BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

CLASIFICACIÓN AUTOMÁTICA DE DOCUMENTOS PROPAGANDÍSTICOS

TESIS PARA OBTENER EL TÍTULO DE:

LICENCIADA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

BEATRIZ AMPARO MERINO REYES

ASESORES:

M.C. YALÚ GALICIA HERNÁNDEZ

DR. DAVID EDUARDO PINTO AVENDAÑO

H. PUEBLA DE ZARAGOZA, FEBRERO 2016

A Dios por darme la fuerza.

A mis padres,

por ser mi motivación;

por su paciencia y apoyo incondicional.

A mis hermanos, a mi familia

por su cariño y ayuda.

A mi asesores de tesis Mtra. Yalú Galicia y Dr. David Pinto,

por su invaluable tiempo, su paciencia y consejos.

Índice general

Capítulo 1

Introducción

Es impresionante como ha crecido la era digital, actualmente ya estamos con la web semántica o web 3.0 donde trabajar con base de datos y recuperación de información es un tema con el que vivimos en todo momento. Diariamente se generan miles de textos digitales como libros, artículos, notas periodísticas e informativas, que en su mayoría es almacenada en los servidores web para que esté disponible a través del internet. Para tener acceso a esta información se requiere de herramientas de búsqueda y recuperación de información, donde se han tenido grandes avances y se han creado diferentes líneas de investigación como extracción de información, búsqueda de respuestas, clasificación de opiniones, clasificación automático de textos, entre otros.

El organizar miles y miles de papeles de una forma manual se ha vuelto obsoleto ya que es muy costoso, complicado y requiere de mucho tiempo. Actualmente, hay una expansión de datos digitales en todo tipo de disciplina humana pero se requieren de métodos algorítmicos cuando trabajamos con enormes colecciones de documentos. Esta información no necesariamente tiene una estructura definida, es decir, no hay un sistema o herramienta única que administre los datos por lo que es necesario contar con trabajos de investigación que puedan ser aplicados en estas tareas.

Para administrar y poder clasificar texto se ha trabajado intensamente en el área de recuperación de información [1], minería de texto [2], extracción de información [3], búsqueda de respuestas [4], clasificación o categorización de textos [5], entre otros. En particular la tarea de clasificación de textos, se basa en asignar un texto a una categoría de acuerdo a ciertas características. Se cuenta con buenos

resultados en estudios de clasificación por su tema o tópico, por ejemplo, si un texto pertenece a deportes, política, sociales. Actualmente se ha incrementado la investigación sobre clasificación no-temática, es decir, encontrar un estilo de redacción para determinar el autor de un texto o clasificar por su género literario. Así podemos decir que existen dos tipos de categorización de textos: la temática, interesada en el qué y la no-temática interesada en el cómo fue escrito el texto.

Este trabajo de tesis se orienta a la clasificación no-temática. Se busca clasificar un texto propagandístico por su contenido, tomando en cuenta la polaridad de las palabras que se usan y los atributos que conducen a una clase adecuada a la que pertenece. Se analiza un texto en español relacionado con la propaganda en México [7], para decir si un texto es positivo o negativo (estable, inestable).

Existe un área exclusiva sobre investigación en lingüística computacional [6], donde se indica qué métodos empíricos o basados en corpus son una opción adecuada para desarrollar sistemas de procesamiento del lenguaje natural eficientes. Entre estos métodos están las formas de aprendizaje automático. Cada idioma enfrenta un problema diferente por lo que existen más trabajos de investigación en unos que otros. En particular, en el idioma inglés hay una gran lista de trabajos realizados en este tema. Las herramientas que llevan a cabo un análisis lingüístico son de vital importancia, por lo que este tipo de tareas es generada por expertos en el área. Un ejemplo de esto es el corpus etiquetado, donde la clasificación de textos debe ser revisada y clasificada por una persona con conocimientos en la materia.

En este trabajo se dan los primeros pasos para tener un corpus en español etiquetado por un experto en materia propagandística. Adicionalmente se cuenta con un diccionario, de verbos que son característicos de cada clase.

Se propone trabajar en clasificación automática de textos con un aprendizaje supervisado.

El “análisis” de un texto propagandístico requiere de un conocimiento político, psicológico, sociólogo, entre otros, para determinar algo más que una opinión acerca de si un texto tiene un enfoque “bueno o malo”, “blanco o negro”, “estable o inestable” o si al dar un discurso es recomendable utilizar ciertas frases o palabras y el porqué. Un resultado así sería lo ideal para un analista en esta área porque facilitaría el trabajo que realiza, este tipo de análisis lo hace el experto a través de sus conocimientos y experiencias en las diversas áreas de estudio. Para que un sistema computacional pudiera dar un resultado de esta dimensión, necesitamos que analice, clasifique y “piense” como lo hace un humano que es experto en propaganda. En este trabajo nos limitamos a definir si el documento propagandístico pertenece a una de las clases previamente definidas.

1.1. Descripción del problema

Antecedentes del Proyecto.

La idea surge porque en la literatura no se cuenta con una herramienta en español que ayude al curioso o al estudiante en política a clasificar documentos propagandísticos, tomando en cuenta que el tema de la propaganda en México es relevante y genera información en todo momento.

Nos enfocamos en la clasificación de textos y en particular documentos propagandísticos en México como tema central de esta tesis. Los recursos lingüísticos, el corpus, la tarea del experto en propaganda y nuestros algoritmos de clasificación son la base para encontrar un método adecuado de clasificación en documentos propagandísticos. Se usarán las categorías *Estable e Inestable* para clasificar un texto con ayuda de los algoritmos que existen en la literatura. Otro recurso importante con el que se cuenta es un diccionario de verbos que son frecuentes en cada una de las clases; éste se ha enriquecido con sus sinónimos.

Se trabajará con clasificación automática de textos con un aprendizaje supervisado. Tomando en cuenta que no existe registrado en la literatura algún corpus sobre propaganda en México, se requiere de la construcción de dicho corpus usando la información propagandística de 3 periódicos nacionales. El corpus construido es en su mayoría texto sobre la campaña política de las elecciones federales 2012 de los periódicos: El Universal, La Jornada y La Razón. Con estos datos se da un primer paso para tener un corpus sobre documentos propagandísticos en México.

Relacionado con este tema se encuentran trabajos de análisis de opinión, análisis de sentimientos, entre otros, que serán de gran ayuda e importancia para los trabajos a futuro.

Nos haremos las siguientes preguntas de investigación:

1. ¿Es posible determinar la polaridad de un texto propagandístico usando la frecuencia de palabras que aparecen en los textos de entrenamiento?
2. ¿Es posible determinar la polaridad de un texto propagandístico usando un diccionario de verbos que caractericen las clases definidas?

1.2. Objetivos

Objetivo general:

- Desarrollar un modelo computacional que permita determinar de forma automática la polaridad (positiva, negativa) de un texto propagandístico.

Objetivos particulares:

- Crear una técnica simple basada en frecuencias para determinar la polaridad de un texto propagandístico.

- Diseñar un modelo basado en técnicas de aprendizaje automático para determinar la polaridad de un texto propagandístico.
- Construir un diccionario de palabras en español con polaridad positiva y negativa, asociados al dominio de texto propagandístico.

1.3 Alcances y limitaciones

Se propone trabajar en el tema de clasificación de textos propagandísticos con la construcción de un corpus en español, donde a través de una fase de entrenamiento el documento es etiquetado en una clase. Se trabajará con una clasificación de textos: “positivos y negativos” o “estables e inestables” como se nombran a las clases definidas.

A partir de esto se pueden definir más clases y subclases, enriqueciendo el corpus con la ayuda de un experto en propaganda, tanto el corpus como el diccionario de verbos pueden ir creciendo. También se puede trabajar a futuro con el tema de análisis de sentimientos para tener un sistema experto con enfoques diferentes.

1.4 Contribuciones

Se busca que esta herramienta de clasificación ayude al curioso, al estudiante de ciencia política y al experto en propaganda política en México.

Se aporta un diccionario de palabras las cuales nombramos como “azules” y “rojas”, las azules son palabras que normalmente se usan más en un texto estable y las rojas en un texto inestable. Este diccionario puede servir al lector para conocer como las palabras se usan en cada contexto y tener su punto de vista o simplemente para ampliar el vocabulario.

1.5 Estructura de la tesis

El contenido de esta tesis está organizado de la siguiente forma:

En el Capítulo 1 se da un panorama general del problema, los objetivos, las propuestas para su solución, las contribuciones y los alcances de este trabajo.

El Capítulo 2 es el marco teórico, donde se explican los conceptos básicos que son necesarios para la comprensión del trabajo realizado, definiendo términos relacionados con la propaganda, polaridad de un texto, aprendizaje automático, la clasificación de textos, los algoritmos de clasificación y métodos de evaluación.

El Capítulo 3 habla de los algunos trabajos relevantes registrados en la literatura, relacionados con aprendizaje automático, la clasificación de textos, análisis de opinión y principalmente donde se ha trabajado con polaridad de textos.

El Capítulo 4 describe las técnicas propuestas en este trabajo de tesis, para determinar la polaridad de textos propagandísticos.

Los resultados experimentales son presentados en el Capítulo 5, donde se discuten cada uno de los experimentos planteados.

Finalmente en el Capítulo 6 las conclusiones y trabajo a futuro.

Capítulo 2

Marco teórico

En este capítulo se definen los conceptos básicos para que el lector esté familiarizado con los términos y herramientas que se usan en el desarrollo de la tesis. En la sección 2.1 se dan algunas definiciones relacionadas con la propaganda política y con la polaridad de textos. En 2.2 se definen los tipos de aprendizaje automático; en 2.3 los algoritmos de aprendizaje; en 2.4 la clasificación de textos; en 2.5 se define la representación vectorial y en 2.6 las medidas de evaluación.

2.1 Definiciones

2.1.1 Acerca de la propaganda

Etimológicamente el término propaganda es originalmente un gerundio del latín, Lourdes Salgado [ver más en 8]. Otros sostienen que deriva de la palabra "propagare" que significa sembrar o difundir. No existe una definición unánime de propaganda. Desde el punto de vista gramatical, es el nombre de toda acción que lleve a difundir o extender el conocimiento de algo. Son propaganda la educación, la publicidad, el intercambio de ideas entre dos personas.

En un diccionario convencional encontramos esta definición: "Acción intensa a favor de una idea, institución o actuación política, destinada a ganarse el apoyo de la opinión pública por medio de los sistemas de difusión de masas" [9]. Esta definición es clara y comprensible en cuanto a la función de la propaganda, puede

ser un poco vaga e imprecisa, la palabra "acción" por ejemplo no nos aclara nada en cuanto a las técnicas, métodos o actividades que se han de realizar. Limita a la propaganda a realizarse en "favor de", cuando la propaganda comprende campañas en "contra", y no especifica si el "ganarse el apoyo" implica solamente que la gente emita una opinión favorable o la obtención de cierta conducta.

De acuerdo con Edmundo Llaca [10] la propaganda es un conjunto de métodos basados principalmente en las materias de la comunicación, la psicología, la sociología y otras, que tienen como propósito influir en que un grupo humano adopte la opinión política de una clase social, la cual se vea reflejada en una determinada conducta.

Según Andrés Valdez Zepeda [11], es el conjunto de técnicas y medios de comunicación social tendientes a influir con fines ideológicos en el comportamiento humano, la propaganda moldea la percepción de la audiencia.

La propaganda es el conjunto de recursos humanos, materiales y técnicos que tienen como propósito, orientar la opinión de una o varias personas hacia un fin específico. La propaganda es anterior a la mercadotecnia y desde sus inicios, se aplicó el término para los temas relacionados con el poder sobre la gente, excluyendo a los productos y servicios. Hasta los años setenta del siglo pasado, comenzó a ser sustituido el término “propaganda” por un neologismo: mercadotecnia política. En la actualidad, algunos grupos insisten en que no es lo mismo “propaganda” que “mercadotecnia política”, aunque lo sean [7].

2.1.2 Definición de jerga

Jerga es el nombre que recibe una variedad lingüística del habla y a veces incomprensible para los hablantes de ésta, usada con frecuencia por distintos grupos sociales con intenciones de ocultar el verdadero significado de sus palabras, a su conveniencia y necesidad. Lenguaje especial y familiar que usan entre sí los individuos de ciertas profesiones y oficios, como los toreros, los estudiantes, entre otros.

Normalmente, los términos usados en la jerga de grupos específicos son temporales (excepto las jergas profesionales), perdiéndose el uso poco tiempo después de ser adoptados.

Profesionales: necesitan de cierto vocabulario que no es común al resto del idioma para ciertos procesos o instrumentos. Por ejemplo, una persona ajena al ámbito docente diría: "Me gusta la forma de enseñar del profesor", mientras que otro docente diría: "Me gusta la didáctica del profesor". Existen diccionarios oficiales para este tipo de jergas.

Sociales: Distintas formas de comunicarse con el propósito de no ser entendido por los demás (por ejemplo en la cárcel) o con intención diferenciadora (de algunos barrios y de adolescentes). En general no hay ningún diccionario que contenga esta jerga debido a la poca perdurabilidad que tiene. [12]

2.1.3 Diccionario

Según la Real Academia Española (RAE) un diccionario es un repertorio en forma de libro o en soporte electrónico en el que se recogen, según un orden determinado, las palabras o expresiones de una o más lenguas, o de una materia concreta, acompañadas de su definición, equivalencia o explicación.

Un diccionario es una obra de consulta de palabras o términos que se encuentran ordenados alfabéticamente. De dichas palabras o términos se proporciona su significado, en el caso de los diccionarios de sinónimos se relacionan con palabras de significado similar. Los más sencillos se limitan a dar una lista de palabras para cada entrada, pero algunos más completos indican además las diferencias de matiz con la palabra buscada, sin llegar a ser un tesauro.

