PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

FACULTAD DE INGENIERÍA

ESCUELA DE INGENIERÍA INFORMÁTICA

MINERÍA DE OPINIÓN Y ANÁLISIS DE SENTIMIENTOS

FELIPE IGNACIO OLIVA VALDEBENITO

INFORME FINAL DE PROYECTO

PARA OPTAR AL TÍTULO PROFESIONAL DE

INGENIERO CIVIL EN INFORMÁTICA

NOVIEMBRE 2014

Dedicatoria

Dedico el presente trabajo de título a mis padres,
Irma y José, quienes me han entregado su apoyo incondicional
en todas las etapas de mi enseñanza y a lo largo de mi vida.

A mis hermanos, Francisco y Matías,
quienes con su alegría y optimismo siempre logran
mostrarme que hay un lado bueno de las cosas.

Al amor de mi vida, Jennifer,
quien me apoya y me motiva día a día a ser cada vez mejor,
por su comprensión, consejos y por sobre todo,
el amor que me ha brindado desde el primer día.

A la familia de Jennifer,
quienes me han acogido como uno más
y apoyado siempre en todo.

A mi profesor guía Rodrigo,
ya que todo lo aprendido y logrado en el tema
es gracias a él, además de permitirme colaborar
en el trabajo para WoPATeC 2014.

Finalmente a mis amigos y compañeros de carrera,
con quienes compartí no solo en las aulas de clases,
sino que también fuera de ellas, en donde descubrí personas excelentes
y de gran corazón, que siempre han estado dispuestos a ayudarme
y apoyarme en todo.

Índice

Resumen

Hoy en día se encuentra disponible una gran cantidad de información a través de distintos medios electrónicos, bibliotecas digitales, colecciones de documentos e Internet. La necesidad de acceder a esta información para su extracción y análisis, así como además poder identificar el comportamiento y el pensamiento de las personas ha sido una importante área que se encuentra joven y en pleno apogeo. Por ello diversas técnicas para manipularla se han ido planteando, dentro de la que se encuentra la clasificación de texto. Sin embargo, el crecimiento constante de la información convierte a esta tarea algo tedioso de realizar, sin dejar de lado lo costoso que resulta, por lo que ésta clasificación busca ser de manera automática. El presente documento abarcara técnicas y enfoques para poder clasificar las opiniones y poder determinar la polaridad de éstas, para ello se plantean distintos escenarios para observar el comportamiento según la representación y clasificador que se utilice.

Lista de Figuras

Lista de Tablas

1 Introducción

Desde el comienzo de los tiempos los humanos han buscado la forma de comunicarse, característica que hace la diferencia con los demás seres vivos al poder pensar de forma racional y poder comunicarse con otros, de aquí nace la escritura. Las personas han plasmado conocimientos, sentimientos y emociones para crear documentos, poesía e incluso verdaderas obras maestras, pero todas ellas, indiferente lo que expresen, contienen un denominador común, información.

Con el paso del tiempo y la incorporación de la tecnología, es que todo el conjunto de documentos han ido creando una gran biblioteca de información digital. Para que estos documentos digitales sean de fácil acceso se han tenido que organizar de tal manera que se permita su recuperación y análisis por medios automáticos. Sin embargo, el problema es complejo y es necesaria la continua investigación en la búsqueda de métodos y representaciones apropiadas. Es por ello que se han creado diversas líneas de investigación para el tratamiento automático de textos, entre las que se encuentran: la recuperación de información, la extracción de información, la búsqueda de respuestas y la clasificación de textos entre otras.

Mucho antes de la aparición de la Internet de modo masivo y abierta al público en general, era común el preguntar a conocidos y amigos sobre la recomendación de un determinado producto, o consultar revistas exclusivas que trataban temas de calidad y comparación de productos para determinar que comprar. Pero, con el surgimiento de Internet ahora existen ciertos lugares donde encontrar este tipo de información, opiniones y recomendaciones, las cuales son dadas por personas comunes de acuerdo a sus experiencias en conformodidad a un producto o bien personas especializadas en temas como por ejemplo críticas de cines y espectáculos.

Blogs, microblogs y redes sociales en general se han convertido hoy en día en un punto de encuentro, donde las personas plasman opiniones, comentarios, sentimientos, anhelos y frustraciones, con esto, la pregunta "¿Qué es lo que las personas piensan?" se puede llegar a responder mediante la extracción, clasificación y análisis de lo que la gente escribe en el mundo digital, por ello empresas han apuntado a esta investigación como proceso clave en la toma de decisiones, el poder saber lo que la gente opina y "siente" respecto a un determinado producto o campaña es indispensable al momento de actuar.

Dentro de la clasificación de opiniones se encuentran tres subtareas principales. Una de estas subtareas consiste en determinar si un texto dado contiene información objetiva o subjetiva. Otra es determinar la orientación de una opinión, es decir, si el texto dado expresa una opinión a favor o en contra (positiva o negativa, respectivamente). Y la otra subtarea es determinar la fuerza o grado de la opinión. En particular, el presente trabajo se centra únicamente en determinar si un texto dado expresa una opinión a favor o en contra.

2 Objetivos

A continuación se presentan los objetivos generales y específicos que se pretenden lograr.

2.1 Objetivo general

Analizar el sentido de los mensajes dentro de un contexto permite clasificar mensajes específicos dentro de ese mismo contexto.

2.2 Objetivos específicos

Analizar documentos de investigaciones previas para generar conocimiento base para el desarrollo de la problemática.

Identificar los elementos que dan origen a la teoría del Análisis de Sentimientos y la Clasificación Automática.

Seleccionar representaciones y clasificadores según la base teórica para desarrollar la problemática.

Plantear escenarios experimentales para comprobar el rendimiento de los clasificadores automáticos.

Determinar la o las representaciones y el o los clasificadores automáticos que obtengan los mejores resultados para el Análisis de Sentimientos.

3 Problemática

Dentro del comercio, quien ofrece un producto o servicio, le gustaría saber si éste cubre las necesidades de la mayoría de los consumidores, o en su caso, tener la retroalimentación adecuada para mejorar dicho producto suscitando en consecuencia un mayor consumo. De igual forma, un personaje público necesita recabar información de su imagen. Para ello es necesario saber qué opina la gente para poder mejorar o cambiar ciertas conductas. En la actualidad gracias a Internet, se puede acceder a opiniones escritas por innumerables personas acerca de diferentes temas, productos, eventos, personas, etc. Estas opiniones se pueden encontrar en correos, editoriales o en sitios especializados para recabar opiniones, entre otros. Un primer paso en el análisis de estas opiniones es determinar su polaridad, es decir, separar aquellas opiniones que expresan algo a favor de las opiniones que expresan algo en contra. Debido al elevado número de documentos disponibles, la tarea de determinar manualmente qué opiniones expresan algo a favor y qué opiniones expresan algo en contra (polaridad), se convierte en una tarea muy costosa, por lo que es necesario el uso de métodos automáticos.

En el presente documento se mostrarán los conceptos básicos que involucran la clasificación automática de textos, pasando por los distintos métodos utilizados en la actualidad y encontrar una aproximación para poder "analizar los sentimientos" expuestos en las redes sociales, específicamente en Twitter, mediante la polaridad de las opiniones vertidas.

3.1 ¿Qué es el Sentimiento?

Uno de los desafíos del Análisis de Sentimientos es la definición de los objetos de estudio de las opiniones y la subjetividad. Originalmente, la subjetividad fue definida por lingüistas, dentro del que destaca, Randolph Quirk [1]. Quirk define un estado privado como algo que no se encuentra abierto a la observación objetiva o verificación. Estos estados privados incluyen emociones, opiniones y especulaciones, entre otros. La definición misma de este estado privado dificulta el análisis del sentimiento. La subjetividad está a menudo implícita en una conversación, además de ser altamente sensible al contexto, y su expresión a menudo es peculiar de cada persona. Sin embargo, esa subjetividad no implica que no sea verdad [2]. Por ejemplo, la frase "Jennifer ama el chocolate" expresa un sentimiento de Jennifer para con el chocolate, pero esto no significa que no sea verdad. Es así, como de esta misma manera no todas las frases objetivas son verdaderas.

3.2 Análisis de Sentimientos

Como campo de investigación, está estrechamente relacionada con (o se puede considerar una parte de) la lingüística computacional, procesamiento del lenguaje natural y la minería de textos. Partiendo por el estudio del estado afectivo (psicología) y el juicio (teoría de la evaluación), este campo tiene por objeto responder a las preguntas estudiadas

durante mucho tiempo en otras áreas sobre el discurso, utilizando nuevas herramientas proporcionadas por la minería de datos y la lingüística computacional.

Análisis de Sentimientos tiene muchos nombres. A menudo, se conoce como análisis de subjetividad, minería de opinión, y extracción de evaluación, con algunas conexiones con la informática afectiva (reconocimiento computacional y la expresión de la emoción) [3]. Este campo por lo general estudia los elementos subjetivos, definidos como "expresiones lingüísticas de los estados particulares en contexto"[2]. Estas suelen ser palabras sueltas, frases u oraciones. A veces, los documentos enteros son estudiados como una unidad de sentimiento, pero es generalmente aceptado que el sentimiento reside en pequeñas unidades lingüísticas [4]. Tanto el sentimiento, como la opinión a menudo se refieren a la misma idea, en este documento se utilizan los términos indistintamente.

Los sentimientos que aparecen en textos se ven de dos formas, la primera es explícitamente, donde la frase subjetiva directamente expresa la opinión (“Es un hermoso día”), mientras que la segunda es implícita, en donde el texto implica una opinión (“Los audífonos se quebraron en dos días”) [5]. La mayoría de los trabajos realizados se han enfocado en el primer tipo de sentimiento, debido a que este es más fácil de analizar.

La polaridad de los sentimientos es una característica particular de los textos. Esta se hace presente regularmente de forma dicotómica, positivo o negativo, a pesar de que también puede ser vista dentro de un rango. Un documento posee varias frases que demuestran opiniones, las cuales podrían tener una polaridad mixta, que es diferente a que estas no tuviesen polaridad. Yendo más lejos, se debe hacer una distinción entre la polaridad del sentimiento y la fuerza que este tiene.

Otra importante parte del sentimiento es el objetivo, pudiendo ser un objeto, un concepto, una persona o cualquier cosa. La mayoría de los trabajos han sido realizados sobre productos o criticas de películas, donde es fácil identificar el tópico del texto. Pero también es útil poner atención a la característica del objeto del cual el escritor se está refiriendo: “¿es la pantalla de la cámara o la duración de la batería el problema que más detectan los consumidores? [6]. Debido a la disponibilidad de datos pertenecientes a comentarios de productos, por ello la extracción de características ha sido altamente estudiada en la década pasada [5]. La mención de estas características en los textos también puede ser explicita (“La duración de la batería es muy corta”) o implícita (“La cámara es muy grande”) [5].

3.3 Metas del Análisis de Sentimientos

El Análisis de Sentimientos para ser llevado a cabo como tal debe considerar ciertas tareas fundamentales.

Según lo descrito por Mejova[6] la primera tarea a realizar es la detección del sentimiento, es decir, el lograr clasificar el texto como objetivo o subjetivo. La forma de llevarse a cabo es analizando los adjetivos que posee una cierta frase, por ejemplo, “Su rostro

es hermoso", puede ser fácilmente identificable su polaridad con el solo hecho de fijarse en el adjetivo que posee.

Lo segundo a lograr es la determinación de la polaridad, quiere decir que a partir de un fragmento de texto se busca clasificar la opinión en dos extremos, positivo y negativo, o al menos localizar la posición en una continuidad entre las dos polaridades [3].

El proceso entonces de clasificación pasa a ser dicotómico, es decir solo tomará dos puntos, positivo y negativo. El texto en general debe pertenecer a una de las dos categorías, en las investigaciones realizadas previamente por algunos autores [6] el nivel de dificultad disminuye al procesar, por ejemplo, críticas de productos, ya que poseen bien definidos el sentido que se da en el texto, sin embargo, existen otros casos en donde no es tan sencillo la clasificación de la polaridad, como es el caso de las noticias, ya que dentro del texto pueden aparecer una serie de sentimientos y el discriminar que es "bueno" o "malo" se torna más difícil de discernir.

Una tercera tarea a lograr, que viene a ser complementario a la clasificación del sentimiento, es descubrir el objetivo de la opinión, es decir, sobre qué está hablando el texto. En frases o textos pequeños suele ser más simple esta tarea, ya que suelen tener solo un objetivo bastante claro y específico, sin embargo, a medida que el tamaño del texto aumenta como en los blogs o páginas web resulta más engorroso, ya que suelen tener más de un objetivo. Otra zona investigada es la extracción de características que nombran en [6], en donde se buscan encontrar palabras claves que logren determinar el sentido u objetivo del texto.

4 Aprendizaje Computacional

Cuando el ser humano adquiere conocimientos, habilidades, actitudes o valores a través del estudio, de la experiencia o la enseñanza, se dice que aprende. Este proceso es fácil para el humano, sin embargo, lograr que una máquina aprenda como lo hacen las personas es una interrogante que existe desde los inicios de las computadoras. Actualmente, no existe una máquina capaz de aprender de la misma manera que lo hace el hombre, sin embargo, se han creado algoritmos eficaces para algunas tareas de aprendizaje [7].

En términos muy generales, se puede decir, que un programa aprende, si el desempeño obtenido para realizar alguna tarea, mejora con la experiencia.

