, 2014).

A continuación, se muestra una imagen del proceso KDD.

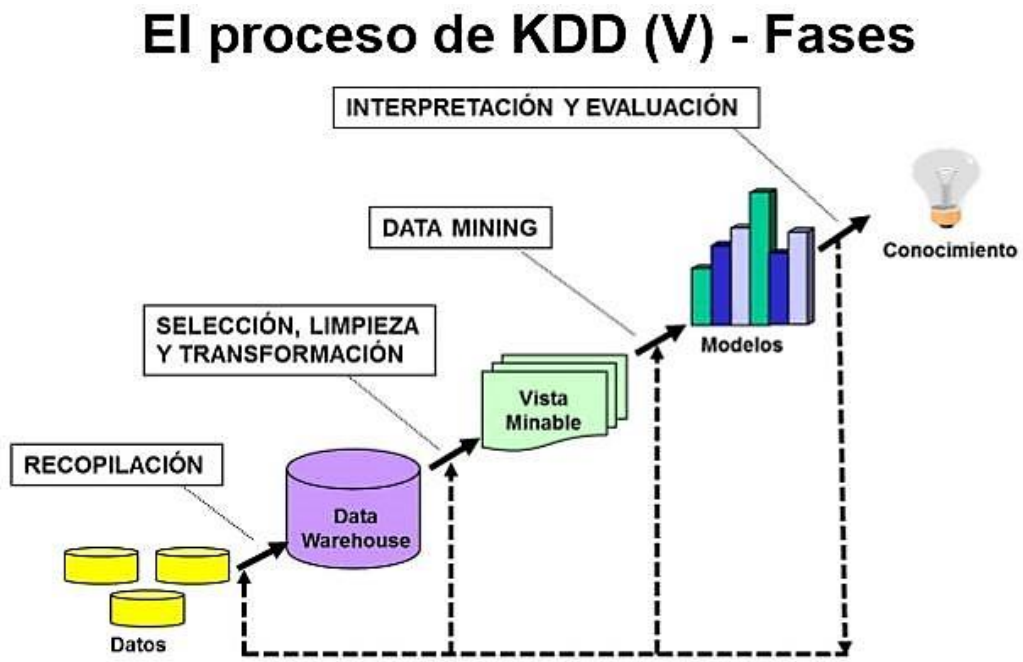

Imagen 9. Proceso de KDD.
Fuente: (Lara, 2014).

2.3.7 Modelo cascada

También conocido como modelo clásico, modelo tradicional o modelo lineal secuencial. Este método, implica un desarrollo rígido que está dividido en una secuencia de actividades (Soloriio, 2013).

Según Galipienso, Botía, Mora y Trigueros (2005) las principales etapas de este modelo se transforman en actividades fundamentales de desarrollo que son:

1. Análisis y definición de requerimientos: Son los servicios y metas que el sistema debe alcanzar.

2. Diseño del sistema y del software: El diseño del sistema, establece una arquitectura completa del sistema. El diseño del software identifica y describe las abstracciones fundamentales del sistema software y sus relaciones.

3. Implementación y prueba de unidades: Durante esta etapa, el diseño del software se lleva a cabo como un conjunto o unidades de programas. La prueba de unidades implica verificar que cada una cumpla su especificación.

4. Integración y prueba del sistema: Las unidades individuales de programas se integran y prueban como un sistema completo para asegurar que se cumplan los requerimientos del software.

5. Funcionamiento y mantenimiento: El sistema se instala y se pone en funcionamiento práctico. El mantenimiento implica corregir errores no descubiertos en las etapas anteriores del ciclo de vida, mejorar la implementación de las unidades del sistema y resaltar los servicios del sistema una vez que se descubren nuevos requerimientos.

2.3.8 El ciberbullying

También conocido como bullying electrónico, bullying en línea, e-bullying, intimidación o acoso en línea. El ciberbullying implica la intimidación, el acoso o la amenaza mediante medios electrónicos (principalmente, vía Internet o teléfono celular), con la intención de dañar (ciberagresor) a otro (cibervíctima) (García, Joffre, Martínez y Llanes, 2011).

Smith, Mahdavi y Carvalho(2006) dividen el ciberbullying en siete subtipos: Mensajes de texto recibidos en el teléfono móvil; fotografías o videos realizados con las cámaras de los móviles, y posteriormente usados para amenazar a la víctima; llamadas acosadoras al teléfono móvil; mensajes de correo electrónico insultantes o amenazantes; salas de chat en las que se arremete contra uno de los participantes o se le excluye socialmente; el acoso mediante los programas de mensajería instantánea; y páginas web donde se difama a la víctima, se descarga información personal a la red o se hacen

concursos en los que se ridiculiza a los demás (como se citó en García et al., 2011, p.119). Es necesario también plasmar la diferencia entre el bullying y el ciberbullying para ello se tomó la siguiente tabla del trabajo titulado Ciberbullying: Forma virtual de intimidación escolar realizada por García et al. (2011).

Tabla 1. Diferencia entre Bullying y Ciberbullying

Bullying	Ciberbullying
Cara a cara	Anónima
En grupo o individual	Individual
Golpes, empujones, agresión verbal o exclusión social	Mensajes, e-mail, imágenes manipuladas
Sólo en horas de escuela	En todos lados y a toda hora
Se limita a la agresión directa	Sin límite de alcance
Sólo audiencia escolar	Audiencia mundial por internet
Certeza de que sólo se encuentra en el ámbito escolar	Incertidumbre de no saber quién ha visto las imágenes o mensajes
Víctima y agresor, víctima-agresor	Cibervíctima-ciberagresor

Fuente: Ciberbullying: Forma virtual de intimidación escolar (García et al., 2011).

2.3.9 La agresividad

La agresividad no puede considerarse un elemento ajeno a la vida cotidiana, no puede ignorarse las expresiones de agresividad que se realizan en forma virtual. Es por ello que es necesario realizar la precisión entre los términos agresividad y agresión. Según García y Ramos (1998), cuando se habla de agresividad, se hace referencia a una disposición o tendencia hostil, mientras que la agresión trata de la concreción de esa capacidad agresiva, por lo que siempre se manifestará en acciones específicas como la acción de la fuerza física o moral (como se citó en López y González, 2013, p.6).

La característica por la que es posible distinguir la agresión en una emisión lingüística es la intencionalidad con la que se realiza el acto, en los comentarios de los usuarios de los blogs esto es algo que queda delineado, aunque no puede afirmarse categóricamente (López y González, 2013). En un acto agresivo o al emitir mensajes agresivos se pretende causar daño o

menoscabar las cualidades de la persona interpelada (López y González, 2013). En García y Ramos (1998) se enfatiza que un acto agresivo, conlleva la idea de que interviene la violencia, puesto que está implícita una fuerza que se aplica hacia otro, restándole valor y utilizando la coerción para lograr sus fines; la violencia en los actos puede generar un patrón ya sea físico o psicológico, es decir, un orden producto de la fuerza, que puede persistir aún después de que esta deja de ejercerse (como se citó en López y González, 2013, p.6). A continuación, se mostrará una tabla que muestra algunos criterios de clasificación de las conductas agresivas, esta tabla se tomó del trabajo titulado: Aspectos conceptuales de la agresión: Definición y modelos explicativos, elaborado por Carrasco y Gonzáles (2006).

Tabla 2. Clasificaciones de las conductas agresivas

Criterio de clasificación	Autor/es	Tipología	Descripción
Naturaleza	Buss (1961); Pastorelli, Barbarelli, Cermak, Rozsa y Caprara (1977); Valzelli (1983)	*Agresión Física*	Ataque a un organismo mediante armas o elementos corporales, con conductas motoras y acciones físicas, el cual implica daños corporales.
		Agresión Verbal	Respuesta oral que resulta nociva para el otro, a través de insultos o comentarios de amenaza o rechazo.
	Galen y Underwood (1997)	*Agresión Social*	Acción dirigida a dañar la autoestima de los otros, su estatus social o ambos, a través de expresiones faciales, desdén, rumores sobre otros o la manipulación de las relaciones interpersonales.
Relación interpersonal	Buss (1961); Ili (1983); Lagerspetz et al. (1988); Björkqvist et al. (1992); Crick y Grotpeter (1995); Grotpeter y Crick (1996); Connor (1998); Crick et al. (1999); Crick, Casas y Nelson (2002)	*Agresión Directa o Abierta*	Confrontación abierta entre el agresor y la víctima, mediante ataques físicos, amenazas verbales, destrucción de la propiedad y comportamiento autolesivo.
		Agresión Indirecta o Relacional	Conductas que hieren a los otros a través de la manipulación de las relaciones con los iguales: Control directo, dispersión de rumores, avergonzar en un ambiente social, alienación rechazo por parte del grupo, e exclusión social.

Motivación	Feshbach (1970); Atkins, Stoff, Osborne y Brown (1993); Kassinove y Sokhodolsky (1995); Berkowitz (1996)	*Agresión Hostil*	Acción intencional encaminada a causar un impacto negativo sobre otro, por el mero hecho de dañarle, sin la expectativa de obtener ningún beneficio material.
		Agresión Instrumental	Acción intencional de dañar por la que el agresor obtiene un objetivo: Ventaja o recompensa, social o material, no relacionada con el malestar de la víctima.
		Agresión Emocional	Agresión de naturaleza fundamentalmente emocional generada no por un estresor externo, sino por el afecto negativo que dicho estresor activa, produciendo ira y tendencias agresivas.

Fuente: Aspectos conceptuales de la agresión: Definición y modelos explicativos (Carrasco y Gonzáles, 2006)

2.3.9.1 Expresiones coprolalias en ambientes virtuales

El uso de lenguaje coprolálico con frecuencia es una manifestación usual de agresividad; esta es una situación que incluye los comentarios emitidos en los blogs (López y González, 2013). Es por ello que en este trabajo se ha elegido hacer un acercamiento en estas expresiones a fin de estudiar la relación entre estos dos elementos, la agresividad y el uso de insultos por medio de palabras altisonantes. Para Martínez (2009) el insulto y el lenguaje coprolálico, se tratan de "formas lingüísticas con valor sociopragmático que tienen un componente etnográfico, con el cual se intenta atacar o humillar a una persona en un momento determinado" (como se citó en López y González, 2013, p.7).

2.3.10 Análisis Léxico

Para poder entender lo que es un análisis léxico, primero se buscó la definición de léxico. Según Pérez y Gardey (2009), definen al léxico como el vocabulario de un idioma, está definido por palabras que hacen referencias a conceptos que pueden ser abstractos o materiales, estas palabras pueden pertenecer a las siguientes clases: Verbos, sustantivos, adverbios y adjetivos. Es por ello que, el análisis léxico consiste en el examen de las palabras que forman parte de la oración (Instituto de ingeniería del conocimiento, 2017).

2.3.11 Teoría de la valoración

López y Gonzáles (2013), utilizaron la Teoría de la Valoración como plataforma de análisis de los comentarios emitidos blogs de los diarios electrónicos. Esta teoría, siguiendo la propuesta de Gales (2011) se sitúa dentro del marco de la Lingüística Sistémico Funcional. Para Martin y Rose (2003) el lenguaje se desarrolla tomando tres niveles que se relacionan estrechamente entre sí, la gramática, el contexto social y el género (como se citó en López y González, 2013, p.8). Esta teoría se divide en tres subsistemas de la valoración: La actitud, el compromiso y la graduación. A continuación, se detallará cada uno de ellas:

1) La actitud: Se clasifica como actitudes todos los enunciados que transmiten una evaluación positiva o negativa, estas actitudes pueden ser explícitas o implícitas; en la actitud explicita, se pueden señalar las palabras o frases que transmiten un sentido positivo o negativo de manera evidente y directa; en la actitud implícita, el análisis es más complejo y la interpretación depende del lector (Kaplan, 2004). Según Kaplan (2004), esta categoría se divide en los subsistemas de afecto, juicio y apreciación.

Afecto es la evaluación mediante la cual el hablante indica su disposición emocional o reporta las repuestas emocionales de terceros hacia personas, cosas, situaciones o eventos. Las emociones se concentran en tres grandes grupos que tiene que ver con la felicidad o la infelicidad; la seguridad o la inseguridad; y la satisfacción o la insatisfacción. El efecto puede expresarse como **(i)** una casualidad, a través de un epíteto se describe a los participantes (ej.: un niño alegre); un atributo (ej.: El niño estaba alegre); o una circunstancia (ej.: el niño jugaba alegremente); **(ii)** un proceso (ej.: afectivo mental: El regalo agradó al niño; afectivo conductual; El niño sonrió); o **(iii)** un comentario (ej.: Felizmente, durmió una larga siesta) (Kaplan, 2004).

Tabla 3. Categorías de afecto: expresiones congruentes

AFECTO		
Categoría	**Ejemplos positivos**	**Ejemplos negativos**
Felicidad/infelicidad	Feliz, alegre, jubiloso/a, optimista	Deprimido/a, triste, miserable, angustiado/a
Seguridad/inseguridad	Confiado/a, seguro/a, tranquilo/a, sereno/a	Ansioso/a, preocupado/a, inseguro/a, intranquilo/a
Satisfacción/insatisfacción	Interesado/a, absorto/a, estar enfrascado/a, gustar	Cansado/a, aburrido/a, exasperado/a, odiar

Fuente: Nuevos desarrollos en el estudio de la evaluación en el lenguaje: La teoría de la valoración (Kaplan, 2004)

Los indicadores lingüísticos del afecto pueden ser verbos (ej.: amar/odiar); adverbios (ej.: alegremente/tristemente); adjetivos (ej.: feliz/triste) y las nominalizaciones, es decir, verbos y adjetivos en sustantivos (ej.: alegría/desesperación) (Kaplan, 2004).

Según Martin (2000), para efectuar una clasificación del afecto, deben tomarse en cuanta las siguientes variables: **(i)** si la cultura construye a los sentimientos como positivos o negativos; **(ii)** si los sentimientos se expresan una manifestación paralingüística (ej.: temblar), o si se experimenta cierto estado mental (ej.: cauteloso), **(iii)** si los sentimientos constituyen como una reacción a un agente específico externo (ej.: al niño le gustaba el maestro**), (iv)** cómo se gradúan los sentimientos, ya que la mayoría de las emociones ofrecen lexicalizaciones que se ubican a lo largo de una escala(ej.: bajo: al niño le agradó el regalo/ medio: al niño le gustó el regalo/alto: al niño le encantó el regalo); y finalmente, (v) si las emociones involucran intensión más que reacción ante estímulos (como se citó en Kaplan, 2004, p.63).

En el **juicio** se puede evaluar la conducta como moral o inmoral, legal o ilegal, socialmente aceptable o inaceptable, encomiable o deplorable, normal o anormal. En función a esto el juicio puede clasificarse en dos grandes tipos: **(i) juicios de estima social,** se evalúa cuán normal es una persona, cuán competente, resuelta y decidida, **(ii) juicios de sanción social**, relacionados con la veracidad y la integridad moral (Kaplan, 2004).

Tabla 4. Categorías de juicios: expresiones congruentes

JUICIO		
ESTIMA SOCIAL		
	Ejemplos positivos (admiración)	Ejemplos negativos (critica sin implicaciones legales)
normalidad	Corriente, común, normal, afortunado/a, moderno/a	Excéntrico/a, extraño/a, raro/a, desafortunado/a, anticuado/a
capacidad	Habilidoso/a, inteligente, intuitivo/a, atlético/a, fuerte	Inhábil, lento/a, tonto/a, torpe, débil
tenacidad	Heroico/a, valiente, confiable, infatigable, perseverante	Cobarde, apresurado/a, no confiable, distraído/a, perezoso/a
SANCIÓN SOCIAL		
	Ejemplos positivos(alabanza)	Ejemplos negativos (condena, puede tener implicaciones legales)
veracidad	Sincero/a, honesto/a, genuino/a, franco/a, directo/a	Deshonesto/a, mentiroso/a, inauténtico/a, manipulador/a
Integridad moral	Moral, bondadoso/a, respetuoso/a de la ley, sensible, justo/a	Inmoral, malvado/a, corrupto/a, cruel, injusto/a

Fuente: Nuevos desarrollos en el estudio de la evaluación en el lenguaje: La teoría de la valoración (Kaplan, 2004)

Kaplan (2004), menciona que los juicios pueden expresarse de manera explícita a través de adverbios (ej.: honestamente), atributos y epítetos (ej.: la gente es conformista e irracional), sustantivos (ej.: un mentiroso), y verbos (ej.: engañar).

Kaplan (2014), indica que la **apreciación** es considera como un sistema mediante el cual los sentimientos humanos hacia productos, procesos y entidades, se evalúan como positivas o negativas. En forma típica con estos valores se evalúan artefacto, textos, planes, políticas, objetos naturales o manufacturados. También las personas pueden ser evaluados mediante la apreciación, cuando se perciben como entidades. Martin (2000) categoriza la apreciación alrededor de tres dimensiones: **(i)** la reacción, **(ii)**la composición, y **(iii)** la valuación (como se citó en Kaplan, 2004, p.66).

Tabla 5. Categorías de apreciación: expresiones congruentes

APRECIACIÓN		
	Positiva	**Negativa**
REACCIÓN: Impacto	Cautivador/a, llamativo/a, atractivo/a, agradable, conmovedor/a.	Aburrido/a, tedioso/a, ascético/a, pedante, soso/a
REACCIÓN: Calidad	Hermoso/a, espléndido/a, encantador/a	Feo/a, repulsivo/a, repugnante
COMPOSICIÓN Balance	balanceado/a, armonioso/a, simétrico/a, proporcionado/a	Desbalanceado/a, discordante, desproporcionado/a, asimétrico/a
COMPOSICIÓN Complejidad	Simple, elegante, detallado/a, preciso/a, intrincado/a	Extravagante, monolítico/a, simplista, impreciso/a
VALUACIÓN	Profundo/a, innovador/a, original, único/a, exigente	Superficial, insignificante, reaccionario/a, conservador/a

Fuente: Nuevos desarrollos en el estudio de la evaluación en el lenguaje: La teoría de la valoración (Kaplan, 2004)

2) El **compromiso**, este sistema tiene que ver con los recursos lingüísticos que pueden utilizarse para posicionar la voz del hablante o del autor en relación con los enunciados comunicados por un texto (Kaplan, 2004). While (2003) menciona que las principales opciones en el sistema de compromiso se ubican en dos dimensiones contrapuestas: la monoglosia y la heteroglosia (como se citó en Kaplan, 2004, p.68). Se considera que los enunciados monoglósicos, equivalentes a las aseveraciones declarativas absolutas, ignoran la diversidad de voces que se ponen en juego en todo acto de comunicación (Kaplan, 2004).

La heteroglosia corresponde a los enunciados que reconocen, de alguna manera, la existencia de otras voces y posturas alternativas; dentro de esta categoría, se distinguen los recursos de extravocalización, que introducen en el texto fuentes de voces externas, y de intravocalización, que corresponden a la voz interna del autor o del hablante, quien asume la responsabilidad por los enunciados emitidos (Kaplan, 2004). A continuación, se plasmará algunos ejemplos tomados de Kaplan (2004) Ejemplos:

- Larry nos ha manifestado que cree que ésta es una mala idea. (extravocalización)

- Considero que ésta es indudablemente una mala idea. (intravocalización)

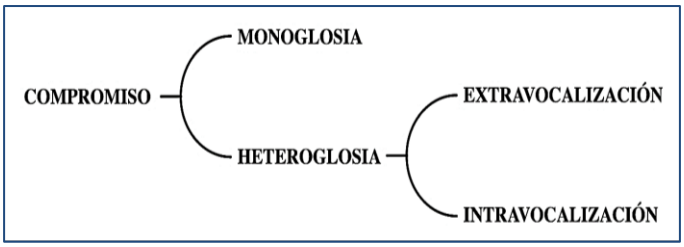

Imagen 10. El sistema de compromiso.
Fuente: (Kaplan, 2004).