2.1.4 Lexicón

En una lengua natural se estudian los diversos productos del mecanismo mental (enunciados). Examinando los enunciados se formulan hipótesis sobre el funcionamiento y estructura de la lengua. Toda lengua consta de formas léxicas, de conjunto de reglas gramaticales (morfológicas y sintácticas) y de un conjunto de reglas fonológicas. El conjunto de formas léxicas que almacena en el cerebro

un hablante constituye su lexicón. La estructura almacena formas léxicas diversas: morfemas, palabras, frases hechas.
Cada forma léxica tiene asociada cierta información:

- Categoría gramatical.
- Significado.
- Organización morfológica.
- Información léxico-sintáctica.
- Representación fonológica.

Una de las disciplinas interesadas en el léxico es la lingüística computacional. En general, la lingüística se ocupa de estudiar:

- Las reglas que estructuran las formas léxicas (morfología).
- Las redes semánticas que forman los significados de las formas léxicas (semántica).
- La composición, evolución y organización de las formas léxicas (vocabulario) de los hablantes (semántica histórica).
- Las características sintácticas de las diversas formas léxica (teoría léxico-sintáctica).

2.1.5 Polaridad de un texto

La polaridad de un texto se refiere a determinar su orientación emocional, a decidir si un texto dado que contiene opiniones expresa una opinión a favor o en contra (positiva o negativa). Al hablar de polaridad podemos abarcar tan profundo como se quiera o se necesite, ya que podemos no solo definir positivo o negativo sino que tan positivo o negativo es un texto, para ello se tiene un estudio en el análisis de sentimientos donde con ayuda de los cuantificadores en el texto se puede calificar las emociones del texto.

2.2 Aprendizaje automático

Han existido diversas opiniones de si una máquina puede ser inteligente o puede llegar a pensar como un ser humano; esto ha sido una interrogante desde el principio de la computación. Lo cierto es que se han desarrollado algoritmos efectivos para desarrollar algunas tareas de aprendizaje.

El ser humano a lo largo de su vida adquiere conocimientos. Al hablar de aprendizaje podemos definirlo como un proceso a través de cual se adquiere conocimientos, habilidades, valores, destreza, actitudes, esto como resultado del estudio, experiencia, razonamiento, observación, entre otros. A esto le llamamos aprendizaje.

El conocimiento a priori es el conocimiento que se adquiere de forma innata y natural, sin necesidad de la experiencia. Podemos decir que un programa aprende si mejora los resultados de alguna tarea en base a la experiencia. El aprendizaje automático es una rama de la inteligencia artificial que estudia los procesos de como “aprender”.

Aprendizaje Automático: es el campo de estudio que da a las computadoras la capacidad de aprender sin ser explícitamente programado según Arthur L. Samuel [13].

Dando una definición matemática podemos decir que una computadora “aprende” a desarrollar la tarea T, si después de proporcionarle la experiencia E, el sistema es capaz de desempeñarse razonablemente bien cuando se vuelva a necesitar que la tarea se ejecute.

“Un programa de ordenador APRENDE:
a partir de una experiencia E
a realizar una tarea T
(de acuerdo con una medida de rendimiento P),
si su rendimiento al realizar T,
medido con P,
mejora gracias a la experiencia E”

Cuando utilizamos aprendizaje automático, el objetivo es que se aprenda a clasificar con base en la experiencia; esto es, a partir de documentos previamente clasificados de forma manual donde un experto sabe que pertenecen a cierta clase, así el sistema va aprendiendo con base en atributos característicos de la clase. Si contamos con clases definidas y los documentos definidos en alguna de esas clases, estaremos usando aprendizaje supervisado.

En el aprendizaje no supervisado (también llamado clustering), las clases son definidas por el sistema de acuerdo a los documentos que va leyendo, es decir, no necesita de un entrenamiento como tal.

El aprendizaje semi-supervisado consiste en tener una parte de entrenamiento y después el sistema sea capaz de crear clases y clasificar los documentos.

2.2.1 Tipos de aprendizaje

En el proceso de aprendizaje se puede tener la opción de supervisar de maneras diferentes dicho proceso, depende de las condiciones, los recursos con lo que se cuente, se elige algún método apropiado. En el aprendizaje supervisado contamos con un experto en el dominio o tema, podemos contar con el aprendizaje no supervisado, también existe el semi- supervisado y aprendizaje por refuerzo.

Aprendizaje supervisado

Para llevar a cabo el aprendizaje supervisado se requiere tener un corpus de entrenamiento previamente clasificado por un experto humano, se intenta aprender de los ejemplos, es decir, que esos documentos son un maestro para el sistema. A través de un algoritmo de clasificación se tienen las características o atributos que definen la clase a la que pertenecen, dicha clase esta previamente definida. Es llamado supervisado por la presencia de los ejemplos de los que se intenta aprender.

El algoritmo produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema. Un ejemplo de este tipo de algoritmo es el problema de clasificación, donde el sistema de aprendizaje trata de etiquetar

(clasificar) una serie de vectores utilizando una entre varias categorías (clases). La base de conocimiento del sistema está formada por ejemplos de etiquetados anteriores. Este tipo de aprendizaje puede llegar a ser muy útil en problemas de investigación biológica, biología computacional y bioinformática.
Algunos algoritmos comúnmente usados por sus buenos resultados son principalmente los bayesianos, pero también están otros como: Máquinas de Soporte Vectorial (SVM), Vecinos cercanos, J48 o C4.5.

Aprendizaje no supervisado

El aprendizaje no supervisado se enfoca en descubrir patrones comunes entre los datos, no se necesita de un maestro, ni de un conjunto de entrenamiento como el caso supervisado. Todo el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formado tan sólo por entradas al sistema, es decir, no se tiene información sobre las categorías de esos ejemplos. Por lo tanto, en este caso, el sistema tiene que ser capaz de reconocer patrones para poder etiquetar las nuevas entradas. De lo cual se predicen características para formar agrupación. Por lo cual también se llama *clustering.*
De los algoritmos comúnmente usados en este tipo de aprendizaje son Kmeans, CobWeb, EM, entre otros.

Aprendizaje semi-supervisado

Este tipo de algoritmos combinan los dos algoritmos anteriores para poder clasificar de manera adecuada. Aquí se aprende con la ayuda de dos conjuntos, uno que contiene datos asociados a su clase y uno que contiene datos no asociados a ninguna clase. Se tiene en cuenta los datos marcados y los no marcados. Algunos algoritmos usados son: Co-training, ASSEMBLE y self-training.

Aprendizaje por refuerzo

El aprendizaje por refuerzo trata de “aprender” a través de la experiencia, pero no a través de un conjunto de ejemplos. El algoritmo aprende observando el mundo

que le rodea. Su información de entrada es el feedback o retroalimentación que obtiene del mundo exterior como respuesta a sus acciones. Por lo tanto, el sistema aprende a base de ensayo-error [18].
Algunas de sus desventajas son que no se sabe si la acción tomada es la mejor posible; requiere explorar las diferentes acciones para aprender cual es la mejor.

2.3 Algoritmos de aprendizaje

Las redes bayesianas, los árboles de decisión y las redes neuronales artificiales son los métodos que son usados en aprendizaje automático en tareas como clasificación de documentos. Una de las ventajas que ofrecen los métodos bayesianos es un análisis no solo cualitativo sino también cuantitativo de los atributos; dan una medida probabilística de la importancia de las variables o atributos en los textos. Es por ello que el aprendizaje basado en redes bayesianas es adecuado y tiene buenos resultados al ser usado en tareas de clasificación de textos.

Algunas características de los métodos bayesianos:

- Cada ejemplo observado va a modificar la probabilidad de que la hipótesis formulada sea correcta aumentando o disminuyendo, es decir, una hipótesis que no concuerda con un conjunto de ejemplos grandes no es desechada por completo, sino que disminuirá la probabilidad estimada por la hipótesis
- Los métodos son robustos al posible ruido presente en los ejemplos de entrenamientos y a la posibilidad de tener entre esos ejemplos de entrenamientos datos incompletos o erróneos.
- Los métodos bayesianos permiten tener en cuenta la predicción de la hipótesis el conocimiento a priori o conocimiento de dominio en probabilidades.

Cualquier sistema de clasificación de patrones se basa en lo siguiente: dado un conjunto de datos (entrenamiento y pruebas) representados por pares <atributo,

valor>, el problema consiste en encontrar una función *f(x)* llamada hipótesis que clasifique los textos.

Al usar el teorema de Bayes en cualquier problema de aprendizaje automático, en especial los de clasificación, podemos estimar las probabilidades a posteriori de cualquier hipótesis consistente con el conjunto de datos de entrenamiento para así escoger la hipótesis más probable. Para estimar estas posibilidades se han propuestos múltiples algoritmos entre ellos los Naïve Bayes.

2.3.1 Naïve Bayes

El clasificador Naïve Bayes [14] es uno de los clasificadores probabilísticos, está construido usando un conjunto de entrenamiento que son los *N* documentos de los cuales se tienen los atributos (palabras) y así estimar la probabilidad de cada clase, dados estos valores se usa el Teorema de Bayes.

$$P(c_j|d) = \frac{P(c_j)P(d|c_j)}{P_d}$$

Como el denominador no distingue entre categorías lo omitimos. Con este método se asume que los atributos son condicionalmente independientes, dadas las clases y de esta forma simplifica los cálculos teniendo.

$$P(c_j|d) = P(c_j)\prod_{i=1}^{M} P(d_i|c_j)$$

$P(c_j)$ Puede ser calculado de la fracción de documentos de entrenamiento que es asignada a la clase c_j

$$\overline{P}(c_j) = \frac{N_j}{N}$$

Donde N_j es el número de documentos de entrenamiento para los cuales la clase es c_j y N es el número total de documentos de entrenamiento. Donde tenemos una estimación $P(d_i|c_j)$ para $\overline{P}(d_i|c_j)$ dada por:

$$\overline{P}(d_i \mid c_j) = \frac{1 + N_{ij}}{M + \sum_{k=1}^{M} N_{kj}}$$

Y donde N_{ij} es el número de veces de la palabra *i* ocurrida dentro de los documentos de la clase c_j en el conjunto de entrenamiento. Para evitar el problema de la probabilidad cero se utiliza Laplace y se agrega un 1. *M* es el número de términos.

2.3.2 K-Vecinos más cercanos

El algoritmo de *k*-vecinos más cercanos (también llamado k-NN, por sus siglas en inglés) para clasificar un nuevo documento *d* selecciona del conjunto de entrenamiento los *k* documentos más semejantes a *d* y asigna el documento a la clase que obtenga una mayor cantidad de muestras en la vecindad.

Se asigna la clase más probable de acuerdo a la búsqueda de ejemplos similares almacenados. El principal cálculo se da cuando se localizan los *k* vecinos más cercanos.

Se usa la distancia Euclidiana entre dos vectores $d = d_1,...,d_n$ y $e = e_1...e_n$ con la ecuación:

$$distancia(d,e) = \sqrt{(d_1 - e_1)^2 + (d_2 - e_2)^2 + ... + (d_n - e_n)^2} = \sqrt{\sum_{i=1}^{n} (d_i - e_i)^2}$$

Para encontrar el coseno tenemos la ecuación:

$$distancia(d,e) = \frac{\sum_{i=1}^{n} (d_n \, x \, e_n)}{\sqrt{\sum_{i=1}^{n} d_n^2 + \sum_{i=1}^{n} e_n^2}}$$

2.3.3 Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial (SVM) tienen un buen desempeño en problemas de clasificación. Es uno de los algoritmos que se recomienda usar en esta tarea de clasificación de textos. SVM se puede ver como el intento de encontrar una superficie (σ_i) que separe los ejemplos positivos de los negativos con una diferencia considerable.

A través de todas las superficies $\sigma_1, \sigma_2, \ldots$ se hace la búsqueda de σ_i que cumple que la distancia mínima entre σ_i y algún ejemplo de entrenamiento sea máxima en el espacio $|A|$ - dimensional que separan a los ejemplos positivos de los negativos en el conjunto de entrenamiento, que son las conocidas superficies de decisión.

2.3.4 LogitBoost

El algoritmo se basa sobre un modelo de regresión logística aditiva de los datos de entrenamiento. El modelo aditivo es una aproximación a la función □(□) de la forma:

$$F(x) = \sum_{m=1}^{M} c_m f_m(x)$$

Donde c_m es la constante a determinar y f_m es la función básica. Si $F(x)$ es el mapeo adecuado que se busca entonces la hipótesis fuerte agregada y f_m es nuestra hipótesis débil, luego se muestra que está ajustado al modelo por la minimización del criterio.

$$J(x) = E(e^{-yf(x)})$$

Donde *y* es la etiqueta de la clase verdadera. Este algoritmo minimiza el criterio mediante el uso de los pasos Newton para adaptar el modelo de regresión logística aditiva para una directa optimización del logaritmo de la verosimilitud binomial.

$$-\log(1 + e^{-2yF(x)})$$

En este algoritmo la variable *y* toma valores de 0 y 1 como resultado final y representa la probabilidad de la forma *y = 1* para *p(x)*, donde *p(x)* se obtiene de la forma:

$$p(x) = \frac{e^{F(x)}}{e^{F(x)} + e^{-F(x)}}$$

2.3.5 Árboles de decisión C4.5

C4.5 es una extensión del algoritmo ID3, este algoritmo genera un árbol de decisión a partir de los datos mediante participaciones realizadas recursivamente. El árbol se construye mediante la estrategia de primero - profundidad (depth - first).

El algoritmo C4.5 utiliza una técnica heurística conocida como proporción de ganancia (gain ratio). Es una medida basada en información que considera diferentes números y diferentes probabilidades de los resultados de las pruebas. El algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de los datos y selecciona la prueba que le haya generado la mayor ganancia de información. Para cada atributo discreto, se considera una prueba con *n* resultados, siendo el número de valores posibles que puede tomar el atributo. Para cada atributo continuo, se realiza una prueba binaria (1,0) sobre cada uno de los valores que toma el atributo en los datos. En cada nodo, el sistema debe decidir que prueba escoge para dividir los datos.