De manera formal, se dice que un programa aprende de la experiencia E con respecto a una clase de tareas T y una medida de desempeño P, si su desempeño en las tareas T, medido con P, mejora con la experiencia E [7].

Se puede decir entonces que el Aprendizaje Computacional estudia los procesos computacionales que hay detrás del aprendizaje en humanos y en las máquinas. Esta disciplina juega un papel importante en muchas áreas de la ciencia.

4.1 Clasificación

La clasificación de textos surge de la necesidad de separar documentos de un tema o clasificación específica de un conjunto de documentos de diferentes temas. Al lograr clasificar los documentos por temas, la búsqueda de información se puede realizar de manera sencilla.

Debido al elevado número de documentos que pueden pertenecer a una colección de documentos, sobretodo en formato electrónico, realizar la clasificación en forma manual, provoca que la tarea sea complicada, costosa y que requiera mucho tiempo, por lo que surge la idea de hacerlo automáticamente.

Así es como surge el área de Clasificación Automática de Textos, en la cual se han utilizado diferentes métodos estadísticos y más recientemente técnicas de Aprendizaje Computacional.

4.2 Tipos de Aprendizaje

Un aspecto importante que influye en el aprendizaje es el grado de supervisión. En algunos casos, un experto en el dominio proporciona al aprendiz retroalimentación acerca de lo que es apropiado para su aprendizaje. En otros casos, a diferencia del aprendizaje supervisado, ésta retroalimentación está ausente, dando lugar al aprendizaje no-supervisado.

Y en otros casos, se combinan el aprendizaje supervisado y el no supervisado, dando lugar al aprendizaje semi-supervisado.

A continuación se describen brevemente en qué consiste cada tipo de aprendizaje.

4.2.1 Aprendizaje supervisado

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. Al conjunto de ejemplos del cual se intenta aprender se le llama conjunto de entrenamiento.

Usando estos datos se construye un modelo de predicción, o aprendiz, el cual permitirá predecir la clase para nuevos objetos no vistos por el aprendiz. La construcción del modelo se logra gracias a un algoritmo de aprendizaje.

Los algoritmos comúnmente utilizados son Naive Bayes, Máquinas de Soporte Vectorial, Vecinos más cercanos, J48 entre otros. En particular, en el área de Clasificación de Textos, los algoritmos típicamente utilizados son Naive Bayes y Máquinas de Soporte Vectorial. Para el propósito del documento no se explicarán todos los algoritmos mencionados, sólo los mayormente usados en el área de Clasificación de Textos, estos se describen en la siguiente sección.

Este tipo de aprendizaje tiene la ventaja de que no es necesario que al aprendiz se le muestren todos los ejemplos existentes, es decir que puede clasificar un ejemplo sin haberlo visto nunca. La desventaja es que a pesar de lo anterior, sí es necesaria una gran cantidad de ejemplos para el entrenamiento.

4.2.2 Aprendizaje no supervisado

Este tipo de aprendizaje no presupone ningún conocimiento previo sobre lo que se quiere aprender. Tampoco existe un maestro que conozca los conceptos a aprender, por esta razón a este tipo de aprendizaje se le denomina Aprendizaje no-supervisado.

A diferencia del aprendizaje supervisado, en el aprendizaje no supervisado, los ejemplos sólo incluyen los atributos, es decir, no se encuentran asociados a una clase. Para este caso la tarea se enfoca en descubrir patrones comunes entre los datos, que permitan separar los ejemplos en clases o jerarquías de clases. De éstas se podrán extraer caracterizaciones, o permitirán predecir características, o deducir relaciones útiles, a lo que se denomina como agrupación (clustering). Algunos de los algoritmos más comunes son:

Cobweb, EM, y Kmeans. Siendo este último el más utilizado en el área de Clasificación de Textos.

Este tipo de aprendizaje tiene la ventaja de que no es necesaria la presencia de un maestro para el aprendizaje o de un conjunto de entrenamiento.

4.2.3 Aprendizaje semi-supervisado

El aprendizaje semi-supervisado es la combinación del aprendizaje supervisado y el no-supervisado. En éste se aprende con la ayuda de dos conjuntos. Uno que contiene datos asociados a una clase, y el otro que contiene datos no asociados a una clase. La idea es aprender con los datos asociados a su clase y asociar una clase a los datos que no contienen asociada una clase. Algunos de los algoritmos más comunes son: Co-training, ASSEMBLE y self-training.

5 Técnicas para la Clasificación Automática

El primer paso para realizar la tarea de Clasificación Automática de Textos utilizando técnicas de Aprendizaje Computacional, consiste en obtener los atributos que describan el texto a clasificar, así como transformarlos a una representación adecuada para ser utilizados por los algoritmos de Aprendizaje Computacional. A este paso previo se le llama extracción de características. En la siguiente sección se explica con mayor detalle cómo se realiza la extracción de características en la Clasificación Automática de Textos.

5.1 Extracción de Características

La extracción de características generalmente consiste en tres etapas:

• Pre-procesamiento

• Indexado

• Reducción de dimensionalidad

5.1.1 Pre-procesamiento

El pre-procesamiento consiste fundamentalmente en eliminar aquellos elementos que generalmente no contienen información para la tarea de la clasificación. Consta de tres posibles fases básicas:

• Eliminación de etiquetas. Si los documentos utilizados contienen algún tipo de etiquetas o cabeceras (ej. etiquetas de html o xml), estás podrán ser removidas, debido a que en algunos casos no proporcionan información útil para la clasificación.

• Eliminación de palabras vacías. Las palabras vacías son palabras que son muy frecuentes y que por lo general no contienen información, por ejemplo: pronombres, preposiciones, conjunciones, artículos, etc.

• Lematización de palabras. Por lematización nos referimos al proceso de remover los sufijos para reducir una palabra a su lema o raíz. Por ejemplo, comprender, comprenderlo y comprendió tienen la raíz comprend.

5.1.2 Indexado

Quizás la representación de documentos más comúnmente usada es la llamada modelo vectorial. En el modelo vectorial, los documentos son representados por vectores de palabras y una colección de documentos son representados por una matriz A (palabra por documento), donde cada entrada representa las ocurrencias de una palabra en un documento.

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}$$

Donde a_{ik} es el peso de la palabra *i* en el documento *k*.

Existen muchos caminos para determinar el peso a_{ik} de la palabra *i* en el documento *k*, pero muchas de las aproximaciones están basadas en dos observaciones empíricas:

• Entre más ocurre una palabra en un documento, más relevante es en el tema del documento.

• Entre más veces ocurre la palabra en los documentos de la colección, será menos relevante.

A continuación se describen los 3 esquemas de ponderado más comúnmente usados.

Ponderado Booleano

Este es el esquema más simple y consiste en asignar 1 a a_{ik} si la palabra ocurre en el documento y 0 en otro caso:

$$a_{ik} = \begin{cases} 1 & si\ f_{ik} > 0 \\ 0 & en\ otro\ caso \end{cases}$$

Donde f_{ik} es la frecuencia de la palabra *i* en el documento *k*.

Ponderado por frecuencia de palabra

Otro esquema simple es usar la frecuencia de la palabra en el documento:

$$a_{ik} = f_{ik}$$

Donde f_{ik} es la frecuencia de la palabra *i* en el documento *k*.

5.1.3 Reducción de Dimensionalidad

Un problema en la clasificación de textos es la alta dimensionalidad en el espacio de atributos, lo que hace que el procesamiento sea extremadamente costoso en términos computacionales. De ahí, que existe la necesidad de reducir el conjunto original de atributos, a este proceso se le llama reducción de dimensionalidad [7]. Existen diversos métodos de reducción de dimensionalidad, a continuación se explican dos de ellos:

Umbral de frecuencia de documento

La frecuencia de documento para una palabra es el número de documentos en los cuales las palabras ocurren. Dado un umbral de frecuencia de documento, se

calcula la frecuencia de documento para cada palabra en el corpus de entrenamiento y se remueven las palabras donde la frecuencia de documento es menor que el umbral determinado. Este método se basa en el supuesto de que las palabras raras generalmente no tienen información para la predicción de categorías.

Ganancia de Información

La Ganancia de Información consiste en medir el número de bits de información para predecir la categoría por medio de la presencia o ausencia de una palabra en el documento.

Sea $c_1,\ldots,c_k$ el conjunto de posibles categorías. La ganancia de información de una palabra *w* es definida como:

$$IG(w) = -\sum_{j=1}^{k} P(c_j)\log\mathrm{P}(c_j) + P(w)\sum_{j=1}^{k} P(c_j|w)\log\mathrm{P}(c_j|w) + P(\overline{w})\sum_{j=1}^{k} P(c_j|\overline{w})\log\mathrm{P}(c_j|\overline{w})$$

Donde $P(c_j)$ es la probabilidad de la clase c_j (fracción de documentos en la colección que pertenece a la clase c_j) y $P(w)$ es la probabilidad de la palabra (fracción de documentos en los cuales la palabra *w* ocurre). $P(c_j|w)$ es la probabilidad de la clase dada la palabra (fracción de documentos de clase c_j que tiene al menos una ocurrencia de la palabra *w*) y $P(c_j|\overline{w})$ es la probabilidad de la clase c_j dada la no ocurrencia de la palabra *w* (fracción de documentos de clase c_j que no contienen la palabra *w*).

La ganancia de información se calcula para cada palabra del conjunto de entrenamiento, y se eliminan aquellas palabras que tengan menor ganancia de información de acuerdo a un umbral.

5.2 N-gramas

La posición de donde se encuentre el término también es fundamental en la representación del documente para el Análisis de Sentimientos. La posición de la palabra determina e incluso puede llegar a invertir la polaridad de la frase [6]. Por ejemplo “En la lata” viene a ser un trigrama, en donde “en” viene a ser una preposición, “la” viene a ser el determinador y finalmente “lata” es el sustantivo, por lo cual se aprecia que los n-gramas se encuentran compuestos de una palabra raíz y una parte de la oración.

5.3 Partes de la oración (POS)

La información de POS (Part of Speech de su sigla en inglés) es explotada muy comúnmente en el análisis de sentimientos, una de las principales razones de aquello es que logra identificar la desambiguación.

Los adjetivos se han utilizado como característica por un gran número de investigadores [3]. Dentro de las primeras propuestas que se realizaron para la predicción basada en la orientación de las palabras fue desarrollado con adjetivos.

A menudo se considera que ciertos adjetivos son buenos indicadores de sentimientos y en algunas ocasiones se han utilizado para la selección de características en la clasificación de sentimientos, principalmente cuando se centran en la polaridad.

5.4 Negaciones

Las negaciones son una pieza fundamental en el análisis de sentimiento, ya que la mayoría de las técnicas individualiza las palabras, con esto frases como "ella me gusta" y "ella no me gusta" las considera similares ya que solo se diferencian por una palabra, por ello es que la negación puede cambiar totalmente la polaridad de una frase. Por ello es que en algunos estudios como muestran en [6] se utiliza una parte de post procesamiento en donde se cambia la polaridad del resultado al ver una negación, otros incluyen el ubicar la negación cercana al adjetivo para así cambiarle el sentido solo al adjetivo. Sin embargo existen casos donde hay negaciones pero no necesariamente implica que el texto posee una polaridad negativa como sería el caso de "No me extraña que todos la amen", para estos caso se han utilizado POS para lograr patrones para identificar la polaridad de estas frases.

5.5 Representaciones

La forma de representar los textos también es fundamental al momento de pasárselos a un algoritmo clasificador, se abordarán los principales que se utilizarán a futuro para el proceso de experimentación.

5.5.1 Unigramas

Es la representación más simple de los N-gramas, en donde cada frase es dejada como tal en un vector de palabras, es decir:

Afuera	hay	un	hermoso	día	soleado

Figura 5.1 Representación mediante unigramas.

5.5.2 Bigramas

Es la siguiente representación de los N-gramas, donde se considera dividir la frase en conjuntos de pares de palabras dentro de un vector, es decir:

Afuera hay	hay un	un hermoso	hermoso día	día soleado

Figura 5.2 Representación mediante bigramas.

5.5.3 Trigramas

Es la última representación de N-gramas que se utilizará para experimentar, ya que se pueden considerar un número mayor para N, para este caso se encuentra formado por conjuntos de 3 palabras dentro de un vector, es decir:

Afuera hay un	hay un hermoso	un hermoso día	hermoso día soleado

Figura 5.3 Representación mediante trigramas.

5.5.4 TF-IDF

Esta representación es la más utilizada para la clasificación de textos, en donde la primera sección *TF* corresponde al valor de la frecuencia del término normalizado, multiplicado por *IDF*, que corresponde a la frecuencia inversa del término en la colección completa N.

$$w(D_i, t_j) = \frac{f_{ij}}{|D_i|} * -log\left(\frac{n_j}{N}\right)$$

En donde n_j corresponde al número de documentos que contienen al término *t* [10].

5.5.5 TF-RFL

Corresponde a la relevancia de la frecuencia de una categoría (etiqueta), el cual es una representación propuesta por [12], la que constituye una nueva representación para el problema de múltiples categorías.

$$tf - rfl_{tdl} = f_{td}\ log_2\left(2 + \frac{a_{t,l}}{\max(1, mean(a_{t,\lambda_{j/l}}))}\right)$$

En donde $mean(a_{t,\lambda_{j/l}})$ es el número promedio de documentos que contienen el término *t* para cada documento clasificado en categorías diferentes a *l*.