3) Para Kaplan (2014), la **gradación**, que es un espacio semántico de escala que está relacionado con la manera en que los hablantes intensifican o disminuyen la fuerza de sus enunciados y gradúan, desdibujando o agudizando, el foco de sus categorizaciones semánticas. Los valores de fuerza se ubican en una escala de intensidad de baja a alta, y se ejemplifican claramente con los adverbios conocidos como "intensificadores", "amplificadores" y "enfáticos" (ej.: ligeramente, muy, en realidad). En estos casos, el valor se expresa explícitamente por medio de un ítem léxico independiente y aislado (ejemplos de intensidad baja, media y alta: en el subsistema de afecto - gustar, amar, adorar; en el subsistema de apreciación - atractivo, hermoso, exquisito) (Kaplan, 2004).

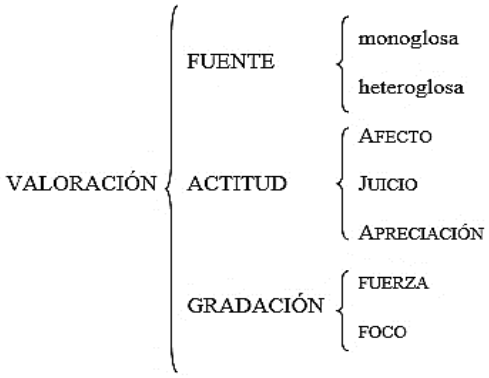

Imagen 11. Sistema de valoración como representa Martin y White (2005).
Fuente: (Como se citó en López, 2008).

2.3.12 La teoría de la relevancia

Para poder entender y definir la teoría de la relevancia, se estudió la tesis doctoral denominada: Los conceptos de la Teoría de la relevancia y su

aplicación al estudio de los marcadores discursivos de Jelena Rajić (2013), a continuación, se plasmará una breve descripción de este trabajo:

La Teoría de la Relevancia surge en el ámbito de las ciencias cognitivas, es por ello que no es una teoría de carácter lingüístico, ya que sus autores consideran la comunicación desde una perspectiva fundamentalmente psicológica: Parten de la teoría modular de la mente e intentan dar una explicación formal y más adecuada de los procesos mentales que dirigen el comportamiento comunicativo humano, tanto el verbal como el no verbal.

Según la teoría de la relevancia, la comunicación consiste en dos procesos diferentes simultáneos: La codificación y descodificación, y otro en la ostensión e inferencia. Son dos aspectos de un mismo fenómeno: La codificación y la ostensión se realizan mediante la emisión lingüística mientras que la descodificación e inferencia corresponden a la interpretación del mensaje.

Un enunciado se considera relevante si con un mínimo de esfuerzo mental produce un máximo de efectos cognitivos, es decir, la nueva información que el interlocutor obtiene a partir de la interpretación del enunciado. La siguiente cuestión que plantea la teoría de la relevancia es cómo se desarrolla el proceso de interpretación. Se ha visto que se da de acuerdo a la descodificación e inferencia, pero ahora nos interesa ver cuáles son los pasos concretos que el interlocutor debe seguir para alcanzar el significado final.

Sperber y Wilson (2007) explican que a partir de un estímulo lingüístico (la forma fonológica, sintáctica y semántica) se obtiene la forma lógica del enunciado (una serie de conceptos relacionados), luego esto se somete a una operación pragmática (como se citó en Rajić, 2013, p.354). Los pasos que se siguen en esta etapa son: Desambiguación o selección de la acepción pertinente de una palabra polisémica; asignación de referentes o determinación de la información situacional y, finalmente, enriquecimiento o precisión del significado de las expresiones vagas, como aun, demasiado, etc.

El resultado de estas tres operaciones es la transformación de la forma lógica en una forma proposicional, que es el verdadero significado del enunciado. Todos los pasos del proceso se realizan de forma automática e

inconsciente. Ahora con la proposición obtenida a partir de la forma lógica codificada por el enunciado, mediante los procesos de desambiguación, asignación y enriquecimiento, recibe el nombre de explicatura (el significado explícito), que en combinación con una serie de datos contextuales se produce una información o varias informaciones.

La explicatura de un enunciado es, pues, la proposición expresada convencionalmente, por medio de la estructura lingüística. A continuación, se plasmará un ejemplo:

a) ¿Te apetece un whisky?

b) No tomo alcohol.

La forma proposicional de “b” no ofrece una respuesta directa, pero hace referencia a la información enciclopédica sobre las bebidas alcohólicas, la cual incluye otra información: El whisky es una bebida alcohólica, dentro de esta última información se produce una implicatura contextual con el siguiente contenido: No me voy a tomar un whisky.

La idea que ha influido de manera más directa en el estudio de los fenómenos lingüísticos y, sobre todo, en el estudio de los marcadores discursivos es el establecimiento de la distinción entre dos tipos de significado: El significado conceptual y el significado procedimental o significado de procesamiento. Las palabras con significado conceptual (nombres, verbos, adjetivos y algunos adverbios) codifican conceptos, es decir, crean representaciones mentales de objetos, acontecimientos, cualidades, etc.; por su parte, las expresiones con significado procedimental codifican instrucciones sobre cómo los hablantes deben tratar la información que proporcionan las palabras con significado conceptual. Se diferencian de las primeras en que no influyen en las condiciones de verdad de las proposiciones en que aparecen, como señalan los siguientes ejemplos:

a) Es español y, por tanto, es buen conductor

b) Es español y, sin embargo, es buen conductor.

Los dos enunciados tienen la misma forma proposicional y cumplen con las mismas condiciones de verdad siempre y cuando se den las situaciones

descritas por las proposiciones Es español y Es buen conductor; sin embargo, no son idénticas desde el punto de vista pragmático, ya que las inferencias que introducen son distintas. En “a”, se infiere que los españoles son buenos conductores, de modo que la proposición Ser buen conductor se entiende como una consecuencia/conclusión de Ser español. Sin embargo, en “b” se implica lo contrario y se infiere que los españoles no son buenos conductores, pero que una persona X, a pesar ser español, es buen conductor.

Significado conceptual y significado de procesamiento

La información procedimental no es un rasgo propio solo de los marcadores, los marcadores constituyen un conjunto de diversas unidades lingüísticas que engloban adverbios, conjunciones, preposiciones, interjecciones, y también locuciones formadas por estas u otras clases de palabras, como son, sin embargo, no obstante, por cierto, de hecho, o sea, etc. El significado de procesamiento lo marcan también otros elementos lingüísticos: Los morfemas flexivos verbales, los determinantes, la entonación, la estructura informativa e incluso palabras con significado conceptual. Por ejemplo, a continuación, se mostrará dos enunciados con un mismo contenido proposicional procede del uso de diferentes formas verbales:

a) El tren salió a las ocho. b) El tren salía a las ocho.

En “a” el pretérito indefinido marca un evento acabado; por otra en “b” el pretérito imperfecto proporciona una información evidencial, quiere decir que su uso se infiere que el hablante no se hace responsable de la veracidad de lo dicho. Con el artículo indeterminado en Juan discutió con una mujer por teléfono, el hablante da a entender que con el sintagma una mujer no se refiere a la mujer de Juan; sin embargo, el artículo determinado en Juan discutió con la mujer remite a que la mujer se pueda referir a la esposa de Juan o bien a una otra mujer identificable para el oyente (por haber sido, por ejemplo, ya mencionada en el discurso), que no sea la mujer de Juan.

Tras el estudio de la teoría de la valoración y la teoría de la relevancia, se optó por elegir la primera, ya que se nos es de mucha utilidad el análisis de comentarios agresivos que se aborda en este trabajo.

2.3.12 Evaluación de Algoritmos

Evaluar el rendimiento de los algoritmos de recomendación no es trivial, ya que diferentes algoritmos pueden ser mejores o peores dependiendo del dataset elegido. Los indicadores de evaluación más usados para los algoritmos de minería de datos son:

MAE: El Error Medio Absoluto (Mean Absolute Error, MAE), es uno de los parámetros de evaluación más utilizado en la evaluación de algoritmos. Este, mide la desviación de las recomendaciones predichas y los valores reales. A menor MAE mejor predice el sistema las evaluaciones de los usuarios. El MAE es un tipo de error estadísticamente muy estudiado y sencillo de comprender (Galán, 2007).

La precisión: Nos puede dar una primera idea de cómo de bueno es el algoritmo para un sistema de recomendación (Galán, 2007). Una precisión detallada por clase consiste en que para cada uno de los valores que puede tomar el atributo de clase este te da el porcentaje de instancias con ese valor que son correctamente predichas (Molina y García, 2012).

Tiempo: Para poder elegir el algoritmo más eficiente, también se debe realizar una comparación del consumo de recursos computacionales como es el caso del tiempo de ejecución que tardan cada uno de ellos en cumplir el objetivo (DECSAI, 2009). Un algoritmo es eficiente cuanto menos en tiempo consume en realizar su tarea (Departamento de Informática Universidad de Valladolid Campus de Segovia, 2007).

CAPITULO III

3. DISEÑO METODOLOGICO

3.1 Hipótesis de investigación

3.1.1 Hipótesis General

El corpus de conocimiento influirá de manera directamente proporcional en el proceso de la clasificación de comentarios de textos agresivos, basadas en técnicas de minería de texto en un análisis léxico.

3.1.2 Hipótesis Específicas

- El corpus de conocimiento influirá directamente sobre la elección del algoritmo de clasificación del grupo de algoritmos clasificadores basadas en las técnicas de minería de textos en un análisis léxico, siendo el elegido IBK por tener el menor Error Absoluto Medio.
- El corpus de conocimiento influirá de manera directa en la precisión de la clasificación de comentarios de textos agresivos basadas en las técnicas de minería de textos en un análisis léxico, siendo esta mayor al 85%.
- El corpus de conocimiento influirá proporcionalmente en el tiempo de generación de los Modelos de Clasificación de comentarios de textos agresivos basadas en las técnicas de minería de textos en un análisis léxico, siendo este no mayor a 20 minutos.
- El corpus de conocimiento influirá proporcionalmente sobre la cantidad de datos óptimos para la clasificación de comentarios de textos agresivos basadas en las técnicas de minería de textos en un análisis léxico, siendo la cantidad óptima 1561 registros de comentarios y 906 atributos (palabras).

3.2 Operacionalización de variables

Tabla 6. Operacionalización de variables

VARIABLES	DEFINICIÓN CONCEPTUAL	DEFINICIÓN OPERACIONAL	DIMENSIONES	INDICADORES
VARIABLE INDEPENDIENTE X: aprendizaje computacional	El aprendizaje computacional se llevó a cabo combinando el corpus de conocimiento con las técnicas de minería de datos. Es corpus de conocimiento, es un conjunto de palabras y comentarios. Las técnicas de minería de datos que se usaron fueron: Naive Bayesina, Árboles de decisión, Reglas de Clasificación y Clasificadores Funcionales.	El corpus de conocimiento está formado por 906 atributos agresivos (palabras), 4 atributos clases o etiquetas (N, A-, A, A+) y 1561 comentarios de textos agresivo. Los algoritmos de las técnicas de minería de datos que se usaron fueron: NaiveBayes, BayesNet, DecisionStump, J48, OneR, PART, SMO, IBK y KStart.	precisión de la clasificación de comentarios de textos agresivos en categorías.	precisión del 0% al 100%.
			tiempo de generación de los Modelos de Clasificación de comentarios de textos agresivos.	Tiempo (segundos)
			elección del algoritmo de clasificación con el menor **MAE**.	Error Absoluto Medio, en un rango de valores de 0 a 0.5.
			cantidad de datos óptimos para la clasificación de comentarios de textos agresivos**.**	- 906 palabras agresivas (atributos) - De 800 a 1561 comentarios de textos.
VARIABLE DEPENDIENTE Y: clasificación de comentarios de textos agresivos.	Es el ordenamiento de los comentarios de textos agresivos en clases o categorías.	Las categorías que se consideran son: muy agresivo, agresivo, poco agresivo y neutro.	Palabras y/o frases del comentario de texto que transmiten sentimientos negativos.	Palabras y/o frases (dependerá del comentario de texto ingresado por el usuario).
			Palabras y/o frases que la cultura constituya como negativos.	País: Perú.
			Intensidad de las palabras y/o frases para graduar los sentimientos trasmitidos en los comentarios de textos agresivos.	La intensidad de las palabras y/o frases se gradúan en las siguientes categorías: muy agresivo, agresivo, poco agresivo y neutro.

Fuente: Elaboración propia.

3.3 Diseño de investigación

Wentz (2014); McLaren (2014); Creswell (2013a), Hernández et al. (2013) y Kalaian (2008) mencionan que el término diseño de investigación se refiere al plan o estrategia concebida para obtener la información que se desea con el fin de responder al planteamiento del problema (como se citó en Hernández, Fernández y Baptista, 2014, p.128). Para ello, en la presente tesis, se eligió un **diseño experimental con posprueba únicamente y grupo de control,** por la manipulación intencional de la variable independiente con el fin de observar y medir los cambios en la variable dependiente. Según Hernández, Fernández y Baptista (2014), "este diseño incluye dos grupos: uno recibe el tratamiento experimental y el otro no (grupo de control). Es decir, la manipulación de la variable independiente alcanza sólo dos niveles: presencia y ausencia". El diseño se diagrama de la siguiente manera:

RG1 X O1

RG2 — O2

Para realizar una adecuada contrastación de resultados, se llevó a cabo varios experimentos que se mencionan a continuación:

Experimento Nro. 01

El primer experimento se realizó con 01 (una) muestra de 906 palabras agresivas (atributos), 4 categorías (atributos clase) y 800 comentarios de textos agresivos.

Experimento Nro. 02

El segundo experimento se realizó con 01 (una) muestra de 906 palabras agresivas (atributos), 4 categorías (atributos clase) y 1200 comentarios de texto.

Experimento Nro. 03

El tercer experimento se llevó a cabo con 01 (una) muestra de 906 palabras agresivas (atributos), 4 categorías (atributos clase) y 1561 comentarios de textos agresivos.

3.4 Población y muestra

3.4.1 Población

La población, está conformada por comentarios de textos emitidos en publicaciones peruanas principalmente en las redes sociales de Twitter, Facebook y Youtube, del 13/02/2015 al 02/03/2017, siendo esta población infinita.

3.4.2 Muestra

La muestra fue generada de manera no probabilística, utilizando el muestreo discrecional, de tal manera que aporten al estudio, es decir, se tomaron comentarios que se consideren agresivos en la sociedad y que son de mayor relevancia, a continuación, se muestra las páginas de las cuales se tomaron las muestras:

- FanPage de Combate (Facebook); publicación: Los desafortunados dichos de Julieta Rodríguez, La modelo argentina que dedicó estos insultos xenófobos contra los peruanos; fecha de publicación: 13/12/2016.
- @OcioLaRepublica (Twitter); publicación: Compañera de Magaly Medina renuncia a noticiero tras viralización de polémico video; fecha de publicación: 02/03/2017.
- @AlanGarciaPeru (Twitter); publicación: LLegué y asistiré a las 9am a la Fiscalía como testigo sobre el gasoducto firmado el 2014; fecha de publicación: 16/02/2017.
- @9CuentameloTodo (Twitter); publicación: Tu puedes ser la afortunada de conquistar el corazón del guapísimo @GutyCarreraT #CuentameloTodo; fecha de publicación: 01/03/2017.
- Valor de la Verdad @ElValordelaVL (Twitter); publicación: ¿Te decía Guty que no lo amabas cuando no tenían relaciones sexuales?; fecha: 20/08/2016.
- @AmorAmorAmorL (Twitter); publicación: ¡Antonio Pavón confiesa estar enamorado de Milett Figueroa!; fecha de publicación: 11/08/2016.
- FanPage de los Oscar 2017 (Facebook); publicaciones de fecha 27/02/2017.

- Frecuencia Latina (youtube); publicación: Tilsa Lozano opacó a Milett Figueroa con 'topless' en Los Olivos; fecha: 13/11/2014.
- 20 Minutos (periódico); publicación: Maluma responde por fin a las críticas sobre su polémica canción machista 'Cuatro babys'; fecha de publicación: 28/02/2017.
- @keikonova2016 (Twitter); publicación: Súmate a la campaña contra la #Repartija Fujimorista del #BCR. Envíanos una foto (Inbox) con el hashtag #RepartijaFujimoristaNoVa; fecha de publicación: 04/11/2016.
- GamerStudioTV (página de transmisión directa de juegos de DOTA); fecha de recolección de comentarios: 29/01/2017.
- @albertofujimori (Twitter); publicación: ¿Qué otro trámite hay que hacer para lograr reconexión suministro de luz Nº 1307500 ATE ? AYUDA PORFAVOR !!; fecha de publicación: 21/12/2016.
- SENSACINE (página de películas); publicación: Cincuenta sombras de Grey; fecha: 13/02/2015.
- @Soy_Veterinario (Twitter); publicación: Unos le llaman asesinato, los taurinos le llaman “arte”; fecha de publicación: 02/09/2014.
- Fanpage de Miss universo 2017 (facebook); publicaciones de fecha: 30/01/2017.
- La informacion.com (página informativa); publicación: top ten de las 10 comidas más asquerosas de España; fecha de publicación: 20/03/2012.
- @canalN_ (Twitter); publicación: Caso #Odebrecht: Allanan vivienda del expresidente #AlejandroToledo en Camacho - La Molina; fecha de publicación: 04/02/2017.

De todos estos links, se obtuvo un total de 1561 comentarios de textos, los cuales fueron separados en tres muestras: 800, 1200 y finalmente el total (1561). Los comentarios de cada uno de los enlaces, fueron divididos proporcionalmente para formar cada uno de las muestras.

3.5 Método de investigación

En la presente tesis, se empleó el **método hipotético deductivo**. Este método consiste en plantear algunas afirmaciones (hipótesis), para posteriormente verificarlas mediante la deducción, de esta manera se puede sacar conclusiones. El procedimiento que sigue este método es el siguiente: Planteamiento del problema, formulación de hipótesis, deducciones de consecuencias de la hipótesis y la verificación (Sabino, 1992).

El método hipotético deductivo, consiste en un procedimiento que intenta dar respuesta a los distintos problemas que se plantea a través de la postulación de hipótesis que se toman como verdaderas. Estás deberán ser contrastadas. Si el resultado es desfavorable, la hipótesis será considerada refutada y en principio deberá ser abandonada. En cambio, si el resultado fuera favorable, la hipótesis será considerada corroborada (Gianella, 1995).

3.6 Técnicas de instrumentos de acopio de datos

El corpus de conocimiento que fue construido, consta de 1561 comentarios de textos extraídos de las redes sociales y 906 palabras agresivas. Con ayuda de un experto (psicólogo), se determinó el grado de agresividad de los comentarios en las siguientes categorías.

Tabla 7. Categorías de clasificación de los comentarios de textos agresivos

Simbología	Categorías	Rango de Valores
(A+)	Muy Agresivo	7-9
(A)	Agresivo	4-6
(A-)	Poco Agresivo	1-3
(NEU)	Neutro	0

Fuente: Elaboración propia.

Nota: para poder saber en qué categoría está un determinado comentario, se analizó las palabras que este contiene, estás palabras tomaron

los siguientes valores: 0; 1; 2; 3; 4; 5; 6; 7; 8; 9 de acuerdo a su grado de agresividad. Por ejemplo, si un comentario contiene palabras de valor 7,8 y/o 9, será clasificado como muy agresivo, si contiene valor(es) de 4, 5 y/o 6 como agresivo.

Los comentarios, fueron transformados al formato **ARFF** (formato de weka), para un adecuado aprendizaje de las técnicas de minería de texto. El archivo ARFF consta de tres partes: @relation, @attribute y @data.