Según Espino (2005), los 3 tipos de pruebas posibles propuestas para el C4.5 son:

- La prueba estándar para las variables discretas, con un resultado y una rama para cada valor posible de la variable.
- Una prueba más compleja, basada en una variable discreta, en donde los valores posibles son asignados a un número variable de grupos con un resultado para cada grupo, en lugar de para cada valor.

- Si una variable A tiene valores numéricos continuos, se realiza una prueba binaria con resultados $A<-Z$ y $A>Z$, para lo cual debe determinar el valor límite Z.

Todas estas pruebas se evalúan observando la ganancia resultante de la división de datos que producen. Ha sido útil agregar una restricción adicional para cualquier división al menos de dos de los subconjuntos C_i deben contener un número razonable de casos. Esta restricción que evita las subdivisiones casi triviales, es tenida en cuenta solamente cuando el conjunto C es pequeño.

2.4 Clasificación automática de textos

La Clasificación automática de textos surge de la necesidad de tener métodos menos costosos que los de forma manual, ya que el número tan elevado de documentos que día a día se tiene hacen que cada vez sea más complicado y costoso en tiempo y dinero.

Un primer paso para clasificar los documentos de manera automática es definir una serie de atributos o palabras que describen el texto a clasificar y de ahí se pone en un formato adecuado que el sistema pueda entender para realizar el análisis.

La clasificación de textos consiste en colocar un documento en su clase correspondiente, para ello necesitamos entrenar al sistema, para que en base a la "experiencia" vaya realizando la tarea de una manera satisfactoria.

Vamos a definir la clase C y el documento D y vamos a clasificar por medio de una función $o: IxC-> \{T,F\}$ llamada el clasificador, donde $C=\{c_1,c_2...,c_3\}$ es un conjunto de categorías previamente definidas y donde I es un conjunto de instancias del problema, comúnmente cada instancia es representada como una lista $A=\{a_1,a_2...,a_3\}$ de valores a los que se nombran como atributos.

Para tener un clasificador que haga esta tarea es necesario un proceso inductivo, un entrenamiento.

Existen diferentes tipos de clasificación y de acuerdo al problema y a las herramientas con las que se cuenta se elige alguna para trabajar. A continuación se mostrarán algunos tipos de clasificación.

Clasificación supervisada

La clasificación supervisada también llamada categorización, parte de la existencia de clases pre-definidas, tiene como objetivo colocar cada documento en su clase correspondiente.

La mayoría de los algoritmos parten la elaboración de un modelo o patrón para cada clase, esta fase se conoce como entrenamiento.

Necesita una colección de documentos ya clasificada manualmente (colección de entrenamiento)

Requiere intervención humana para la clasificación de la colección de entrenamiento y para la revisión y refinamiento de resultado

Clasificación no supervisada

También llamada clustering, donde no tenemos clases pre-establecidas y el propio sistema establece las clases o “clusters” de forma automática.

Clasificación semi-supervisada

El aprendizaje supervisado es aquel en donde se intenta aprender de ejemplos como si estos fueran un maestro. Se asume que cada uno de estos ejemplos incluye características o atributos que especifican o definen a qué categoría o clase pertenece, de un conjunto de categorías o clases predefinidas, de esta manera cada ejemplo se asocia con su clase.

Este tipo de aprendizaje es llamado supervisado por la presencia de los ejemplos para guiar el proceso de aprendizaje.

2.5 Representación de documentos o extracción de características.

La clasificación automática de textos tiene como objetivo categorizar documentos dentro de un número fijo de clases. Al utilizar la herramienta de aprendizaje automático, esté aprende a clasificar a partir de ejemplos que permitan elegir su categoría o clase de forma automática.

Para trabajar con los textos y asignarlos a su clase, debemos ponerlos en un formato que pueda ser leído por la computadora, esto consiste en obtener los atributos que describan el texto a clasificar y representarlos en vectores para ser utilizados por los algoritmos de aprendizaje, a este paso previo se le llama extracción de características.

En la extracción de características vamos a trabajar con:

- Pre-procesamiento
- Indexado

Pre-procesamiento

Consiste en trabajar con la colección de documentos eliminando aquellas palabras que no representan información útil a nuestro clasificador, podemos trabajar con las siguientes fases para este paso:

- Eliminar etiquetas: en esta parte se eliminan las etiquetas, caracteres como llaves, sinos de puntuación y todos aquellos símbolos que no aporten información.
- Eliminar palabras vacías: son todas las palabras que comúnmente se usarían en cualquier texto, como los artículos, preposiciones, pronombres, conjunciones, etc.
- Lematización de palabras: se usa algún truncador o lematizador para identificar las raíces y remover los sufijos para reducir una palabra por

su lema o raíz. Ejemplo de truncador es el Porter y de lematizador el Tree Tagger.

Indexado

Una vez eliminadas las palabras y características que no son útiles al clasificador necesitamos que los documentos sean interpretados por el clasificador, debemos transformarlos a un modelo adecuado.

El indexado denota la actividad de hacer el mapeo de un documento d_j en una forma compacta de su contenido. El más usado es el modelo vectorial que consiste en representar cada documento como un vector de palabras y así la colección de documentos la podemos representar como una matriz. Es decir, si tenemos nuestra colección de documentos *A*, donde cada entrada representa las ocurrencias de una palabra en un documento.

$$A = \begin{pmatrix} a_{11} \ldots a_{1m} \\ a_{n1} \ldots a_{nm} \end{pmatrix}$$

Donde cada vector es un documento con m atributos o palabras y n representa el número de documentos en la colección, a_{ij} es el número de ocurrencias de la palabra j en el documento i y existen varias formas para determinar este número al cual definimos como el peso de la palabra j en el documento i.

A continuación se describen 3 métodos de ponderado más usados para determinar el peso a_{ij}.

- **Ponderado Booleano**

  Asigna el peso 1 en caso de que la palabra ocurra o exista en documento y 0 en otro caso.

$$a_{ij} = \begin{cases} 1 \quad \text{si} \quad f_{ij} > 0 \\ 0 \textit{ en otro caso} \end{cases}$$

- **Ponderado por frecuencia de término**

  Consiste en asignar el número de veces que la palabra ocurre en documento.

$$a_{ij} = f_{ij}$$

- **Ponderado TFxIDF**

  En los casos anteriores no se toma en cuenta la frecuencia de términos en toda la colección de documentos, en el ponderado por frecuencia es solo por documento. El Ponderado TFxIDF (Term Frequency x Inverse Document Frequency) el cual asigna el peso a la palabra j en el documento i en proporción al número de ocurrencias de la palabra el documento, y en proporción inversa al número de documentos en la colección en la que cada palabra ocurre al menos una vez.

$$a_{ij} = f_{ij} x \log\left(\frac{N}{n_i}\right)$$

Donde f_{ij} es la frecuencia de la palabra i en documento j, N es el número de documentos en toda la colección, n_i es el número de documentos en los que la palabra i ocurre.

2.6 Medidas de evaluación en los clasificadores

Para analizar si un clasificador o método es efectivo, no solo debemos decir si se clasificó de forma correcta o no, hay que tomar en cuenta otras medidas de rendimiento que evalúan el desempeño de los clasificadores.

Sean:

- *a* el número de documentos correctamente clasificados (verdadero positivo).
- *b* el número de documentos incorrectamente clasificados (falso positivo).
- *c* el número de documentos incorrectamente rechazados (falso negativo).
- *d* el número de documentos correctamente rechazados (verdadero negativo).

La Precisión (□), el Recuerdo (□), son medidas ampliamente usadas para saber el desempeño del clasificador o del método usado.

La Exactitud nos da el porcentaje de los documentos correctamente clasificados.

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

La Precisión es la probabilidad de que un documento que es etiquetado en la clase c_1 en efecto pertenece a la clase c_1.

$$Precision = \frac{a}{a+b}$$

El Recuerdo es la probabilidad de que un documento que pertenece a la clase c_2 es etiquetado en la clase c_2.

$$Recuerdo = \frac{a}{a+c}$$

F-Measure (f) es una medida que engloba en una sola el Recuerdo y la Precisión descrita por

$$f = 2 * \frac{precision * recuerdo}{precision + cobertura}$$

Validación cruzada

El método de validación cruzada es un método de que se usa para estimar errores de predicción, el cual se basa en dividir un conjunto de documentos en k subconjuntos, y se repite este proceso como prueba y entrenamiento k veces. En cada proceso se utiliza una partición diferente como datos de entrenamiento y prueba, al terminar estos resultados son promediados.

Por ejemplo si a k le asignamos un valor de 5 podemos ver la representación en la tabla 2.1

	1	2	3	4	5
1ª	Prueba	Entrenamiento	Entrenamiento	Entrenamiento	Entrenamiento
2ª	Entrenamiento	Prueba	Entrenamiento	Entrenamiento	Entrenamiento
3ª	Entrenamiento	Entrenamiento	Prueba	Entrenamiento	Entrenamiento
4ª	Entrenamiento	Entrenamiento	Entrenamiento	Prueba	Entrenamiento
5ª	Entrenamiento	Entrenamiento	Entrenamiento	Entrenamiento	Prueba

Tabla 2.1 Ejemplo de validación cruzada con 5 pliegues

Capítulo 3

Estado del arte

En este capítulo se describen algunos trabajos previos registrados en la literatura que son relacionados a la tesis que se presenta, como es clasificación de documentos determinando la polaridad de textos, clasificación de opiniones y subjetividad de un texto. Algunos utilizando métodos de aprendizaje supervisado, no supervisado y en otros casos aprendizaje semi-supervisado.

3.1 Polaridad de textos

En referencia al trabajo presentado por Priego, Pinto y Mejri [19], se observa que se trabaja con la evaluación de la polaridad para unidades fraseológicas verbales (VPU). Este tipo de unidades lingüísticas son frases constituidas por dos o más palabras que se caracterizan por tener un cierto grado de fijación, además de tener la característica de que al menos una de estas palabras es un verbo que juega el rol de predicado. Se dice que las VPU no aparecen tan frecuentemente en las frases de los recursos léxicos que se han construido manualmente y por tanto, este problema de cobertura podría impactar el rendimiento para diferentes tareas en las cuales se involucra el procesamiento del lenguaje natural. Así, la construcción de sistemas de entendimiento automático para este tipo de estructura lingüística es muy importante ya que son una forma estándar de expresar un concepto o idea. En el artículo mencionado presentan un conjunto de experimentos para la identificación automática de la polaridad de VPU. Los

autores reportan un rendimiento máximo de 80% para esta tarea en particular cuando la información contextual de una unidad fraseológica está considerada, en comparación con un 62% cuando solamente se usa la VPU, es decir, sin su información conceptual. Estos resultados destacan la importancia de analizar automáticamente este tipo de estructuras lingüísticas.

Jain, Harshit, Mogadala, Aditya, Varma, Vasudeva [20] presentan el trabajo “Feature Analysis and Polarity Classification of Expressions from Twitter and SMS Data” donde utilizan frases de Twitter y mensajes SMS para el análisis de polaridad. Twitter es un microbloging de los más utilizados en las redes sociales donde los usuarios comparten sus opiniones personales importantes de eventos o sucesos y pueden libremente escribir opiniones cortas, utilizando de 1 a 140 caracteres. La polaridad del contexto de la frase donde aparece cierta palabra puede ser muy diferente al de la polaridad de la palabra. Las palabras positivas se utilizan en frases que expresan sentimientos negativos y viceversa. Muy a menudo, las palabras que son positivas y negativas fuera de contexto son neutrales en contexto, lo que significa que no están siendo utilizados para expresar un sentimiento. En una técnica con un gran léxico de palabras marcadas previamente, “mal” tendría una puntuación negativa, haciendo toda la frase negativa. Con el análisis a nivel de frase usando "muy mal", “muy” solo actúa como un intensificador para la palabra “mal” y toda la oración sigue estando marcada como negativa, solo cuando vemos la frase más allá, con otros argumentos podemos ver algún sentimiento positivo. El objetivo es la clasificación contextual de las expresiones de los tweets y SMS. Se trabajó con un método supervisado clasificando en tres clases: positiva, negativa y neutra; con 24939 expresiones. Los resultados de Precisión, Recuerdo y F-Measure se muestran en la tabla 3.1 en a) las frases en Twitter y en b) las frases en SMS. Se tienen resultados alentadores aunque se recomienda más investigación con un conjunto de datos equilibrado.

Clase	Precisión	Recuerdo	F-Measure
Positiva	0.8120	0.8120	0.8120
Negativa	0.6477	0.7073	0.6762
Neutral	0.3333	0.0375	0.0674

3.1 a) Expresiones en twitter

Clase	Precisión	Recuerdo	F-Measure
Positiva	0.6823	0.8263	0.7475
Negativa	0.7520	0.6947	0.7222
Neutral	0.0588	0.0063	0.0114

3.1 b) Expresiones en SMS

En “Automatic Construction Of Polarity-Tagged Corpus From HTML Documents” [22], NobuhiroKaji, Masaru Kitsuregawa presentan un trabajo en donde una de las cuestiones importantes fue determinar la polaridad u orientación semántica de un texto HTML, decidir si el texto transmite un contenido positivo o negativo. Se toma un enfoque estadístico donde a través de un método supervisado se tiene que aprender de un corpus hecho con textos como frases, palabras y documentos previamente etiquetados con la polaridad a la que corresponden. En este trabajo, los autores proponen un nuevo método para la construcción de la polaridad utilizando un corpus de documentos en HTML, la idea fue utilizar ciertas estructuras de diseño y modelo lingüístico. El corpus y el desarrollo de este trabajo es el idioma japonés. Mediante su uso se puede extraer de una forma fácil y automática frases que expresan opinión de los documentos HTML, esto es debido a que este método es totalmente automático y puede ser aplicado de manera arbitraria a los documentos HTML. Esto no sucedía con los métodos anteriores. Se construyó un corpus que consta de 126,610 oraciones o frases y para validar la calidad del corpus, dos expertos humanos evaluaron una parte del corpus y encontraron que el 92% de las oraciones fueron correctamente clasificadas. Se usó Naïve Bayes para el entrenamiento y fue probado en 3 conjuntos de datos. Los resultados mostraron que el clasificador logra más de un 80% de precisión en cada conjunto de datos.