6 Algoritmos de Aprendizaje Computacional

La clasificación automática de textos consiste en asignar automáticamente una o más categorías predefinidas a documentos de textos libres. En años recientes se han aplicado diversas técnicas de Aprendizaje Computacional a la clasificación automática de textos.

A continuación se describen algunos de los algoritmos de Aprendizaje Computacional que han sido propuestos para la clasificación de textos y que se utilizarán para el propósito que se desea lograr.

Se define la siguiente notación:

Sea $d = d_1,\ldots,d_M$ el vector de documentos y $c_1,\ldots,c_K$ las posibles categorías o clases. Se asume que se tiene un conjunto de entrenamiento que consiste de N vectores de documentos $d_1,\ldots,d_N$ con las clases $y_1,\ldots,y_N$.

6.1 Naive Bayes

El clasificador Naive Bayes (Bayes Ingenuo) se construye usando el conjunto de entrenamiento para estimar la probabilidad de cada clase dados los valores de atributos (palabras) del documento de una nueva instancia. Usando el Teorema de Bayes para estimar las probabilidades:

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

El denominador en la ecuación anterior no distingue entre categorías y puede ser eliminado. Este método asume que los atributos son condicionalmente independientes, dada la clase. Esto simplifica los cálculos.

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

Una estimación $\hat{P}(c_j)$ para $P(c_j)$ puede ser calculada de la fracción de documentos de entrenamiento que es asignada a la clase c_j:

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

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.

Una estimación $\hat{P}(d_i|c_j)$ para $P(d|c_j)$ está dada por:

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

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 (agregar un 1). *M* es el número de términos en el vocabulario.

A pesar de que la suposición de independencia condicional es generalmente falsa para la aparición de la palabra en documentos, el clasificador Naive Bayes es sorprendentemente efectivo.

6.2 Máquinas de Soporte Vectorial

Máquinas de Soporte Vectorial han mostrado un buen desempeño en general en una gran variedad de problemas de clasificación, más recientemente en clasificación de textos.

En términos geométricos, el problema que resuelve las SVM (Support Vector Machine) es identificar una frontera de decisión linear entre dos clases, a través de una línea que los separe, maximizando el espacio del hiperplano. Sin embargo, las SVM incluyen una función llamada kernel, la cual permite realizar separaciones no lineales de los datos, proyectando la información a un espacio de características de mayor dimensión. Esto se logra cambiando la representación de la función, mapeando el espacio de entradas D a un nuevo espacio de características $F = \{\varphi(d) | \, d \in D\}$. Esto es:

$$d = \{d_1, d_2, \ldots, d_n\} \rightarrow \varphi(d) = \{\varphi(d)_1, \varphi(d)_2 \ldots, \varphi(d)_n\}$$

En la figura 6.1 se muestra un mapeo de un espacio de entradas de dos dimensiones a un nuevo espacio de características de dos dimensiones, donde la información no puede ser separada por una máquina lineal mientras que en el nuevo espacio de características esto resulta sencillo.

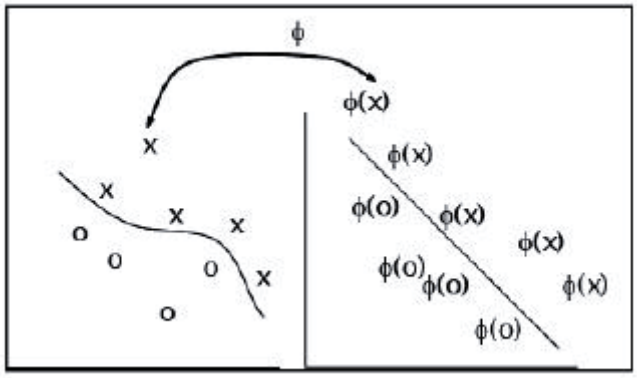

Figura 6.4 Mapeo del espacio de entradas a un espacio de mayor dimensión.

La función real φ no necesita ser conocida, es suficiente tener una función kernel *k*, la cual hace posible realizar el mapeo de la información de entrada al espacio de características de forma implícita y entrenar a la máquina lineal en dicho espacio.

6.3 Árboles de Decisión (C4.5)

C4.5 es un algoritmo usado para generar un árbol de decisión desarrollado por Ross Quinlan[8], el cual es una exterior del algoritmo ID3, creado por él mismo.

C4.5 construye árboles de decisión desde un grupo de datos de entrenamiento usando el concepto de entropía de información. Los datos de entrenamiento se consideran $S=s_1, s_2,\ldots$ de datos ya clasificados, cada uno $s_i = x_1, x_2,$ es un vector, donde x_i representan los atributos o características de los datos. Luego los datos de entrenamiento son aumentados con un vector $C=c_1, c_2,\ldots$, donde c_i representan la clase a la cual pertenecen.

En cada nodo que posee el árbol se escoge un atributo de los datos que discrimina de mejor manera el conjunto, dividiéndolo así en subconjuntos pertenecientes a una clase u otra, de esta forma, una vez entrenado los datos se procede a clasificar los nuevos datos a partir de las decisiones que tenga que ir tomando en cada nodo, llegando así a determinar a qué clase debe pertenecer.

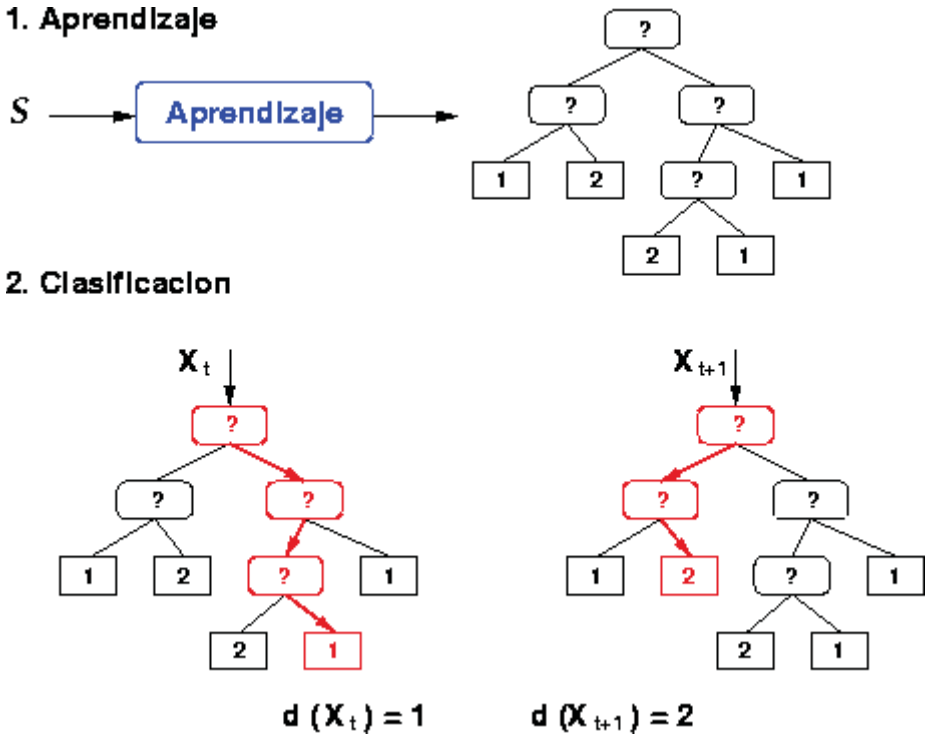

Figura 6.5. Esquema Algoritmo C4.5 [8]

Existe una implementación open source de este algoritmo en el lenguaje de programación java, denominado J48.

7 Evaluación

Otro punto importante y más bien final, es la evaluación de los algoritmos, ya que no siempre entregan una efectividad del 100%, por ende se necesitan métricas capaces de dar información relevante al desempeño de la clasificación.

7.1 Matriz de Confusión

En el campo de la inteligencia artificial, una matriz de confusión es una herramienta visual que se utiliza en el aprendizaje supervisado, cada columna que posee representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias de la clase real. Uno de los principales beneficios de las matrices de confusión es que facilitan ver si el sistema se confunde entre dos clases.

	Clasificado como A	Clasificado como B
Real A	Verdadero Positivo (VP)	Falso Negativo (FN)
Real B	Falso Positivo (FP)	Verdadero Negativo (VN)

Tabla 7.1. Matriz de Confusión.

A partir de la matriz de confusión es que se pueden obtener métricas para la evaluación del clasificador.

7.2 Sensibilidad (Recall)

La sensibilidad indica la capacidad del estimador para dar como casos positivos los casos que realmente lo son. La sensibilidad viene a ser la fracción de los verdaderos positivos.

Para calcularla se utiliza la siguiente fórmula:

$$Recall = \frac{VP}{VP + FN}$$

7.3 Precisión

La precisión es el cociente entre los verdaderos positivos y la suma de los verdaderos positivos y los falsos positivos, como se muestra a continuación:

$$Precision = \frac{VP}{VP + FP}$$

7.4 Valor F

El valor F es la medida de la precisión que tiene una clasificación y utiliza los valores obtenidos de la precisión y recall.

Para calcularse se utiliza:

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

8 Corpus

Para poder crear un clasificador es necesario, primero que todo, tener un conjunto de datos que sirva para la fase de entrenamiento y posterior pruebas, a esto se le denomina corpus.

Corpus en general se refiere a un conjunto de datos que se utilizarán como base para algún tipo de prueba investigativa, sin embargo, en el tema que se está abordando se precisa en definir lo que es el corpus lingüístico, este es un conjunto, generalmente muy amplio de ejemplos reales del uso de una determinada lengua, estos pueden ser textos o muestras orales, para este caso se consideran solo la parte escrita ya que se trabajaran con mensajes escritos por usuarios de la red social Twitter.

8.1 Creación de los datos

Los datos que se utilizarán para llevar a cabo la fase experimental son provenientes de los tuits emitidos por 67.000 personas con menciones a las empresas Ripley y Falabella entre el mes de Octubre del año 2013 y el mes de Enero de 2014, proporcionados por la empresa Analitic S.A., sin embargo estos datos fueron extraídos en bruto, por lo que previamente se deben eliminar las frases que se encuentran repetidas, esto por el efecto de los re-tuits o RT como se denominan a los mensajes que son reenviados entre los distintos usuarios de la red social, por ello es que al aparecer concursos por parte de las empresas es que deben ser eliminados cerca de 40.000 frases, reduciendo el total de mensajes. Luego existen tuits que por alcances de nombres se encuentran escritos en portugués y otros en inglés, por lo que estos deben ser eliminados ya que solo se trabajará con idioma español. Finalmente los datos definitivos con los que se trabajará son 1800 mensajes, divididos en partes iguales, es decir, 600 positivos, 600 negativos y 600 neutros, estos últimos, como se explicó previamente no son mensajes que no posean polaridad, sino que, poseen polaridad mixta y no pueden ser catalogados en ninguno de los dos polos, por ello es que se realizarán experimentos tomando en cuenta dos y tres sentidos, además cada uno de los sentidos se encuentran divididos en partes iguales para las menciones de las empresas Ripley y Falabella.

Para el proceso preliminar de clasificación por sentido se debe realizar de manera manual, es decir, analizar las frases que se tienen y catalogarlas en uno de los tres sentidos, para ello no debe ser realizado por una sola persona, ya que se considera que un solo juicio no necesariamente puede estar en lo correcto.

Una buena práctica para llevar a cabo la clasificación es que se realice por al menos dos personas [9], de tal modo que uno puede ir dando el sentido y el segundo va corroborando, de esta forma se deja de lado la imparcialidad que pudiese tener el primer encargado de clasificar, no obstante, existe otro punto importante a considerar, el cual es el caso de un desacuerdo entre las partes, para ello se da la posibilidad de incluir una tercera persona que actúe como árbitro o de su apoyo a uno de los dos, esto por ejemplo se ha realizado en algunos otros corpus.

En este punto en particular se contó con cinco personas encargadas de llevar a cabo un mismo proceso, clasificar las frases en positivo, negativo y neutro, por lo cual el modo de decidir el sentido se realizó según el porcentaje de acuerdo de todos los integrantes. Si se tenía una clasificación donde 4 o 5 de los integrantes daban con el mismo resultado se pasaba directamente a la sección de los datos a utilizar, además dependiendo de la cantidad de los datos que se desean utilizar también son considerados algunos que poseen 3 preferencias.

9 Experimentos

Los experimentos que se llevaron a cabo, fueron planteados a partir de 4 hipótesis, como se cuentan con mensajes pertenecientes a empresas del área del retail como Ripley y Falabella y separados en partes iguales se procedieron a crear 3 escenarios distintos.

Para la primera hipótesis se plantea que si se clasifican marcas de una determinada industria, como el retail, es posible luego clasificar marcas que se encuentren dentro del mismo rubro.

En la segunda hipótesis, hace referencia a la mejora del desempeño del clasificador si se realiza un proceso de training y testing a cada marca de manera independiente, esto debería ser posible ya que el proceso se realiza de manera más preciso y enfocado en una marca en particular.

Para el tercer escenario se propone observar las marcas de una determinada industria, en donde sí se clasifica para una, es posible utilizar ese mismo clasificador para otra marca.

Finalmente luego de recrear los tres escenarios se buscará probar si la forma de representar los textos influye en el desempeño que obtienen los clasificadores, esto conllevaría a decir si importa o no la forma que se le dé al texto para el correcto funcionamiento del clasificador.