@relation:

Es una cadena de caracteres que indica lo que se está tratando.

@attribute:

En esta sección se incluyó los atributos (palabras) y su respectivo tipo (en este caso numérico). Así también se declararán los atributos clase (**A+, A, A-, NEU**).

@data:

Aquí, se encuentran los comentarios de textos transformados al formato ARFF de weka. Los comentarios fueron convertidos del tipo nominal a numérico, además se les asignó una clase (neutro, poco agresivo, agresivo o muy agresivo). A continuación, se muestra un pequeño ejemplo:

'eres una persona fea y sin gusto' --- > (0,5,0,0,0,0,0,0,0,0, A)

Para la generación de los números, se buscó cada uno de los atributos en el comentario. Si el atributo se encontraba, se ponía su valor previamente designado (1, 2, 3, 4, 5, 6, 7, 8, 9); caso contrario se ponía 0 (cero). Cuando se terminó de recorrer todos los atributos, se buscó el mayor valor numérico que nos indicó la clase a la que pertenece dicho comentario (neutro, poco agresivo, agresivo o muy agresivo).

La última letra "**A"**, en el caso del ejemplo, es el atributo clase al cual pertenece este comentario de texto.

La estructura de un fichero con formato ARFF es la siguiente:

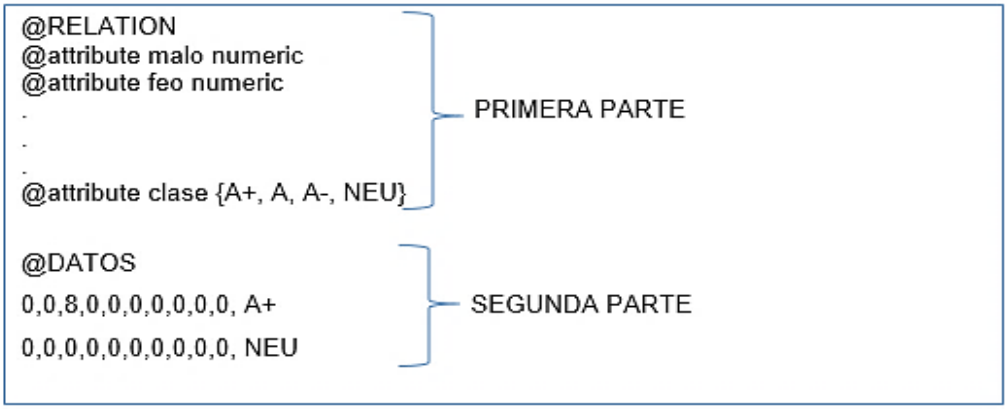

Imagen 12. Fichero con formato ARFF.
Fuente: Elaboración propia.

3.7 Técnicas de análisis de datos

Una vez que las técnicas de minería de textos (algoritmos), hayan aprendido del corpus de conocimiento, se hizo experimentos con otro conjunto de datos de prueba. Es allí donde se vio la precisión y los errores con los que se clasifica un comentario de texto. Para el análisis de los datos se usó la **estadística implementada en weka** como: Kappa statistic, Mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error; pero se prestó especial interés en el Error Absoluto Medio (MAE). Cabe recalcar que weka es una herramienta netamente estadística y es la que se usó para el análisis de datos.

Como lo define Arriagada y Peralta (2008), el Error Absoluto, es la diferencia entre el valor de la medida y el valor tomado como exacto (valor verdadero o valor probable), este error puede ser positivo o negativo, y dependerá de la medida del valor real. Mientras que el Error Absoluto Medio, es el promedio de todos los Errores Absolutos (WordPress, 2017). Es con este último que se determinó que algoritmo es el más apto para la tarea.

CAPITULO IV

PRESENTACIÓN, ANÁLISIS E INTERPRETACIÓN DE RESULTADOS

4.1 Descripción de la solución desarrollada

Para poder saber cómo el aprendizaje computacional basado en técnicas de minería de textos, influye en la clasificación de comentarios de textos agresivos, nos apoyamos de la metodología KDD. Esta metodología nos permitió organizar en fases las tareas que realizamos para llegar a los objetivos. Las etapas que se siguieron fueron:

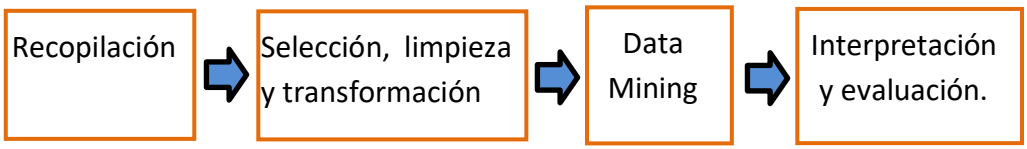

Imagen 13. Diagrama de bloques del proceso KDD.
Fuente: Elaboración propia.

A continuación, se detallan cada uno de estos pasos:

4.1.1 Recolección

En esta primera etapa, se realizó la recolección de comentarios de textos, especialmente agresivos de las principales redes sociales: facebook, twitter y youtube. Se logró recolectar 1561 comentarios de textos de diversos temas: Política, cine, moda, deportes, espectáculos, programas de televisión. Dando prioridad a los comentarios emitidos por los cibernautas peruanos. A continuación, se muestran algunos enlaces de las páginas web de las cuales se extrajo los comentarios:

- FanPage de Combate (Facebook); publicación: Los desafortunados dichos de Julieta Rodríguez, La modelo argentina que dedicó estos insultos xenófobos contra los peruanos; fecha de publicación: 13/12/2016.

- @OcioLaRepublica (Twitter); publicación: Compañera de Magaly Medina renuncia a noticiero tras viralización de polémico video; fecha de publicación: 02/03/2017.
- @AlanGarciaPeru (Twitter); publicación: LLegué y asistiré a las 9am a la Fiscalía como testigo sobre el gasoducto firmado el 2014; fecha de publicación: 16/02/2017.
- @9CuentameloTodo (Twitter); publicación: Tu puedes ser la afortunada de conquistar el corazón del guapísimo @GutyCarreraT #CuentameloTodo; fecha de publicación:01/03/2017.
- Valor de la Verdad @ElValordelaVL (Twitter); publicación: ¿Te decía Guty que no lo amabas cuando no tenían relaciones sexuales?; fecha: 20/08/2016.
- @AmorAmorAmorL (Twitter); publicación: ¡Antonio Pavón confiesa estar enamorado de Milett Figueroa!; fecha de publicación: 11/08/2016.
- FanPage de los Oscar 2017 (Facebook); publicaciones de fecha 27/02/2017.
- Frecuencia Latina (youtube); publicación: Tilsa Lozano opacó a Milett Figueroa con 'topless' en Los Olivos; fecha: 13/11/2014.
- 20 Minutos (periódico); publicación: Maluma responde por fin a las críticas sobre su polémica canción machista 'Cuatro babys'; fecha de publicación: 28/02/2017.
- @keikonova2016 (Twitter); publicación: Súmate a la campaña contra la #Repartija Fujimorista del #BCR. Envíanos una foto (Inbox) con el hashtag #RepartijaFujimoristaNoVa; fecha de publicación: 04/11/2016.
- GamerStudioTV (página de transmisión directa de juegos de DOTA); fecha de recolección de comentarios: 29/01/2017.
- @albertofujimori (Twitter); publicación: ¿Qué otro trámite hay que hacer para lograr reconexión suministro de luz Nº 1307500 ATE ? AYUDA PORFAVOR !!; fecha de publicación: 21/12/2016.
- SENSACINE (página de películas); publicación: Cincuenta sombras de Grey; fecha: 13/02/2015.
- @Soy_Veterinario (Twitter); publicación: Unos le llaman asesinato, los taurinos le llaman “arte”; fecha de publicación: 02/09/2014.
- Fanpage de Miss universo 2017 (facebook); publicaciones de fecha: 30/01/2017.

- La informacion.com (página informativa); publicación: top ten de las 10 comidas más asquerosas de España; fecha de publicación: 20/03/2012.
- @canalN_ (Twitter); publicación: Caso #Odebrecht: Allanan vivienda del expresidente #AlejandroToledo en Camacho - La Molina; fecha de publicación: 04/02/2017.

4.1.2 Selección, limpieza y transformación

Esta es la etapa del preprocesamiento de comentarios. Como se sabe, los comentarios de las redes sociales, son muchas veces escritas de manera incorrecta; también se hace uso de emoticones y caracteres como: @, #, ¡!, ¿?, “”, etc. Que perjudican el estudio adecuado de los datos, para ello es necesario e importante la limpieza de cada uno de ellos. Esta limpieza se realizó de forma manual. A continuación, se muestra un ejemplo:

Comentario antes de su limpieza:

@miky1604 @danilo_camil @confioenti @albertofujimori no decías q era un héroe y valiente y esta llorando como una nena. Lloron y cobarde q se fuge x un tunel.

Comentario después de su limpieza:

@miky1604 @danilo_camil @confioenti @albertofujimori no decías que era un héroe y valiente y está llorando como una nena. Llorón y cobarde que se fugó por un túnel.

Para el uso de las técnicas de minería de textos, los comentarios fueron transformados al formato weka. Weka es una herramienta de minería de datos que posee un conjunto de técnicas para el análisis de textos. Todos los comentarios fueron pasados del tipo nominal al tipo numérico, y puesto a una determinada clase, según la agresividad de las palabras que estes poseen. El formato de weka se denomina ARFF (Attribute-Relation File Format) y está compuesto por una estructura claramente diferenciada en tres partes: cabecera (se define el nombre de la relación), declaraciones de atributos (se declaran los atributos que compondrán el archivo junto a su tipo) y sección de datos (datos que componen la relación, separados entre comas).

La transformación de todos los cometarios al formato weka, nos servirá para entrenar y evaluar a los algoritmos de clasificación. A continuación, se verá una parte pequeña del formato ARFF usado.

```
@RELATION comentarios

@ATTRIBUTE aburrida INTEGER
@ATTRIBUTE aburridas INTEGER
@ATTRIBUTE aburrido INTEGER
@ATTRIBUTE aburridos INTEGER
@ATTRIBUTE acomplejada INTEGER
@ATTRIBUTE acomplejadas INTEGER
@ATTRIBUTE acomplejado INTEGER
@ATTRIBUTE acomplejados INTEGER
.
.
.
@ATTRIBUTE zángano INTEGER
@ATTRIBUTE zánganos INTEGER
@ATTRIBUTE zoquete INTEGER
@ATTRIBUTE zoquetes INTEGER
@ATTRIBUTE zorra INTEGER
@ATTRIBUTE zorras INTEGER
@ATTRIBUTE zorrita INTEGER
@ATTRIBUTE zorritas INTEGER
@ATTRIBUTE zorro INTEGER
@ATTRIBUTE class {NEUTRO, POCOAGRESIVO, AGRESIVO, MUYAGRESIVO}
@DATA
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,8,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
```

Imagen 14. Formato ARFF construido.
Fuente: Elaboración propia.

4.1.3 Data Mining

En esta etapa se utilizó el conjunto de librerías de java que posee la herramienta weka para la extracción de conocimiento desde la base de datos (corpus de conocimiento).

Weka realiza tareas de clasificación, regresión, clustering, asociación y visualización. El presente proyecto se basa en la clasificación, es por ello que se usó los algoritmos de clasificación que nos brinda esta herramienta.

A continuación, se plasma cómo se empleó y aplicó las diferentes técnicas (algoritmos) de Weka al corpus de conocimiento.

- **Primer paso:** Se cargó los datos ya transformados (formato ARFF) en la herramienta weka.

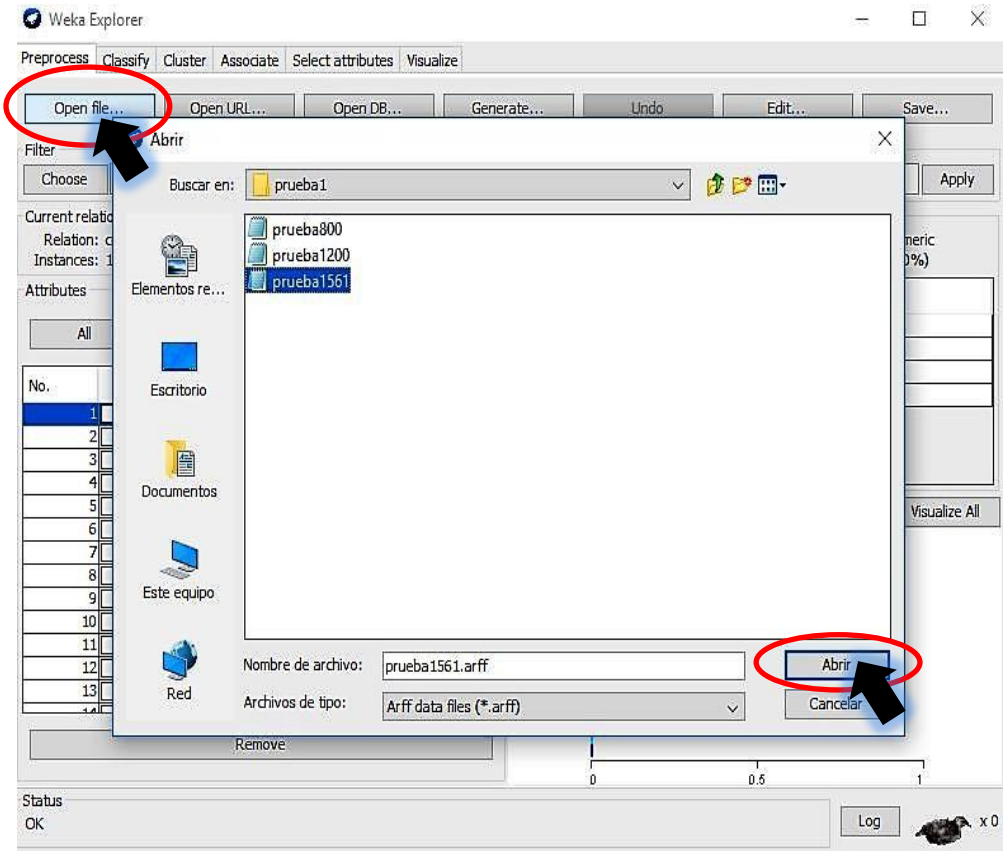

Imagen 15. Cargado del archivo ARFF en la herramienta weka.
Fuente: Elaboración propia.

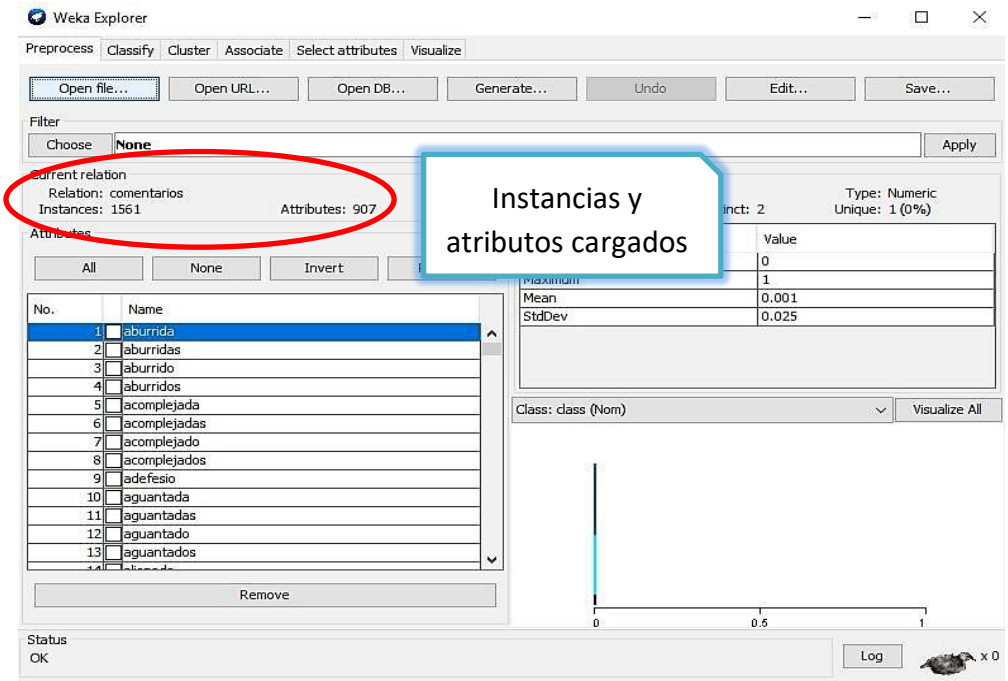

Imagen 16. Verificación de la cantidad de instancias y atributos cargados.
Fuente: Elaboración propia.

- **Segundo paso:** Se eligió la de tarea a realizar, en este caso fue el de clasificación. Luego la técnica y algoritmo a evaluar.

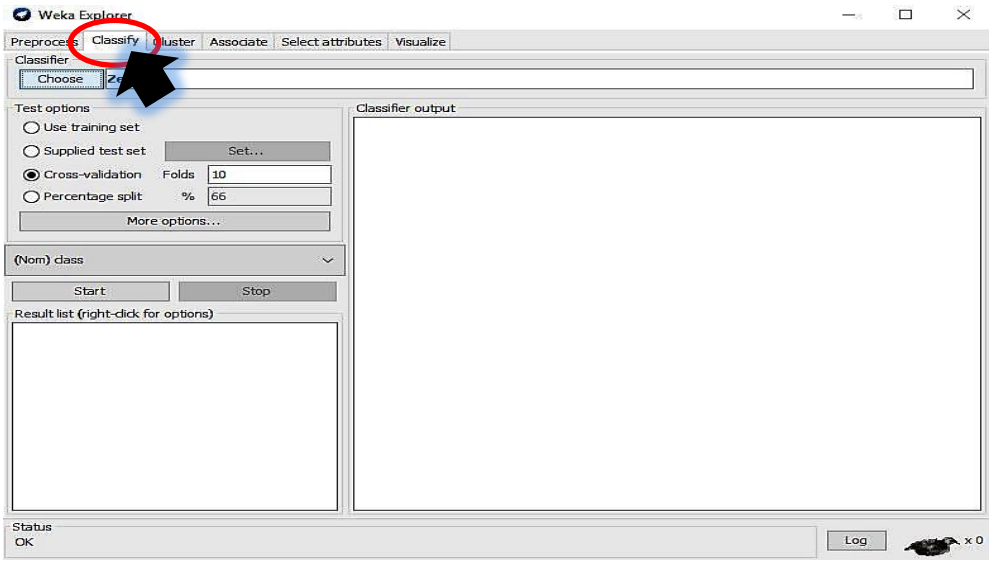

Imagen 17. Elección de la tarea que se realizó (clasificación).
Fuente: Elaboración propia.

