UNIVERSIDAD DE CONCEPCIÓN
Facultad de Ingeniería
Departamento de Ing. Civil Informática y
Ciencias de la Computación

Profesor Patrocinante:
Dr. John Atkinson A.

Comisión:
Dr. Gonzalo Rojas
Dr. Leo Ferres

Análisis automático de opiniones para sistemas de micro-blogging

Diego Felipe Caro Alarcón

Informe de Memoria de Título
para optar al título de
Ingeniero Civil Informático

Abril de 2010

Resumen

En este trabajo se presenta el desarrollo y construcción de un prototipo para el análisis automático de opinión en mensajes de micro-blogging. Para ello se han estudiado dos enfoques: el análisis de opinión a nivel de documento y a nivel de frase. Ambos enfoques estudiados provienen del área de análisis de sentimiento, y utilizan diversas técnicas de aprendizaje supervisado y no supervisado. El problema fue afrontado utilizando técnicas de clasificación automática de texto, combinando ambos enfoques de análisis de opinión. El prototipo diseñado posee dos fases, la primera se encarga de detectar los mensajes que poseen una opinión, para en una segunda fase etiquetar la polaridad asocioda a cada uno de ellos. El dominio de aplicación utilizado en el desarrollo de este prototipo son los mensajes referentes a los candidatos a la presidencia de Chile 2010 del sistema de micro-blogging Twitter.

Agradecimientos

A mi familia, por ayudarme a ser quien soy y regalarme una educación *sin deuda*.

A mis amigos de la universidad y del laboratorio, por ayudarme a hacer realidad esta memoria, por aclararme dudas y proponer ideas cuando lo necesitaba.

A mis profesores, por darme las herramientas para ser un profesional.

Tabla de contenidos

Lista de tablas

Lista de figuras

Capítulo 1

Introducción

En este capítulo se presenta una introducción general al tema tratado durante el desarrollo de esta memoria de título, junto con los objetivos que se ha propuesto desarrollar y la organización general de este informe.

1.1 Introducción general

Desde tiempos remotos ha sido necesario saber lo que las personas opinan sobre diversos temas, como una forma de reflexión frente a lo vivido o para crear conciencia sobre temas de interés común. Históricamente esta tarea ha sido relegada a los medios de comunicación, dónde críticos periodistas plasman en periódicos su impresión sobre diversas temáticas, tanto en el área social como productiva y tecnológica, intentando determinar por ejemplo, que tan popular o bien recibidos son nuevos planes de gobierno, nuevas leyes o nuevos productos. Sin embargo, durante el último tiempo las nuevas tecnologías han permitido que personas que no pertenecen a estos medios, también tengan un espacio para dar opinión. Un ejemplo de esto son los llamados *"blogs"*, sitios webs que permiten la publicación de artículos en la web sin costo para el usuario, donde las personas pueden compartir opiniones en cualquier tópico.

Hoy en día el instrumento más común para conocer la opinión de la gente es la encuesta, pues con ella el experto puede medir y comprobar cuantitativamente lo que necesita conocer sobre algún producto o servicio. Sin embargo, es imposible desconocer el desfase temporal que posee esta herramienta desde el momento en que se realiza, hasta que se obtienen los resultados del estudio, problema que es inherente a la forma en que se realizan (entrevista persona a persona), que requiere de una coordinación y tiempo de planificación. Aprovechando las nuevas tecnologías, las encuestas también se han llevado a la Web, reduciendo los costos operativos y disminuyendo el desfase temporal. Sin embargo aún poseen el problema de que se debe incentivar a las personas a contestarlas.

Utilizando el fenómeno de expansión de la Web, tanto en páginas disponibles, como en el número de usuarios conectados y sirviendo contenidos, en los últimos años se han creado una serie de aplicaciones que utilizan técnicas vinculadas al lenguaje natural que permiten analizar las opiniones o sentimientos de las personas, utilizando los artículos y documentos disponibles tanto en blogs como en sitios personales [1]. Éstas aplicaciones básicamente intentan obtener la polaridad de una frase, un párrafo o un documento con respecto a un tema, con el fin de obtener una visión global de la percepción con respecto a un objeto o entidad (ej. productos, servicios, etc.) [2].

Uno de los problemas de algunas aplicaciones que analizan blogs, es que obtienen una estadística acumulada sobre las opiniones o sentimientos. Es decir, no se puede medir la variación que ha tenido la opinión de los usuarios durante algun periodo de tiempo, por ejemplo, la evaluación semanal de la aprobación a la *"Presidenta Michelle Bachelet"*.

En esta memoria se desarrolló e implementó un prototipo para analizar los mensajes de usuarios de un sistema de red social Web derivado de los blogs llamado micro-blogging, un tipo de servicio que permite a sus usuarios enviar y publicar mensajes breves (alrededor de 140 caracteres, aproximadamente 22 palabras), que generalmente sólo incluyen texto. Estos mensajes se muestran en la página de perfil del usuario, y también son enviados de forma inmediata a otros usuarios que han elegido la opción de recibirlas. Los usuarios que reciben estos mensajes también pueden responder o continuar la conversación, escribiendo otro mensaje al sistema, lo que permite una interacción a tiempo casi real de lo que "está pasando" en la Web.