Para dejarlo explicado de modo ilustrativo se puede apreciar en la figura 9.3.

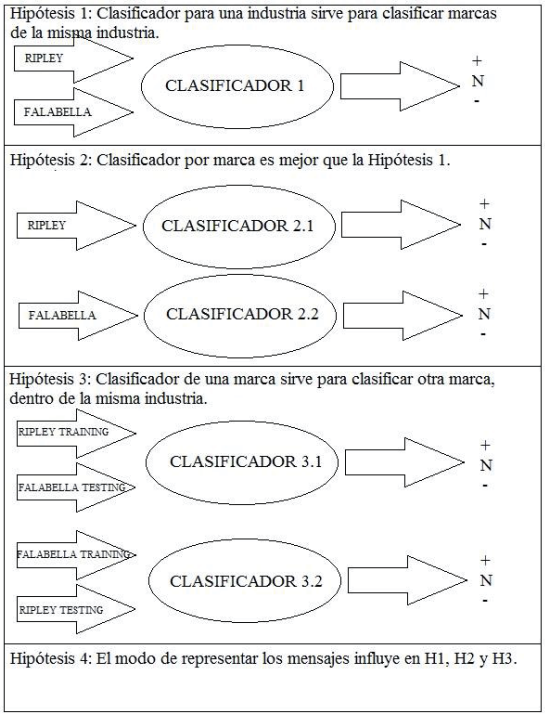

Figura 9.6 Escenarios para las hipótesis planteadas.

Para el proceso de representación de los mensajes de Twitter se utilizarán las 5 ya mencionadas previamente, Unigramas, Bigramas, Trigramas, TF-IDF y TF-RFL.

En la clasificación se realizará mediante la herramienta WEKA, por ende, se debió realizar programas codificados en lenguaje C para lograr las representaciones de N-gramas, así como también la de TF-RFL.

Los algoritmos encargados de llevar a cabo el proceso de clasificación automática serán Naive Bayes, SVM y J48, la última es la implementación en WEKA para los árboles de decisión.

10 Resultado de las pruebas

Para la evaluación de los clasificadores se utilizaron las tres métricas ya mencionadas previamente, precisión, recall y F_1. Los valores numéricos en tablas pueden ser encontrados en el anexo de este trabajo.

10.1 Hipótesis 1

Para el primer caso, la idea es utilizar los mensajes de forma total, de las dos empresas de retail, para así generar un único clasificador que obtenga un buen desempeño para las empresas que componen la industria.

Los resultados obtenidos para el caso de utilizar dos sentidos, la precisión se observa en la figura 10.7, en donde se aprecia el claro desempeño de los clasificadores, siendo TF-RFL quien logra la precisión más alta logrando sobre un 92% con todos los algoritmos.

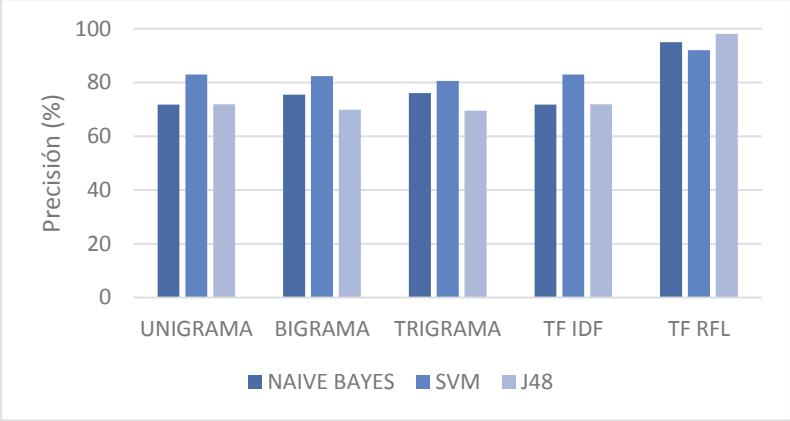

Figura 10.7. Gráfico Precisión para dos sentidos Hipótesis 1.

En la figura 10.8 se muestran los resultados obtenidos para dos sentidos para el recall, donde al igual que el gráfico anterior se observa la supremacía que obtiene la representación TF-RFL por sobre las demás representaciones.

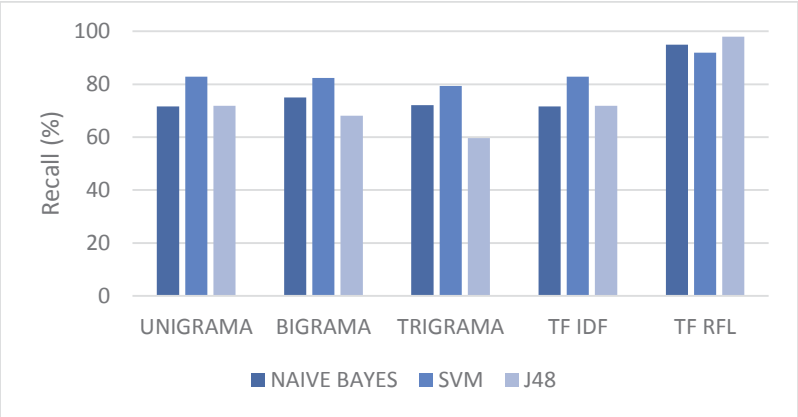

Figura 10.8 . Gráfico Recall para dos sentidos Hipótesis 1.

Finalmente en la figura 10.9 se muestran los resultados obtenidos por F_1, el cual se utilizarán finalmente para probar la Hipótesis 4, en el gráfico se observa al igual que las dos métricas anteriores como la representación de TF-RFL logra el mejor desempeño, con valores por sobre el 91.9% para todos los clasificadores, mientras que las SVM son quienes obtienen los mejores resultados en las distintas representaciones, obteniendo en todas valores cercanos al 80%.

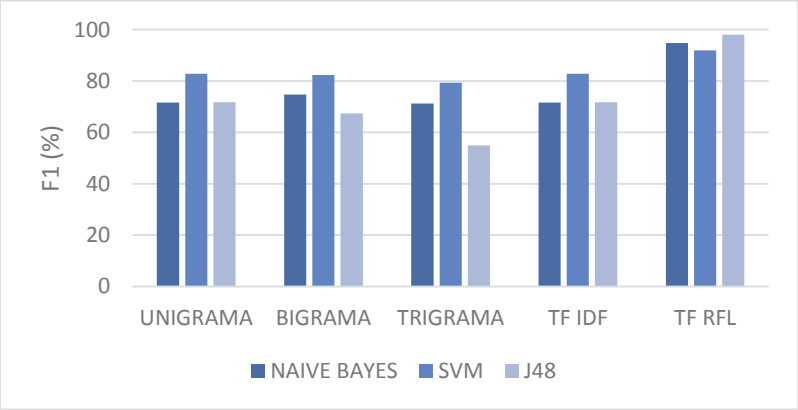

Figura 10.9. Gráfico F_1 para dos sentidos Hipótesis 1.

Luego de mostrar el desempeño obtenido utilizando sólo dos sentidos, se procederá a mostrar los resultados utilizando tres sentidos, positivo, negativo y neutro, ya que al poseer más clases se busca ver cómo funcionan tanto las representaciones como los clasificadores.

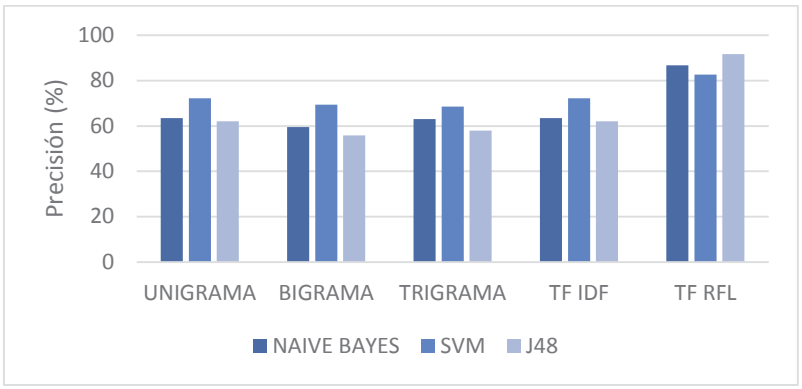

Figura 10.10. Gráfico Precisión para tres sentidos Hipótesis 1.

En la figura 10.10 se aprecian los valores obtenidos para la métrica de la precisión cuando se utilizan tres sentidos, se rescata como TF-RFL logra buenos resultados a pesar de haber incluido una nueva categoría a los datos, sin embargo, también se aprecia una caída en los porcentajes de alrededor de un 10% en comparación a realizar la clasificación con dos sentidos.

En la figura 10.11 y 10.12 se ven los resultados para recall y F_1 respectivamente, los cuales obtienen valores similares a la precisión, los mejores resultados para todas las representaciones se ven en las SVM, mientras que para TF-RFL en particular los árboles de decisión obtienen un muy buen desempeño, logrando para F_1 un valor de 90.9%, alto si se compara con representaciones de N-gramas, quienes van desde un 48% a un 61%.

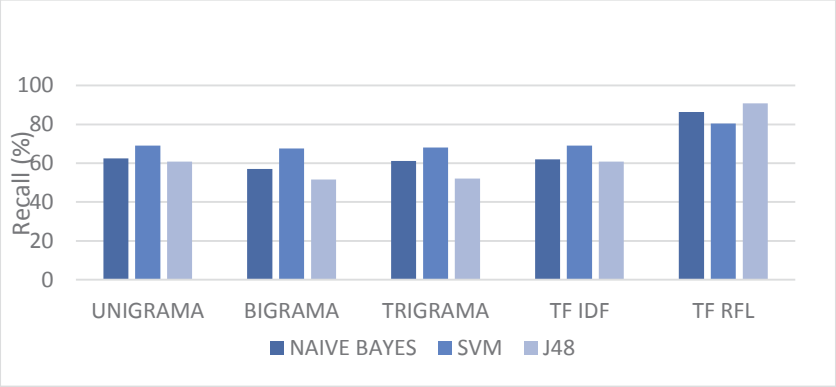

Figura 10.11. Gráfico Recall para tres sentidos Hipótesis 1.

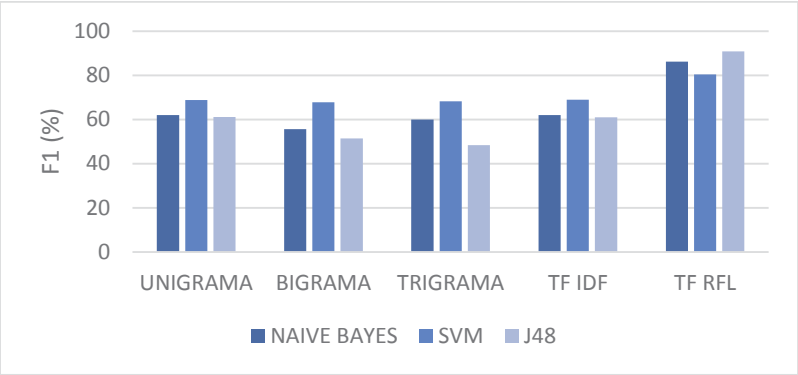

Figura 10.12. Gráfico F_1 para tres sentidos Hipótesis 1.

De los resultados presentados se puede concluir que la Hipótesis 1 es correcta, clasificar para la industria o rubro sirve para clasificar las marcas de aquella industria, en este caso clasificar mensajes de Falabella y Ripley como conjunto de entrenamiento sirven para probar aquellas marcas y dan buenos resultados, tanto en dos o tres sentidos y utilizando la representación de TF-RFL se logran valores por sobre el 90%.

10.2 Hipótesis 2

A continuación se presentan los resultados obtenidos para probar la Hipótesis 2, esto es crear clasificadores de manera independiente para cada marca, es decir Falabella por si sola y Ripley por si sola.

En primer lugar se mostrarán los valores del clasificador para la empresa de retail Falabella.

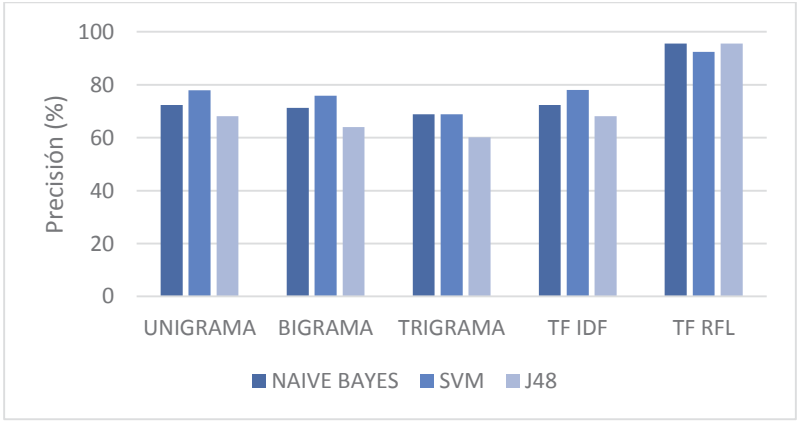

Figura 10.13. Precisión para dos sentidos, para clasificador de Falabella en Hipótesis 2.

Como se aprecia en la figura 10.13, los resultados son bastante similares a los obtenidos de la Hipótesis 1, en donde se aprecia como las SVM obtienen buenos desempeños en diferentes representaciones, mientras que la mejor representación sería TF-RFL.