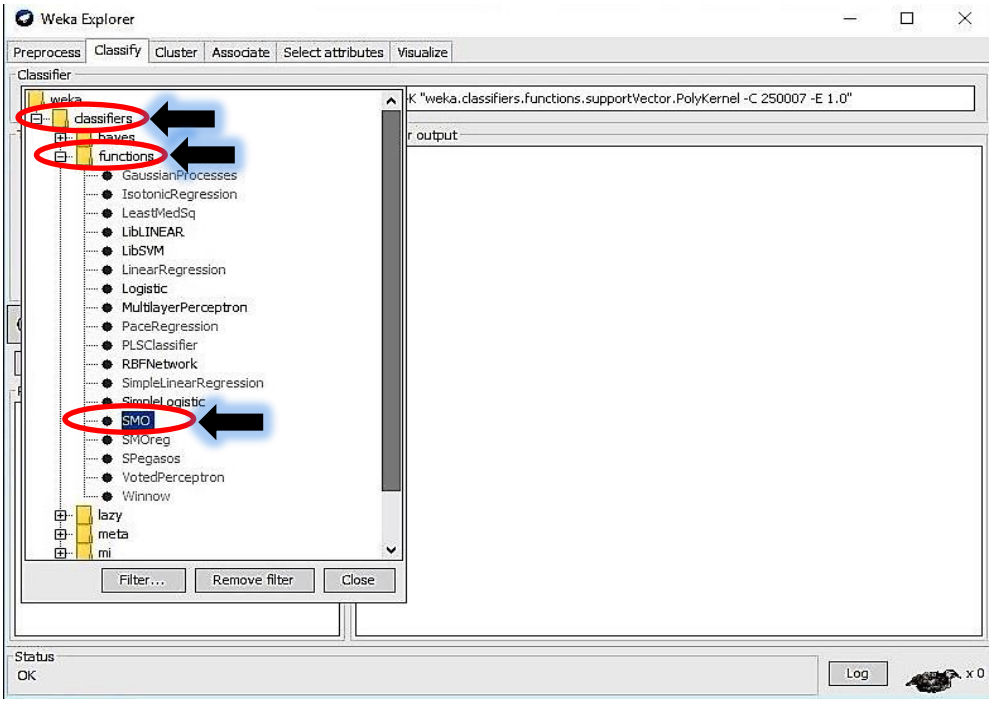

Imagen 18. Técnicas y algoritmos que se evaluaron.
Fuente: Elaboración propia.

- **Tercer paso**: Se eligió el tipo de prueba, es este caso fue Crossvalidation de tamaño 10.

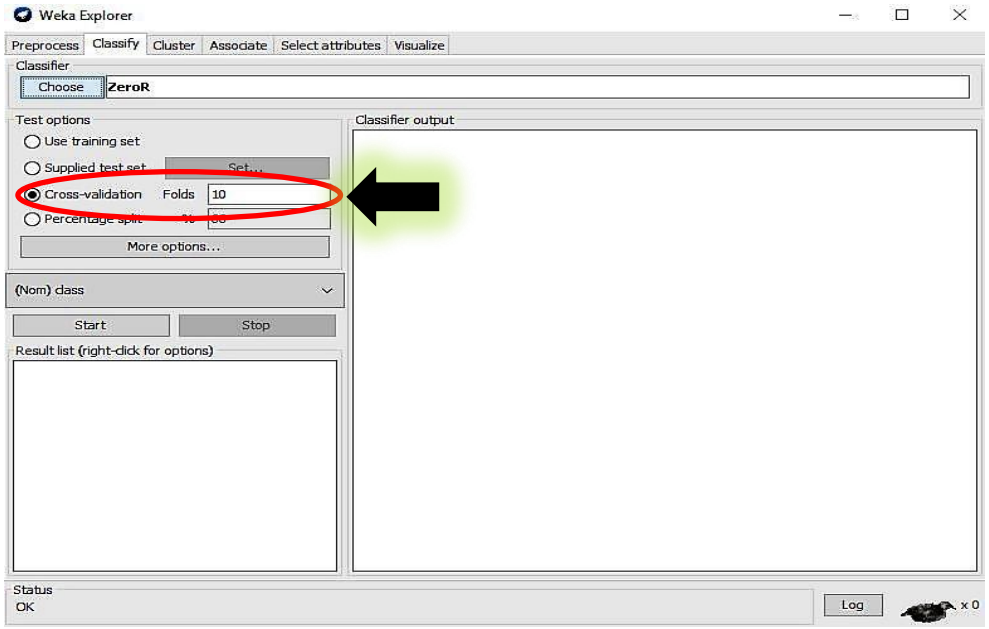

Imagen 19. Prueba Crossvalidation de tamaño 10.
Fuente: Elaboración propia.

- **Cuarto paso:** Se inició de la construcción y validación del modelo de clasificación.

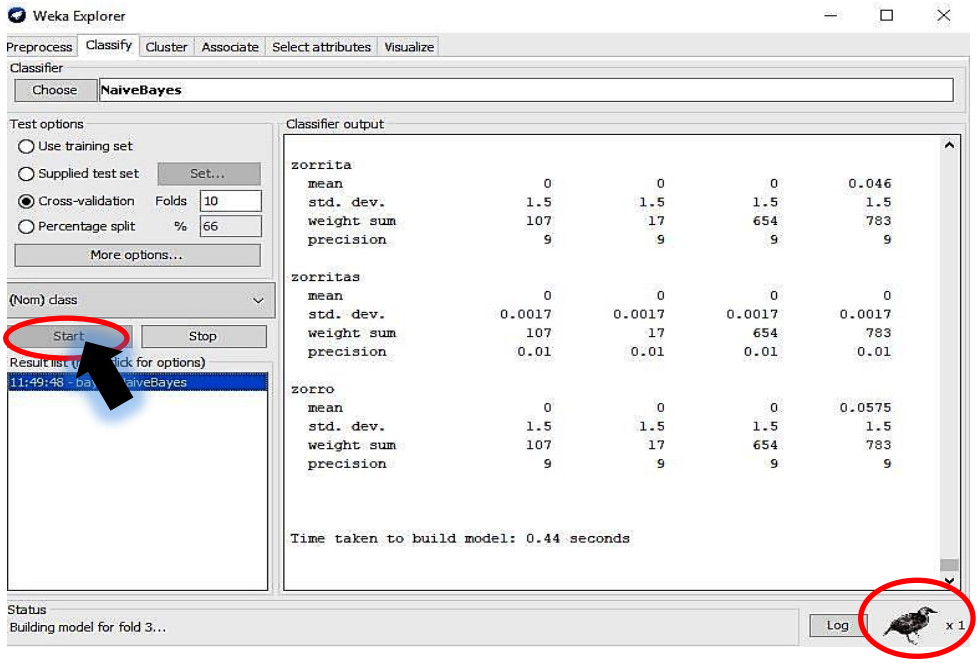

Imagen 20. Construcción del modelo de clasificación.
Fuente: Elaboración propia.

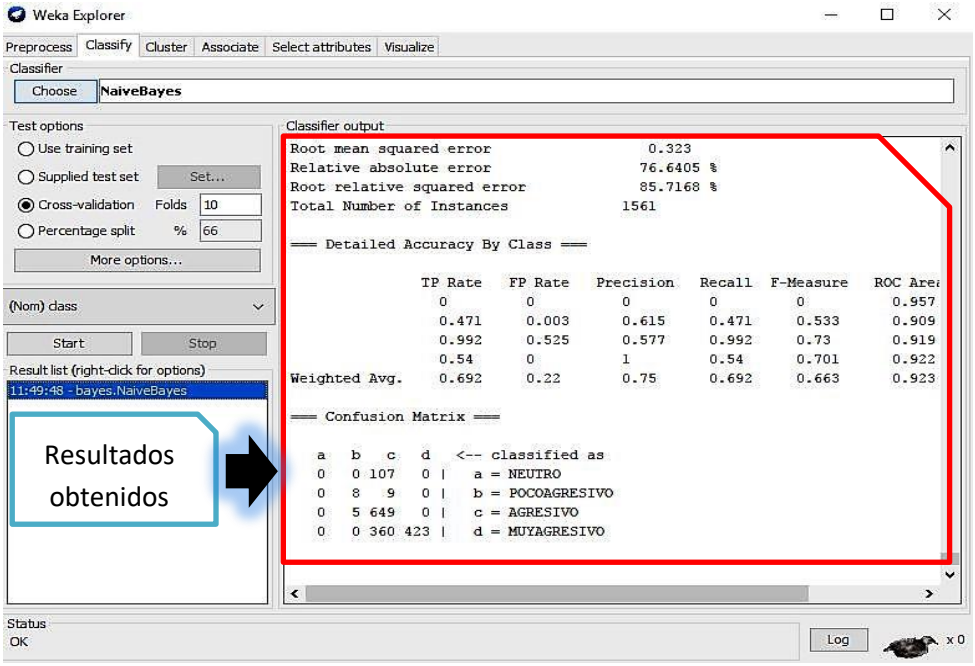

Imagen 21. Validación del modelo de clasificación.
Fuente: Elaboración propia.

- **Quinto Paso:** Se guardó el modelo de clasificación para la implementación de servicio web.

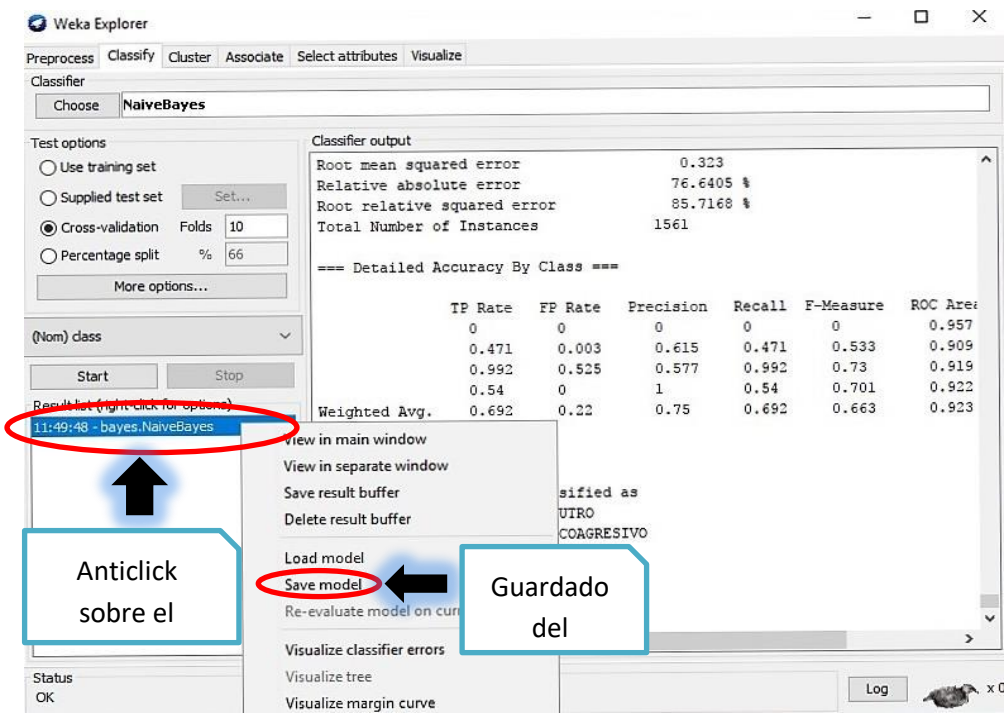

Imagen 22. Guardado del modelo de clasificación.
Fuente: Elaboración propia.

4.1.4 Interpretación y evaluación

Como se vio en la etapa anterior (Data Mining), los algoritmos que fueron evaluados para el corpus de conocimiento son: NaiveBayes, BayesNet, DecisiónStump, J48, OneR, Part, SMO, IBK y KStart. Para la evaluación de estos se usó la prueba estadística del tipo Crossvalidation 10 que nos ofrece weka, este tipo de prueba hace que la muestra original (corpus de conocimiento) se divida aleatoriamente en 10 submuestras. De las 10 submuestras, una de ellas se retiene como datos de validación para probar el modelo y el resto como datos de formación del modelo.

IBK fue el algoritmo que dio mejores resultados frente al corpus de conocimiento, la clasificación que realiza su modelo es más preciso en comparación a los demás.

4.2 Presentación de resultados

Resultados del experimento Nro.01

A continuación, se presenta los resultados obtenidos de los algoritmos NaiveBayes, BayesNet, DecisionStump, J48, OneR, PART, SMO, IBK y KStart con una muestra de 800 comentarios.

Tabla 8. Resultados del experimento nro.01

Algoritmos	Naive Bayes	Bayes Net	Decision Stump	J48	OneR	Part	SMO	IBK	KStart
Error Absoluto Medio	0.21	0.24	0.26	0.25	0.22	0.24	0.27	0.14	0.24
Precisión	71.6%	50.7%	30.3%	58.2%	61.7%	57.6%	79.2%	88.8%	89.4%
Tiempo de generación del modelo	3 s	4 s	2 s	4 s	4 s	3 s	8 s	5 s	2 min

Fuente: Elaboración propia.

Seguidamente, se muestra la gráfica estadística de los Errores Medios Absolutos de todos los algoritmos analizados:

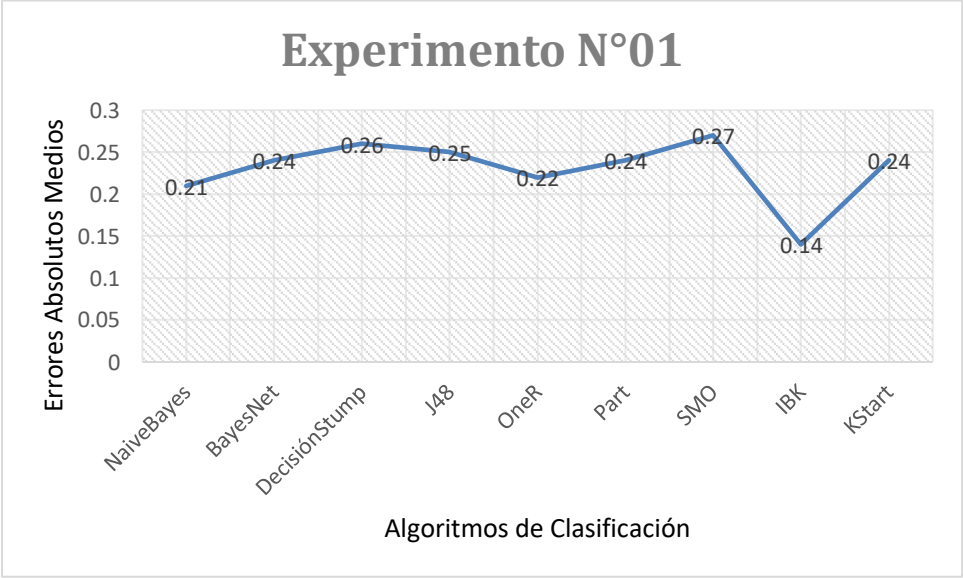

Imagen 23. Gráfica estadística del Error Absoluto Medio (experimento 01).
Fuente: Elaboración propia.

Resultados del experimento Nro.02

A continuación, se presenta los resultados obtenidos de los algoritmos NaiveBayes, BayesNet, DecisionStump, J48, OneR, PART, SMO, IBK y KStart con una muestra de 1200 comentarios.

Tabla 9. Resultados del experimento nro.02

Algoritmos	Naive Bayes	Bayes Net	Decision Stump	J48	OneR	Part	SMO	IBK	KStart
Error Absoluto Medio	0.22	0.23	0.27	0.25	0.24	0.25	0.27	0.11	0.38
Precisión	77.2%	72.9%	26.4%	70.8%	52.4%	70.8%	80.3%	89%	5%
Tiempo de generación del modelo	4 s	5 s	3 s	6 s	5 s	7 s	17 s	9 s	6 min 5 s

Fuente: Elaboración propia.

Seguidamente, se muestra la gráfica estadística de los Errores Medios Absolutos de todos los algoritmos analizados:

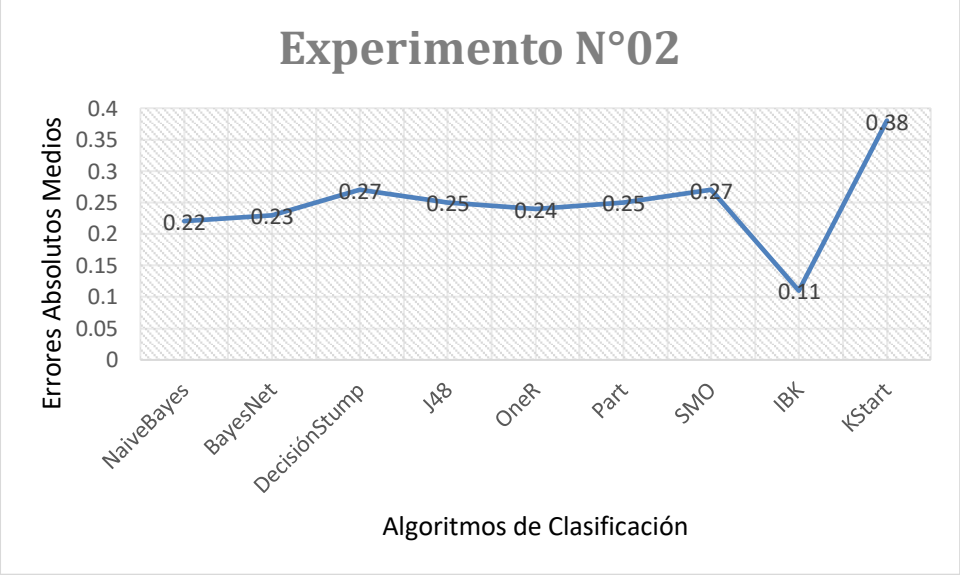

Imagen 24. Gráfica estadística del Error Absoluto Medio (experimento 02).
Fuente: Elaboración propia.

Resultados del experimento Nro.03

A continuación, se presenta los resultados obtenidos de los algoritmos NaiveBayes, BayesNet, DecisionStump, J48, OneR, PART, SMO, IBK y KStart con una muestra de 1561 comentarios.

Tabla 10. Resultados del experimento nro.03

Algoritmos	Naive Bayes	Bayes Net	Decision Stump	J48	OneR	Part	SMO	IBK	KStart
Error Absoluto Medio	0.22	0.22	0.26	0.24	0.25	0.24	0.27	0.09	0.38
Precisión	75%	73.2%	54.8%	71.1%	57.3%	71%	82.1%	89.3%	5%
Tiempo de generación del modelo	6 s	6 s	4 s	8 s	6 s	16 s	26 s	14 s	12 min 24 s

Fuente: Elaboración propia.

Seguidamente, se muestra la gráfica estadística de los Errores Medios Absolutos de todos los algoritmos analizados:

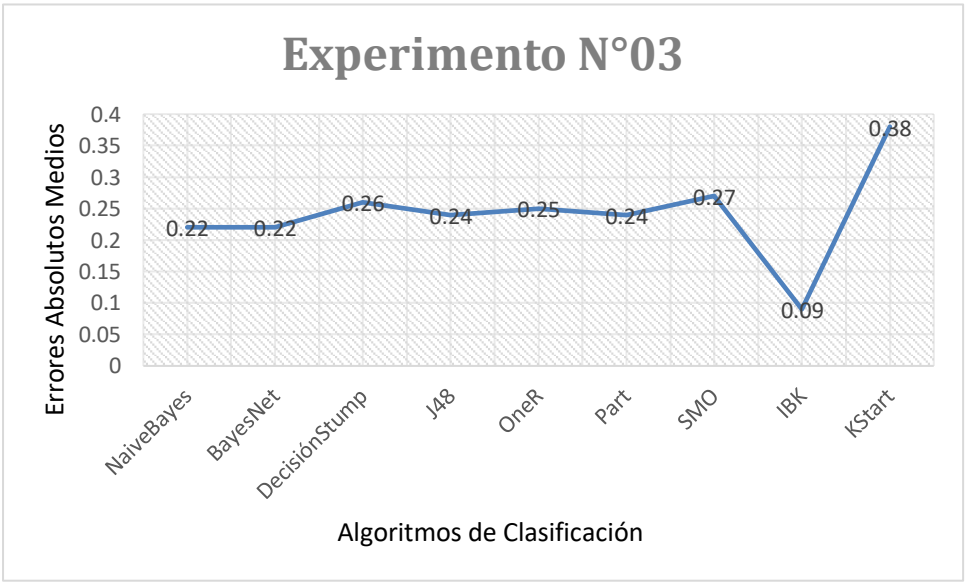

Imagen 25. Gráfica estadística del Error Absoluto Medio (experimento 03).
Fuente: Elaboración propia.