Si bien algunos usuarios de micro-blogging, intentan responder a la pregunta ¿Qué estás haciendo? [1], usualmente se responde a muchas otras preguntas, pensamientos y conversaciones entre los usuarios, lo que permite responder subjetivamente a algunos temas de interés colectivo como política, calidad de productos o servicios, etc. Es importante destacar que estas opiniones llevan implícitamente un contexto asociado, es decir, son válidas en un determinado dominio y en un determinado tiempo.

El desafío que presenta esta memoria es desarrollar un prototipo para el análisis automático de opiniones en mensajes de micro-blogging, con el fin obtener la opinión global de un concepto sobre un dominio particular, en mensajes que hayan sido enviados en una ventana de tiempo: diario, semanal y mensual. En concreto, se analizarán los mensajes escritos en la plataforma de micro-blogging *Twitter*, que traten sobre las elecciones presidenciales 2009-2010 de Chile, y que contengan conceptos relacionados a los candidatos.

1.2 Objetivos

Objetivo general

El objetivo general de esta memoria de título es desarrollar un prototipo capaz de analizar la opinión global sobre un concepto, en un dominio y tiempo determinado, en textos de sistemas de micro-blogging.

[1] El 19 de noviembre de 2009, el servicio de micro-blogging Twitter cambió la pregunta a ¿Qué está pasando? http://blog.twitter.com/2009/11/whats-happening.html

Objetivos específicos

1. Investigar las técnicas y métodos de detección y clasificación de opinión en documentos de texto, para el análisis automático de opinión en mensajes de micro-blogging.

2. Desarrollar un método de clasificación de opinión, sobre un conjunto de mensajes de servicios de micro-blogging.

3. Construir un prototipo que implemente el método de clasificación de opinión y muestre la evolución de la opinión global a través de una ventana de tiempo (evolución diaria, semanal o mensual).

4. Evaluar el prototipo construido y analizar los resultados obtenidos.

1.3 Organización

Este documento se organiza de la siguiente forma: en el capítulo 2 se presenta un marco teórico sobre técnicas utilizadas para el análisis automático de opinión, el capítulo 3 presenta y explica el modelo construido para el prototipo de análisis de opinión, en el capítulo 4 se describen y discuten los resultados obtenidos de la aplicación del prototipo de análisis de opinión en mensajes de micro-blogging, y finalmente en el capítulo 5 se concluye acerca de los resultados y se proponen trabajos futuros respecto al tema desarrollado.

Capítulo 2

Marco Teórico

En este capítulo se presenta los conceptos y teorías que respaldan las decisiones y diseños utilizados en esta memoria de título. Además, se revisan los trabajos existentes sobre minería de opinión y análisis de sentimientos.

2.1 Fundamentos

La información textual disponible en el mundo puede ser categorizada entre dos tipos: hechos y opiniones. Los hechos son expresiones objetivas acerca de entidades, eventos y sus propiedades. En cambio, las opiniones usualmente son expresiones subjetivas que describen los sentimientos de las personas hacia entidades u objetos y sus propiedades [3].

En general, las opiniones pueden expresar sentimientos sobre cualquier cosa, por ej. productos, servicios, un evento o un tema. Algunos autores utilizan el término *objeto* para referirse a la entidad a la cual hace referencia una opinión. Asimismo, un objeto puede tener un conjunto de *componentes* (o *partes*), cada uno de los cuales posee un conjunto de *atributos*, definiendo una jerarquía basada en una relación *es-parte-de*. De esta forma, se ha asumido que las opiniones pueden ser estructuradas como un marco compuesto de los siguientes constituyentes [4]:

Opinion holder: Persona u organización que expresa una opinión (el autor de una columna periodística o una carta al director).

Objeto: Entidad que puede ser un producto (servicio, eventos, un tema de conversación, etc...) o una clase particular de algún dominio (ej.: modelo de un automóvil).

Componente: Miembro o parte del objeto que está siendo evaluado (motor, interior de un vehículo).

Atributo: Característica de un objeto respecto del cual se realiza una opinión (tamaño, color, diseño, etc.).

Opinión: Es el punto de vista (positivo o negativo), emoción o actitud que describe el *Opinion holder* sobre un componente o atributo del objeto en cuestión.

Orientación de una Opinión: Orientación de una opinión (polaridad u orientación semántica) sobre un componente o atributo de un objeto (*positiva*, *negativa* o *neutra*).

Para clarificar el rol de estos componentes, considere el siguiente ejemplo [4]: *"Hace algunos días compré una cámara Powershot. Tomé algunas fotografías con la cámara, creo que los colores son buenos, incluso con la utilización del flash"*. De esta opinión es posible observar que el objeto es la cámara *Powershot*, el componente son las fotografías, el atributo es el color, la opinión es buena (*"**buenos** colores'*) y la orientación es positiva.