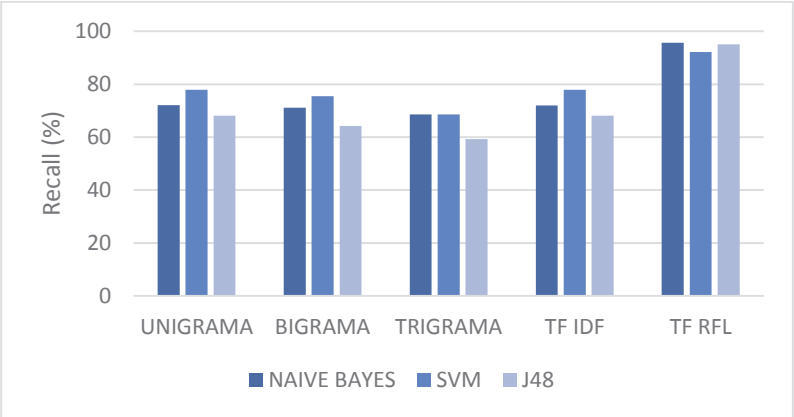

Figura 10.14. Recall para dos sentidos, para clasificador de Falabella en Hipótesis 2.

El recall obtenido que se puede observar en la figura 10.14, muestra como el clasificador de J48 obtiene bajo desempeño en 4 representaciones, mientras que al utilizarlo en TF-RFL es quien presenta de los valores más alto, con un 95.1%, esto hace denotar que el tipo de representación que se utilice influye en gran medida en el desempeño del clasificador, sin embargo, eso se verá al resolver la Hipótesis 4.

En la figura 10.15, el F1, muestra los valores más globalizados de entre las tres métricas, por lo que se utilizará en pasos posteriores, aquí se puede apreciar como la representación de TF-RFL, obtiene valores por sobre el 90%, lo que quiere decir que del total de textos clasificados, en la fase de training obtiene un muy buen rendimiento, en donde solo unos poco mensajes fueron mal clasificados en alguna categoría que no lo fuese, mientras que para otras representaciones, son las SVM quienes funcionan como mejor clasificador al dar valores muy cercanos al 80%.

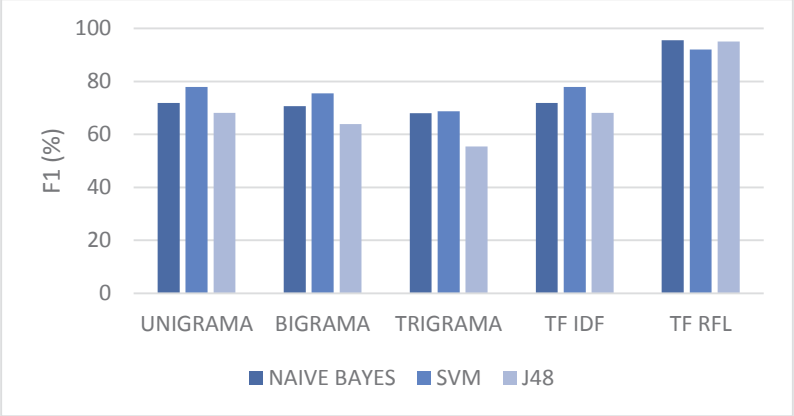

Figura 10.15. F_1 para dos sentidos, para clasificador de Falabella en Hipótesis 2.

Luego de presentar los resultados para dos sentidos en la empresa Falabella, se pasan a mostrar los valores obtenidos al utilizar tres sentidos, positivo, negativo y neutro.

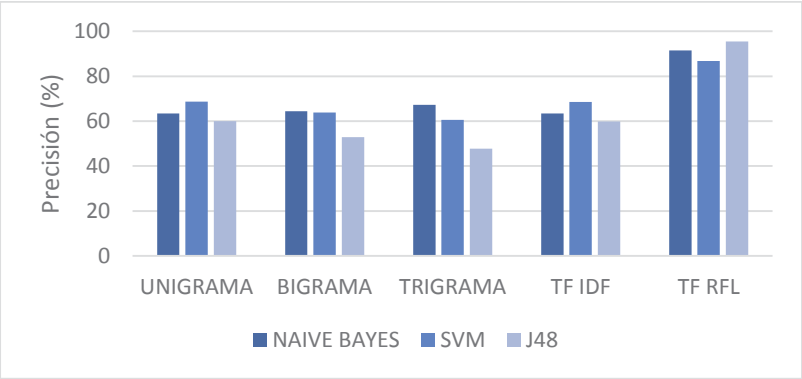

Figura 10.16. Precisión para tres sentidos, para clasificador de Falabella en Hipótesis 2.

La precisión obtenida bajaron en promedio cerca de un 10% en comparación l experimento con dos sentidos, ello debido a que aumenta la complejidad para el clasificador, sin embargo la representación TF-RFL, quién se encarga de dar valores de acuerdo a la importancia que tiene una palabra en cada categoría obtiene un buen desempeño, logrando valores por sobre el 90% para Naive Bayes y J48, mientras que para SVM obtiene un 86.8%, resultados bastante buenos, más aún en donde J48 casi no decae frente a utilizar dos sentidos, ya que solo presenta un 0.1% de diferencia en precisión, lo anterior se aprecia en la figura 10.16.

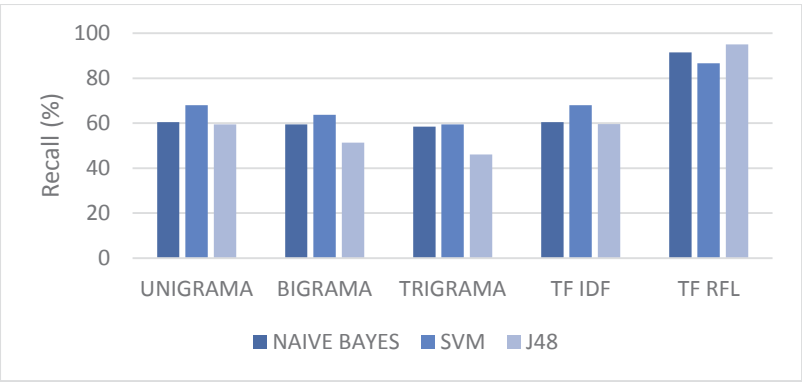

Figura 10.17. Recall para tres sentidos, para clasificador de Falabella en Hipótesis 2.

El recall presenta valores similares a los de la figura 10.16, en 10.17 se aprecia que las representaciones de Naive Bayes se mantiene casi constante cercano al 60% en 4 representaciones, mientras que en TF-RFL logra un muy buen desempeño con un 91.5%. TF-RFL muestra ser la mejor representación, mientras que SVM el mejor clasificador.

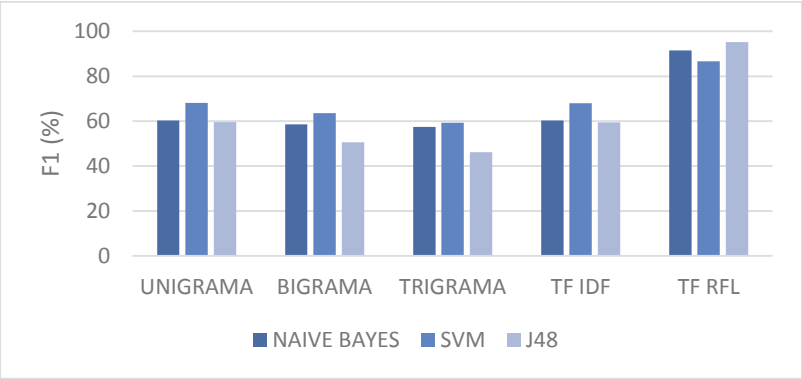

Figura 10.18. F_1 para tres sentidos, para clasificador de Falabella en Hipótesis 2.

Finalmente, el recall en la figura 10.18 muestra los resultados para el clasificador de Falebella utilizando tres sentidos, en donde TF-RFL muestra ser la mejor representación, sobrepasando incluso en más de un 30% a otras representaciones como el caso del clasificador J48.

Luego de mostrar los resultados para Falabella en dos y tres sentidos, se expondrán los resultados para la otra empresa, Ripley, en primer lugar en dos sentidos y finalmente utilizando tres sentidos.

En la figura 10.19 se encuentra la precisión para el clasificador construido únicamente para la empresa Ripley, los valores son bastante buenos para todas las representaciones y superando al clasificador de la Hipótesis 1, en donde J48 solo se equivocó en clasificar 1 tuit de entre el total puesto a prueba.

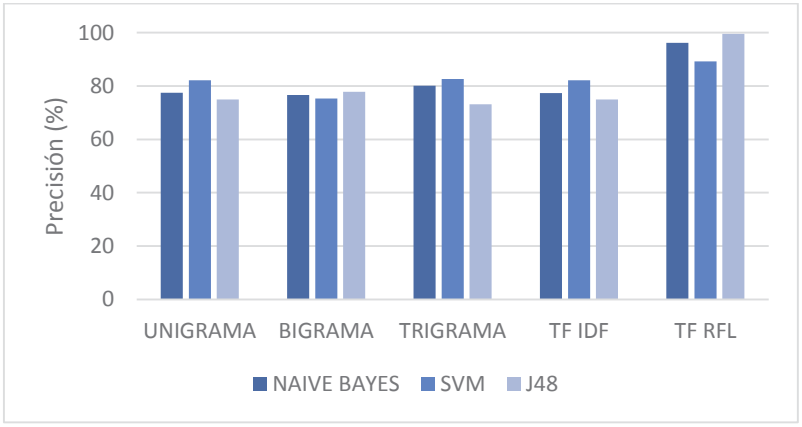

Figura 10.19. Precisión para dos sentidos, para clasificador de Ripley en Hipótesis 2.

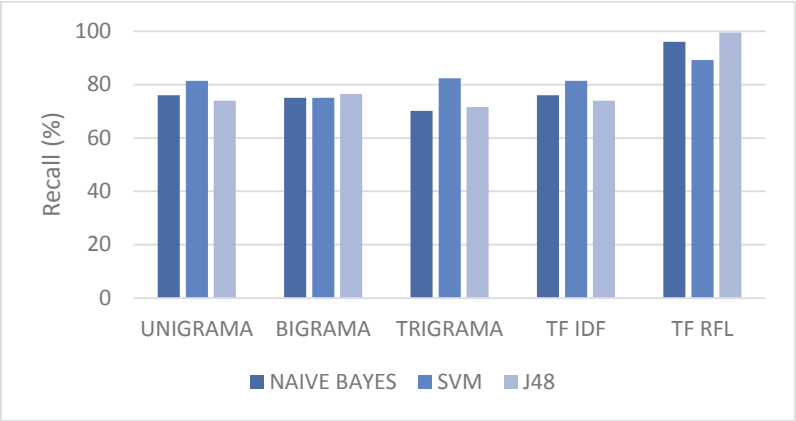

Figura 10.20. Recall para dos sentidos, para clasificador de Ripley en Hipótesis 2.

Recall de la figura 10.20 obtiene valores similares, representaciones como los bigramas obtienen resultados homogéneos para los distintos clasificadores, mientras las representaciones de Unigramas y TF-IDF obtienen resultados bastantes similares, en donde tomar cualquiera de éstas dos arroja similares valores.

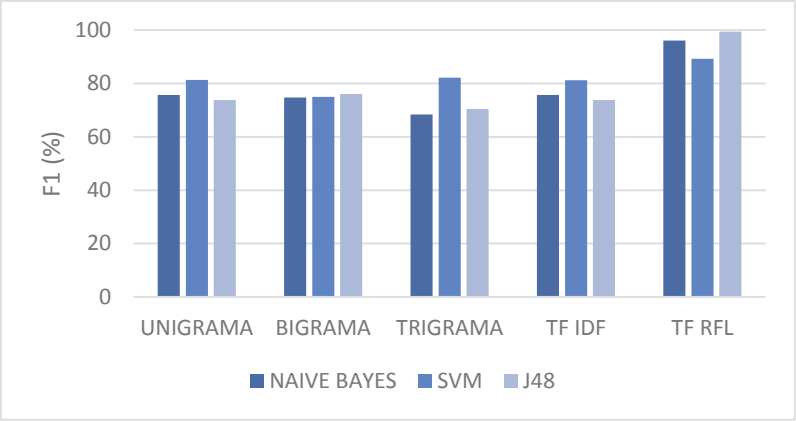

Figura 10.21. F_1 para dos sentidos, para clasificador de Ripley en Hipótesis 2.

Finalmente en la figura 10.21 se observan los valores de F_1, en donde se puede ver como TF-RFL, pasa a ser la representación con los valores más altos, en donde J48 obtiene un 99.5% de rendimiento, muy cercano Naive Bayes con un 96.1% y en tercer lugar SVM, con un 89.2%, sin embargo en general todos presentan un muy buen rendimiento con ésta representación, llegando a ser mejor que el clasificador de la Hipótesis 1.

A continuación se muestran los valores obtenidos para el clasificador de Ripley en tres sentidos, en primer lugar se aprecia la precisión en la figura 10.22, donde nuevamente la representación de TF-RFL es quien se lleva los mejores resultados con valores siempre por sobre el 80%, en donde J48 se encuentra con 94.2%, en donde solo una decena de tuits fueron mal clasificados en la parte de testing.