Resumen de los resultados del experimento 01, 02 y 03

A continuación, se presenta el resumen de los resultados obtenidos en el experimento 01, 02 y 03 de los algoritmos NaiveBayes, BayesNet, DecisionStump, J48, OneR, PART, SMO, IBK y KStart.

Tabla 11. *Tabla con el resumen de resultados de los experimentos 01, 02 y 03*

Algoritmos	Naive Bayes	Bayes Net	Decision Stump	J48	OneR	Part	SMO	IBK	KStart
Error Absoluto Medio con 800 comentarios	0.21	0.24	0.26	0.25	0.22	0.24	0.27	0.14	0.24
Error Absoluto Medio con 1200 comentarios	0.22	0.23	0.27	0.25	0.24	0.25	0.27	0.11	0.38
Error Absoluto Medio con 1561 comentarios	0.22	0.22	0.26	0.24	0.25	0.24	0.27	0.09	0.38

Fuente: Elaboración propia.

Seguidamente, se muestra la gráfica estadística con el resumen de los Errores Medios Absolutos obtenidos en los experimentos 01, 02 y 03.

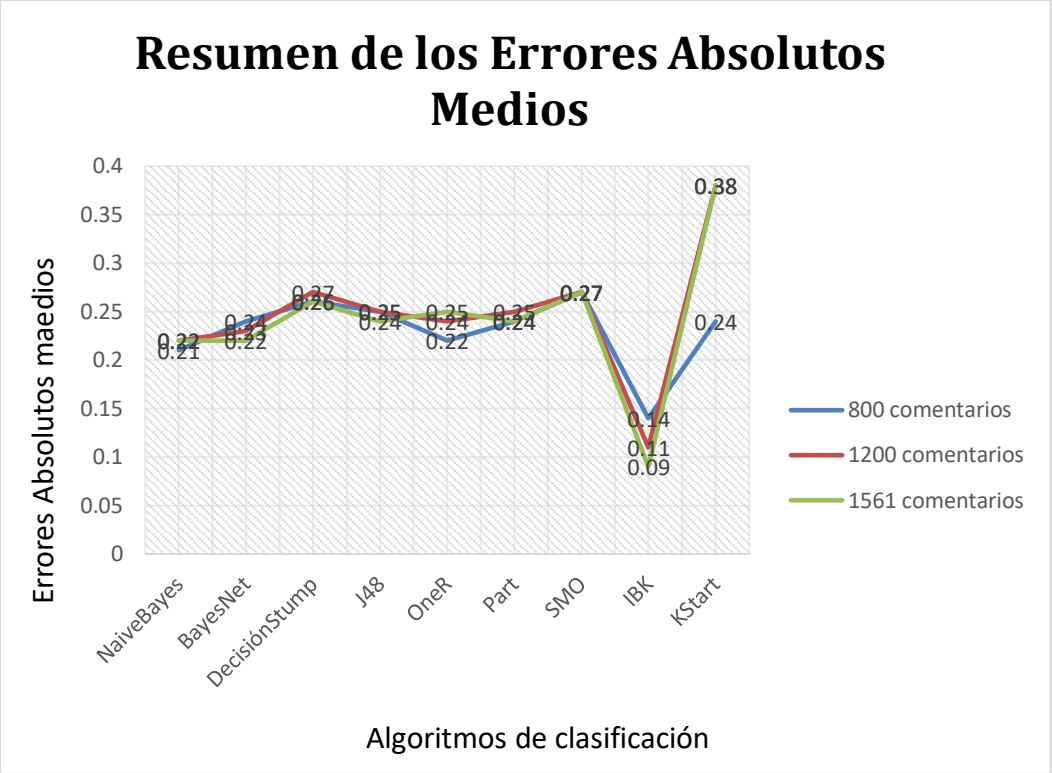

Imagen 26. Gráfica estadística con el resumen de los Errores Absolutos Medios de los tres experimentos. Fuente: Elaboración propia.

4.3 Análisis y discusión de resultados

4.3.1 Resultados del experimento Nro.01

El experimento se realizó con una muestra de 800 comentarios tomados principalmente de las redes sociales de Facebook, Twitter y Youtube. Los algoritmos que participaron en el experimento fueron: NaiveBayes, BayesNet, DecisionStump, J48, OneR, PART, SMO, IBK, KStart. Se realizó una prueba estadística del tipo Crossvalidation 10 que nos ofrece weka, este tipo de prueba hace que la nuestra original se divida aleatoriamente en 10 submuestras. De las 10 submuestras, una de ellas se retiene como datos de validación para probar el modelo y el resto como datos de formación del modelo, esto quiere decir que se realizó 10 validaciones cruzadas en 10 hojas cada una, lo cual implica que el algoritmo de aprendizaje se ejecutó 100 veces. En el experimento se comparó los Errores Medios Absolutos con un nivel alfa (significancia) del 0.05. El algoritmo IBK es

el que tiene el menor Error Absoluto Medio, siendo este 0.14; mientras que en el resto de los algoritmos el Error Absoluto Medio es mayor a 2 como se puede observar en la tabla 8. Por lo tanto, el modelo de clasificación generado por IBK junto al corpus de conocimiento es el más eficiente para la clasificación de cometarios de textos agresivos.

4.2.2 Resultados del experimento Nro.02

Este experimento se realizó con una muestra de 1200 comentarios, tomados también de las redes sociales de Facebook, Twitter y Youtube. Los algoritmos que participaron en este experimento, fueron los mismos del experimento nro.01. Del mismo modo, se realizó la prueba estadística Crossvalidation 10. Al comparar los Errores Medios Absolutos de los algoritmos con un nivel alfa (significancia) del 0.05, nuevamente IBK fue el de menor Error Absoluto Medio, siendo este ahora 0.11. Mientras que el algoritmo KStar incrementó su error significativamente a un 0.38, los resultados se muestran en la tabla 9.

Al igual que el experimento 01, el modelo de clasificación generado por el algoritmo IBK junto al corpus de conocimiento, es el más eficiente para la clasificación de comentarios de textos agresivos.

4.2.3 Resultados del experimento Nro.03

Para este experimento se utilizó una muestra de 1561 comentarios, tomados también de las redes sociales de Facebook, Twitter y Youtube. Los algoritmos que se evaluaron fueron los mismos del experimento 01 y 02. Se empleó la misma prueba estadística que los dos experimentos anteriores (Crossvalidation 10). Al comparar los Errores Medios Absolutos de los algoritmos con un nivel alfa (significancia) del 0.05, IBK fue nuevamente el algoritmo que obtuvo menor Error Absoluto Medio, siendo este 0.09, como se muestra en la tabla 10.

Del mismo modo que el experimento 01 y 02, el modelo de clasificación generado por el algoritmo IBK junto al corpus de conocimiento, es el más eficiente para la clasificación de comentarios de textos agresivos.

4.3 Comprobación de hipótesis

4.3.1 Comprobación de las hipótesis específicas

Hipótesis específica 1

H_1: El corpus de conocimiento influirá de manera directa en la precisión de la clasificación de comentarios de textos agresivos basados en las técnicas de minería de textos, con una aproximación mayor al 85%.

De acuerdo a los resultados obtenidos en el experimento 01, 02 y 03, el algoritmo más eficiente para el corpus de conocimiento construido es IBK, que pertenece a los clasificadores LAZY. El corpus de conocimiento influye de manera proporcional en la clasificación de comentarios de textos agresivos, ya que con una muestra de 800 comentarios la precisión es de 88.8%; con una muestra de 1200 comentarios la precisión es de 89%; mientras que con una muestra de 1561 comentarios la precisión es de 89. 3%, lo cual indica que, a más comentarios de textos en el corpus de conocimiento, mayor es la precisión de la clasificación. Por lo tanto, la hipótesis H_1 es verdadera.

Hipótesis específica 2

H_2: El corpus de conocimiento influirá proporcionalmente en el tiempo de generación de los modelos de clasificación de comentarios de textos agresivos basados en las técnicas de minería de textos, siendo esta no mayor a 20 minutos.

El corpus de conocimiento influye proporcionalmente en el tiempo de generación de los modelos de clasificación de cometarios de textos agresivos. Veamos los casos de las construcciones de los modelos de clasificación generados con el algoritmo IBK y el corpus de conocimiento en diversos tamaños.

Para una muestra de 800 cometarios que conforman el corpus de conocimiento el tiempo fue de 5 s, para 1200 comentarios fue 9 s, mientras que para 1561 comentarios fue de 14 s. Cabe mencionar también que el algoritmo que empleó mayor tiempo en la construcción de su modelo con una data de 1561 cometarios, fue KStart, siendo este 12 min con 5s. Esto quiere decir que a mayor corpus de conocimiento mayor es el tiempo que se emplea en la

generación del modelo de clasificación. Por lo tanto, la hipótesis H_2 es verdadera.

Hipótesis específica 3

H_3: El corpus de conocimiento influirá directamente sobre la elección del algoritmo de clasificación del grupo de algoritmos clasificadores basados en las técnicas de minería de textos, siendo el elegido IBK por tener el menor Error Absoluto Medio.

Como se puede visualizar en los experimentos 01, 02 y 03; se analizó diversos algoritmos del grupo de clasificadores de la herramienta weka junto al corpus de conocimiento construido.

Combinando cada algoritmo con el corpus de conociendo, se obtuvo un modelo de clasificación, que sometida a una prueba del tipo Crossvalidation 10 con un nivel alfa(significancia) del 0.05, nos mostró los diversos Errores Absolutos Medios.

En el experimento 01, IBK tuvo un Error Absoluto Medio de 0.14, siendo este el más bajo de todos (ver tabla 8). En el experimento 02, IBK nuevamente obtuvo el menor error siendo este 0.11 (ver tabla 9). Mientras que en el experimento 03, IBK adquirió un error de 0.09, siendo este otra vez el más inferior de todos (ver tabla 10). Por lo tanto, la hipótesis H_3 es verdadera.

Hipótesis específica 4

H_4: El corpus de conocimiento influirá proporcionalmente sobre la cantidad de datos óptimos para la clasificación de comentarios de textos agresivos basadas en las técnicas de minería de textos, siendo la cantidad óptima 1561 registros de comentarios y 906 atributos (palabras).

Como IBK es el algoritmo de menor Error Absoluto Medio, veamos como el corpus de conocimiento influyó sobre este. En el experimento 01, se obtuvo una precisión del 88.8%, al trabajar con una muestra de 800 comentarios y 906 atributos. Incrementando la cantidad de comentarios a 1200 y manteniendo la cantidad de atributos en el experimento 02, la precisión subió a un 89%. Y finalmente en el experimento 03, se acrecentó los comentarios a 1561, obteniéndose una precisión del 89.3%, siendo esta muy buena para la

clasificación de comentarios de textos agresivos. Por lo tanto, la hipótesis H_4 es verdadera.

4.3.1 Comprobación de la hipótesis general

H_0: El corpus de conocimiento influirá de manera directamente proporcional en la clasificación de comentarios de textos agresivos, basados en las técnicas de minería de textos.

Con los resultados obtenidos en el experimento 01, 02 y 03, vemos que el corpus de conocimientos influye de manera directa en el tiempo, precisión, elección del algoritmo de clasificación y en la cantidad de datos óptimos para la clasificación de comentarios de textos agresivos. Por lo tanto, la hipótesis H_0 es verdadera.

CAPITULO V

IMPLEMENTACIÓN DEL SERVICIO WEB

Después de haber conocido qué modelo es el más eficiente para la clasificación de comentarios de textos agresivos, se prosiguió a implementar el servicio web. Para organizar las actividades desarrolladas, se usó el modelo cascado. Los pasos que se siguieron fueron: análisis, diseño, implementación y pruebas. A continuación, se detalla cada uno de ellos:

Análisis:

En esta etapa, se definió la función que debe realizar el servicio web. El objetivo fundamental del servicio web, es la clasificación de comentarios de textos en categorías nominales (muy agresivo, agresivo, poco agresivo o neutro). Los actores involucrados en este sistema son los cibernautas y el servicio web.

Diseño:

Para poder desarrollar el servicio web, se eligió NetBeans IDE 8.2 como entorno de desarrollo, GlassFish 4.1.1 como servidor y Jelastic como plataforma de servicio. También se realizó un diagrama de componentes para poder identificar los archivos y clases java que debía generarse, y ver la dependencia que existe entre ellas. A continuación, se muestra el diagrama de componentes.

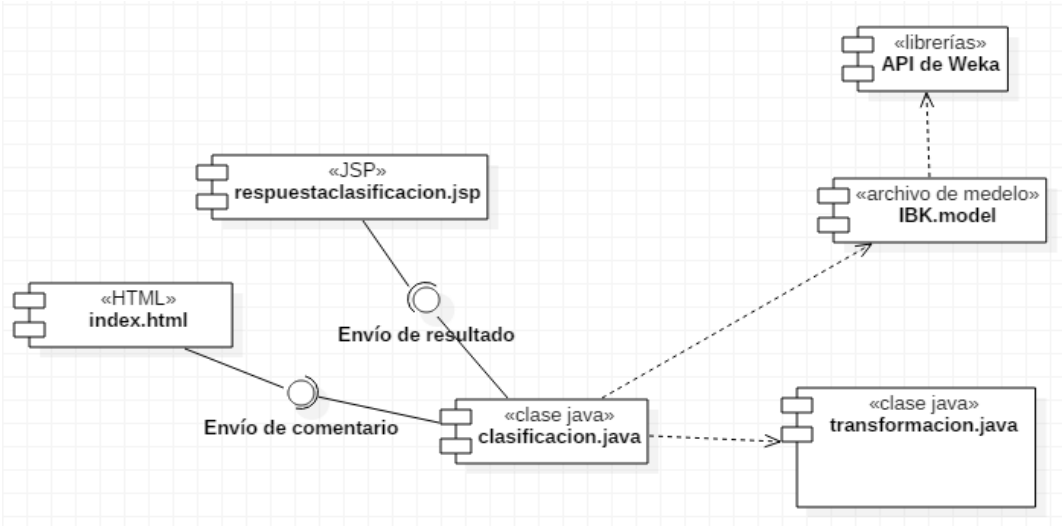

Imagen 27. Diagrama de componente.
Fuente: Elaboración propia.

Así mismo, se definió las tareas de cada uno de estos componentes.

- **API WEKA:** Se usó la colección de algoritmos de la versión 8.2 de weka para el análisis de datos y la construcción del modelo de clasificación cuya extensión es
  ". model".
- **IBK.model:** Es un archivo generado por la API de WEKA. El modelo que se incorporó al servicio web para la clasificación de comentarios agresivos fue el de IBK por haber obtenido mejores resultados (ver capítulo IV).
- **transformacion.java:** Esta clase fue implementada en NetBeans, realiza la transformación de un comentario de texto ingresado por el usuario al formato ARFF para ser reconocido por el modelo de clasificación.
- **clasificacion.java:** Es la clase que incorpora al modelo de IBK y el producto generado por transformacion.java, para darnos como resultado la clasificación en un comentario.
- **index.html:** Es la página que se muestra al usuario para que pueda ingresar un comentario de texto y visualizar el resultado de su clasificación.
- **respuestaclasificacion.jsp:** Aquí, se muestra al usuario el resultado de la clasificación de su comentario.

Implementación: En esta fase, se prosiguió a programar a nivel local el servicio web en el IDE de NetBeans. Luego se generó el archivo de aplicación web (war) para ser cargada y desplegada en la plataforma como servicio Jelastic. A continuación, se muestra un diagrama de despliegue en el cual se visualiza el lugar donde se implementan y ejecutan cada componente.

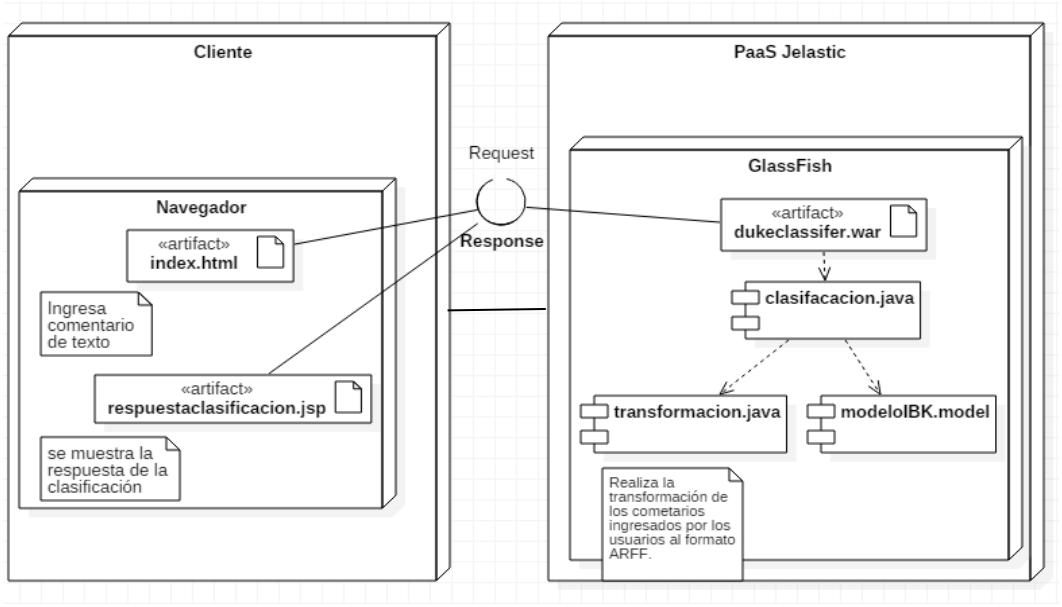

Imagen 28. Diagrama de despliegue.
Fuente: Elaboración propia**.**

Enseguida, se muestra los resultados que generaron cada uno de los compontes y artefactos.

- **Index.html:** En esta interfaz, se podrá ingresar el comentario de texto que desea clasificar.

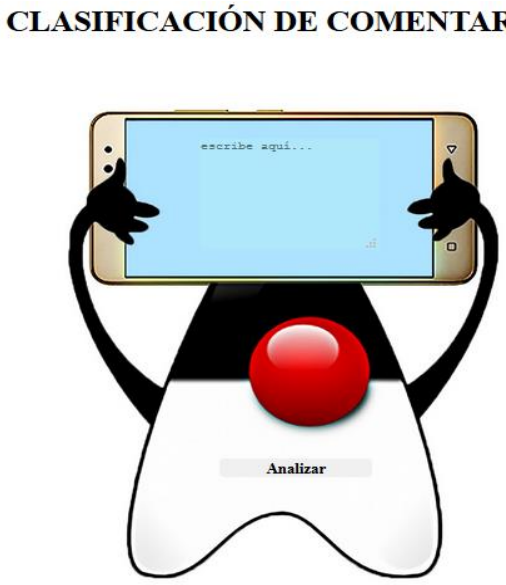

Imagen 29. Interfaz que se muestra al usuario.
Fuente: Elaboración propia.

- **modeloIBK.model:** Será utilizado por la clase clasificacion.java para clasificar nuevos comentarios ingresados por los usuarios. A continuación, se muestra la información del modelo.

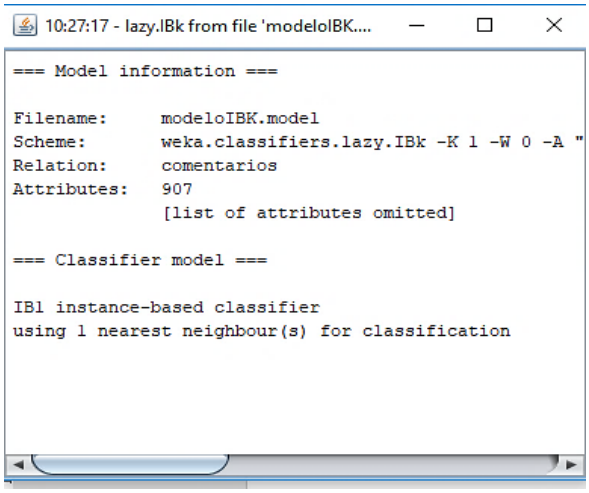

Imagen 30. Información del modelo de clasificación.
Fuente: Elaboración propia.