Es importante señalar que en algunos dominios específicos (ej.: foros de discusión en la web, blogs, etc.), las personas suelen utilizar una jerarquía más pequeña al momento de opinar sobre un *objeto* en particular. Por ejemplo, en la revisión de algún producto en blogs especializados, se denomina *características* del producto tanto a los atributos, como a los componentes. También se utiliza el término tema (topic) o aspect (aspecto) [5] para referirse a las "características" en los casos donde el objeto de estudio es un tema (topic).

Opinión explícita e implícita

En cierto tipo de frases es posible encontrar directamente el tipo de característica que está siendo evaluada, es decir, existe un concepto que relaciona directamente la característica del objeto con la opinión del sujeto. A este tipo de opinión se les denomina *opinión explícita*. En los otros casos, donde la opinión no contiene directamente el tipo de característica evaluada, se le conoce como *opinión implícita* [3].

Por ejemplo, la frase *"La duración de la batería de mi iPhone es muy corta"*, indica que la característica evaluada del objecto "iPhone" es la "duración de la batería". Sin embargo, en otras frases como "Mi teléfono es muy grande" no es posible encontrar el tipo de característica evaluada, aunque claramente hace referencia al tamaño del teléfono (uso del adjetivo "grande").

Polaridad y orientación de una opinión

Las emociones han sido estudiadas en diversas áreas de la ciencia (por ejemplo, psicología, biología, etc.) [3]. Sin embargo, aún no existe un acuerdo acerca de cuáles son las básicas. Generalmente se presentan 6 emociones primarias [6]: amor, alegría, sorpresa, ira, tristeza y miedo, cada una con un nivel de intensidad distinto, y algunas de ellas están íntimamente relacionadas (alegría-tristeza).

Se pueden distinguir dos nociones importantes relacionados con las emociones. Por un lado está la noción de estado mental que poseen las personas, y por otro, las expresiones del lenguaje utilizadas para representar dicho estado mental. En el lenguaje usado para transmitir los sentimientos, es posible encontrar palabras o expresiones con un sentido negativo o positivo. A esto se le llama *polaridad* y el sentido que contengan éstas expresiones dependen del contexto donde se realice la opinión. No obstante, una expresión descrita utilizando palabras distintas con misma polaridad, pueden tener la misma orientación.

Algunos investigadores [3] han definido la polaridad (u orientación de la opinión) en el área de análisis de opinión como positiva, negativa o neutra con respecto a algún objeto. Por ejemplo, observe las siguientes frases:

- *"Hulk es una película basada en una excelente serie, está llena de acción,* ***me gustó mucho*** *la forma como la plantearon"*, es una **opinión *positiva***.
- *"La película Hulk no se parece en nada a la serie, creo que ha sido* ***una pérdida de dinero*** *venir al cine"*, es una **opinión *negativa***.

En ambas frases la "película Hulk" es el *objeto* al cuál se le realiza la opinión. Las palabras que definen la polaridad para la primera frase es "me gustó mucho" (positiva), y para la segunda lo son "una pérdida de dinero".

2.2 Trabajos Relacionados

Mucha de la investigación basada en análisis de sentimientos y minería de opiniones se basa en la clasificación de documentos que expresan una opinión positiva o negativa [2]. Este tipo enfoque conocido como *"clasificación de sentimientos a nivel de documento"*, considera todo el documento como una unidad, asumiendo que contiene una opinión. En este ámbito se han realizado estudios con técnicas de aprendizaje automático cuya principal diferencia es la utilización de distintas configuraciones para los features y el tipo de clasificador utilizado (Naive Bayes o de Support Vector Machine) [2], y también técnicas basadas en aprendizaje no supervisado [7] que utilizan algoritmos de información mutua.

Otro tipo de enfoque se encarga de analizar las frases presentes en un documento, asumiendo que no todas poseen una opinión. A este tipo de enfoque se le conoce como *"clasificación de sentimientos a nivel de frase"*, y básicamente intenta clasificar aquellas frases que contienen opinión (frases subjetivas), para posteriormente obtener la polaridad del documento [7].

2.2.1 Clasificación de opiniones a nivel de documento

La clasificación de opiniones a nivel de documento busca determinar si un documento presenta una opinión positiva o negativa con respecto a un objeto. Para esto se asume la

siguiente afirmación: "un documento con opinión d (por ejemplo, un review acerca de una película) expresa la opinión sobre un **objeto** y además es realizada por sólo un **opinion holder**". Ésta afirmación también se ha utilizado para analizar reviews de productos y servicios, pero puede no mantenerse para foros o blogs, pues normalmente estos textos presentan comparaciones entre varios productos (varios *objetos*) y pueden ser escritos por varios autores (varios *opinion holders*).