Álvarez Romero, en su tesis “Clasificación Automática de Textos usando Reducción de Clases basada en Prototipos” [23], se enfoca en el estudio del desempeño que pueden alcanzar dos clasificadores que son los más usados en el tema de clasificación de Textos (Naïve Bayes y SVM), si se reduce el problema inicial multi-clase a un problema donde el clasificador sólo tenga que distinguir entre dos clases, un problema binario. Para la reducción se propone un esquema de prototipo para representar a las clases, que a diferencia de otros, asigna un peso a cada atributo de acuerdo a la importancia que éste tiene para cada clase. Como resultado de los experimentos se realizó un método de Clasificación Automática de Textos usando reducción de clases basada en prototipos. Este método resultó tener mejor desempeño comparable al que resulta de aplicar la clasificación de manera tradicional. Se tiene un procedimiento para el cálculo de prototipos de las clases a diferencia de otros métodos que dan un valor a la palabra de acuerdo al valor que tenga para cada clase, se tiene una medida de similitud basada en una intersección pesada de palabras que junto con el cálculo propuesto de prototipo llegan a tener desempeños comparables al mejor método de clasificación basado en prototipos.

Abu-Jbara, Jefferson Ezra, Dragomir Radev en “Towards NLP-based Bibliometrics” [24] nos hablan sobre la bibliometría que es una parte de la cienciometría que aplica métodos matemáticos y estadísticos a toda la literatura de carácter científico y a los autores que la producen, con el objetivo de estudiar y analizar la actividad científica. Para ello se ayuda de leyes bibliométricas, basadas en el comportamiento estadístico regular que a lo largo del tiempo han mostrado los diferentes elementos que forman parte de la Ciencia. Los instrumentos utilizados para medir los aspectos de este fenómeno social son los indicadores bibliométricos, medidas que proporcionan información sobre los resultados de la actividad científica en cualquiera de sus manifestaciones. Entonces las medidas bibliométricas son comúnmente usadas para estimar la popularidad y el impacto de investigación publicada. Estas medidas bibliométricas existentes proporcionan

indicadores cuantitativos de lo bien que un artículo es publicado. Esto no necesariamente representa la calidad de los trabajos presentados en el documento, por ejemplo si esto se calcula para un investigador, todas las citas son tratadas por igual, sin hacer caso de que algunas citas podrían ser negativas. En este trabajo los autores proponen el uso del PLN para agregar un aspecto cualitativo a la bibliometría. Se analiza el texto que acompaña a las citas en artículos científicos. Se proponen métodos supervisados para identificar el texto que se está citando y analizarlo para determinar el efecto (autor intención) y polaridad (autor sentimiento).

3.2 Análisis de sentimientos

Ulli Waltinger en su tesis “A Lexical Resource for German Sentiment Analysis” [25] evalúa una opción en inglés y tres para alemán. El análisis de sentimiento y la clasificación de polaridad se han estudiado ampliamente en los diferentes niveles de documentos (como frases y oraciones), sin embargo pocos exploraron el efecto de una función de selección de polaridad basada en la subjetividad para el idioma alemán. Este trabajo evalúa cuatro diferentes recursos de sentimientos con el clasificador SVM, de forma comparativa y combinando características basadas en polaridad. Usando un enfoque de traducción semi-automática, fueron capaces de construir tres recursos diferentes para un análisis de sentimiento para el idioma alemán. El diccionario llamado “German Polarity Clues” finalizado ofrece 10,141 características de polaridad, asociados a tres clases, determinando la dirección positiva, negativa y neutral de características específicas. Los resultados muestran que el tamaño de los diccionarios se relaciona con las características, no así con la exactitud. Usando una polaridad basada en selección de características, considerando una cantidad mínima de obtiene para ambos idiomas el mejor rendimiento.

Meng, Sistla, Clement Yu, Dragut, Wang presentan “Polarity Consistency Checking for Sentiment Dictionaries” [26], donde destacan que la polaridad de palabras es importante para aplicaciones como Minería de opiniones y Análisis de sentimientos. Un número de sentimientos de palabras y oraciones en el diccionario han sido construidos manual o semi-automáticamente. Los diccionarios tienen inexactitudes sustanciales, además de casos obvios, donde la misma palabra aparece con diferentes polaridades en diferentes diccionarios. Los diccionarios exhiben casos complejos que no pueden ser detectados por mera inspección. Se introduce el concepto de consistencia de la polaridad de las palabras/sentidos en diccionarios de sentimientos en este documento. Los autores reducen el problema de la consistencia de polaridad a un problema de satisfacibilidad (SAT) y utilizan un algoritmo Fast para detectar incoherencias de un diccionario de sentimientos. Se llevaron a cabo experimentos en cuatro diccionarios de sentimientos y en WordNet. Las opiniones expresadas en la Web como blogs, periódicos, etc., son un importante criterio para el éxito de un producto o de una política de estado. Por ejemplo, un producto consistentemente con buenas críticas es probable que se venda así. El enfoque general es resumir la polaridad semántica (positivo, negativo) de las oraciones/documentos mediante el análisis de las orientaciones de las palabras individuales. Los diccionarios de sentimientos se utilizan para facilitar la integración. Existen numerosos trabajos que, dado un sentimiento léxico analizan la estructura de una frase/documento para inferir su orientación, el titular de una opinión, el sentimiento de la opinión, etc.; varios dominios independientemente del diccionario de sentimientos han sido manualmente o semi-automáticamente creados. Los autores presentan un sistema de clasificación de la polaridad de los textos usando patrones. Un número de sentimientos en las palabras y oraciones en el diccionario han sido manualmente o semi automáticamente construidos.

3.3. Clasificación de opiniones

Nadia Araujo se enfoca en la tarea de clasificación de opiniones [27], donde se aborda el problema de determinar la polaridad de opiniones, es decir clasificar aquellas opiniones que expresan algo a favor o en contra, a nivel de oración, bajo un enfoque de Aprendizaje Computacional utilizando características léxicas. Una de las contribuciones de este trabajo es la caracterización de opiniones y la creación de un corpus con un enfoque de aprendizaje semi-supervisado de clasificación de textos de opinión. Caracterizar un objeto, en este caso un texto o un fragmentos e texto, consiste en extraer un conjunto de atributos que describan el objeto. En este caso se busca una descripción que permita distinguir las opiniones positivas de las negativas y por otro lado que dicha descripción sea fácil de extraer o construir, es decir, que sea concisa y que el número de atributos no sea excesivamente grande. En una primera prueba se utilizó Naïve Bayes como método de aprendizaje computacional; se utilizó como conjunto de atributos la bolsa de palabras usando secuencias de una y dos palabras (uni-gramas y bi-gramas) y el uso de Secuencias Frecuentes Maximales para caracterizar los datos, usando validación cruzada en 10 pliegues con la herramienta Weka. La mejor exactitud fue de 63.27%. Se eliminaron las instancias con *0* en los atributos logrando una exactitud del 66.69 con validación cruzada. La siguiente prueba fue con Secuencias Frecuentes Maximales para determinar el umbral de frecuencia apropiado, donde se tuvo el mejor resultado con un umbral de frecuencia de 4 con un 72.09% de exactitud. Posteriormente el método supervisado basado en self-training tuvo un 73% de exactitud.

La tesis “Clasificación Automática de Textos considerando el Estilo de Redacción” presentada por Coyotl Morales [28], propone métodos que permiten determinar los rasgos léxicos para caracterizar el estilo de redacción de los textos contribuyendo a las tareas de atribución de autoría y clasificación por género. Los métodos propuestos tienen como objetivo determinar el conjunto de colocaciones propias de un estilo; éstas se expresan a través de secuencias de elementos léxicos:

palabras y signos de puntuación. El primer método caracteriza los documentos a partir de las secuencias frecuentes maximales de la colección. Se determina un umbral de frecuencia, de esta manera el desempeño del método depende en su mayoría de la buena definición del umbral, por lo cual es necesario probar hasta obtener el más adecuado. El segundo método (iterativo), ensambla un conjunto de características al combinar las secuencias frecuentes maximales extraídas con diferentes umbrales de frecuencia; este conjunto es el resultado de un proceso iterativo y de la definición de una condición de paro.

Capítulo 4

Clasificación Automática de Documentos Propagandísticos

La clasificación automática de textos propagandísticos es una disciplina con la que no se tienen trabajos registrados en la literatura en México. Dentro de la clasificación automática de textos podemos identificar varias subtareas, como encontrar la subjetividad del texto, la fuerza o grado de opinión, polaridad u orientación de un texto. En esta última se enfocan los resultados realizados en la presente tesis, clasificar los textos en dos clases definidas: estable e inestable. Aunque esta clasificación se puede realizar con una palabra, una oración y un documento, en este trabajo de tesis nos enfocamos al análisis de los documentos.

Se describe en 4.1 el corpus en español; en el 4.2 se habla del diccionario de verbos el cual se enriqueció con sinónimos; en 4.3 con el uso de técnica de frecuencias, los algoritmos de clasificación y el corpus se desarrolla un método para el entrenamiento.

4.1 Corpus en español

Se define un corpus como una colección de textos en un leguaje natural, elegido para caracterizar un estado o variedad de un lenguaje. Dentro de los tipos de

corpus que existen destaca el corpus del lenguaje escrito y corpus del lenguaje hablado. En cualquiera de los dos tipos el corpus actúa como un repositorio de información el cual puede ser manipulado para extraer el conocimiento.

Se construyó un corpus relativamente pequeño pero suficiente para nuestro objetivo en esta tesis. Se dan los primeros pasos para el primer corpus en español sobre el tema propagandístico registrado en la literatura. Se tomaron las noticias de las elecciones federales 2012 donde se eligieron 3 periódicos nacionales por su nota periodística: La Razón, La Jornada y El Universal. Se recolectaron 409 documentos, de los cuales el 70% fueron etiquetados como ambiguos con la ayuda del experto en el tema de propaganda para su clasificación manual, se recomienda enriquecer este corpus en español para futuros trabajos.

En la tabla 4.1 se presentan la cantidad de documentos y los medios periodísticos donde se obtuvo la colección.

Periódico nacional	**La Razón**	**El Universal**	**La Jornada**
No. De documentos	110	132	167

Tabla 4.1 Documentos recolectados para el corpus

De la colección de documentos, 271 son etiquetados ambiguos, 75 inestables y 63 estables. En la tabla 4.2 podemos apreciar la cantidad de documentos y a que medio pertenecen, fue el corpus utilizado para el entrenamiento y con el que se hicieron las pruebas.

Periódico nacional	**La Razón**	**El Universal**	**La Jornada**
No. De documentos	45	46	47

Tabla 4.2 Documentos utilizados para entrenamiento y pruebas

4.2 Diccionario

Se inició con una lista de verbos, la cual fue separada por el experto en dos grupos: “azules” y “rojos”, las palabras azules son verbos característicos de la clase estable y las rojas de la clase inestable. Esta lista tenía 166 verbos, 71 azules y 95 rojos, con la que se trabajó buscando los sinónimos para así obtener un “Diccionario de Sinónimos”. Después de depurar la lista de verbos nos quedaron 850 verbos azules y 978 rojos, teniendo un total de **1828** verbos.

En la tabla 4.3 se muestra la lista inicial de verbos “azules”, que caracterizan a la clase Estable. Son 71 verbos en infinitivo.

Palabras Azules			
Levantar	Explorar	Dialogar	Conocer
Admirar	Confiar	Equilibrar	Cumplir
Considerar	Precisar	Consensuar	Estabilizar
Beneficiar	Enaltecer	Confirmar	Orientar
Proteger	Calmar	Persuadir	Honrar
Cuidar	Disculpar	Sostener	Realizar
Pronosticar	Respetar	Liberar	Decir
Equilibrar	Erigir	Legitimar	Clarificar
Conseguir	Enfrentar	Acceder	Puntualizar
Facilitar	Compartir	Remediar	Anticipar
Defender	Concluir	Estandarizar	Autentificar
Destapar	Conjuntar	Tranquilizar	Concursar
Perdonar	Pacificar	Formalizar	Sanar
Limpiar	Restituir	Hacer	Normalizar
Transparentar	Salvaguardar	Entusiasmar	Acrisolar
Construir	Dirigir	Obtener	Matizar
Contener	Disuadir	Defender	Justificar
Evaluar	Informar	Liderar	Acertar

Tabla 4.3 Palabras que se etiquetaron en el diccionario como azules

En la tabla 4.4 se muestran las palabras que llamamos “rojas” y que caracterizan a la clase Inestable. Son 95 verbos en infinitivo.