- **tranformacion.java:** Transforma el comentario de texto ingresado por el usuario al formato arff, para poder ser reconocido y clasificado por el modelo de clasificación.

```
run:
Comentario:
esta es una casa hermosa
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,,?
```

Comentario sin clasificación

Imagen 31. Resultado obtenido de la clase transformacion.java
Fuente: Elaboración propia.

- **clasificacion.java:** Se encarga de cargar el modelo de clasificación (model.IBK) para darle una clase al comentario transformado por tranformacion.java.

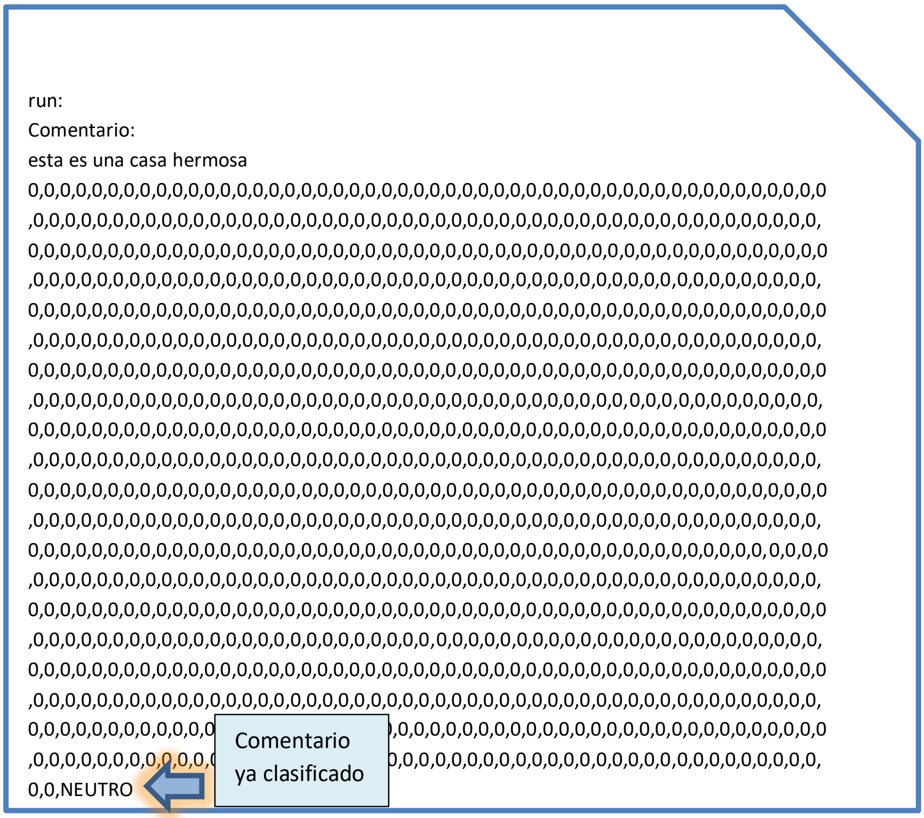

Imagen 32. Resultado obtenido de la clase clasificacion.java
Fuente: Elaboración propia.

- **respuestaclasificacion.jsp:** Muestra el resultado de la clasificación al usuario.

Pruebas: Una vez implementado el servidor, se verificó que todas sus clases realicen las tareas de manera correcta, de esta manera se pudo obtener sin problemas el resultado final de la clasificación, razón para la cual fue construida.

CONCLUSIONES

1. Con los resultados obtenidos en el experimento 01, 02 y 03, vemos que el corpus de conocimientos influye de manera directa, en la precisión, tiempo, elección del algoritmo de clasificación y en la cantidad de datos óptimos para la clasificación de comentarios de textos agresivos.

2. El algoritmo que tiene menor Error Absoluto Medio (0.09) para el corpus de conocimiento construido (1561 comentarios) es IBK. Por lo tanto, su modelo, es el más eficiente para la clasificación de comentarios agresivo.

3. El corpus de conocimiento, influye de manera directamente proporcional en la precisión de la clasificación de comentarios de textos. El algoritmo IBK que es el más óptimo para la clasificación de comentarios de textos agresivos, con una data de 800 comentarios, tuvo una precisión del 88.8%; con una data de 1 200 comentarios un 89%, mientras que con 1 561 comentarios un 89.3%. Esto quiere decir que cuanto más grande sea la data del corpus de conocimiento, mayor será la precisión de la clasificación.

4. El corpus de conocimiento, influye de manera proporcional en el tiempo de generación de los modelos de clasificación. El algoritmo IBK con una data de 800 comentarios, generó el modelo de clasificación en 5 s; con una data de 1 200 comentarios, lo realizó en 9 s; mientras que con una data de 1561 se tardó 14 segundos. Es decir, cuanto más grande sea el corpus de conocimiento, mayor será el tiempo empleado en construir los modelos de clasificación.

5. La cantidad de datos óptimos para la clasificación de comentarios de textos agresivos, es de 906 atributos (palabras) y 1561 comentarios, ya que, con esta cantidad, se obtiene una precisión del 89.3%.

RECOMENDACIONES

Se plantean algunas recomendaciones con el fin de mejorar los resultados obtenidos en la presente investigación. Pero sobre todo con un objetivo mayor que es el de mitigar el crecimiento del ciberbullying a través de comentarios de textos agresivos. Estas recomendaciones son:

- Seguir incrementando el tamaño del corpus de conocimiento, cuanto mayor data mejor serán los resultados obtenidos por el algoritmo IBK.

- Realizar el análisis de comentarios de textos, incluyendo emoticones que son símbolos que reflejan estados de ánimo.

- Ampliar el estudio de palabras vulgares más usadas en países de habla hispana.

- Implementar un corrector de abreviaturas agresivas usadas con mayor frecuencia.

- Realizar la clasificación de comentarios de textos agresivos en otros idiomas diferentes del español.

- Emplear el método stemming para reducir palabras a su raíz y disminuir los atributos del corpus de conocimiento.

- Usar la herramienta de minería de datos Weka, por ser de licencia GNU-GPL.

BIBLIOGRAFÍA

Abelleira, M. A. P., y Cardoso, C. A. (2010). Minería de texto para la categorización automática de documentos. Carnegie Mellon University, Madrid, España.

Advanced Tech Computing Group UTPL (2008, 14 de abril). *Clasificación supervisada y no supervisada.* [Mensaje en un blog]. Recuperado de https://advancedtech.wordpress.com/2008/04/14/clasificacion-supervisada-y-no-supervisada/

Alfonso M. I., Botía A., Mora F. y Trigueros J. P. (2005). *Ingeniería del software: Ingeniería del software (7ma ed.)*. España: Pearson. Recuperado de https://ingenieriasoftware2011.files.wordpress.com/2011/07/ingenieria-de-software-ian-sommerville-7ma-edicion-prentice-hall.pdf

Aníbal G. (2009, 11 de agosto). *CRISP-DM, Una metodología para proyectos de Minería de Datos.* [Mensaje en un blog]. Recuperado de https://anibalgoicochea.com/2009/08/11/crisp-dm-una-metodologia-para-proyectos-de-mineria-de-datos/

Apaza Delgado, S. H. (2016). *MODELO COMPUTACIONAL DE MINERÍA DE MICROBLOGS PARA EL ANÁLISIS DEL COMPORTAMIENTO DEL CONSUMIDOR DE TELEFONÍA CELULAR.* (Tesis doctoral). Pontificia Universidad Católica del Perú, Perú. Recuperado de http://tesis.pucp.edu.pe/repositorio/handle/123456789/7515

Arriagada E. y Victor.D (2008). *ERROR ABSOLUTO ERROR RELATIVO*. Recuperado de http://colabora.inacap.cl/sitios/merlot/Materiales%20MerlotChile/mlcastro/Ciencias%20y%20Tecnolog%C3%ADa/F%C3%ADsica/Manuales%20F%C3%ADsica%20Mec%C3%A1nica/Error%20absoluto%20error%20relativo.pdf

Bordignon, F. (2007, 06 de junio). *Clasificacíón de textos por el método Naive Bayes* [Información de un Blog]. Recuperado de http://ferbor.blogspot.pe/2007/06/clasificacn-de-textos-por-el-mtodo.html

Boris, A. y Cebrián, Z. (2017). *Inteligencia en redes de comunicaciones Practica WEKA Diagnóstico cardiología.* Recuperado de file:///D:/12mem.pdf

Carrasco Ortiz, M. A. & González Calderón J. (2006). Aspectos conceptuales de la agresión: definición y modelos explicativos. Psicología, 2(4), 7-38. Recuperado de http://revistas.uned.es/index.php/accionpsicologica/article/viewFile/478/417

Cataldi, L. (2013). BULLYING: SEGUN LAS ULTIMAS ESTADISTICAS HAY RELACION ENTRE LOS CASOS DE ACOSO ESCOLAR Y EL SUICIDIO JUVENIL [Información de un Blog]. Recuperado de http://mediacionyviolencia.com.ar/bullying-segun-las-ultimas-estadisticas-hay-realacion-entre-los-casos-de-acoso-escolar-y-el-suicidio-juvenil/

Cornejo, R. (2014, 17 de octubre). Ciberbullying, una amenaza que crece entre los escolares en el Perú. *Perú21*. Recuperado de http://peru21.pe/actualidad/cedro-ciberbullying-amenaza-que-crece-entre-escolares-2201471

Cortizo J. C. (2011, 13 de mayo). *Minería de Opiniones, o Análisis del Sentimiento.* [Mensaje en un blog]. Recuperado de http://www.baquia.com/emprendedores/2011-05-13-mineria-de-opiniones-o-analisis-del-sentimiento

Departamento de Ciencias de la Computación e I.A. (2009). *La eficiencia de los algoritmos Análisis y Diseño de Algoritmos.* Recuperado de http://elvex.ugr.es/decsai/algorithms/slides/2%20Eficiencia.pdf

Departamento de Informática Universidad de Valladolid Campus de Segovia (2007). *Complejidad Algorítmica.* Recuperado de https://www2.infor.uva.es/~jvalvarez/docencia/tema5.pdf

Dorronsoro Ibero, J. R. (2013). Algoritmos SVM para problemas sobre big data. (Tesis de maestría). Universidad Autónoma de Madrid, Escuela Politécnica Superior - Departamento de Ingeniería Informática, Madrid, España.

Dubiau, L. (2013). *Procesamiento de Lenguaje Natural en Sistemas de Análisis de Sentimientos*. (Tesis). Universidad de Buenos Aires, Facultad de Ingeniería, Argentina. Recuperado de http://materias.fi.uba.ar/7500/Dubiau.pdf

Duda, R, Hart, P. y Stork, D. (2012). *Pattern classification (2da ed.)*. John Wiley & Sons. Recuperado de https://books.google.com.pe/books?hl=es&lr=&id=Br33IRC3PkQC&oi=fnd&pg=PR3&dq=Duda,+R.+%26+Hart,+P.+(1973).+Pattern+classification+and+scene+analysis.+New+York&ots=2wFPLsbbKu&sig=2QZwb4XLCdg4s3Qlxqwqe9qU9c4#v=onepage&q&f=false

ESET (2013, 23 de octubre). Semana de la No Violencia: ¿cuál es el panorama respecto al cyberbullying en Latinoamérica? [Información de un Blog]. Recuperado de http://www.somoseset.com/2013/10/23/semana-no-violencia-panorama-cyberbullying-latinoamerica/#more-19323

Explorable.com (2009, 21 de septiembre). Investigación Empírica [Información de un Blog]. Recuperado de https://explorable.com/es/investigacion-empirica

Fenghua Huang y and Luming Yan (2014). Combined Kernel-Based BDT-SMO Classification of Hyperspectral Fused Images. Recuperado de https://www.hindawi.com/journals/tswj/2014/738250/

Ferri C. (2012, 27 de mayo). *Mi página de Weka.* [Mensaje en un blog]. Recuperado de http://users.dsic.upv.es/~cferri/weka/

Gala García, Y. (2013). *Algoritmos SVM para problemas sobre big data*. (Tesis). Universidad Autónoma de Madrid, departamento de ingeniería informática. Recuperado de https://repositorio.uam.es/bitstream/handle/10486/14108/66152_Yvonne_Gala_Garcia.pdf?sequence=1

Galán Nieto, A. M. (2007). *Filtrado Colaborativo y Sistemas de Recomendación.* Recuperado de http://www.it.uc3m.es/jvillena/irc/practicas/06-07/31.pdf

García J. y Molina J. M. (2012). *Técnicas de Análisis de Datos*. Recuperado de http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos/libroDataMiningv5.pdf

García Maldonado G., Joffre Velázquez V. M., Martínez Salazar G. J., Llanes Castillo A. (2011). Ciberbullying: forma virtual de intimidación escolar. Salud, Psicología y Educación ,1(40), 115-130. Recuperado de http://www.scielo.org.co/pdf/rcp/v40n1/v40n1a10.pdf

Gianella A. E. (1995). *Introducción a la epistemología y a la metodología de la Ciencia*. España: Universidad Nacional de La Plata (pp. 39-128). Recuperado de http://miel.unlam.edu.ar/recursos/contenido/2403-B/El-Metodo-Hipotetico-Deductivo.pdf

Jiménez, A., Márquez, V., López M. (2015). R. M. Zapata, E. Soriano, A. J. Gonzales, V.V. Márquez y M. López. *Educación y salud en una sociedad globalizada* (pp.213-213). Almería, España: Universidad de Almería. Recuperado de https://books.google.com.pe/books?id=VjwwBwAAQBAJ&lpg=PA213&dq=Seg%C3%BAn%20Smith%20(2010)%2C%20el%20ciberbullying%20es%20un%20acto%20agresivo%20e%20intencionado%20llevado%20a%20cabo%20de%20manera%20repetida%20y%20constante%2C%20mediante%20el%20uso%20de%20formas%20de%20contacto%20electr%C3%B3nicas%20por%20parte%20de%20un%20grupo%20o%20de%20un%20individuo%20contra%20una%20v%C3%ADctima%20que%20no%20puede%20defenderse%20f%C3%A1cilmente&hl=es&pg=PA213#v=onepage&q&f=false

Gran Diccionario de la Lengua Española. (2016). [Información de una página web]. Recuperado de http://es.thefreedictionary.com/clasificar

Grupo de RPP NOTICIAS. (2013, 05 de junio). Escolares de Chiclayo son víctimas del 'ciberbullying'. *RPP NOTICIAS* Recuperado de http://rpp.pe/peru/actualidad/escolares-de-chiclayo-son-victimas-del-ciberbullying-noticia-601611

Grupo de RPP NOTICIAS. (2015, 26 de julio). Escolares que sufren ciberbullying están más expuestos a las drogas. *RPP NOTICIAS*. Recuperado de

http://rpp.pe/lima/actualidad/escolares-que-sufren-ciberbullying-estan-mas-expuestos-a-las-drogas-noticia-820571

Grupo La República. (2013, 18 de septiembre). El 12.1% de escolares en el Perú sufre de ciberbullying. *La República*. Recuperado de http://larepublica.pe/18-09-2013/el-121-de-escolares-en-el-peru-sufre-de-ciberbullying

Han, J. y Micheline K. (2006). *Data Mining: Concepts and Techniques (2da ed.)*. Estados Unidos: Morgan Kaufmann. Recuperado de https://web.engr.illinois.edu/~hanj/bk2/toc.pdf

Hernández R., Fernández C., Baptista MP., Metodología de la Investigación. México: McGRAW-HILL / INTERAMERICANA EDITORES, S.A. DE C.V.; 2014, p. 128 y p. 140-145.

Hernández J., Ramírez M. J. y Ferri C. (2004). *Introducción a la minería de datos*. España: Pearson.

Infortelecom (2016, 09 de junio). *Jelastic PaaS.* [Mensaje en un blog]. Recuperado de https://infortelecom.es/blog/que-es-jelastic-paas/

Instituto de ingeniería del conocimiento (2017, 17 de octubre). Procesamiento del lenguaje natural ¿qué es? [Información de una página web]. Recuperado de http://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/

Kaplan N. (2011). Nuevos desarrollos en el estudio de la evaluación en el lenguaje: la teoría de la valoración. Lingüística, *1*(22), 52-78. Recuperado de http://www.redalyc.org/articulo.oa?id=34702203

Lage García, L. (2014). *Herramienta para el análisis de la opinión en tweets periodísticos*. (Tesis de maestría). Universidad Pompeu Fabra Barcelona, facultad ingeniería. Recuperado de https://repositori.upf.edu/bitstream/handle/10230/22939/LageGarcia_2014.pdf?sequence=1

Lara J. A. (2014). *Minería de datos*. España: CEF.

López Ferrero C. (2008). La valoración y la emoción en español en discursos especializados. Literatura, 1(1), 1-14. Recuperado de http://elvira.lllf.uam.es/clg8/actas/pdf/paperCLG65.pdf

López Ireta, L. T. & Gónzales Arias, C. (2013). Las expresiones agresivas en los comentarios de los usuarios de blogs periodísticos: un análisis por medio de la teoría de la valoración. Instituto de Literatura y Ciencias del Lenguaje, *1*(1), 1-18. Recuperado de http://www.um.es/tonosdigital/znum24/secciones/estudios-15-_expresiones_agresivas_en_blogs.htm

López Puga, J., García García. J, de la Fuente Sánchez. L y de la Fuente Solana. E. Las redes bayesianas como herramientas de modelado en psicología. *Anales de psicología, 23(2),* 307-316. Recuperado de http://www.redalyc.org/html/167/16723218/

Martínez E., Martín M. T., Perea J.M., Ureña L. A. (2011). Técnicas de clasificación de opiniones aplicadas a un corpus en español. *Procesamiento del Lenguaje Natural,* 47, 163-170. Recuperado de http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/958/711

Marquez C. (2016, de mayo). *Las Bases De Datos.* [Mensaje en un blog]. Recuperado de http://gorditosexy2.blogspot.pe/2016/05/las-bases-de-datos.html

Mendoza L. (2016, 8 de abril). *MySQL es un sistema de gestión de bases de datos relacional.* [Mensaje en un blog]. Recuperado de https://prezi.com/by-blbw5vlix/mysql-es-un-sistema-de-gestion-de-bases-de-datos-relacional/

Microsoft (2017). *Modelos de minería de datos (Analysis Services - Minería de datos).* Recuperado de https://msdn.microsoft.com/es-es/library/cc645779.aspx

Molina J.M. y García J. (2011). *Técnicas de Minería de Datos basadas en Aprendizaje Automático.* Recuperado de https://santiagozapatakdd.files.wordpress.com/2011/03/curso-kdd-full-cap-3.pdf

Morales Pizarro, A. (2016). *Diseño de un estimador de ángulo de balanceo de un vehículo industrial mediante redes neuronales.* (Tesis). Universidad Carlos III de Madrid, Ingeniería Mecánica. Recuperado de https://e-archivo.uc3m.es/bitstream/handle/10016/24416/TFG_Alberto_Morales_Pizarro_2016.pdf

Pérez C. y Santín D. (2006). *Data Mining Soluciones con Enterprise Miner.* México: ALFAOMEGA.