Técnicas de aprendizaje supervisado

A nivel de documento es posible encontrar dos enfoques para detectar la polaridad. Por un lado la utilización de técnicas de aprendizaje supervisado (Bayes, SVM, etc.) ha logrado una precisión de 83 % [2]. En la formulación del análisis de opinión como un problema de clasificación, se han definido dos clases: positivo y negativo, cada una representando una opinión positiva o negativa respectivamente.

Comúnmente, al analizar textos presentes en medios digitales como la Web, se utiliza un conjunto de términos o features para representarlo. Algunas de las investigaciones utilizan una representación matemática de los documentos, basada en un vector que contiene características, tales como frecuencia de palabras. Algunos tipos de representaciones básicas de documentos incluyen:

Frecuencia de palabras Al igual que en las técnicas usuales de recuperación de información, en el área de análisis de sentimientos es posible representar un documento como un vector de términos individuales. Tradicionalmente se ha utilizado el modelo *TF-IDF* (*Term Frequency Inverse Document Frequency*) [8], que consiste en un vector de pesos que indica la importancia de un término en relación a su distribución dentro de un conjunto de documentos.

En algunos enfoques [2] los mejores resultados se obtuvieron utilizando la presencia de términos, que corresponde a un vector que indica si un término está presente dentro de un documento. La mejora se reflejó en todos los clasificadores usados en la medición, con una precisión del 82 % para el clasificador SVM.

N-gramas y posición Un n-grama es una subsecuencia de n términos presentes en algún texto. Algunas investigaciones han reportado buenos resultados en el uso de presencia de unigramas para la clasificación de polaridad en el dominio de reviews de películas,

puesto que al realizar una opinión sólo basta utilizar al menos una palabra para cambiar el sentido de una oración [2].

Frase	**Bigramas**	**Trigramas**
Mi mamá me mima	{Mi, mamá}, {mamá, me}, {me, mima}	{Mi, mamá, me}, {mamá, me, mima}
El día está soleado	{El, día}, {día, está}, {está, soleado}	{El, día, está}, {día, está, soleado}

Tabla 2.1: Ejemplo de bigramas y trigramas

En algunos estudios también se ha incluido información acerca de la posición del n-grama dentro del documento (por ejemplo, mitad del documento o últimos párrafos), pues esto potencialmente podría tener efectos importantes si el documento posee algún tipo de estructura formal, como por ejemplo: introducción, desarrollo y conclusión [2] [9].

Por otro lado, el manejo de la negación es clave pues en algunos casos puede cambiar la orientación semántica de una palabra (*"esto es bueno"* versus *"esto no es muy bueno"*). Para tratar esta situación algunas investigaciones también sugieren la captura de negación a través del uso de bigramas y trigramas [2]. Otras técnicas utilizan clasificadores automáticos [10], donde se etiquetan previamente aquellas palabras que representan negación, para posteriormente cambiar el sentido de la opinión.

Técnicas de aprendizaje no supervisado

En análisis de opiniones también se han aplicado técnicas de aprendizaje no supervisado en la clasificación de sentimientos a nivel de documento. Por ejemplo, se ha desarrollado un algoritmo que se basa en la selección de frases subjetivas [7], para posteriormente obtener la orientación semántica de éstas usando una medida de información mutua con respecto a las conceptos "excellent" y "poor". En el estudio se obtuvo que un 84 % de los documentos del dominio de automóviles se le asignó correctamente la polaridad. Para el dominio de películas se obtuvo que un 65 % de los documentos obtuvo correctamente la polaridad (tanto positiva como negativa).

La selección de frases subjetivas (frases que contienen opiniones) se realiza a través de un filtro de aquellas que contienen adjetivos y adverbios, pues son buenos indicadores de subjetividad u opiniones [11, 12, 13]. Posteriormente se extraen dos palabras consecutivas,

	Primera palabra	**Segunda palabra**	**Palabra contexto (no extraída)**
1.	Adjetivo	Sustantivo	Cualquiera
2.	Adjetivo	Adjetivo	no Sustantivo
3.	Sustantivo	Adjetivo	no Sustantivo

Tabla 2.2: Reglas para extraer frases subjetivas escritas en inglés, extraídas desde [7]

donde alguna es un adverbio o un adjetivo, y la otra es una palabra del contexto. A éstas dos palabras se les asigna una etiqueta léxica del tipo Part of Speech (POS), conforme a una serie de patrones (ver figura 2.2.1). Por ejemplo la regla 2 indica que la primera palabra debe ser un adjetivo y que la segunda palabra debe ser un adjetivo, pero la tercera palabra (que no es extraída) no puede ser un sustantivo. Por ejemplo, en la oración *"This camera produces beautiful pictures"*, el segmento"beautiful pictures" sería extraído por la regla número 1 [3]. En la segunda etapa se estima la polaridad de las frases subjetivas dentro del documento, utilizando una medida de información mutua *PMI* (Pointwise Mutual Information) como la siguiente:

$$PMI(term_1, term_2) = log_2 \frac{Pr(term_1 \ \& \ term_2)}{Pr(termi_1)Pr(term_2)} \tag{2.1}$$

donde, $Pr(term_1 \ \& \ term_2)$ es la probabilidad de co-ocurrencia de los términos $term_1$ y $term_2$, y $Pr(term_1)Pr(term_2)$ es la probabilidad que los dos términos co-ocurran si es que son estadísticamente independientes. La división presente entre éstas dos medidas indica el grado de dependencia que poseen estadísticamente. Luego la Orientación de la Opinión (OO) de una frase se calcula en base a la asociación entre una palabra de referencia positiva (palabra "excellent") y otra de referencia negativa (palabra "poor"), pues en los sistemas de clasificación de productos o películas, es común evaluar mediante una escala de cinco niveles (usualmente como una serie de estrellas), donde la peor valoración corresponde a una estrella, que es representada en la palabra "poor", mientras que la mejor valoración corresponde a cinco estrellas, representada en la palabra "excellent". Así la orientación se define como:

$$OO(\text{ frase }) = PMI(\text{ frase , "excellent" }) - PMI(\text{ frase , "poor" }) \tag{2.2}$$

Otros métodos [7] utilizan un motor de búsqueda en la web (ej.: Altavista) para determinar la orientación de la opinión. En este tipo de enfoque se utiliza el operador HITS, que recupera documentos que coinciden con una consulta, y el operador NEAR, que filtra

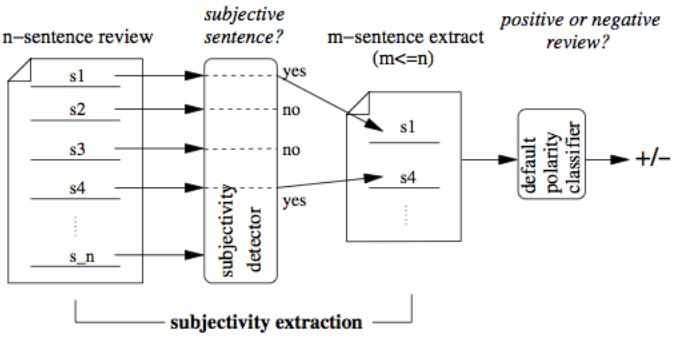

Figura 2.1: Arquitectura de análisis a nivel de frase, extraída desde [17]

documentos que contienen dentro de un radio de 10 términos, algún otro término. Tomando $hits(query)$ como el número de documentos recuperados, la ecuación 2.2 puede ser reescrita como:

$$OO(\text{ frase }) = log_2(\frac{hits(\text{ frase “excellent”})hits(\text{“poor”})}{hits(\text{frase NEAR “poor”})hits(\text{“excellent”})}) \tag{2.3}$$

Finalmente, el algoritmo de aprendizaje no supervisado etiqueta un review como *“recomendado”* si el promedio de los valores de la Orientación de la Opinión (OO) es positiva, o como *“no recomendado”* en otro caso.

2.2.2 Análisis a nivel de frase

La clasificación de polaridad a nivel de frase asume que sólo las frases subjetivas (FS) son portadoras de sentimiento u opinión [14] [15]. También se asume que una frase está escrita por solo un **opinion holder** y que la opinión se realiza sobre un **objeto**.

Este análisis se realiza en varias etapas (ver figura 2.1). Primero se seleccionan aquellas frases que son subjetivas, y luego se analiza la polaridad presente en las frases seleccionadas, para finalmente obtener la polaridad del texto. Para diferenciar entre una frases subjetiva y una objetiva, considere el siguiente ejemplo [16]:

Frase subjetiva *“Hulk es una película basada en una excelente serie, está llena de acción, me gustó mucho la forma como la plantearon”.*

Frase objetiva *“Hulk es una película de 120 minutos de duración. Tiene aproximadamente*

4 escenas de real acción donde el hombre verde pelea muy rudo".

En este ejemplo, ambas frases están referenciando a la película *Hulk*. Sin embargo, la primera constituye una percepción personal acerca de la película, mientras que la oración "me gustó mucho" muestra una opinión positiva. La segunda frase, en cambio, no entrega información respecto de quien realiza la opinión. Así, cuando un texto sólo contiene frases objetivas, no se puede concluir nada con respecto a la emoción implícita en la frase [16].

Detección de frases subjetivas

Análisis de frases basado en grafos Una técnica utilizada para abordar la detección de frases subjetivas es utilizar una combinación de algoritmos de optimización junto a técnicas de clasificación tradicionales. Algunos enfoques aplican algoritmos de minimización de costos sobre un grafo dirigido cuyos nodos representan las distintas frases presentes en un documento, y donde las aristas representan una función de costo asociada a etiquetar una frase como subjetiva u objetiva. El algoritmo minimiza una función de costo en las arista y una función que cuantifica que tan importante es etiquetar una frase en las dos clases (subjetiva y objetiva), obteniendo la ruta que contiene las frases de una clase [17]. Para la clasificación de la polaridad se utiliza un clasificador Naive Bayes a nivel de documento [2], obteniendo un 86,4 % de accuracy al seleccionar una frase subjetiva y posteriormente calcular su polaridad, usando como función de costo las probabilidades estimadas Naive Bayes de que una frase pertenezca a la clase subjetiva.