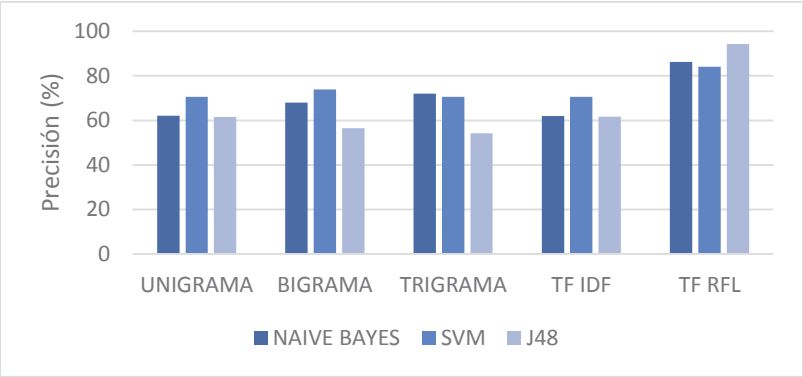

Figura 10.22. Precisión para tres sentidos, para clasificador de Ripley en Hipótesis 2.

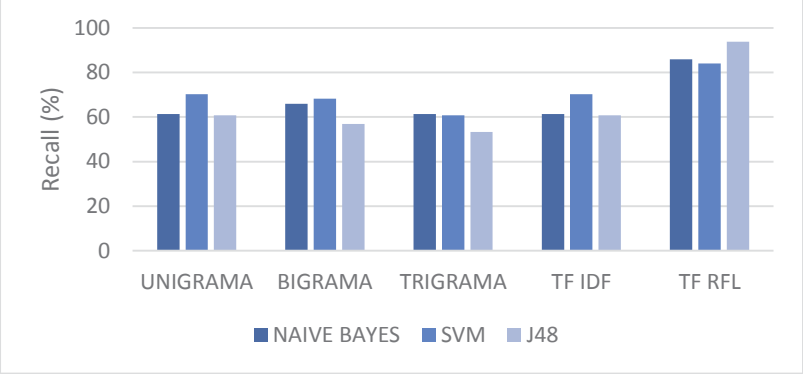

Figura 10.23. Recall para tres sentidos, para clasificador de Ripley en Hipótesis 2.

Para el caso del recall de la figura 10.23, 4 representaciones se encuentran por bajo el 70%, lo cual disminuye su rendimiento, sin embargo, TF-RFL se mantiene con altos valores siempre sobre el 80%, debido a la forma que posee de representar los mensajes escritos funciona de muy buena manera con multiclases.

Finalmente en la figura 10.24 se presenta el F_1 para el clasificador de Ripley en tres sentidos, donde como era de esperar, TF-RFL muestra los mejores valores, mediante el clasificador J48 obtiene el valor más alto con un 93.8%, muy por sobre otras representaciones como los trigramas en donde solo obtiene un 52.9%, lo cual es un valor bastante bajo.

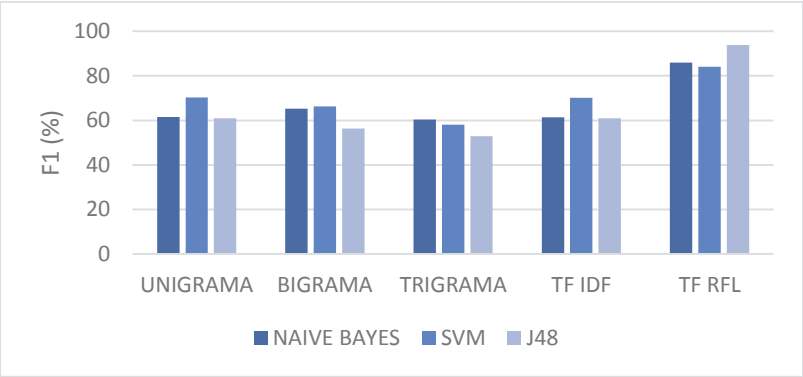

Figura 10.24. F_1 para tres sentidos, para el clasificador de Ripley en Hipótesis 2.

Por lo tanto, se puede concluir de la Hipótesis 2 que también es correcta, clasificar para una marca en específico suele ser mejor que el caso general de clasificar para toda la industria.

10.3 Hipótesis 3

La tercera Hipótesis busca demostrar si se clasifica para una marca es posible utilizar ese clasificador en otra marca dentro de la misma industria, para ese proceso se utilizaron mensajes de las empresas Falabella y Ripley. En primer lugar se presentarán los resultados de entrenar (training) con Ripley y probar (testing) con Falabella, en dos y luego en tres sentidos.

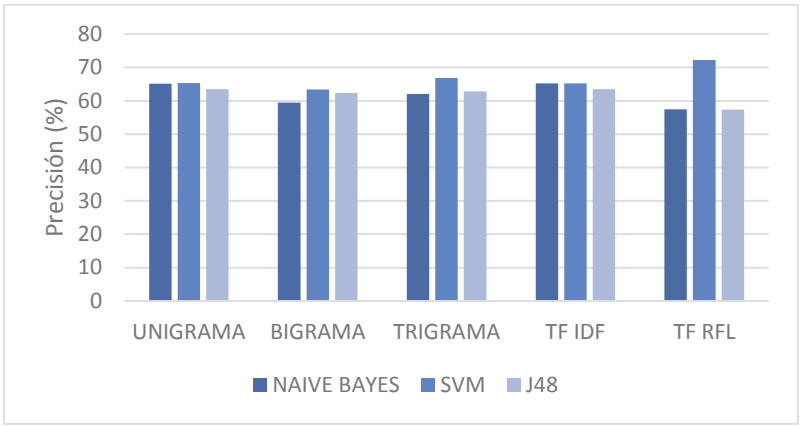

Figura 10.25. Precisión para dos sentidos, training Ripley, testing Falabella en Hipótesis 3.

Los valores decaen si se compara con las primeras dos hipótesis, e inclusive TF-RFL quien presentaba los mayores valores, en este escenario se queda por debajo de las demás representaciones, a excepción cuando se utiliza en conjunto a las SVM.

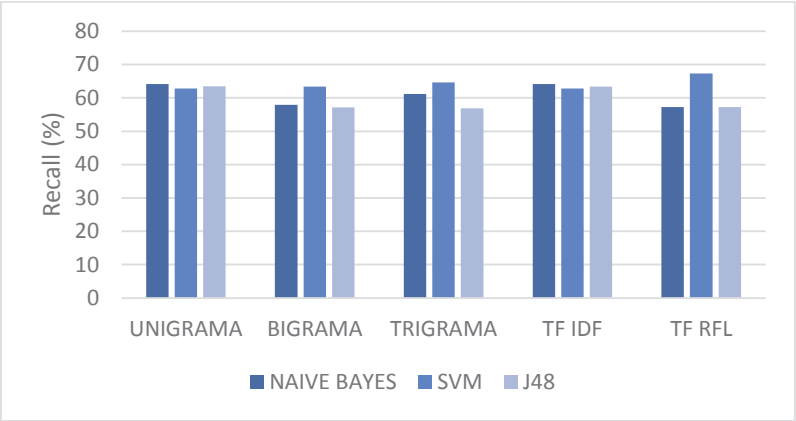

Figura 10.26. Recall para dos sentidos, training Ripley, testing Falabella en Hipótesis 3.

Para el recall que se aprecia en la figura 10.26, los valores tienden a ser más homogéneos y las variaciones entre una y otra representación solo varían en algunos puntos, sin embargo todos se encuentran cercanos al 65%, lo cual es bajo para el rendimiento de un clasificador.

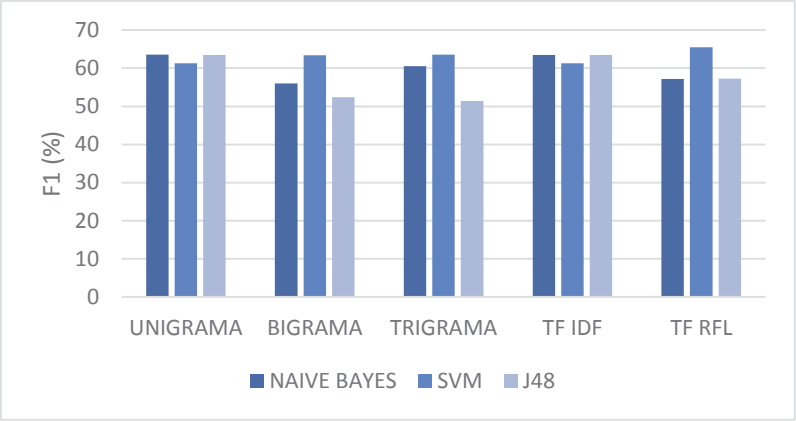

Figura 10.27. F_1 para dos sentidos, training Ripley, testing Falabella en Hipótesis 3.

Finalmente para F_1 de la figura 10.27 observamos la misma tónica de las métricas anteriores, donde los resultados son bajos y hace pensar que se debería rechazar la hipótesis.

Luego de presentar el escenario de entrenar con Ripley y testear con Falabella en dos sentidos, se procede a mostrar las pruebas al utilizar tres sentidos.

En primer lugar, tenemos la precisión en la figura 10.28 para tres sentidos, donde TF-RFL se encuentra bajo en comparación a las anteriores hipótesis, y esta vez los valores se encuentran cercanos a solo un 50%.

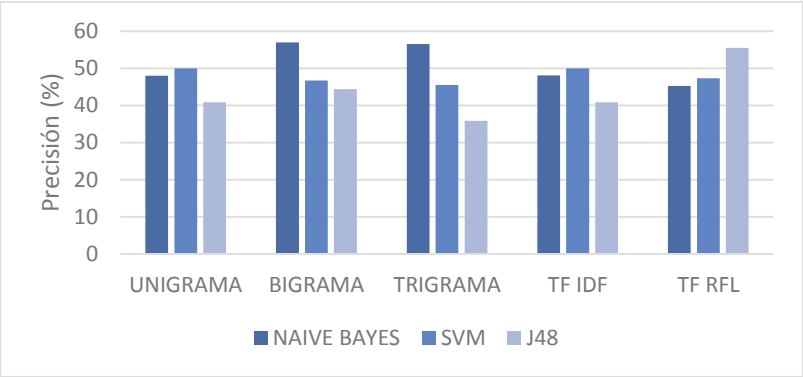

Figura 10.28. Precisión para tres sentidos, training Ripley, testing Falabella en Hipótesis 3.

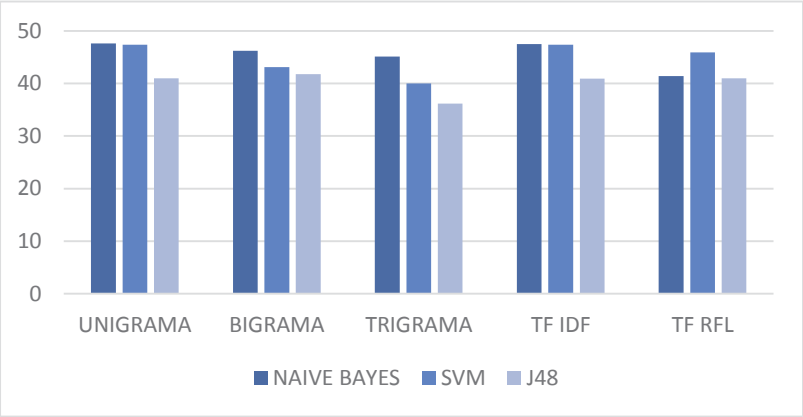

Figura 10.29. Recall para tres sentidos, training Ripley, testing Falabella en Hipótesis 3.

Al igual que la métrica anterior, ahora en recall de la figura 10.29, se ven que los valores bajan aún más, llegando todos a estar bajo el 50%.

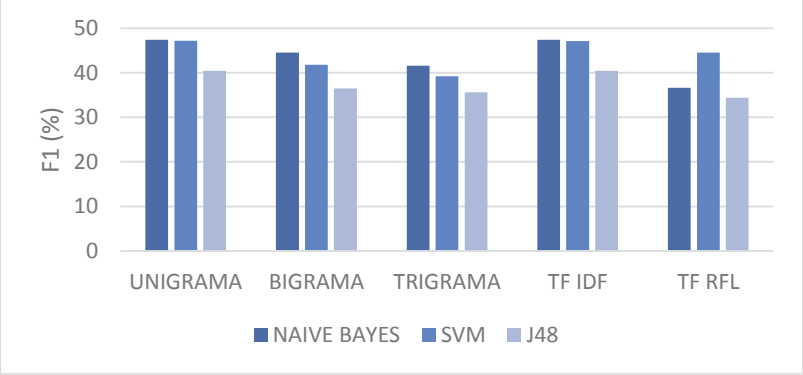

Figura 10.30. F_1 para tres sentidos, training Ripley, testing Falabella en Hipótesis 3.

De la figura 10.30 solo se puede concluir lo mismo que las métricas anteriores, donde no se logra ni un 50% de rendimiento, mostrando que en trabajando en tres sentidos los resultados son considerablemente bajos.

A continuación se presenta el escenario inverso, en donde se procede a entrenar con Falabella y probar con Ripley el clasificador, para ver si existen mejoras en comparación al escenario recién propuestos.