Pérez C. y Santín D. (2006). *Minería de Datos Técnicas y Herramientas.* Madrid, España: Paraninfo. ALFAOMEGA

Pérez J. y Merino M. (2008). *Definición de tiempo*. Recuperado de http://definicion.de/tiempo/

Pérez P. J. y Gardey A. (2009). Definición de léxico. [Información de una página web]. Recuperado de https://definicion.de/lexico/

Pérez Porto J. y Gardey A. (2010). Definición de comentario. [Mensaje en un blog]. Recuperado de http://definicion.de/comentario/

Perkins J. (2010). *Python Text Processing with NLTK 2.0 Cookbook.* Birmingham, UK.: Packt Publishing. Recuperado de http://caio.ueberalles.net/ebooksclub.org__Python_Text_Processing_with_NLTK_2_0_Cookbook.pdf

Rajić Jelena (2013). *LOS CONCEPTOS DE LA TEORÍA DE LA RELEVANCIA Y SU APLICACIÓN AL ESTUDIO DE LOS MARCADORES DISCURSIVOS.* (Tesis doctoral). Universidad de Belgrado Serbia, Serbia. Recuperado de https://colindancias.uvt.ro/index.php/colindancias/article/view/30/27

Rodríguez D. y Doblado J. (2010). *Redes Bayesianas en la Ingeniería del Software.* Recuperado de http://www.cc.uah.es/drg/b/RodriguezDolado.BBN.2007.pdf

Romero P. (2016, 26 de mayo). *REDES SOCIALES Mujer contra mujer... en Twitter.* [Mensaje en un blog]. Recuperado de http://www.elespanol.com/ciencia/20160526/127737256_0.html

Sabino, C. (1992). *El proceso de investigación* (pp.21). Bogotá, Colombia: Ed. Lumen. Recuperado de https://metodoinvestigacion.files.wordpress.com/2008/02/el-proceso-de-investigacion_carlos-sabino.pdf

Sancho F. (2016, 15 de mayo). Clasificación Supervisada y No Supervisada. [Mensaje en un blog]. Recuperado de http://www.cs.us.es/~fsancho/?e=77

Rojano Ramos, S. (2014). *Instrumentación y control en instalaciones de proceso, energía y servicios auxiliares* (pp. 130-131). Málaga, Colombia: Innovación y Cualificación S.L. Recuperado de https://books.google.com.pe/books?id=aqenBAAAQBAJ&pg=PT18&lpg=PT18&dq=%C2%BFqu%C3%A9+es+la+precision?+Es+la+cercan%C3%ADa+que+el+resultado+de+una+medici%C3%B3n+est%C3%A1+del+valor+verdadero&source=bl&ots=5-OOHXbuK9&sig=nidfAMdEwA7YE0DvOTGpgW4YQj4&hl=es&sa=X&ved=0ahUKEwjxm7_estfXAhXhQd8KHVyjDhIQ6AEIaTAG#v=onepage&q=%C2%BFqu%C3%A9%20es%20la%20precision%3F%20Es%20la%20cercan%C3%ADa%20que%20el%20resultado%20de%20una%20medici%C3%B3n%20est%C3%A1%20del%20valor%20verdadero&f=false

Soloriio M. (2013, 16 de abril). Metodología en cascada. [Información de una página web]. Recuperado de http://metodologiaencascada.blogspot.pe/

Tapia M. R., Ruiz O. N, Chirinos C. A. (2014). Modelo de Clasificación de Opiniones Subjetivas en Redes Sociales. *Ingeniería: Ciencia, Tecnología e Innovación,* 1(1), 1-15. Recuperado de https://www.google.com.pe/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwiVidGeyd3XAhUjY98KHYJNA90QFggIMAA&url=http%3A%2F%2Frevistas.uss.edu.pe%2Findex.php%2FING%2Farticle%2Fdownload%2F107%2F210&usg=AOvVaw1WwG81Ca7KZy5o0FGHIYO0

Universia. (2014, 22 de mayo). Cyberbullying: 1 de cada 4 casos de bullying es a través de las redes. *Uni>ersia España*. Recuperado de http://noticias.universia.es/cultura/noticia/2016/09/21/1143827/cyberbullying-1-cada-4-casos-bullying-traves-redes.html

Vilares, D., Alonso, M. A., y Gómez C. (2013). Una aproximación supervisada para la minería de opiniones sobre tuits en español en base a conocimiento lingüístico. *Procesamiento del Lenguaje Natural,* 51, 127-134. Recuperado de https://docs.google.com/viewerng/viewer?url=https://rua.ua.es/dspace/bitstream/10045/30627/1/PLN_51_14.pdf

Villamil Torres J. A. y Delgado Rivera J. A. (2007). Entrenamiento de una red neuronal multicapa para la tasa de cambio euro - dólar (EUR/USD), 3(27). Recuperado de http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-56092007000300012

WordPress (2017, 31 de octubre). *Error absoluto y error absoluto medio.* [Mensaje en un blog]. Recuperado de http://www.statisticshowto.com/absolute-error/

Wikidot (2010, 12 de diciembre). *Redes Bayesianas.* [Mensaje en un blog]. Recuperado de http://redbay.wikidot.com/

ANEXOS

ANEXO 1: Glosario de términos de la investigación

Término	Definición
ARFF	ARFF (Attribute-Relation File Format), es el formato de la herramienta de minería de datos weka.
BayesNet	Son modelos que pueden tener diversas aplicaciones para la clasificación, predicción, diagnostico, etc. Además, pueden dar información en cuanto a cómo se relacionan las variables del dominio, las cuales pueden ser interpretadas como causa-efecto (wikidot, 2010).
Ciberbullying	El ciberbullying implica la intimidación, el acoso o la amenaza mediante medios electrónicos (principalmente, vía Internet o teléfono celular), con la intención de dañar (ciberagresor) a otro (cibervíctima) (García, Joffre, Martínez y Llanes, 2011).
Corpus de conocimiento	Son datos utilizados en el entrenamiento de los algoritmos de minería de datos. Estos datos ya llevan etiquetas, es decir son ejemplos ya clasificados (Álvarez, *et al,* 2007, p.112).
Crossvalidation	Es una prueba estadística que nos ofrece la herramienta weka, este tipo de prueba divide la muestra original (corpus de conocimiento) en submuestras. De las cuales, una de ellas se retiene como datos de validación para probar el modelo y el resto como datos de formación del modelo.

DecisionStump	Decision Stump es un algoritmo sencillo que genera un árbol de decisión de un único nivel, a pesar de su simplicidad, en algunos problemas llega a conseguir resultados interesantes; admite atributos numéricos como simbólicos y clases de ambos tipos también García *et al*. (2012).
IBK	Es un algoritmo basado en instancias, por ello consiste únicamente en almacenar los datos presentados. Cuando una nueva instancia es encontrada, un conjunto de instancias similares relacionadas es devuelto desde la memoria y usado para clasificar la instancia consultada (Boris y Cebrián, 2012).
J48	Es un algoritmo J48 que forma parte de los algoritmos basados en árboles de decisión; la característica fundamental de este algoritmo es que incorpora una poda del árbol en la clasificación, esto es útil para cortar aquellas ramas del árbol con menor capacidad predictiva (Boris y Cebrián, 2012).
MAE	Es el Error Absoluto Medio, que nos da la diferencia entre el valor de la medida y el valor tomado como exacto (valor verdadero o valor probable), este error puede ser positivo o negativo, y dependerá de la medida del valor real (Arriagada y Peralta, 2008).
Naive Bayes	Es un método de clasificación probabilístico, que es utilizado para la clasificación de nuevas instancias de un documento dentro de un conjunto finito de clases predeterminadas (Bordignon, 2007).

OneR	Es uno de los algoritmos más simples de reglas de clasificación, genera un árbol de decisión de un nivel expresado mediante reglas (García et al., 2012).
PaaS Jelastic	Jelactic Paas (Plataform As A Service), permite crear de forma sencilla entornos de trabajo para desplegar todo tipo de proyectos web (Infortelecom, 2016).
PART	PART (Generación de reglas por árboles de decisión parcial), es un sistema que se basa en las dos estrategias básicas para la inducción de reglas: el covering y la generación de reglas a partir de árboles de decisión; toma la estrategia del covering dado que genera una regla de decisión, elimina los ejemplares que dicha regla cubre y continúa generando reglas hasta que no queden ejemplos por clasificar (García et al., 2012).
SMO	SMO (sequential minimal optimization), es un algoritmo utilizado para entrenar máquinas de vectores de soporte usando un kernel polinomial, es un método supervisado que ha demostrado buenas propiedades para la categorización de documentos (Pérez y Cordoso, 2010).
SVM	Las Máquinas de Vectores Soporte, constituyen un método basado en aprendizaje para la resolución de problemas de clasificación. Esta resolución se basa en una primera fase de entrenamiento (donde se les informa con múltiples ejemplos ya resueltos, en forma de pares {problema, solución}) y una segunda fase de uso para la resolución de problemas. En ella, las SVM se convierten en una "caja negra" que proporciona una respuesta (salida) a un problema dado (entrada) (Divisadero, 2016).

ANEXO 2: Resultados de los modelos de clasificación del algoritmo NaiveBayes

Modelo generado con 800 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         608               76      %
Incorrectly Classified Instances       192               24      %
Kappa statistic                          0.5275
Mean absolute error                      0.2108
Root mean squared error                  0.3192
Relative absolute error                 75.5238 %
Root relative squared error             85.5237 %
Total Number of Instances              800

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.925    NEUTRO
                 0.444     0          1         0.444     0.615      0.899    POCOAGRESIVO
                 0.68      0.083      0.826     0.68      0.746      0.916    AGRESIVO
                 0.918     0.417      0.729     0.918     0.813      0.911    MUYAGRESIVO
Weighted Avg.    0.76      0.26       0.716     0.76      0.728      0.913

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0   6  51 |   a = NEUTRO
   0   4   0   5 |   b = POCOAGRESIVO
   0   0 200  94 |   c = AGRESIVO
   0   0  36 404 |   d = MUYAGRESIVO
```

Modelo generado con 1200 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         901               75.0833 %
Incorrectly Classified Instances       299               24.9167 %
Kappa statistic                          0.555
Mean absolute error                      0.2179
Root mean squared error                  0.3236
Relative absolute error                 76.8086 %
Root relative squared error             85.9533 %
Total Number of Instances             1200

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.944    NEUTRO
                 0.462     0.002      0.75      0.462     0.571      0.831    POCOAGRESIVO
                 0.994     0.415      0.621     0.994     0.764      0.919    AGRESIVO
                 0.667     0.002      0.998     0.667     0.8        0.921    MUYAGRESIVO
Weighted Avg.    0.751     0.169      0.772     0.751     0.727      0.921

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0  84   0 |   a = NEUTRO
   0   6   7   0 |   b = POCOAGRESIVO
   0   2 484   1 |   c = AGRESIVO
   0   0 205 411 |   d = MUYAGRESIVO
```

Modelo generado con 1561 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1080               69.1864 %
Incorrectly Classified Instances       481               30.8136 %
Kappa statistic                          0.4518
Mean absolute error                      0.2179
Root mean squared error                  0.323
Relative absolute error                 76.6405 %
Root relative squared error             85.7168 %
Total Number of Instances             1561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.957    NEUTRO
                 0.471     0.003      0.615     0.471     0.533      0.909    POCOAGRESIVO
                 0.992     0.525      0.577     0.992     0.73       0.919    AGRESIVO
                 0.54      0          1         0.54      0.701      0.922    MUYAGRESIVO
Weighted Avg.    0.692     0.22       0.75      0.692     0.663      0.923

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0 107   0 |   a = NEUTRO
   0   8   9   0 |   b = POCOAGRESIVO
   0   5 649   0 |   c = AGRESIVO
   0   0 360 423 |   d = MUYAGRESIVO
```

ANEXO 3: Resultados de los modelos de clasificación del algoritmo BayesNet

Modelo generado con 800 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         432               54      %
Incorrectly Classified Instances       368               46      %
Kappa statistic                          0.1321
Mean absolute error                      0.2424
Root mean squared error                  0.3464
Relative absolute error                 86.8353 %
Root relative squared error             92.8075 %
Total Number of Instances              800

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.577    NEUTRO
                 0.444     0.001      0.8       0.444     0.571      0.61     POCOAGRESIVO
                 0.497     0.375      0.435     0.497     0.463      0.628    AGRESIVO
                 0.641     0.492      0.614     0.641     0.627      0.651    MUYAGRESIVO
Weighted Avg.    0.54      0.408      0.507     0.54      0.522      0.637

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0  29  28 |   a = NEUTRO
   0   4   3   2 |   b = POCOAGRESIVO
   0   1 146 147 |   c = AGRESIVO
```

Modelo generado con 1200 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         738               61.5    %
Incorrectly Classified Instances       462               38.5    %
Kappa statistic                          0.3303
Mean absolute error                      0.2321
Root mean squared error                  0.3384
Relative absolute error                 81.7983 %
Root relative squared error             89.8871 %
Total Number of Instances             1200

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.601    NEUTRO
                 0.462     0.003      0.667     0.462     0.545      0.621    POCOAGRESIVO
                 0.994     0.644      0.513     0.994     0.677      0.666    AGRESIVO
                 0.403     0          1         0.403     0.574      0.695    MUYAGRESIVO
Weighted Avg.    0.615     0.261      0.729     0.615     0.575      0.676

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0  84   0 |   a = NEUTRO
   0   6   7   0 |   b = POCOAGRESIVO
   0   3 484   0 |   c = AGRESIVO
```

Modelo generado con 1561 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1004               64.3177 %
Incorrectly Classified Instances       557               35.6823 %
Kappa statistic                          0.3673
Mean absolute error                      0.2234
Root mean squared error                  0.3319
Relative absolute error                 78.5899 %
Root relative squared error             88.0782 %
Total Number of Instances             1561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.62     NEUTRO
                 0.118     0.003      0.333     0.118     0.174      0.689    POCOAGRESIVO
                 0.994     0.61       0.54      0.994     0.7        0.69     AGRESIVO
                 0.45      0          1         0.45      0.62       0.724    MUYAGRESIVO
Weighted Avg.    0.643     0.255      0.732     0.643     0.606      0.702

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0 107   0 |   a = NEUTRO
   0   2  15   0 |   b = POCOAGRESIVO
   0   4 650   0 |   c = AGRESIVO
```

ANEXO 4: Resultados de los modelos de clasificación del algoritmo DecisionStump

Modelo generado con 800 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         440               55      %
Incorrectly Classified Instances       360               45      %
Kappa statistic                          0
Mean absolute error                      0.2612
Root mean squared error                  0.3614
Relative absolute error                 93.565  %
Root relative squared error             96.826  %
Total Number of Instances              800

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.526    NEUTRO
                 0         0          0         0         0          0.491    POCOAGRESIVO
                 0         0          0         0         0          0.558    AGRESIVO
                 1         1          0.55      1         0.71       0.567    MUYAGRESIVO
Weighted Avg.    0.55      0.55       0.303     0.55      0.39       0.56

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0   0  57 |   a = NEUTRO
   0   0   0   9 |   b = POCOAGRESIVO
   0   0   0 294 |   c = AGRESIVO
   0   0   0 440 |   d = MUYAGRESIVO
```

Modelo generado con 1200 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         616               51.3333 %
Incorrectly Classified Instances       584               48.6667 %
Kappa statistic                          0
Mean absolute error                      0.2655
Root mean squared error                  0.3644
Relative absolute error                 93.587  %
Root relative squared error             96.8014 %
Total Number of Instances             1200

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.526    NEUTRO
                 0         0          0         0         0          0.464    POCOAGRESIVO
                 0         0          0         0         0          0.555    AGRESIVO
                 1         1          0.513     1         0.678      0.566    MUYAGRESIVO
Weighted Avg.    0.513     0.513      0.264     0.513     0.348      0.558

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0   0  84 |   a = NEUTRO
   0   0   0  13 |   b = POCOAGRESIVO
   0   0   0 487 |   c = AGRESIVO
   0   0   0 616 |   d = MUYAGRESIVO
```

Modelo generado con 1561 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         781               50.032  %
Incorrectly Classified Instances       780               49.968  %
Kappa statistic                          0.1183
Mean absolute error                      0.2634
Root mean squared error                  0.3629
Relative absolute error                 92.6616 %
Root relative squared error             96.3083 %
Total Number of Instances             1561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.536    NEUTRO
                 0         0          0         0         0          0.481    POCOAGRESIVO
                 0.899     0.774      0.456     0.899     0.605      0.564    AGRESIVO
                 0.246     0.1        0.712     0.246     0.366      0.574    MUYAGRESIVO
Weighted Avg.    0.5       0.375      0.548     0.5       0.437      0.566

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0  97  10 |   a = NEUTRO
   0   0  15   2 |   b = POCOAGRESIVO
   0   0 588  66 |   c = AGRESIVO
   0   0 590 193 |   d = MUYAGRESIVO
```

ANEXO 5: Resultados del modelo de clasificación del algoritmo J48

Modelo generado con 800 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         464               58      %
Incorrectly Classified Instances       336               42      %
Kappa statistic                          0.0906
Mean absolute error                      0.2502
Root mean squared error                  0.3543
Relative absolute error                 89.6253 %
Root relative squared error             94.921  %
Total Number of Instances              800

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.596    NEUTRO
                 0         0          0         0         0          0.548    POCOAGRESIVO
                 0.129     0.028      0.731     0.129     0.22       0.631    AGRESIVO
                 0.968     0.894      0.57      0.968     0.717      0.634    MUYAGRESIVO
Weighted Avg.    0.58      0.502      0.582     0.58      0.475      0.629

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0   0  57 |   a = NEUTRO
   0   0   0   9 |   b = POCOAGRESIVO
   0   0  38 256 |   c = AGRESIVO
   0   0  14 426 |   d = MUYAGRESIVO
```

Modelo generado con 1200 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         670               55.8333 %
Incorrectly Classified Instances       530               44.1667 %
Kappa statistic                          0.2356
Mean absolute error                      0.2468
Root mean squared error                  0.3513
Relative absolute error                 87      %
Root relative squared error             93.3359 %
Total Number of Instances             1200

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.567    NEUTRO
                 0         0          0         0         0          0.503    POCOAGRESIVO
                 1         0.743      0.479     1         0.648      0.614    AGRESIVO
                 0.297     0          1         0.297     0.458      0.632    MUYAGRESIVO
Weighted Avg.    0.558     0.302      0.708     0.558     0.498      0.619

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0  84   0 |   a = NEUTRO
   0   0  13   0 |   b = POCOAGRESIVO
   0   0 487   0 |   c = AGRESIVO
   0   0 433 183 |   d = MUYAGRESIVO
```

Modelo generado con 1561 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         906               58.0397 %
Incorrectly Classified Instances       655               41.9603 %
Kappa statistic                          0.2609
Mean absolute error                      0.2409
Root mean squared error                  0.3472
Relative absolute error                 84.7535 %
Root relative squared error             92.1456 %
Total Number of Instances             1561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.6      NEUTRO
                 0         0          0         0         0          0.541    POCOAGRESIVO
                 1         0.722      0.5       1         0.666      0.635    AGRESIVO
                 0.322     0          1         0.322     0.487      0.649    MUYAGRESIVO
Weighted Avg.    0.58      0.303      0.711     0.58      0.523      0.639

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0 107   0 |   a = NEUTRO
   0   0  17   0 |   b = POCOAGRESIVO
   0   0 654   0 |   c = AGRESIVO
   0   0 531 252 |   d = MUYAGRESIVO
```

ANEXO 6: Resultados del modelo de clasificación del algoritmo OneR

Modelo generado con 800 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         453               56.625  %
Incorrectly Classified Instances       347               43.375  %
Kappa statistic                          0.0453
Mean absolute error                      0.2169
Root mean squared error                  0.4657
Relative absolute error                 77.6891 %
Root relative squared error            124.7681 %
Total Number of Instances              800

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.5      NEUTRO
                 0         0          0         0         0          0.5      POCOAGRESIVO
                 0.054     0.006      0.842     0.054     0.102      0.524    AGRESIVO
                 0.993     0.956      0.56      0.993     0.716      0.519    MUYAGRESIVO
Weighted Avg.    0.566     0.528      0.617     0.566     0.431      0.519