Técnicas de aprendizaje supervisado Trabajos recientes [18] utilizan técnicas de aprendizaje supervisado para seleccionar frases subjetivas, proponiendo un enfoques para obtener mensajes subjetivos y otro para obtener la polaridad en los mensajes que tienen opinión. El primer enfoque se basa en similaridad, explorando la hipótesis de que dentro de un contexto, las opiniones serán más similares a otras, utilizando la herramienta *SimFinder*, que realiza clustering por frases similares dentro de documentos de texto [19]. El segundo enfoque utilizado para obtener frases subjetivas involucra entrenar un clasificador Naive Bayes, usando un lexicón de sentimiento que almacena features adicionales como etiquetas léxicas del tipo Part of Speech (POS tags) y sentimiento en documentos con ejemplos de opiniones. Como features se incluyen palabras, bigramas y trigramas, además de POS tags y la orientación de cada palabra presente en la frase. Para la clasificación de polaridad se propone la utilización de un clasificador Naive Bayes múltiple $C_1, C_2, C_3, ..., C_m$, entrenando por separando

cada clasificador por feature $F_1, F_2, ..., F_m$. Las features son las mismas que las propuestas en el 2do método de detección de subjetividad (palabras, bigramas, trigramas, POS tags y polaridad). Se obtuvo entre un 80 % y un 90 % de precisión al obtener opiniones y un 50 % para clasificar textos objetivos.

Métodos basados en Lexicones La detección de subjetividad en frases también se ha estudiado a través de diccionarios de palabras etiquetadas por expertos que incluyen su polaridad. Uno de los trabajos propone la creación de patrones subjetivos a partir de palabras etiquetadas como subjetivas, débilmente subjetivas u objetivas, para posteriormente crear diccionarios semánticos tipo WordNet [20]. El método se basa en la hipótesis de que las frases objetivas o subjetivas se corresponden con algún patrón, lo que normalmente ocurre en textos formalmente escritos. En una primera etapa, se clasifican las frases presentes en un documento como objetiva o subjetiva utilizando un clasificador de alta precisión, que considera listas de palabras que han mostrado que poseen un alto grado de subjetividad [20], etiquetando una frase como subjetiva, si contiene dos o más palabras seleccionadas como fuertemente subjetivas. En cambio, para la clasificación de frases objetivas se verificó que una frase no contiene palabras etiquetadas como fuertemente subjetivas. Luego de etiquetar una frase como subjetiva se extrae un patrón subjetivo, que consiste en una regla sintáctica que incluye distintos tipos de etiquetas léxicas que comúnmente representan sentimiento. En una segunda etapa, se reentrena el sistema de detección con los nuevos patrones, obteniendo una precisión entre 71 % y un 85 % para identificar frases subjetivas. Usando el mismo enfoque, se ha explorado el uso de sustantivos para la creación de patrones [21]. En primer lugar se considera un diccionario inicial que se construye utilizando sustantivos, indicando si son fuerte o débilmente subjetivos, para posteriormente utilizar un clasificador del tipo Naive Bayes, logrando una precisión del 81 % en la identificación de las frases subjetivas.

Detección de polaridad

Análisis basado en grafos La utilización de adjetivos para obtener la polaridad (positiva o negativa) de una oración ha generado buenos resultados en la detección de opinión a nivel de frase [22]. Un método propuesto se basa en las relaciones léxicas de un término presente en WordNet [23]. Para esto, se utiliza la teoría diferencial semántica semántica de Osgood [24], como medida de distancia para evaluar la polaridad de una palabra.

Básicamente se define un grafo de adjetivos usando una relación de sinonimia entre ellos.

En el grafo, los autores definieron una medida de distancia $d(t_1, t_2)$ entre los términos t_1 y t_2, que entrega la distancia de la ruta más corta que conecta ambos términos (con $d(t_1, t_2) = \infty$ si t_1 y t_2 no están conectados). La orientación de un término está determinada por la distancia relativa a los términos bueno (orientación positiva) y malo (orientación negativa), definida por la función:

$$EVA(t) = \frac{d(t, malo) - d(t, bueno)}{d(bueno, malo)} \tag{2.4}$$

donde un adjetivo t se clasificará como positivo, si $EVA(t)$ entrega un valor positivo, y como negativo en caso contrario. El valor absoluto de la función EVA determina el nivel de fuerza que posee el adjetivo. La constante $d(bueno, malo)$ se utiliza para normalizar los valores de fuerza entre $[-1; +1]$. El resultado obtenido con esta técnica indica que en un 67,3 % de los casos evalúa correctamente la polaridad de una palabra.