Palabras Rojas			
Abatir	Enloquecer	Juzgar	Rumorar
Abominar	Enturbiar	Lacerar	Sofocar
Abusar	Exagerar	Ladear	Sollozar
Accidentar	Fallar	Largar	Someter
Afectar	Fanatizar	Llover	Sospechar
Agredir	Ganar	Maldecir	Sugestionar
Asesinar	Gobernar	Maniatar	Tajar
Augurar	Golpear	Maquillar	Tambalear
Burlar	Henchir	Mentir	Tantear
Burocratizar	Hostigar	Moler	Temer
Calumniar	Humillar	Novatear	Tergiversar
Chismorrear	Ignorar	Obcecar	Tiranizar
Colapsar	Imitar	Obedecer	Tirar
Condenar	Impedir	Obsequiar	Tironear
Contagiar	Incumplir	Observar	Trozar
Corromper	Inestabilidad:	Opacar	Trucar
Criminalizar	Infartar	Perjudicar	Turbar
Delinquir	Inflar	Prescribir	Ubicar
Demoler	Influir	Proferir	Ultimar
Desprestigiar	Intoxicar	Purificar	Urdir
Desviar	Joder	Retocar	Usurpar
Distorsionar	Jorobar	Retrasar	Utilizar
Embutir	Jugar	Revocar	
Enfermar	Jurar	Robar	

Tabla 4.4 Palabras que se etiquetaron en el diccionario como rojas

Al buscar los sinónimos de estos 166 verbos, nos enfrentamos con un primer problema donde había verbos repetidos en la lista de azules y rojas, este problema fue resuelto con ayuda del experto en propaganda para definir un solo grupo o eliminar la palabra de ambos grupos. En la tabla 4.5 se muestran algunos verbos que tuvieron conflicto ya que se repetían en los dos grupos, el grupo nuevo es donde quedaron asignados o si fueron eliminados.

Palabra con conflicto	Grupo anterior	Grupo nuevo
PACIFICAR	Ambas	Se eliminó
FACILITAR	Ambas	Se eliminó
IGNORAR	Ambas	Roja
PRONOSTICAR	Ambas	Azul
BENEFICIAR	Ambas	Azul

Tabla 4.5 Ejemplo de algunos verbos con conflicto

Por mencionar un ejemplo de los problemas con palabras repetidas en ambas clases, fue el caso de “pronosticar” y “augurar” que en el diccionario los podemos encontrar como sinónimos, pero se usa pronosticar (azul) cuando se tiene alguna base científica de algo que va a suceder y augurar (roja) es cuando se dice sin algún sustento, este tipo de cosas no las puede definir la computadora sin ayuda del humano experto.

El esquema 4.1 se considera inestable, redactado por el experto en análisis de propaganda; usando el diccionario encontramos 6 palabras rojas que aparecen subrayadas y 2 azules que están resaltadas en negrita.

*Se comienza a **integrar** un expediente contra el director de dicha oficina: entre otras lindezas, se le investigará por: humillar al personal, hostigarlo; zaherir a quienes no le **pagan** sus constantes borracheras; maldecir a los miembros del sindicato que se resisten a sus burlas; se dice que tarde o temprano los chismorreos que acostumbra detonar a sus enemigos se volverán contra él mismo.*

Esquema 4.1 Ejemplo de párrafo inestable

El esquema 4.2 es estable de acuerdo con el experto en propaganda, usando el diccionario podemos encontrar 8 palabras azules resaltadas en negrita y 2 rojas que se muestran subrayadas.

***Perdonar** es una virtud propia de todas las religiones, **enseñanza** y remedio a la vez; empero, **levantar** sanciones injustas es un hecho admirable en estos tiempos donde se **considera** que el derecho es más importante que la moral. Por ello, ser **considerado** con el más débil, el menos educado, el menos **favorecido** socialmente hablando, no solamente **beneficia** al que es sensible y al que recibe dicha sensibilidad; también **protege** la impronta de nuestros hijos.*

Esquema 4.2 Ejemplo de párrafo estable

Es importante mencionar que el diccionario de verbos está en infinitivo por lo cual fue necesario usar un lematizador y así convertir las palabras o verbos conjugados en su forma infinitiva, para un mejor uso del diccionario.

4.3 Clasificación automática supervisada de un texto propagandístico de acuerdo a su contenido

Con la ayuda de un sistema desarrollado en java se tuvieron las opciones para poder elegir la forma de representar los atributos en el modelo vectorial: ponderado booleano, ponderado por frecuencia de término o ponderado TF-IDF; así como diferentes opciones para poder quitar las palabras vacías y la opción de poder usar un lematizador de texto [ver Apéndice A], de esta forma se realizaron experimentos para después obtener los atributos, eligiendo la técnica de ponderado por frecuencia.

Utilizando la herramienta de aprendizaje automático Weka[1], se eliminaron etiquetas o signos, antes de aplicar los algoritmos de clasificación.

En la tabla 4.8 se muestran los signos de puntuación y dígitos que se eliminaron en la colección de documentos.

Etiquetas eliminadas
()[]{}
,.;:"'?!
123456789

Tabla 4.8 Signos eliminados en los textos antes de usar el clasificador.

Se obtuvieron 1666 atributos del corpus con el que se aplicaron 7 clasificadores que se detallan en el capítulo 5.

1 http://www.cs.waikato.ac.nz/ml/weka/

En la tabla 4.9 se tiene la representación general de la matriz de confusión la posición de los True Positive (Verdadero Positivo), False Positive (Falso Positivo), False Negative (Falso Negativo), True Negative (Verdadero Negativo). Con los resultados de la matriz de confusión nos basamos para tener medidas de precisión como Exactitud, Recuerdo, Precisión y F-Measure. Las columnas representan la clase de predicción y las filas la clase real.

		Clase predictiva	
Clase real		Inestable	Estable
	Inestable	Verdadero positivo (VP)	Falso positivo (FP)
	Estable	Falso Negativo (FN)	Verdadero Negativo (VN)

Tabla 4.9 Matriz de confusión

Para realizar la parte de entrenamiento y pruebas usamos la misma colección de documentos, aplicando validación cruzada con 10 pliegues para las diferentes caracterizaciones con las herramientas que ofrece.

Capítulo 5

Resultados experimentales

En este capítulo se muestran los resultados con el diccionario de sinónimos. Por otra parte, uno de los objetivos de esta tesis es la implementación de un método supervisado para la clasificación de documentos propagandísticos, los resultados se muestran también en este capítulo.

5.1 Basada en diccionario

Para este experimento primero se tuvo que lematizar el corpus, ya que el diccionario solo usa verbos en infinitivo, usamos el Tre Tagger en español para procesar el texto del corpus y así obtener documentos con verbos en su tiempo infinitivo.

En la Tabla 5.1 se puede observar que después de ejecutar el algoritmo sin expansión por sinónimos se encontró una exactitud promedio de 55.79%, lo cual significa que fueron clasificadas correctamente 54 noticias estables y 23 inestables de un total de 138 noticias, mientras que el resto fueron clasificadas incorrectamente.

Tipo de documento	Número de documentos	Clase correcta. Sin sinónimos		Clase incorrecta. Sin sinónimos	
Azul	63	54	85.71%	9	14.28%
Rojo	75	23	30.66%	52	69.33%
Comportamiento Promedio	**138**	**77**	**55.79%**	**61**	**44.20%**

Tabla 5.1 Resultados de los experimentos usando clasificación por palabras sin expansión con sinónimos

Por otro lado, en la Tabla 5.2 se puede observar que cuando se usó el diccionario de palabras con los sinónimos, se obtuvo una exactitud promedio del 64.49%, clasificando correctamente 50 noticias estables y 39 inestables, de un total de 138 noticias. El resto de las noticias fueron clasificadas incorrectamente.

Tipo de documento	Número de documentos	Clase correcta. Con sinónimos		Clase incorrecta. Con sinónimos	
Azul	63	50	79.36%	13	20.63%
Rojo	75	39	52%	36	48%
Comportamiento Promedio	**138**	**89**	**64.49%**	**49**	**35.51%**

Tabla 5.2 Resultados de los experimento usando clasificación por la palabras con expansión con sinónimos

Discusión

El uso de un diccionario de palabras rojas (clase inestable) y azules (clase estable) resulta altamente efectivo, siempre y cuando su cobertura sea amplia. Es claro que este tipo de recursos léxicos representan el conocimiento real de un experto en el área y puede ser usado para la tarea específica de determinar la estabilidad de un texto. Sin embargo, su aplicación se encuentra supeditada a la magnitud y calidad del recurso lingüístico. Así, entre mejor y mayor sea el

diccionario, mejores serán los resultados obtenidos. La desventaja se encuentra en el costo necesario para construir este tipo de recursos léxicos.

5.2 Resultados de los algoritmos de clasificación

Los algoritmos bayesianos son de los que comúnmente tienen buenos resultados en el área de clasificación, también máquinas de soporte vectorial, vecinos más cercanos, entre otros, por lo que se utilizaron 7 algoritmos de clasificación. Para la evaluación se tomó en cuenta el tiempo de ejecución de cada algoritmo, la cantidad de documentos que fueron correctamente e incorrectamente clasificados y con la matriz de confusión se obtienen las medidas de rendimiento Exactitud, Precisión, Recuerdo y F-Measure. A continuación se muestran los resultados obtenidos.

Clasificador Naïve Bayes

En la tabla 5.2 se puede observar el desempeño del clasificador Naïve Bayes con un tiempo de ejecución de 0.19 segundos, el cual clasificó de manera correcta un 72.46% que son 100 documentos y el 27.54% que son 38 documentos de manera incorrecta, de un total de 138 documentos. Se usó la técnica de validación cruzada con 10 pliegues.

NAÏVE BAYES		
Tiempo de ejecución	Correctamente clasificados	Incorrectamente clasificados
0.19 segundos	100 documentos 72.463 %	38 documentos 27.536 %

Tabla 5.2 Tiempo de ejecución y exactitud de Naïve Bayes

En la tabla 5.3 se tienen los resultados de la matriz de confusión, 59 documentos inestables de 75 fueron clasificados en la clase correcta y 41 documentos estables de 63 son los clasificados correctamente.

	Inestable	Estable
Inestable	59	16
Estable	22	41

Tabla 5.3 Matriz de confusión del clasificador Naïve Bayes

Con la matriz de confusión medimos la Precisión, el Recuerdo y el F-Measure del algoritmo Naïve Bayes, obtenemos los resultados mostrados en la tabla 5.4

	Precisión	Recuerdo	F-Measure
Inestable	0.728	0.787	0.756
Estable	0.719	0.651	0.683
Promedio ponderado	**0.724**	**0.725**	**0.723**

Tabla 5.4 Los resultados de Precisión, Recuerdo, F-Measure

Naïve Bayes Multinomial

El clasificador Naïve Bayes Multinomial considera el número de apariciones de cada término para evaluar la contribución de su probabilidad condicional dada la clase del documento. En la tabla 5.5 podemos apreciar los resultados con un 84.057% de documentos clasificados en la clase correcta y 15.942% clasificados en la clase incorrecta con un tiempo de ejecución de 0.05 segundos.

NAÏVE BAYES MULTINOMIAL		
Tiempo de ejecución	Correctamente clasificados	Incorrectamente clasificados
0.05 segundos	116 documentos 84.058 %	22 documentos 15.942 %

Tabla 5.5 Tiempo de ejecución y exactitud de Naïve Bayes Multinomial

En la tabla 5.6 podemos apreciar los resultados en la matriz de confusión, de los 116 clasificados de manera correcta 59, son inestables y 57 estables, de los 22 clasificados en la clase incorrecta 16 se clasificaron como estables y 6 inestables.

	Inestable	**Estable**
Inestable	59	16
Estable	6	57

Tabla 5.6 Matriz de confusión con el clasificador Naïve Bayes Multinomial

Con la matriz de confusión se tienen las medidas de rendimiento de este clasificador, donde se representan los resultados en la tabla 5.7 con los resultados de Precisión, Recuerdo y F-Measure.

	Precisión	**Recuerdo**	**F-Measure**
Inestable	0.908	0.787	0.843
Estable	0.781	0.905	0.838
Promedio ponderado	**0.850**	**0.841**	**0.841**

Tabla 5.7 Resultados, Precisión, Recuerdo, F-Measure

Máquinas de soporte vectorial

El algoritmo de clasificación Máquinas de soporte vectorial (Support Vector Machines, SVMs) lo podemos encontrar en Weka en la clasificación de los lineales como SMO.

Los resultados de este algoritmo de clasificación los podemos apreciar en la tabla 5.8, donde el 84.058% fueron clasificados de manera correcta y el 15.942% en la clase incorrecta.

MÁQUINAS DE SOPORTE VECTORIAL		
Tiempo de ejecución	Correctamente clasificados	Incorrectamente clasificados
0.19 segundos	116 documentos 84.058 %	22 documentos 15.942 %

Tabla 5.8 Tiempo de ejecución y exactitud

En la matriz de confusión que se muestra en la tabla 5.9 podemos ver que 64 documentos de los 116 correctamente clasificados son inestables y 52 estables. De los 22 clasificados en la clase incorrecta 11 fueron clasificados en la clase estable y 11 en la inestable.

	Inestable	**Estable**
Inestable	64	11
Estable	11	52

Tabla 5.9 Matriz de confusión con el clasificador Máquinas de Soporte Vectorial

Con la matriz de confusión se tienen los resultados de Precisión, Recuerdo y F-Measure del algoritmo Máquinas de Soporte Vectorial, los cuales se muestran en la tabla 5.10

	Precisión	Recuerdo	F-Measure
Inestable	0.853	0.853	0.853
Estable	0.825	0.825	0.825
Promedio ponderado	**0.841**	**0.841**	**0.841**

Tabla 5.10 Resultados de Precisión, Recuerdo, F-Measure

Algoritmo Vecinos más Cercanos

El algoritmo Vecinos más cercanos (k-nearest neighbours) KNN lo encontramos como IBK en Weka dentro de la clasificación lazy.

Al aplicar este algoritmo de clasificador se puede notar en la tabla 5.11 un tiempo de ejecución de 0.02 segundos, que es muy eficiente en costo de tiempo, con un 61.594% de documentos clasificados en la clase correcta y un 38.405% clasificados en la clase incorrecta.