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0   0  57 |   a = NEUTRO
   0   0   0   9 |   b = POCOAGRESIVO
   0   0  16 278 |   c = AGRESIVO
   0   0   3 437 |   d = MUYAGRESIVO
```

Modelo generado con 1200 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         627               52.25   %
Incorrectly Classified Instances       573               47.75   %
Kappa statistic                          0.0262
Mean absolute error                      0.2388
Root mean squared error                  0.4886
Relative absolute error                 84.1478 %
Root relative squared error            129.805  %
Total Number of Instances             1200

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.5      NEUTRO
                 0         0          0         0         0          0.5      POCOAGRESIVO
                 0.053     0.021      0.634     0.053     0.098      0.516    AGRESIVO
                 0.976     0.955      0.519     0.976     0.677      0.51     MUYAGRESIVO
Weighted Avg.    0.523     0.499      0.524     0.523     0.388      0.512

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0   0  84 |   a = NEUTRO
   0   0   0  13 |   b = POCOAGRESIVO
   0   0  26 461 |   c = AGRESIVO
   0   0  15 601 |   d = MUYAGRESIVO
```

Modelo generado con 1561 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         794               50.8648 %
Incorrectly Classified Instances       767               49.1352 %
Kappa statistic                          0.0163
Mean absolute error                      0.2457
Root mean squared error                  0.4957
Relative absolute error                 86.4189 %
Root relative squared error            131.5268 %
Total Number of Instances             1561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.5      NEUTRO
                 0         0          0         0         0          0.5      POCOAGRESIVO
                 0.024     0.006      0.762     0.024     0.047      0.509    AGRESIVO
                 0.994     0.979      0.505     0.994     0.67       0.507    MUYAGRESIVO
Weighted Avg.    0.509     0.494      0.573     0.509     0.356      0.508

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0   0 107 |   a = NEUTRO
   0   0   0  17 |   b = POCOAGRESIVO
   0   0  16 638 |   c = AGRESIVO
   0   0   5 778 |   d = MUYAGRESIVO
```

ANEXO 7: Resultados del modelo de clasificación del algoritmo PART

Modelo generado con 800 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         463               57.875  %
Incorrectly Classified Instances       337               42.125  %
Kappa statistic                          0.0884
Mean absolute error                      0.2407
Root mean squared error                  0.3475
Relative absolute error                 86.2311 %
Root relative squared error             93.11   %
Total Number of Instances              800

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.613    NEUTRO
                 0         0          0         0         0          0.566    POCOAGRESIVO
                 0.129     0.03       0.717     0.129     0.219      0.651    AGRESIVO
                 0.966     0.894      0.569     0.966     0.716      0.657    MUYAGRESIVO
Weighted Avg.    0.579     0.503      0.576     0.579     0.474      0.651

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0   0  57 |   a = NEUTRO
   0   0   0   9 |   b = POCOAGRESIVO
   0   0  38 256 |   c = AGRESIVO
   0   0  15 425 |   d = MUYAGRESIVO
```

Modelo generado con 1200 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         670               55.8333 %
Incorrectly Classified Instances       530               44.1667 %
Kappa statistic                          0.2356
Mean absolute error                      0.2468
Root mean squared error                  0.3513
Relative absolute error                 87      %
Root relative squared error             93.3359 %
Total Number of Instances             1200

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.567    NEUTRO
                 0         0          0         0         0          0.503    POCOAGRESIVO
                 1         0.743      0.479     1         0.648      0.614    AGRESIVO
                 0.297     0          1         0.297     0.458      0.632    MUYAGRESIVO
Weighted Avg.    0.558     0.302      0.708     0.558     0.498      0.619

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0  84   0 |   a = NEUTRO
   0   0  13   0 |   b = POCOAGRESIVO
   0   0 487   0 |   c = AGRESIVO
   0   0 433 183 |   d = MUYAGRESIVO
```

Modelo generado con 1561 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         902               57.7835 %
Incorrectly Classified Instances       659               42.2165 %
Kappa statistic                          0.2566
Mean absolute error                      0.2415
Root mean squared error                  0.3478
Relative absolute error                 84.9558 %
Root relative squared error             92.2812 %
Total Number of Instances             1561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.589    NEUTRO
                 0         0          0         0         0          0.536    POCOAGRESIVO
                 1         0.727      0.498     1         0.665      0.633    AGRESIVO
                 0.317     0          1         0.317     0.481      0.648    MUYAGRESIVO
Weighted Avg.    0.578     0.304      0.71      0.578     0.52       0.636

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0 107   0 |   a = NEUTRO
   0   0  17   0 |   b = POCOAGRESIVO
   0   0 654   0 |   c = AGRESIVO
   0   0 535 248 |   d = MUYAGRESIVO
```

ANEXO 8: Resultados del modelo de clasificación del algoritmo SMO

Modelo generado con 800 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         652               81.5    %
Incorrectly Classified Instances       148               18.5    %
Kappa statistic                          0.6633
Mean absolute error                      0.2725
Root mean squared error                  0.346
Relative absolute error                 97.6152 %
Root relative squared error             92.6946 %
Total Number of Instances              800

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.503    NEUTRO
                 0.444     0.001      0.8       0.444     0.571      0.721    POCOAGRESIVO
                 0.966     0.271      0.675     0.966     0.794      0.848    AGRESIVO
                 0.827     0.028      0.973     0.827     0.894      0.9      MUYAGRESIVO
Weighted Avg.    0.815     0.115      0.792     0.815     0.79       0.85

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0  57   0 |   a = NEUTRO
   0   4   4   1 |   b = POCOAGRESIVO
   0   1 284   9 |   c = AGRESIVO
   0   0  76 364 |   d = MUYAGRESIVO
```

Modelo generado con 1200 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1000               83.3333 %
Incorrectly Classified Instances       200               16.6667 %
Kappa statistic                          0.697
Mean absolute error                      0.2708
Root mean squared error                  0.3434
Relative absolute error                 95.4312 %
Root relative squared error             91.237  %
Total Number of Instances             1200

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0          0          0.504    NEUTRO
                 0.462     0.003      0.667     0.462      0.545      0.729    POCOAGRESIVO
                 0.979     0.265      0.716     0.979      0.827      0.858    AGRESIVO
                 0.839     0.014      0.985     0.839      0.906      0.913    MUYAGRESIVO
Weighted Avg.    0.833     0.115      0.803     0.833      0.807      0.86

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0  84   0 |   a = NEUTRO
   0   6   6   1 |   b = POCOAGRESIVO
   0   3 477   7 |   c = AGRESIVO
   0   0  99 517 |   d = MUYAGRESIVO
```

Modelo generado con 1561 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1341               85.9065 %
Incorrectly Classified Instances       220               14.0935 %
Kappa statistic                          0.7432
Mean absolute error                      0.2682
Root mean squared error                  0.3396
Relative absolute error                 94.3434 %
Root relative squared error             90.1263 %
Total Number of Instances             1561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.505    NEUTRO
                 0.706     0.003      0.706     0.706     0.706      0.851    POCOAGRESIVO
                 0.982     0.229      0.755     0.982     0.854      0.876    AGRESIVO
                 0.877     0.009      0.99      0.877     0.93       0.936    MUYAGRESIVO
Weighted Avg.    0.859     0.101      0.821     0.859     0.832      0.88

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
   0   0 107   0 |   a = NEUTRO
   0  12   5   0 |   b = POCOAGRESIVO
   0   5 642   7 |   c = AGRESIVO
   0   0  96 687 |   d = MUYAGRESIVO
```

ANEXO 9: Resultados del modelo de clasificación del algoritmo IBK

Modelo generado con 800 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         582               72.75    %
Incorrectly Classified Instances       218               27.25    %
Kappa statistic                          0.5798
Mean absolute error                      0.1347
Root mean squared error                  0.3568
Relative absolute error                 48.261  %
Root relative squared error             95.5796 %
Total Number of Instances              800

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.244      0.239     1         0.386      0.918    NEUTRO
                 0.444     0          1         0.444     0.615      0.539    POCOAGRESIVO
                 0.67      0.073      0.842     0.67      0.746      0.796    AGRESIVO
                 0.736     0          1         0.736     0.848      0.888    MUYAGRESIVO
Weighted Avg.    0.728     0.044      0.888     0.728     0.775      0.852

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
  57   0   0   0 |   a = NEUTRO
   5   4   0   0 |   b = POCOAGRESIVO
  97   0 197   0 |   c = AGRESIVO
  79   0  37 324 |   d = MUYAGRESIVO
```

Modelo generado con 1200 comentarios de textos:

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         944               78.6667 %
Incorrectly Classified Instances       256               21.3333 %
Kappa statistic                          0.6621
Mean absolute error                      0.1067
Root mean squared error                  0.3156
Relative absolute error                 37.6203 %
Root relative squared error             83.8438 %
Total Number of Instances             1200

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.168      0.309     1         0.472      0.942    NEUTRO
                 0.462     0.002      0.75      0.462     0.571      0.519    POCOAGRESIVO
                 0.801     0.093      0.855     0.801     0.827      0.85     AGRESIVO
                 0.753     0          1         0.753     0.859      0.901    MUYAGRESIVO
Weighted Avg.    0.787     0.049      0.89      0.787     0.816      0.879

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
  84   0   0   0 |   a = NEUTRO
   7   6   0   0 |   b = POCOAGRESIVO
  95   2 390   0 |   c = AGRESIVO
  86   0  66 464 |   d = MUYAGRESIVO
```

Modelo generado con 1561 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1273               81.5503 %
Incorrectly Classified Instances       288               18.4497 %
Kappa statistic                          0.7026
Mean absolute error                      0.0903
Root mean squared error                  0.2872
Relative absolute error                 31.756  %
Root relative squared error             76.222  %
Total Number of Instances             1561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.133      0.357     1         0.526      0.956    NEUTRO
                 0.706     0.003      0.75      0.706     0.727      0.728    POCOAGRESIVO
                 0.835     0.1        0.857     0.835     0.846      0.883    AGRESIVO
                 0.777     0          1         0.777     0.874      0.922    MUYAGRESIVO
Weighted Avg.    0.816     0.051      0.893     0.816     0.837      0.906

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 107   0   0   0 |   a = NEUTRO
   5  12   0   0 |   b = POCOAGRESIVO
 104   4 546   0 |   c = AGRESIVO
  84   0  91 608 |   d = MUYAGRESIVO
```

ANEXO 10: Resultados del modelo de clasificación del algoritmo KStart

Modelo generado con 800 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         584               73      %
Incorrectly Classified Instances       216               27      %
Kappa statistic                          0.5836
Mean absolute error                      0.2345
Root mean squared error                  0.3202
Relative absolute error                 84.0081 %
Root relative squared error             85.797  %
Total Number of Instances              800

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.248      0.237     1         0.383      0.922    NEUTRO
                 0.444     0          1         0.444     0.615      0.933    POCOAGRESIVO
                 0.66      0.063      0.858     0.66      0.746      0.917    AGRESIVO
                 0.748     0          1         0.748     0.856      0.947    MUYAGRESIVO
Weighted Avg.    0.73      0.041      0.894     0.73      0.779      0.934

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
  57   0   0   0 |   a = NEUTRO
   5   4   0   0 |   b = POCOAGRESIVO
 100   0 194   0 |   c = AGRESIVO
  79   0  32 329 |   d = MUYAGRESIVO
```

Modelo generado con 1200 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          84               7      %
Incorrectly Classified Instances      1116              93      %
Kappa statistic                          0
Mean absolute error                      0.375
Root mean squared error                  0.433
Relative absolute error                132.1693 %
Root relative squared error            115.0324 %
Total Number of Instances             1200

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1          0.07      1         0.131      0.5      NEUTRO
                 0         0          0         0         0          0.5      POCOAGRESIVO
                 0         0          0         0         0          0.5      AGRESIVO
                 0         0          0         0         0          0.5      MUYAGRESIVO
Weighted Avg.    0.07      0.07       0.005     0.07      0.009      0.5

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
  84   0   0   0 |   a = NEUTRO
  13   0   0   0 |   b = POCOAGRESIVO
 487   0   0   0 |   c = AGRESIVO
 616   0   0   0 |   d = MUYAGRESIVO
```

Modelo generado con 1561 comentarios de textos:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         107               6.8546 %
Incorrectly Classified Instances      1454              93.1454 %
Kappa statistic                          0
Mean absolute error                      0.375
Root mean squared error                  0.433
Relative absolute error                131.9099 %
Root relative squared error            114.9036 %
Total Number of Instances             1561

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1          0.069     1         0.128      0.5      NEUTRO
                 0         0          0         0         0          0.5      POCOAGRESIVO
                 0         0          0         0         0          0.5      AGRESIVO
                 0         0          0         0         0          0.5      MUYAGRESIVO
Weighted Avg.    0.069     0.069      0.005     0.069     0.009      0.5

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 107   0   0   0 |   a = NEUTRO
  17   0   0   0 |   b = POCOAGRESIVO
 654   0   0   0 |   c = AGRESIVO
 783   0   0   0 |   d = MUYAGRESIVO
```

ANEXO 11: Resultados de los errores medios absolutos de los algoritmos

Error medio absoluto de 800 comentarios de textos:

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Mean_absolute_error
Datasets:   1
Resultsets: 9
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       10/09/17 04:11 PM
```

Dataset	(1) bayes.NaiveBa	(2) bayes.Baye	(3) trees.Deci	(4) trees.J48	(5) rules.OneR	(6) rules.PART	(7) functions.	(8) lazy.IBk '	(9) lazy.KStar
comentarios (100)	0.21(0.01)	0.24(0.01) v	0.26(0.01) v	0.25(0.01) v	0.22(0.01)	0.24(0.01) v	0.27(0.00) v	0.14(0.02) *	0.24(0.01) v
(v/ /*)		(1/0/0)	(1/0/0)	(1/0/0)	(0/1/0)	(1/0/0)	(1/0/0)	(0/0/1)	(1/0/0)

Error medio absoluto de 1200 comentarios de textos:

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -V -result-matrix "weka.experiment.ResultMatrixPlainText -mean-prec 2 -stddev-prec 2 -col-
Analysing:  Mean_absolute_error
Datasets:   1
Resultsets: 9
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       10/09/17 04:28 PM

Dataset                  (1) bayes.NaiveBa | (2) bayes.Baye (3) trees.Deci (4) trees.J48  (5) rules.OneR (6) rules.PART (7) functions. (8) lazy.IBk ' (9) lazy.KStar
-----------------------------------------------------------------------------------------------------------------------------------------------------------------
comentarios              (100)   0.22(0.01) |   0.23(0.01) v   0.27(0.01) v   0.25(0.01) v   0.24(0.01) v   0.25(0.01) v   0.27(0.00) v   0.11(0.02) *   0.38(0.00) v
-----------------------------------------------------------------------------------------------------------------------------------------------------------------
                                   (v/ /*) |        (1/0/0)        (1/0/0)        (1/0/0)        (1/0/0)        (1/0/0)        (1/0/0)        (0/0/1)        (1/0/0)
```

Error medio absoluto de 1561 comentarios de textos:

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Mean_absolute_error
Datasets:   1
Resultsets: 9
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       10/09/17 07:38 PM

Dataset                  (1) bayes.NaiveBa | (2) bayes.Baye (3) trees.Deci (4) trees.J48  (5) rules.OneR (6) rules.PART (7) functions. (8) lazy.IBk ' (9) lazy.KStar
-----------------------------------------------------------------------------------------------------------------------------------------------------------------
comentarios              (100)   0.22(0.01) |   0.22(0.01) v   0.26(0.01) v   0.24(0.01) v   0.25(0.00) v   0.24(0.01) v   0.27(0.00) v   0.09(0.01) *   0.38(0.00) v
-----------------------------------------------------------------------------------------------------------------------------------------------------------------
                                   (v/ /*) |        (1/0/0)        (1/0/0)        (1/0/0)        (1/0/0)        (1/0/0)        (1/0/0)        (0/0/1)        (1/0/0)
```

ANEXO 12: MATRIZ DE CONSISTENCIA

TÍTULO: INFLUENCIA DEL APRENDIZAJE COMPUTACIONAL BASADO EN TÉCNICAS DE MINERÍA DE TEXTOS EN LA CLASIFICACIÓN DE COMENTARIOS DE TEXTOS AGRESIVOS**.**

PROBLEMA	OBJETIVO	HIPÓTESIS	VARIABLES	INDICARES	ESCALAS
Problema General ¿De qué manera influye el corpus de conocimiento en el proceso de la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico?	**Objetivo General** Determinar la influencia del corpus de conocimiento en el proceso de la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico.	**Hipótesis General** El corpus de conocimiento influirá de manera directamente proporcional en el proceso de la clasificación de comentarios de textos agresivos, basadas en técnicas de minería de textos en un análisis léxico.	**Variable Independiente:** Aprendizaje computacional. **Variable Dependiente:** Clasificación de comentarios de textos agresivos.	-Palabras y cometarios de texto. -Categorías o clases	-Unidades -Muy agresivo, agresivo, poco agresivo y neutro
Problema Específico 1 ¿De qué manera influye el corpus de conocimiento en la elección del mejor algoritmo de clasificación del grupo de algoritmos clasificadores de tal manera que tenga el menor **Error Absoluto Medio** en la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico?	**Objetivo Específico 1** Determinar la influencia del corpus de conocimiento en la elección del algoritmo de clasificación del grupo de algoritmos clasificadores de tal manera que tenga el menor **Error Absoluto Medio** en la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico.	**Hipótesis Específica 1** El corpus de conocimiento influirá directamente sobre la elección del algoritmo de clasificación del grupo de algoritmos clasificadores basadas en las técnicas de minería de textos en un análisis léxico, siendo el elegido IBK por tener el menor Error Absoluto Medio.	Elección del algoritmo.	Algoritmo con menor Error Absoluto Medio	Numérica
Problema Específico 2 ¿De qué manera influye el corpus de conocimiento en la precisión de la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico?	**Objetivo Específico 2** Determinar la influencia del corpus de conocimiento en la precisión de la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico.	**Hipótesis Específica 2** El corpus de conocimiento influirá de manera directa en la precisión de la clasificación de comentarios de textos agresivos basadas en las técnicas de minería de textos en un análisis léxico, siendo esta mayor al 85%.	Precisión de la clasificación de comentarios de textos agresivos en categorías.	Precisión	%
Problema Específico 3 ¿De qué manera influye el corpus de conocimiento en el tiempo de generación de los Modelos de Clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico?	**Objetivo Específico 3** Determinar la influencia del corpus de conocimiento en el tiempo de generación de los Modelos de Clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico.	**Hipótesis Específica 3** El corpus de conocimiento influirá proporcionalmente en el tiempo de generación del Modelo de Clasificación de comentarios de textos agresivos basadas en las técnicas de minería de textos en un análisis léxico, siendo este no mayor a 20 minutos.	Tiempo de generación del Modelo de Clasificación de comentarios de textos agresivos.	Tiempo	Segundos
Problema Específico 4 ¿De qué manera influye el corpus de conocimiento en la cantidad de datos óptimos para la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico?	**Objetivo Específico 4** Determinar la influencia del corpus de conocimiento en la cantidad de datos óptimos para la clasificación de comentarios de textos agresivos, basadas en las técnicas de minería de textos en un análisis léxico.	**Hipótesis Específica 4** El corpus de conocimiento influirá proporcionalmente sobre la cantidad de datos óptimos para la clasificación de comentarios de textos agresivos basadas en las técnicas de minería de textos en un análisis léxico, siendo la cantidad óptima 1561 registros de comentarios y 906 atributos (palabras).	Cantidad de datos óptimos para la clasificación de comentarios de textos agresivos**.**	Cantidad de palabras y mensajes	Unidades