Métodos basados en construcciones gramaticales Otra forma de determinar el sentimiento presente en un texto, es obtener la polaridad a partir de los pares de adverbios/adjetivos presentes dentro de una frase, para posteriormente usarlos en la clasificación de orientación según el tipo de conjunciones que los relacionan (por ej.: "pero", "y", etc...) [25]. Este enfoque [26] se basa en la identificación de adjetivos y adverbios para clasificar la polaridad de una frase, pues cuantifican la fuerza con la cuál actúa un adjetivo. El método clasifica adjetivos que pertenecen a cinco categorías: (1) adverbios de afirmación, (2) adverbios de duda, (3) adverbios de intensificación fuerte a adjetivos, (4) adverbios de intensificación débil y (5) adverbios de negación y minimización, las cuales poseen un puntaje asociado según el adverbio afirme al adjetivo. Finalmente la calidad de la clasificación se basa en el puntaje obtenido al aplicar una serie de reglas a los adjetivos presentes en una frase.

Métodos basados en Lexicones Una forma de detectar la polaridad es utilizar el conocimiento de algunos expertos para clasificar la intencionalidad de una palabra. En [15] se propone utilizar palabras etiquetadas por personas como positivas o negativas, con el fin de expandir ambas clases a través de las relaciones de antonimia y sinonimia presentes en WordNet. Esta técnica se basa en la hipótesis de que un sinónimo de una palabra conserva su orientación semántica, y que el antónimo invierte su polaridad. A este método se le aplica posteriormente, clasificadores como Naive Bayes o SVMs, logrando una exactitud de 66 %.

Otra forma de detectar la polaridad de una frase es asociar el contexto de una palabra con el posible significado que esta pueda tener. Así algunos proponen [27] la utilización de un lexicón con una taxonomía asociada a cada palabra, con el fin de clasificar el sentido o polaridad de un término en relación a las palabras que le acompañan. La hipótesis es que las "unidades atómicas" de las expresiones no son palabras individuales, sino que un grupo de ellas (por ej.: *"extremadamente aburrido"* o *"realmente no muy bien"*). Los atributos propuestos para evaluar términos o expresiones incluyen: Actitud (afecto, aprecio, juicio), graduación (fuerza, foco), Orientación (positiva, negativa), polaridad (marcada, no marcada), y forman parte de las features utilizadas para obtener la orientación semántica. El método obtuvo un accuracy de un 90.2 % utilizando las evaluaciones obtenidas por actitud y orientación, junto a la frecuencia de palabras, en la identificación de polaridad.

2.2.3 Análisis de múltiples niveles

Uno de los problemas presentes en la detección de polaridad a nivel de documento [2] o frase [17], es que no existe un método para desambiguar textos (por ej.: un párrafo) que contienen opiniones de diferente polaridad, pero que en su totalidad corresponden a una orientación semántica positiva (o negativa). Considere el siguiente párrafo [28]:

"Este es el primer reproductor Mp3 que uso ... Creo que suena bien ... Después de algunas semanas empezó a tener problemas con el conector de audífonos ... No quiero comprarme otro". (*Review de un reproductor Mp3 en Amazon.com*)

El extracto en general presenta una opinión negativa. Sin embargo, no todas las frases lo son, por ejemplo la segunda frase indica que el producto funciona de buena forma ("suena bien"). Las técnicas usuales a nivel de frase [17] no consideran el contexto en el que se desarrollan las opiniones, sólo realizan un análisis a nivel de oración, lo que no permite incluir el sentido del conjunto de frases en el párrafo .

El análisis de múltiples niveles [26, 28] utiliza la información disponible en la estructura implícita construida por la persona que realiza la opinión para de desambiguar a un nivel más general (por ej. a nivel de párrafo), utilizando los niveles más bajos (por ej.: a nivel de frase). Uno de los trabajos propuestos [28] se basa en el análisis de dos niveles, reuniendo la polaridad presente en todas las frases subjetivas de un documento, para posteriormente obtener una evaluación global utilizando dichos resultados. El método propuesto utiliza un grafo no dirigido para calcular la polaridad de un documento, cuyos nodos representan las

frases que son subjetivas, asignándole la polaridad correspondiente. La obtención de la polaridad a nivel de documento se realiza a través de una búsqueda de las etiquetas de polaridad que maximizan una función de puntaje sobre cada uno de los cliques presentes en el grafo. El modelo logra detectar correctamente la polaridad un 82,8 % de las veces.

2.3 Clasificación automática de texto

Algunas de las técnicas estudiadas en el campo de la detección de sentimiento utilizan clasificadores para detectar las opiniones presente en un documento [3]. Éstas técnicas se basan en una representación vectorial que caracteriza un documento para asignarle una etiqueta asociada al tipo de sentimiento (positivo o negativo) presente en la opinión. Las representaciones comúnmente usadas consisten en un vector de frecuencia de términos, entre los que se pueden incluir palabras, etiquetas léxicas, u otro tipo de representaciones tales como bigramas [2].