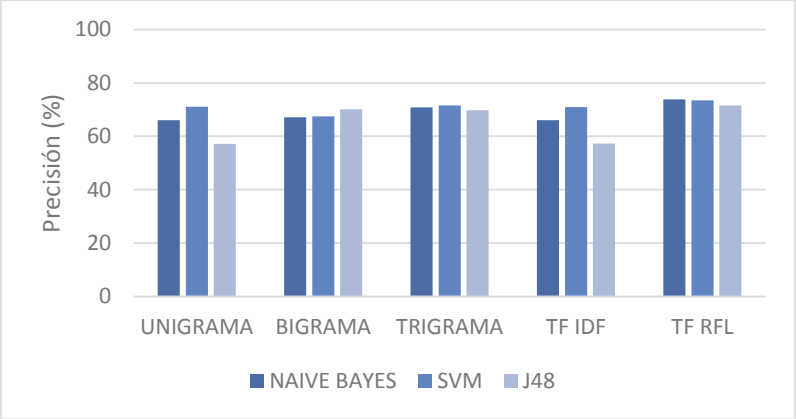

Figura 10.31 Precisión en dos sentidos, training Falabella, testing Ripley en Hipótesis 3.

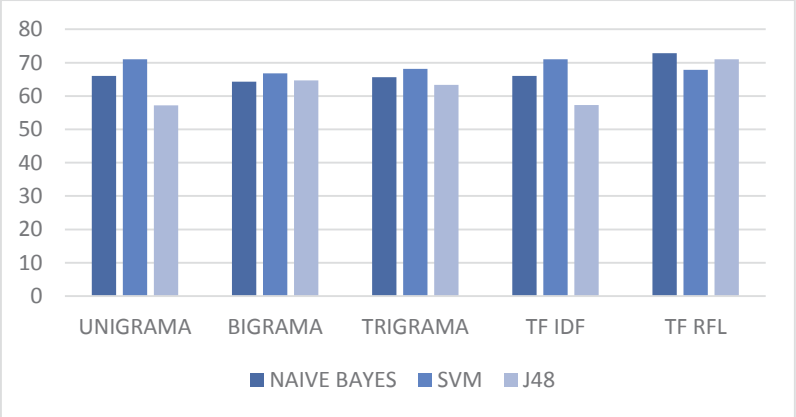

Figura 10.32 Recall en dos sentidos, training Falabella, testing Ripley en Hipótesis 3.

En las figuras 10.31, 10.32 y 10.33 se observa el conjunto de las tres evaluaciones realizadas a los clasificadores en dos sentidos, donde se aprecia que los resultados se encuentran muy cercanos al 70%, bajo si se compara con los resultados obtenidos de las Hipótesis 1 y 2, en donde habían representaciones y algoritmos que alcanzaban valores por sobre el 90% e inclusive por sobre el 95%, por lo cual este escenario no tiende a ser demasiado favorable para los clasificadores.

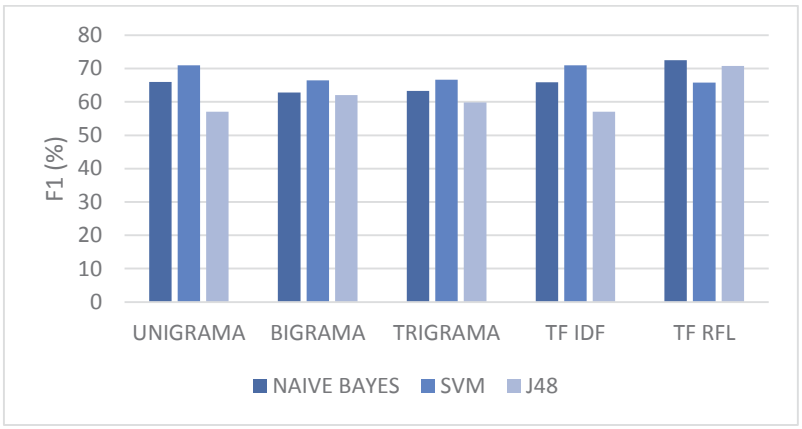

Figura 10.33. F_1 en dos sentidos, training Falabella, testing Ripley en Hipótesis 3.

Finalmente para terminar la Hipótesis 3, se presentan los valores de los resultados obtenidos por entrenar con Falabella y testear con Ripley utilizando tres sentidos.

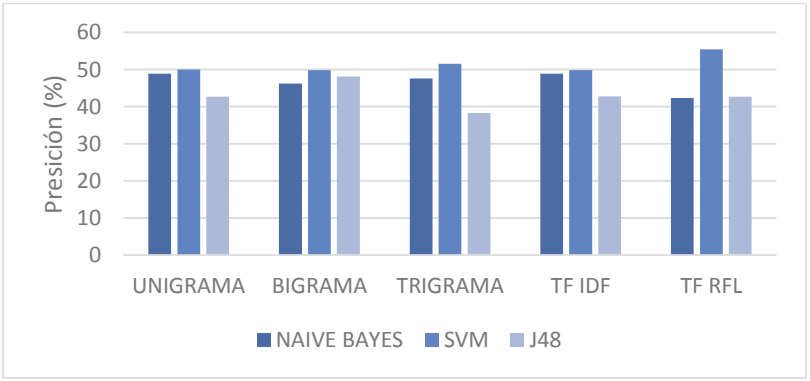

Figura 10.34. Precisión en tres sentidos, training Falabella, testing Ripley en Hipótesis 3.

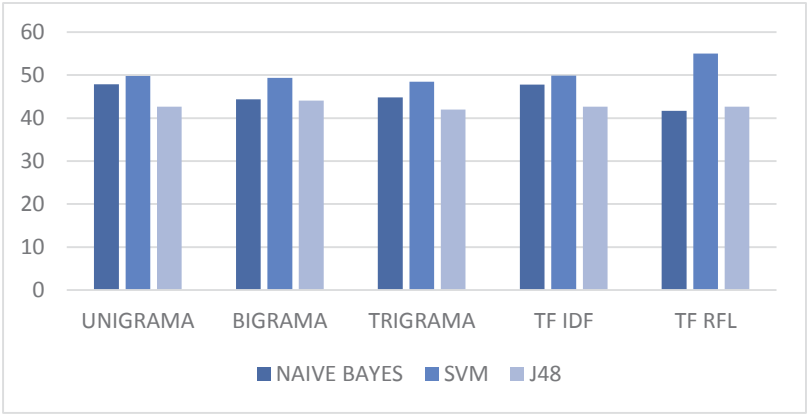

Figura 10.35. Recall en tres sentidos, training Falabella, testing Ripley en Hipótesis 3.

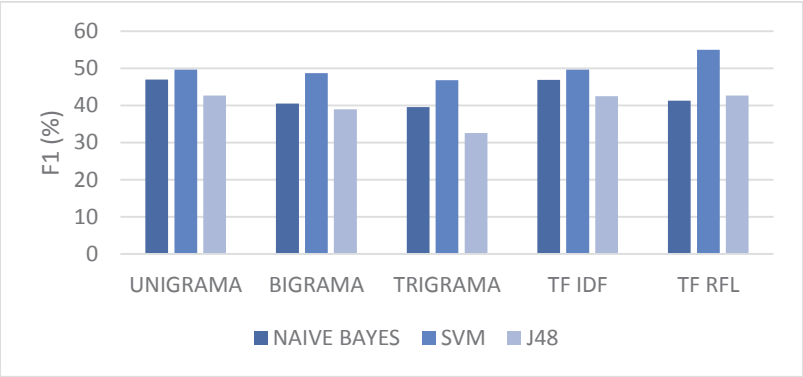

Figura 10.36. F_1 en tres sentidos, training Falabella, testing Ripley en Hipótesis 3.

En las figuras 10.34, 10.35 y 10.36 se observa el último escenario de la Hipótesis 3, en donde al igual que el caso de entrenar con Ripley y testear con Falabella se observan rendimientos por bajo el 50%, lo que hace descartar la posibilidad de esta Hipótesis, por lo que finalmente se puede concluir que no es conveniente y no obtiene buenos resultados el crear un clasificador para una determinada marca y utilzarlo para otra, por lo que la Hipótesis 3 se rechaza, al menos para las representaciones y clasificadores utilizados en estos experimentos.

10.4 Hipótesis 4

La hipótesis 4 plantea si el modo de representar los mensajes de texto importa para el clasificador que se utilice, para ello se muestra los valores F_1 obtenidos para los distintos clasificadores y representaciones. En la tabla 9.2 se muestran los resultados obtenidos cuando se prueban con dos sentidos.

	F1 / HIPOTESIS 1			F1 / HIPOTESIS 2						F1 / HIPOTESIS 3						
	NB	SVM	J48	NB	SVM	J48	NB	SVM	J48	NB	SVM	J48	NB	SVM	J48	PROMEDIO
UNIGRAMAS	71,6	82,9	71,8	71,9	77,9	68,1	75,7	81,3	73,8	63,6	61,3	63,5	66	71	57,1	70,50
BIGRAMAS	74,8	82,3	67,4	70,7	75,5	63,9	74,7	75	76,1	56	63,4	52,4	62,8	66,5	62,1	68,24
TRIGRAMAS	71,3	79,3	54,9	68	68,7	55,4	68,4	82,2	70,4	60,5	63,6	51,4	63,3	66,7	59,9	65,60
TF-IDF	71,6	82,9	71,8	71,9	77,9	68,1	75,8	81,2	73,8	63,5	61,3	63,5	65,9	71	57,1	70,49
TF-RFL	94,8	91,9	98	95,6	92,1	95,1	96,1	89,2	99,5	57,2	65,5	57,3	72,5	65,8	70,8	82,76

Tabla 9.2. Valores de F_1 para los clasificadores en sus distintas representaciones.

Se aprecian claras diferencias entre unos y otros, tanto por separado como en su promedio, en donde se ve claramente que la representación de TF-RFL obtiene los mejores valores y por ende el mejor desempeño, sin embargo para lograr apreciar con mayor claridad la veracidad de la hipótesis se realizó un test estadístico utilizando la prueba t-Student entre las distintas representaciones, ellas se ven reflejadas en la tabla 9.3.

	UNIGRAMA	BIGRAMA	TRIGRAMA	TF-IDF	TF-RFL
UNIGRAMA		0,0184	0,0035	0,1887	0,0015
BIGRAMA			0,1343	0,0187	0,00013
TRIGRAMA				0,0036	0,00040
TF-IDF					0,00151
TF-RFL					

Tabla 9.3. Prueba t-Student para dos colas en las representaciones utilizadas.

Como se aprecia en la tabla 9.3, las pruebas estadísticas concluyen que las diferentes representaciones obtienen desempeños diferentes estadísticamente significativos. Con esto se puede concluir que la representación si influye en el desempeño de los clasificadores.

De forma más gráfica, en la figura 10.37 se aprecian los valores promedios obtenidos para F1 en los distintos escenarios, en donde se ve la clara diferencia de utilizar TF-RFL en las primeras dos hipótesis, mientras que para el tercer escenario baja considerablemente, igualándose a las demás representaciones.

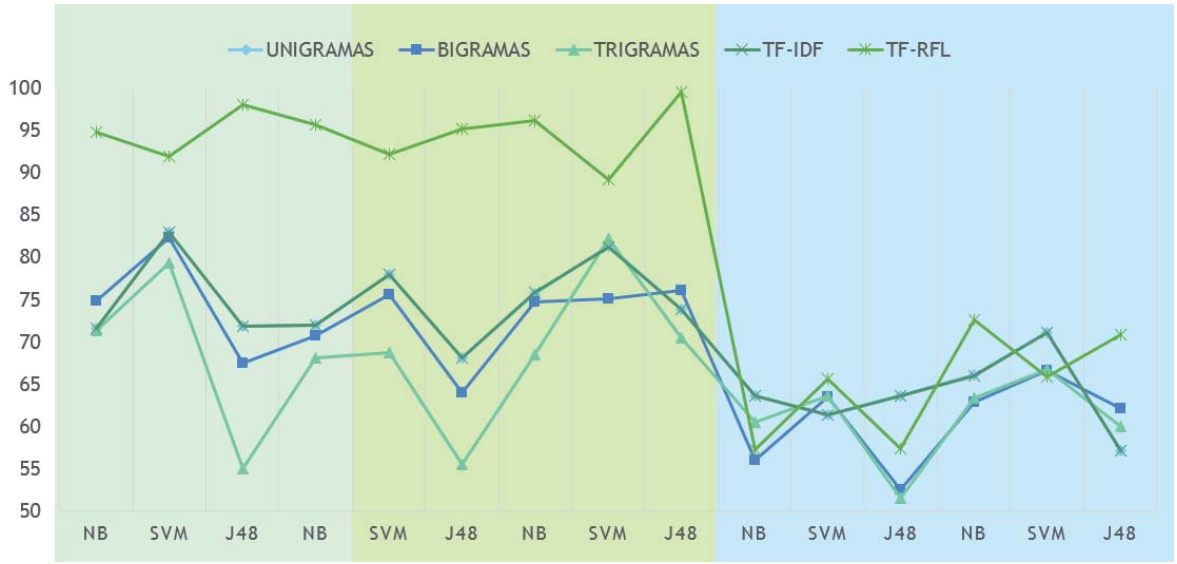

Figura 10.37. Gráfico de promedios F1 en los distintos escenarios.

11 Conclusión

En el presente trabajo se presentaron los fundamentos teóricos y conceptos básicos para enfrentar la problemática planteada, la clasificación automática de textos de opinión a través de su polaridad.

La clasificación automática es una herramienta potente, que ha ido facilitando el proceso de la extracción y clasificación manual, disminuyendo costos y tiempos.