VECINOS MÁS CERCANOS		
Tiempo de ejecución	Correctamente clasificados	Incorrectamente clasificados
0.02 segundos	85 documentos 61.594 %	53 documentos 38.405 %

Tabla 5.11 Exactitud y tiempo de ejecución de Vecinos más cercanos

La matriz de confusión nos muestra en la tabla 5.12 que de los 85 documentos clasificados de manera correcta, 69 fueron inestables y 16 estables. De los 53 clasificados de manera incorrecta, 47 fueron clasificados en la clase inestable y 6 en la clase estable.

	Inestable	Estable
Inestable	**69**	**6**
Estable	**47**	**16**

Tabla 5.12 Matriz de confusión con el clasificador Vecinos más cercanos

Con la matriz de confusión tenemos las medidas de rendimiento del algoritmo de clasificación Vecinos más cercanos, Precisión, el Recuerdo y F-Measure. Los resultados se muestran en la tabla 5.13

	Precisión	Recuerdo	F-Measure
Inestable	0.595	0.920	0.723
Estable	0.727	0.254	0.376
Promedio ponderado	**0.655**	**0.616**	**0.565**

Tabla 5.13 Precisión, Recuerdo, F-Measure. Vecinos más cercanos

LogitBoost

El algoritmo LogitBoost se basa en un modelo regresión logística. Los resultados obtenidos en tiempo de ejecución fueron de 2.93 segundos, un 60.144% fueron correctamente clasificados y un 39.855% incorrectamente clasificados, como se muestra en la tabla 5.14

LOGITBOOST		
Tiempo de ejecución	Correctamente clasificados	Incorrectamente clasificados
2.93 segundos	83 documentos 60.1449 %	55 documentos 39.8551 %

Tabla 5.14 Exactitud y tiempo de ejecución de LogitBoost

La matriz de confusión se nos muestra en la tabla 5.15. De los 83 documentos clasificados de manera correcta, 43 fueron inestables y 40 estables; de los 55 clasificados de manera incorrecta 23 fueron clasificados en la clase inestable y 32 en la clase estable.

	Inestable	**Estable**
Inestable	43	32
Estable	23	40

Tabla 5.15 Matriz de confusión con el clasificador LogitBoost

Con la matriz de confusión obtenemos las medidas de rendimiento del algoritmo de clasificación LogitBoost, Precisión, el Recuerdo y F-Measure. Los resultados se muestran en la tabla 5.16

	Precisión	**Recuerdo**	**F-Measure**
Inestable	0.652	0.573	0.610
Estable	0.556	0.635	0.593
Promedio ponderado	**0.608**	**0.616**	**0.602**

Tabla 5.16 Resultados Precisión, Recuerdo, F-Measure de LogitBoost

Decision Table Majority (DMT)

Es un algoritmo de decisión simple basado en mayorías. Su representación tiene dos componentes, un *esquema* formado por características que se incluyen en la tabla que son los atributos y un *cuerpo* definido por las características del esquema que son las reglas.

En la tabla 5.17 podemos apreciar el tiempo de ejecución del algoritmo Decision Table Majority de 2.93 segundos así como un 68.115% de documentos correctamente clasificados y un 31.884% incorrectamente clasificados.

DECISION TABLE MAJORITY		
Tiempo de ejecución	Correctamente clasificados	Incorrectamente clasificados
2.93 segundos	94 documentos 68.1159 %	44 documentos 31.8841 %

Tabla 5.17 Exactitud y tiempo de ejecución de Decision Table Majority

La matriz de confusión se nos muestra en la tabla 5.18. De los 94 documentos clasificados de manera correcta, 60 fueron inestables y 34 estables; de los 44 clasificados de manera incorrecta 29 fueron clasificados en la clase inestable y 34 en la clase estable.

	Inestable	**Estable**
Clase Inestable	60	15
Clase Estable	29	34

Tabla 5.18 Matriz de confusión. Decision Table Majority

Con la matriz de confusión tenemos las siguientes medidas de rendimiento del algoritmo de clasificación Decision Table Majority, como la Precisión, el Recuerdo y F-Measure. Los resultados se muestran en la tabla 5.19

	Precisión	**Recuerdo**	**F-Measure**
Inestable	0.674	0.800	0.732
Estable	0.694	0.540	0.607
Promedio ponderado	**0.683**	**0.681**	**0.675**

Tabla 5.19 Precisión, Recuerdo, F-Measure. Decision Table Majority

Algoritmo Árboles de Decisión C4.5

C4.5 es un algoritmo usado para generar un árbol de decisión usado para clasificación. En Weka lo encontramos en la clasificación de Trees como J48; es una implementación open source en lenguaje de programación Java del algoritmo C4.5 en la herramienta Weka de minería de datos.

En la tabla 5.20 podemos apreciar el tiempo de ejecución del algoritmo Árboles de Decisión C4.5 de 0.67 segundos, así como un 68.115% de documentos correctamente clasificados y un 31.884% incorrectamente clasificados.

C4.5		
Tiempo de ejecución	Correctamente clasificados	Incorrectamente clasificados
0.67 segundos	94 documentos 68.1159 %	44 documentos 31.8841 %

Tabla 5.20 Exactitud y tiempo de ejecución de C4.5

La matriz de confusión se nos muestra en la tabla 5.21, de los 94 documentos clasificados de manera correcta 55 fueron inestables y 39 estables, de los 44 clasificados de manera incorrecta 24 fueron clasificados en la clase inestable y 20 en la clase estable.

	Inestable	**Estable**
Clase Inestable	55	20
Clase Estable	24	39

Tabla 5.21 Matriz de confusión con el clasificador C4.5

Con la matriz de confusión obtenemos las medidas de rendimiento del algoritmo de clasificación C4.5, Precisión, el Recuerdo y F-Measure. Los resultados se muestran en la tabla 5.22

	Precisión	Recuerdo	F-Measure
Inestable	0.696	0.733	0.714
Estable	0.661	0.619	0.639
Promedio ponderado	**0.680**	**0.681**	**0.680**

Tabla 5.22 Los resultados Precisión, Recuerdo, F-Measure

En la tabla 5.23 podemos ver una comparativa de los 7 algoritmos, en función del tiempo de ejecución, los archivos que fueron clasificados en la clase correcta y los que fueron clasificados en la incorrecta, también se muestran el desempeño de los clasificadores con las medidas de evaluación en la Precisión, Recuerdo y F-Measure. Como se puede ver en la tabla los mejores resultados son Naïve Bayes Multinomial con un tiempo de ejecución de 0.05 segundos, un 84.06 % de documentos correctamente clasificados, 0.850 de Precisión, 0.841 de Recuerdo y .841 de F-Measure. Máquinas de Soporte Vectorial con un tiempo de ejecución mayor a Naïve Bayes Multinomial, pero con resultados muy similares con 84.06% de documentos correctamente clasificados y 0.841 de Precisión, Recuerdo y F-Measure.

Clasificador	Tiempo de ejecución	Correctamente clasificados	Incorrectamente clasificados	Precisión Promedio	Recuerdo Promedio	F-Measure Promedio
Naïve Bayes	0.19 segundos	72.46%	27.54%	0.724	0.725	0.723
Naïve Bayes Multinomial	0.05 segundos	84.06%	15.94%	0.850	0.841	0.841
Máquinas de Soporte	0.19 segundos	84.06%	15.94%	0.841	0.841	0.841
Vecinos más cercanos	0.02 segundos	61.59%	38.41%	0.655	0.616	0.565
LogitBoost	2.93 segundos	60.14%	39.86%	0.608	0.601	0.602
Decision Table Majority	4 segundos	68.12%	31.88%	0.683	0.681	0.675
Árboles de decisión c4.5	0.67 segundos	68.12%	31.88%	0.680	0.581	0.680

Tabla 5.23 Resultados comparativos evaluando 7 clasificadores

Al comparar el tiempo de ejecución que realizó cada uno de los algoritmos en la colección de documentos, podemos apreciar en la figura 5.1 que Naïve Bayes Multinomial tiene el tiempo más rápido con 0.05 segundos y el algoritmo Decision Table Majority fue el que consumió más tiempo con 4 segundos.

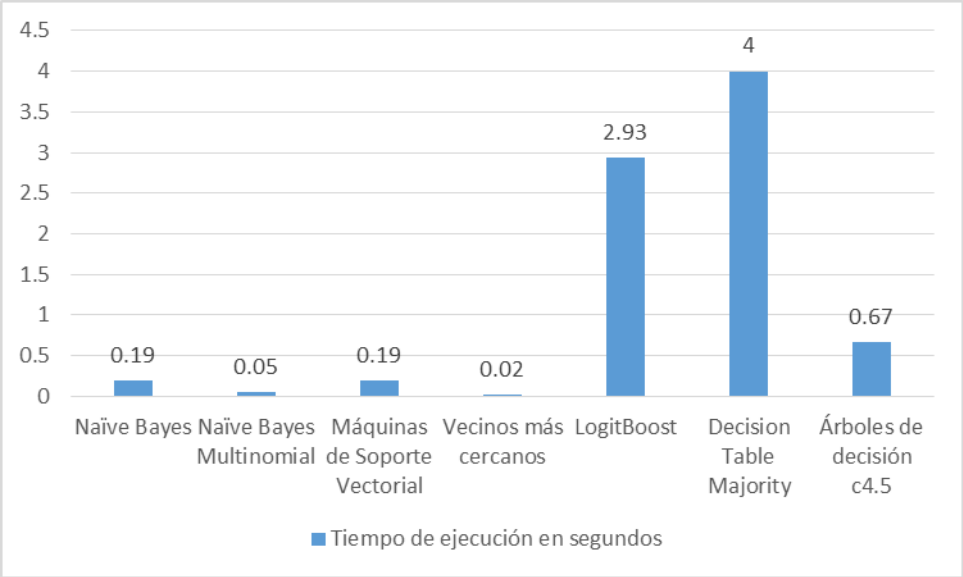

Figura 5.1 Tiempo de ejecución de los algoritmos de clasificación

La colección fue de 138 documentos, 75 inestables y 63 estables. En la figura 5.2 podemos apreciar el porcentaje de documentos correctamente clasificados por cada algoritmo donde Naïve Bayes Multinomial y Máquinas de Soporte Vectorial obtuvieron el mismo porcentaje de 84.04.

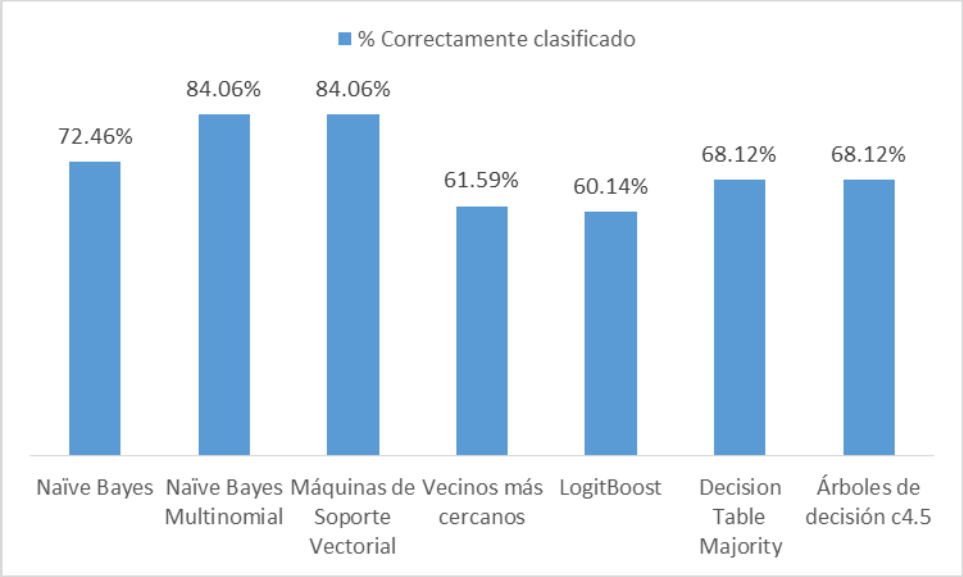

Figura 5.2 Documentos correctamente clasificados con los diferentes algoritmos de clasificación

En la figura 5.3 podemos apreciar el desempeño de los algoritmos de clasificación con los resultados de F-Measure, donde Naïve Bayes Multinomial y Máquinas de Soporte Vectorial tienen un valor de 0.841

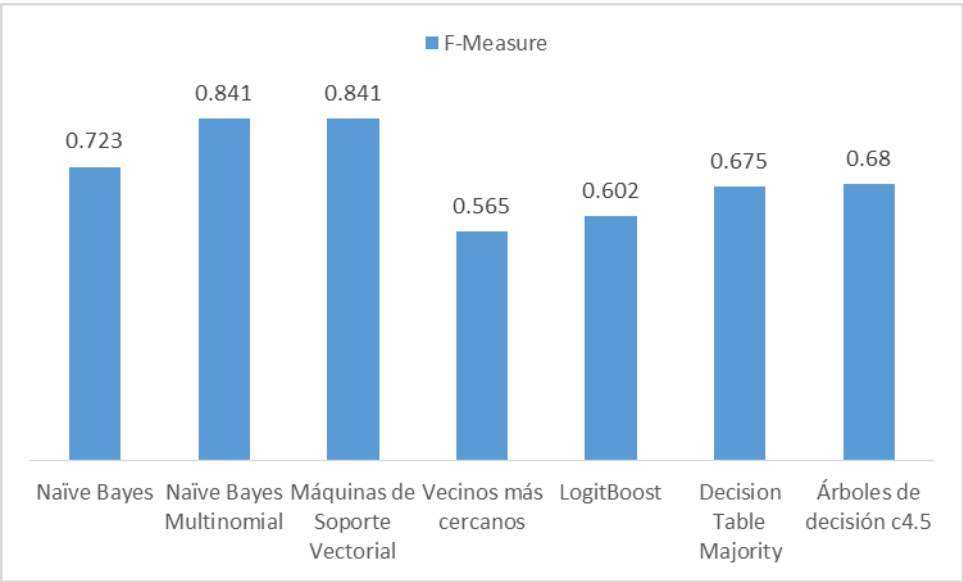

Figura 5.3 Medida de rendimiento F-Measure que se tuvo con los algoritmos de clasificación.

Capítulo 6

Conclusiones y trabajo a futuro

En el capítulo 1 se plantearon las siguientes preguntas de investigación:

1. ¿Es posible determinar la polaridad de un texto propagandístico usando un diccionario de verbos que caractericen las clases definidas?
2. ¿Es posible determinar la polaridad de un texto propagandístico usando la frecuencia de palabras que aparecen en los textos de entrenamiento?

Con respecto a la primera pregunta de investigación hemos llevado a cabo experimentos tomando en cuenta la utilización única de un diccionario de palabras positivas y negativas. El diccionario consistió de un conjunto estático de verbos. Las evaluaciones indican que dado un texto para el cuál se desconoce si pertenece a la clase estable o inestable, aplicando únicamente las palabras del diccionario, podemos determinar con un 64.49% de exactitud a que clase pertenece. Los resultados no son tan altos como esperábamos, sin embargo contestando esta pregunta de investigación podemos decir que SI es posible determinar la polaridad de un texto propagandístico usando un diccionario de verbos, enriqueciendo y refinando este diccionario para así incrementar el porcentaje de exactitud.

Acerca de la segunda pregunta de investigación, recolectamos una serie de documentos propagandísticos, posteriormente fueron etiquetados manualmente para identificar su clase, documento estable o inestable; esto se conoce como un

corpus de entrenamiento. Se utilizaron algoritmos de clasificación con una técnica de frecuencia de palabras donde el mejor resultado fue una exactitud del 84.06%, es decir, que ese porcentaje de documentos fueron clasificados correctamente. Para responder la pregunta decimos que SI es posible determinar la polaridad de un texto propagandístico usando la frecuencia de palabras que aparecen en los textos de entrenamiento y puede mejorar ampliando el corpus de entrenamiento.

Después de comparar con 7 algoritmos de clasificación se observa que el clasificador de Naïve Bayes Multinomial obtiene el mejor comportamiento, clasificando correctamente el 84.06% de los documentos, con una Precisión de 0.850 un Recuerdo y F-Measure de 0.841. El tiempo de ejecución de este algoritmo fue de 0.05 segundos.

El algoritmo de Máquinas de Soporte Vectorial también clasificó el 84.06% de los documentos en la clase correcta, y tuvo una Precisión, Recuerdo y F-Measure de 0.841. Estos resultados son muy similares a los obtenidos por el algoritmo de Naïve Bayes Multinomial pero con tiempo de ejecución de 0.19 segundos.

Bajo este análisis, podemos decir que el mejor algoritmo de clasificación en este experimento de documentos propagandísticos es Naïve Bayes Multinomial.

Trabajos a futuro

1. El 70% de la colección de documentos clasificados por el experto en propaganda resultaron ambiguos, por lo cual se recomienda ampliar el corpus.

2. En este proyecto se usaron solamente 2 clases (estable e inestable), sin embargo, es recomendable definir un porcentaje del grado de estabilidad o inestabilidad que puede tener un documento.

3. Empleando técnicas de análisis de sentimientos y de opinión es posible tener conclusiones de un documento, desde el punto de vista del psicólogo,

politólogo, sociólogo, entre otros expertos, esto será de gran ayuda al analista en textos propagandísticos.

Referencias

[1] R. Baeza and B. Ribeiro. Modern Information Retrieval. Addison Wesley, 1999.

[2] R. Feldman and J. Sanger, The Text Mining Handbook. Cambridge University Press, 2007.

[3] J. Cowie and W. Lehnert. Information extraction. Communications of the ACM, 39(1):80-91, 1996.

[4] L. Eikvil, Information extraction from world wide web – A survey. Technical Report, Norwegian Computing Center, 1999.

[5] W. Cohen and Y. Singer. Context-sensitive learning methods for text categorization. ACM Transactions on Information Systems

[6] Dark marketing : propaganda oscura, contradicción e irracionalidad en la guerra política contemporánea. Mauricio Saldaña Rodríguez, Jorge David Cortés Moreno. Dirección de Fomento Editorial. BUAP 2009

[8] Lourdes Salgado Martín. Marketing Político: Arte y Ciencia de la persuación en democracia. Paidos. España 2002

[10] Edmundo González Llaca. Teoría y práctica de la propaganda. Editorial Grijalbo 1981.

[11] http://maestraencomunicacionpolitica.blogspot.com.ar/2009/12/propaganda-politicadefinicion.html

[12] Cleary, Linda M. 1993. A profile of Carlos: Strengths of a Nonstandard Dialect Writer. In Linda M. Cleary and Michael D. Linn, eds., Linguistics for Teachers. NY: McGraw-Hill.

[13] J. Turno. Information extraction, multilinguality and portability. Revista Iberoamericana de Inteligencia Artificial. 22:57-78, 2003.

[14] R. Feldman and J. Sanger, The Text Mining Handbook. Cambridge University Press, 2007.

[19] Evaluating Polarity for Verbal Phraseological Units. Belém Priego Sánchez, David Pinto y Salah Mejri . Université Paris y Benemérita Universidad Autónoma de Puebla. 191-200 Micai 2014.

[20] Jain, Harshit, Mogadala, Aditya, Varma, Vasudeva. Feature Analysis and Polarity Classification of Expressions from Twitter and SMS Data, 2013, pp, 525-529.

[21] Sylvain Kahane. Polarized Grammars. In proceddings Annual Meeting of the Association of Computational Linguistics 2006. Nanterre France. pp, 137-412.

[22] Nobuhiro Kaji, Masaru Kitsuregawa. Automatic Construction Of Polarity-Tagged Corpus From HTML Documents. Proceddings of Annual Meeting of the Association of Computational Linguistics, 2006, pp, 452-459.

[23] Juan de Dios Álvarez Romero. Clasificación Automática de Textos usando Reducción de Clases basada en Prototipos. INAOE 2009.

[24] Abu-Jbara, Jefferson Ezra, Dragomir Radev. Purpose and Polarity of Citation: Towards NLP-based Bibliometrics. Proceedings of NAACL-HLT. Pages 596–606, Atlanta, Georgia. June 2013.

[25] Ulli Waltinger. A Lexical Resource for German Sentiment Analysis. Text Technology, Bielefeld University , Germany 2010.

[26] Dragut, Wan, Yu, Sistla, Meng. Polarity Consistency Checking for Sentiment Dictionaries. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 997–1005, Jeju, Republic of Korea, July 2012.

[27] Nadia Patricia Araujo Arredondo. Método Semisupervisado para la Clasificación Automática de Textos de Opinión. INAOE Febrero 2010.

[28] Rosa María Coyotl Morales. Clasificación Automática de Textos considerando el Estilo de Redacción. INAOE 2007.

Anexo A. Herramientas usadas

El presente apéndice muestra una herramienta desarrollada para el pre-procesamiento de texto y para las pruebas hechas con el diccionario. La cual se realizó con java 7.

A continuación se muestran algunas pantallas del sistema realizado.
En la figura A.1 se muestra una carpeta que se ha cargado a la cual se le puede aplicar una lista de opciones como quitar acentos, quitar signos de puntuación y aplicar el lematizador Tree Tagger. Tenemos la opción de limpiar lista y volver a elegir.

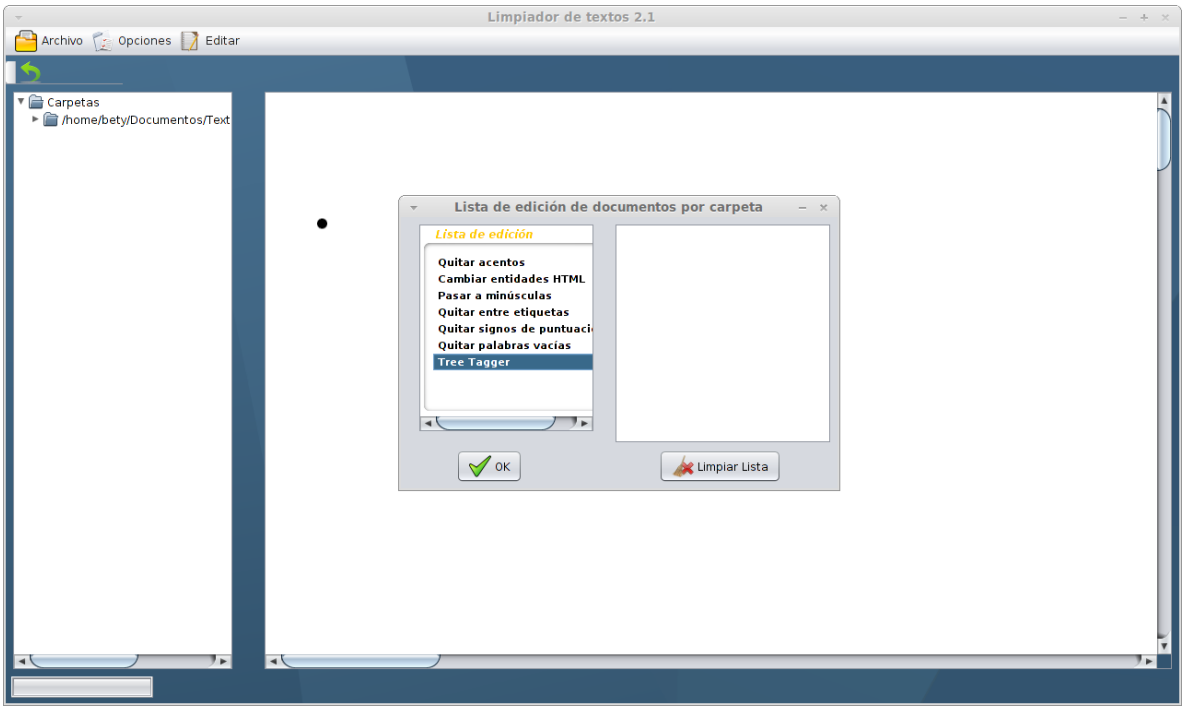

Figura A.1

En la figura A.2 se muestra una pantalla después de aplicar la lista de edición, donde el resultado es una matriz con las palabras encontradas en los documentos.

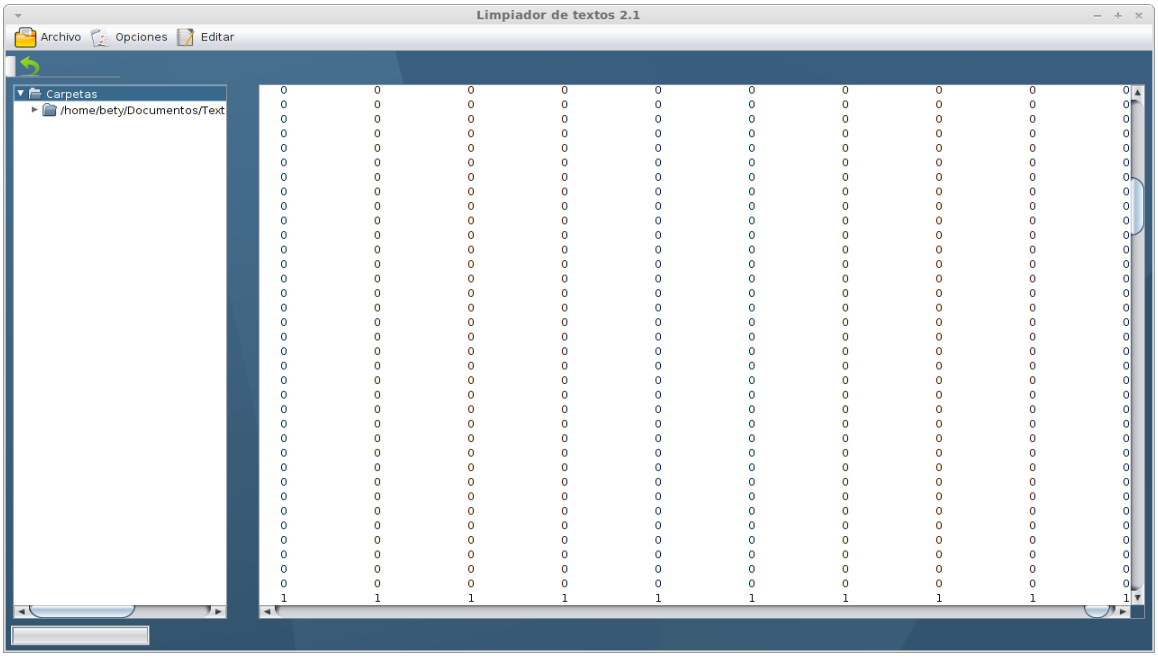

Figura A.2

En la figura A3 se muestran algunas palabras que son resultado del pre-procesamiento aplicado a los documentos de la carpeta que se cargo.