En la actualidad, el avance tecnológico en el que el mundo se ve envuelto empuja a las personas a expresarse en medios sociales digitales, como es el caso de las redes sociales, el decir que tan agradable, bueno, malo o disconforme se encuentran con un determinado producto, personaje o tema en particular hace que sea de vital importancia rescatar esa información, ya sea por empresas para realizar su proceso de decisión, o campañas políticas que buscan saber si se encuentran en la dirección correcta.

Se ha mostrado y definido en qué consiste un sentimiento, su análisis y como este se puede catalogar de manera automática.

Se vio cómo funciona el aprendizaje computacional y los tipos que existen en la actualidad, con ellos se realizan una aproximación a los algoritmos mayormente utilizados en el proceso de aprendizaje y clasificación, los cuales sirven como una base sólida en la formación de conocimiento para enfrentar futuros desafíos.

La etapa de clasificación manual es fundamental que se encuentre correcta, para que así los algoritmos puedan funcionar de manera correcta, por ello es que se trabajó con una mayor cantidad de personas para lograr mejores resultados.

La etapa de pre-procesamiento, al igual que en el proceso de minería de datos, forma un papel principal para la obtención de buenos resultados. Transformar los mensajes, realizar limpieza a los datos, convertirlos para que el clasificador entienda el texto es un punto fundamental.

Con respecto a los resultados y las hipótesis se puede concluir que, para el primer escenario los valores obtenidos mediante la representación de TF-RFL logra muy buenos resultados, siempre por sobre el 90%, mientras que si se quiere optar por otra representación se debe preferir como algoritmo clasificador a las SVM, ya que predominan con los mejores resultados. Por ello es que la primera hipótesis se logra demostrar, clasificar para una industria sirve para las marcas que la componen y además se obtienen buenos resultados.

Para la segunda hipótesis, se obtienen también buenos resultados, por lo que clasificar para una marca en particular sirve para clasificar aquella misma marca a futuro, mostrando una leve mejoría en desempeño que el primer escenario.

Para la tercera hipótesis, se mostró que con las representaciones y algoritmos utilizados no se obtienen buenos resultados, por lo que no es recomendable utilizar este escenario.

Y finalmente la última hipótesis, sobre la importancia y relevancia de la representación se puede decir que es verdadera, ya que influye en el desempeño de los clasificadores, según el algoritmo que se utilice, existirá una representación que más se le acomode y obtenga mejores resultados. A modo general se opta por elegir a TF-RFL como representación de los mensajes y a SVM como clasificador.

12 Referencias

[1] Quirk R., Greenbaum S, G. L. and Svartvik, J. (1985). A comprehensive grammar of theEnglish language. Longman.

[2] Wiebe, J. M., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. Computational Linguistics, 30:277–308.

[3] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. Foundation and Trends in Information Retrieval, 2(1-2):1–135.

[4] Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS), 21(4):315–346.

[5] Liu, B. (2006). Web Data Mining, chapter Opinion Mining. Springer.

[6] Mejova Y. (2010), Sentiment Analysis: An Overview.

[7] Araujo N. (2009), Método Semisupervisado para la Clasificación Automática de Textos de Opinión.

[8] Breiman, L., Friedman, J., Olshen, R., and Stone, C., (1984), Classification and Regresion Trees, Wadsworth International Group.

[9] Colorado, F. (2007), Metodología, construcción y explotación de corpus anotados semántica y anafóricamente.

[10] Holts A., Riquelme C., Alfaro R. (2010), Automated Text Binary Classification using Machine Learning Approach. XXIX International Conference of the Chilean Computer Science Society.

[11] Alfaro R., Allende H. (2010), A new input representation for multi-label text classification. 4th International Conference on Intelligent Information Technology Application (IITA)

[12] Alfaro R., Allende H. (2010), Text Representation in Multi-label Classification: Two New Input Representations 10th International Conference on Adaptive and Natural Computing Algorithms (ICANNGA'11).

Anexos

Anexo A: Tablas de Resultados utilizados para gráficos

PRECISION			
DOS SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	71,8	82,9	71,9
BIGRAMA	75,5	82,4	69,8
TRIGRAMA	76	80,6	69,5
TF IDF	71,8	82,9	71,9
TF RFL	95	92	98,1

Tabla A.4 Precisión Hipótesis 1 en dos sentidos.

PRECISION			
TRES SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	63,5	72,2	62
BIGRAMA	59,5	69,4	55,9
TRIGRAMA	63,1	68,5	58
TF IDF	63,5	72,2	62
TF RFL	86,7	82,6	91,7

Tabla A.5 Precisión Hipótesis 1 en tres sentidos.

RECALL			
DOS SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	71,6	82,8	71,8
BIGRAMA	75	82,4	68,1
TRIGRAMA	72,1	79,4	59,6
TF IDF	71,6	82,8	71,8
TF RFL	94,9	91,9	98

Tabla A.6 Recall Hipótesis 1 en dos sentidos.

RECALL			
TRES SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	62,4	69,1	60,9
BIGRAMA	57	67,6	51,6
TRIGRAMA	61,1	68,1	52,1
TF IDF	62	69,1	60,9
TF RFL	86,3	80,4	90,8

Tabla A.7 Recall Hipótesis 1 en tres sentidos.

F			
DOS SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	71,6	82,9	71,8
BIGRAMA	74,8	82,3	67,4
TRIGRAMA	71,3	79,3	54,9
TF IDF	71,6	82,9	71,8
TF RFL	94,8	91,9	98

Tabla A.8 F_1 Hipótesis 1 en dos sentidos.

F			
TRES SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	62,1	68,9	61,1
BIGRAMA	55,6	67,8	51,5
TRIGRAMA	60	68,2	48,4
TF IDF	62,1	69	61
TF RFL	86,3	80,5	90,9

Tabla A.9 F_1 Hipótesis 1 en dos sentidos.

PRECISION			
DOS SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	72,3	77,9	68,1
BIGRAMA	71,2	75,8	64
TRIGRAMA	68,8	68,8	60,1
TF IDF	72,3	78	68,1
TF RFL	95,6	92,4	95,5

Tabla A.10 Precisión Hipótesis 2 para Falabella en dos sentidos.

PRECISION			
TRES SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	63,4	68,7	60
BIGRAMA	64,4	63,9	52,9
TRIGRAMA	67,3	60,6	47,8
TF IDF	63,4	68,6	59,9
TF RFL	91,5	86,8	95,4

Tabla A.11 Precisión Hipótesis 2 para Falabella en tres sentidos.

RECALL			
DOS SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	72,1	77,9	68,1
BIGRAMA	71,1	75,5	64,2
TRIGRAMA	68,6	68,6	59,3
TF IDF	72	77,9	68,1
TF RFL	95,6	92,2	95,1

Tabla A.12 Recall Hipótesis 2 para Falabella en dos sentidos.

RECALL			
TRES SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	60,5	68	59,5
BIGRAMA	59,5	63,7	51,3
TRIGRAMA	58,5	59,5	46,1
TF IDF	60,5	68	59,6
TF RFL	91,5	86,6	95,1

Tabla A.13 Recall Hipótesis 2 para Falabella en tres sentidos.

F			
DOS SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	71,9	77,9	68,1
BIGRAMA	70,7	75,5	63,9
TRIGRAMA	68	68,7	55,4
TF IDF	71,9	77,9	68,1
TF RFL	95,6	92,1	95,1

Tabla A.14 F_1 Hipótesis 2 para Falabella en dos sentidos.

F			
TRES SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	60,3	68,1	59,6
BIGRAMA	58,6	63,6	50,6
TRIGRAMA	57,5	59,3	46,2
TF IDF	60,3	68	59,5
TF RFL	91,5	86,6	95,1

Tabla A.15 F_1 Hipótesis 2 para Falabella en tres sentidos.

PRECISION			
DOS SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	77,5	82,2	75
BIGRAMA	76,7	75,3	77,9
TRIGRAMA	80,1	82,6	73,2
TF IDF	77,4	82,2	75
TF RFL	96,2	89,2	99,5

Tabla A.16 Precisión Hipótesis 2 para Ripley en dos sentidos.

PRECISION			
TRES SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	62,1	70,6	61,5
BIGRAMA	68	73,9	56,5
TRIGRAMA	72	70,5	54,2
TF IDF	62	70,6	61,6
TF RFL	86,2	84,1	94,2

Tabla A.17 Precisión Hipótesis 2 para Ripley en tres sentidos.

RECALL			
DOS SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	76	81,4	74
BIGRAMA	75	75	76,5
TRIGRAMA	70,1	82,4	71,6
TF IDF	76	81,4	74
TF RFL	96,1	89,2	99,5

Tabla A.18 Recall Hipótesis 2 para Ripley en dos sentidos.

RECALL			
TRES SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	61,4	70,3	60,8
BIGRAMA	66	68,3	56,9
TRIGRAMA	61,4	60,8	53,3
TF IDF	61,4	70,2	60,7
TF RFL	85,9	84	93,8

Tabla A.19 Recall Hipótesis 2 para Ripley en tres sentidos.

F			
DOS SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	75,7	81,3	73,8
BIGRAMA	74,7	75	76,1
TRIGRAMA	68,4	82,2	70,4
TF IDF	75,7	81,2	73,8
TF RFL	96,1	89,2	99,5

Tabla A.20 F_1 Hipótesis 2 para Ripley en dos sentidos.

F			
TRES SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	61,5	70,3	60,9
BIGRAMA	65,3	66,2	56,4
TRIGRAMA	60,4	58,1	52,9
TF IDF	61,4	70,2	60,9
TF RFL	86	84	93,8

Tabla A.21 F_1 Hipótesis 2 para Ripley en tres sentidos.

PRECISION			
DOS SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	65,1	65,3	63,5
BIGRAMA	59,5	63,4	62,4
TRIGRAMA	62,1	66,9	62,8
TF IDF	65,2	65,2	63,5
TF RFL	57,5	72,2	57,4

Tabla A.22 Precisión Hipótesis 3 training Ripley en dos sentidos.

PRECISION			
TRES SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	48	50	40,9
BIGRAMA	57	46,7	44,4
TRIGRAMA	56,6	45,5	35,9
TF IDF	48,1	50	40,9
TF RFL	45,3	47,3	55,5

Tabla A.23 Precisión Hipótesis 3 training Ripley en tres sentidos.

RECALL			
DOS SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	64,2	62,8	63,5
BIGRAMA	57,9	63,4	57,2
TRIGRAMA	61,2	64,7	56,9
TF IDF	64,2	62,8	63,4
TF RFL	57,3	67,3	57,3

Tabla A.24 Recall Hipótesis 3 training Ripley en dos sentidos.

RECALL			
TRES SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	47,6	47,4	41
BIGRAMA	46,2	43,1	41,8
TRIGRAMA	45,1	40	36,2
TF IDF	47,5	47,4	40,9
TF RFL	41,4	45,9	41

Tabla A.25 Recall Hipótesis 3 training Ripley en tres sentidos.

F			
DOS SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	63,6	61,3	63,5
BIGRAMA	56	63,4	52,4
TRIGRAMA	60,5	63,6	51,4
TF IDF	63,5	61,3	63,5
TF RFL	57,2	65,5	57,3

Tabla A.26 F_1 Hipótesis 3 training Ripley en dos sentidos.

F			
TRES SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	47,4	47,2	40,4
BIGRAMA	44,5	41,8	36,5
TRIGRAMA	41,6	39,2	35,6
TF IDF	47,4	47,1	40,4
TF RFL	36,6	44,5	34,4

Tabla A.27 F_1 Hipótesis 3 training Ripley en tres sentidos.

PRECISION			
DOS SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	66	71,1	57,2
BIGRAMA	67,1	67,5	70,1
TRIGRAMA	70,8	71,6	69,8
TF IDF	66,1	71	57,3
TF RFL	73,9	73,5	71,6

Tabla A.28 Precisión Hipótesis 3 training Falabella en dos sentidos.

PRECISION			
TRES SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	48,9	50	42,7
BIGRAMA	46,2	49,9	48,1
TRIGRAMA	47,6	51,6	38,3
TF IDF	48,9	49,9	42,8
TF RFL	42,4	55,5	42,7

Tabla A.29 Precisión Hipótesis 3 training Falabella en tres sentidos.

RECALL			
DOS SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	66	71	57,2
BIGRAMA	64,3	66,8	64,7
TRIGRAMA	65,6	68,1	63,3
TF IDF	66	71	57,3
TF RFL	72,8	67,8	71

Tabla A.30 Recall Hipótesis 3 training Falabella en dos sentidos.

RECALL			
TRES SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	47,9	49,8	42,7
BIGRAMA	44,4	49,4	44,1
TRIGRAMA	44,8	48,5	42
TF IDF	47,8	49,9	42,7
TF RFL	41,7	55	42,7

Tabla A.31 Recall Hipótesis 3 training Falabella en tres sentidos.

F			
DOS SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	66	71	57,1
BIGRAMA	62,8	66,5	62,1
TRIGRAMA	63,3	66,7	59,9
TF IDF	65,9	71	57,1
TF RFL	72,5	65,8	70,8

Tabla A.32 F_1 Hipótesis 3 training Falabella en dos sentidos.

F			
TRES SENTIDOS	NAIVE BAYES	SVM	J48
UNIGRAMA	47	49,7	42,7
BIGRAMA	40,5	48,7	39
TRIGRAMA	39,6	46,8	32,6
TF IDF	46,9	49,7	42,5
TF RFL	41,3	55	42,7

Tabla A.33 F_1 Hipótesis 3 training Falabella en tres sentidos.