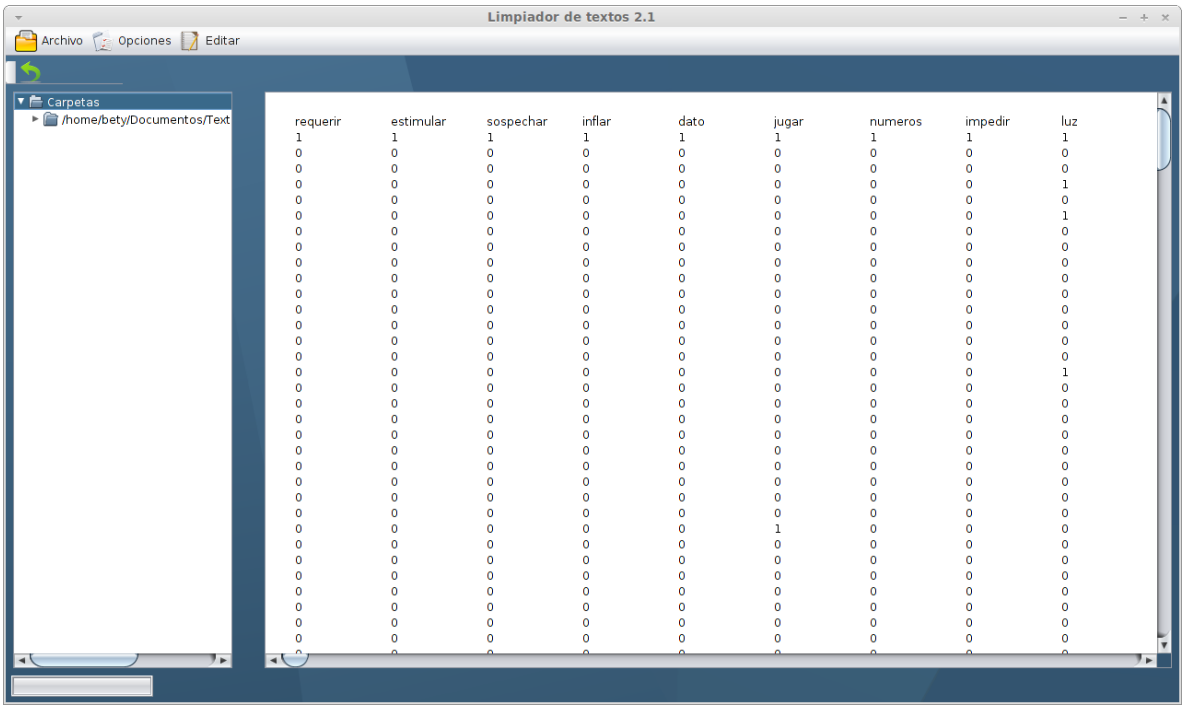

Figura A.3

En el sistema realizado para el apoyo al pre-procesamiento también se realizó para las pruebas hechas con el diccionario de palabras rojas y azules, donde como resultado nos dice cuantas palabras azules y cuantas rojas tiene cada documento. En el diccionario tenemos los verbos y sus sinónimos. En la figura A.4 podemos apreciar el nombre del documento, cuantos verbos azules y cuantos rojos aparecen el documento, del lado izquierdo aparecen los verbos que son de primer nivel es decir los que se tienen en una primera lista sin sinónimos y los que aparecen a la izquierda son los verbos que están como sinónimos de la primera lista.

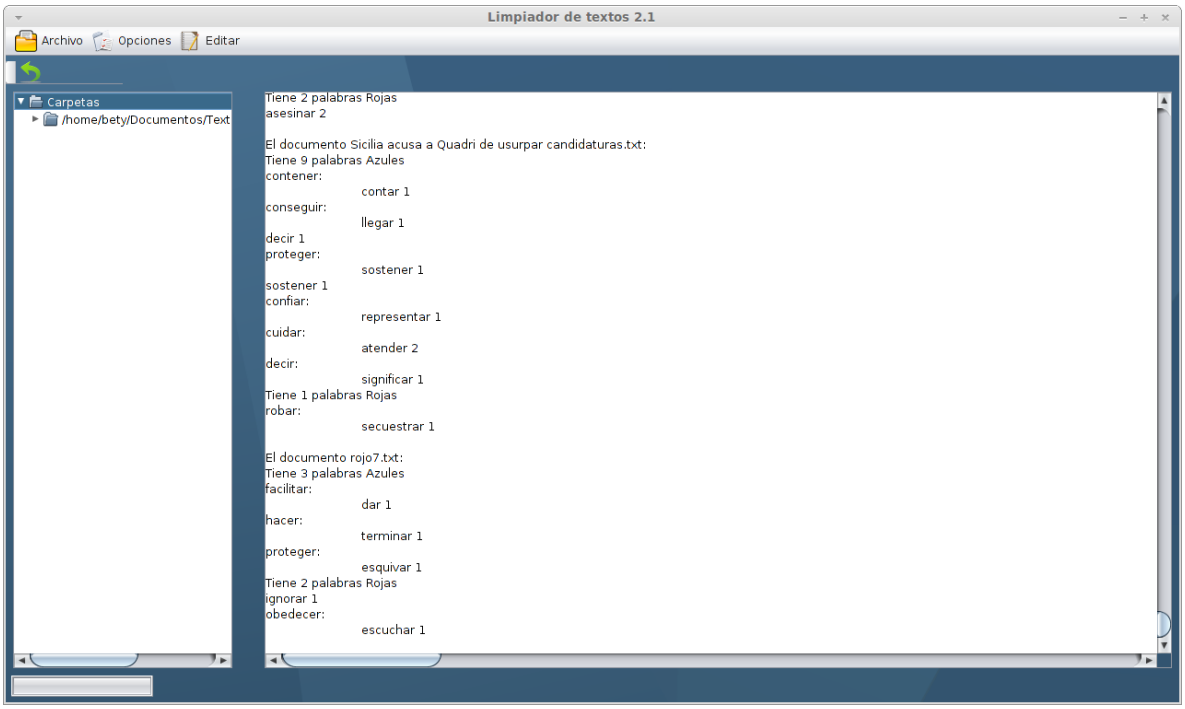

Figura A.4

En el sistema se tiene la opción de guardar con la extensión .csv y con .arff que es la que usa Weka, en la pantalla A.5 se observa que se guardó con la extensión arff.

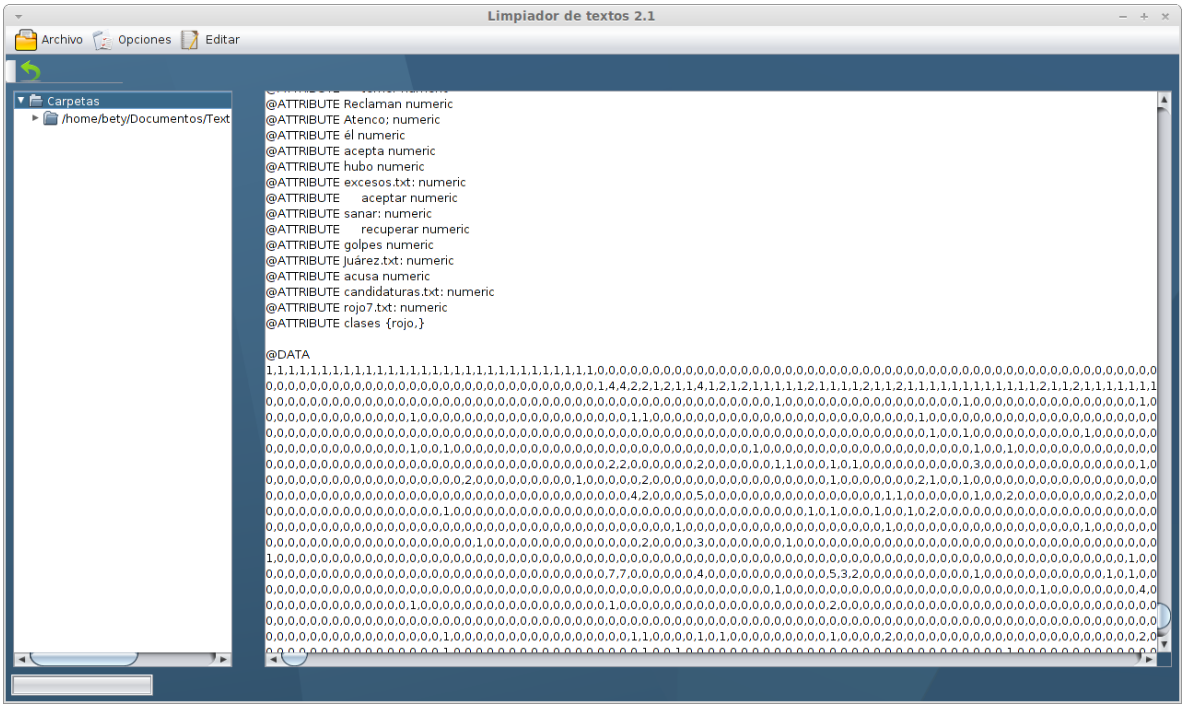

Figura A.5

Tenemos también la opción de guardar por ocurrencia de palabra en el documento, se elige por ponderado booleano, por frecuencia o TFxIDF. En la figura A.6 se muestra el resultado de guardar por frecuencia de palabra.

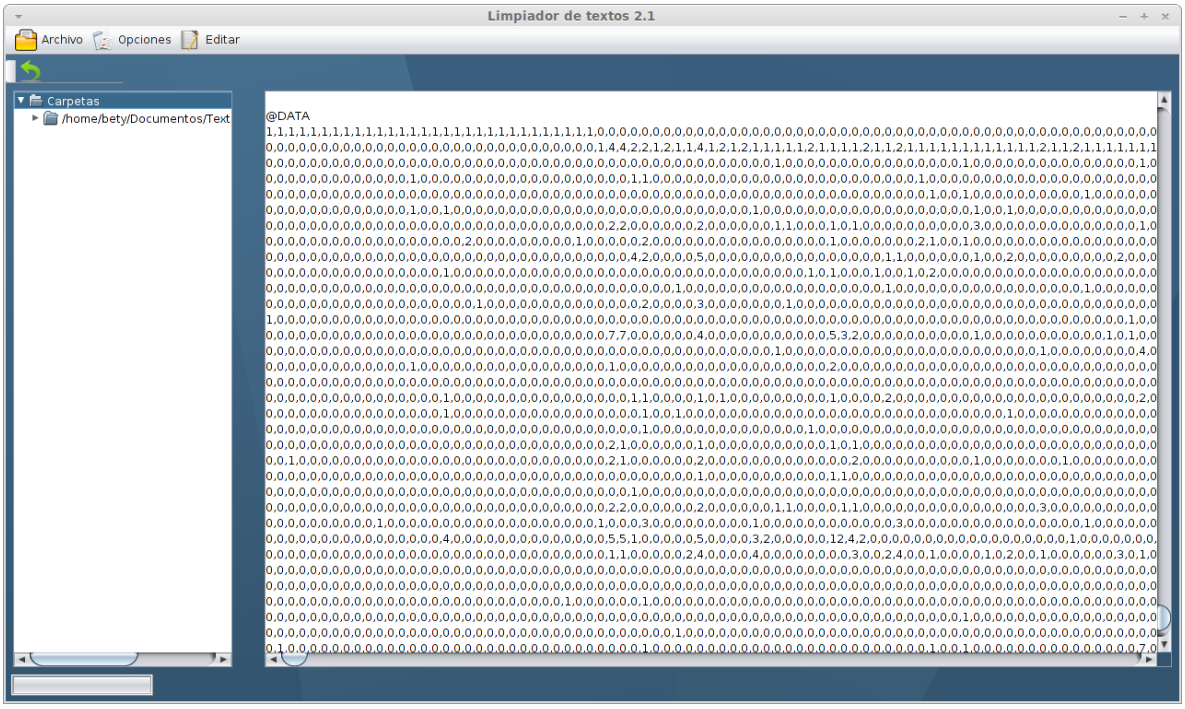

Figura A.6

En la figura A.8 se muestran algunos de los atributos; resultado de guardar como arff.

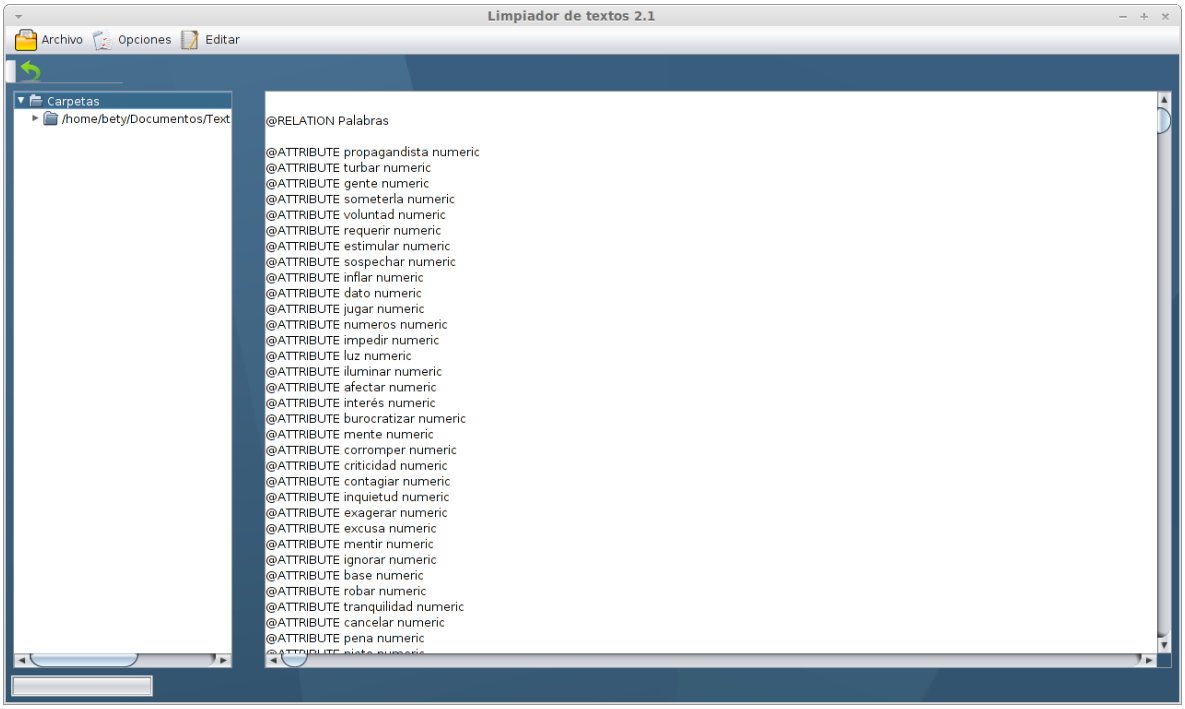

Figura A.8