La clasificación de opinión se realiza utilizando el conocimiento adquirido a través de documentos con opinión previamentes etiquetados con una clase (positiva o negativa), mediante un modelo de conocimiento generado por alguna técnicas de clasificación automática. Algunas de las técnicas de clasificación automática utilizadas en la detección de opinión y polaridad son Naive Bayes, que utiliza la teoría de probabilidades de Bayes para asignar una etiqueta a un documento con una opinión, o las Support Vector Machines, que asignan una etiqueta mediante algoritmos de optimización [30].

2.3.1 Naive Bayes

Naive Bayes es un clasificador de tipo probabilístico que utiliza la teoría de Bayes [30] mediante la suposición de que las variables de los vectores a clasificar son independientes entre sí.

Sea C una variable aleatoria que puede tomar n valores $C_1, C_2, ..., C_n$, donde C_i denota la clase i y un vector m dimensional x, que consiste en la representación vectorial de un mensaje. Además, asumiendo que las probabilidades *a priori* $P(C_i)$ y que la probabilidad condicional $p(x|C_i)$ son conocidas y se distribuyen bajo una curva normal, al observar una variable aleatoria x, la probabilidad *a posteriori* $P(C_i|x)$ queda definida por:

$$P(C_i|x) = \frac{P(C_i)p(x|C_i)}{p(x)} \tag{2.5}$$

dónde $p(x)$ representa la evidencia (las features del mensaje) o conocimiento que se extrae desde un corpus, que se define de la siguiente manera:

$$p(x) = \sum_{i=1}^{n} p(x|C_i)P(C_i) \tag{2.6}$$

Luego, como $p(x)$ representa la evidencia o conocimiento acerca de las variables y además no depende de C, son consideradas una constante al calcular la probabilidad de que un mensaje pertenezca a una clase u otra. Posteriormente, aplicando repetidamente la definición de probabilidad condicional y asumiendo que la variables x_i son independientes entre sí, es posible aplicar la siguiente propiedad:

$$P(A|B) = P(A) \tag{2.7}$$

con lo que la ecuación 2.6 puede ser reescrita como:

$$P(C_i|x) = P(C_i) \prod_{j=1}^{n} p(x_j|C_i) \tag{2.8}$$

Posteriormente el clasificador se combina con una regla de decisión. Una comúnmente usada se basa en asignar la clase que posee un mayor valor de probabilidad *a posteriori*, como por ejemplo:

$$g(x_1, x_2, ..., x_m) = \arg \max_{i \in 1,..,n} P(C_i|x_1, x_2, ..., x_m) \tag{2.9}$$

2.3.2 Support Vector Machines

Uno de los clasificadores con mejor desempeño en el área de análisis de sentimientos es el conformado por las Máquinas de Soporte Vectorial (Support Vector Machines, SVMs) [17]. Las SVMs están conformadas por un conjunto de algoritmos de aprendizaje supervisado, definiendo un clasificador de tipo binario. El objetivo es encontrar un hiperplano que maximice el margen (o distancia) entre los vectores de las muestras más cercanas de dos clases [30]. Para

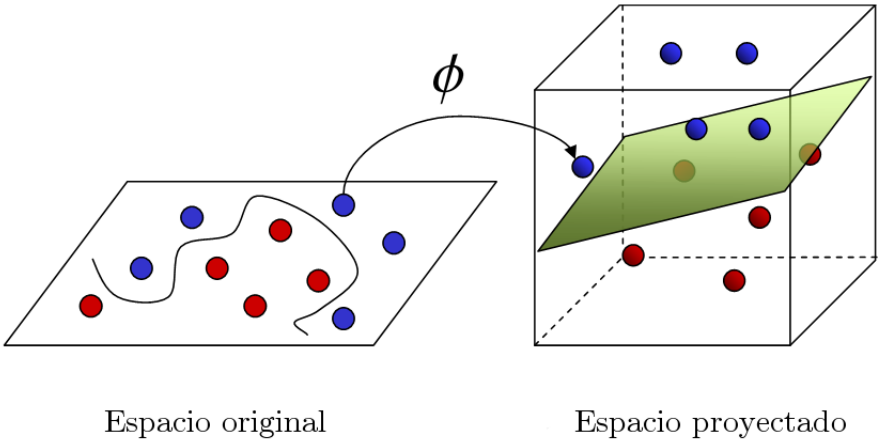

Figura 2.2: Proyección vectorial SVM

clasificar, la SVM asigna una etiqueta $y \in \{-1, 1\}$ para la clase 1 y clase 2 respectivamente, a una variable n-dimensional x, la que corresponde al resultado de la función discriminante definida por:

$$g(X) = w^t \phi(X) + b \tag{2.10}$$

donde $\phi(x) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ es una función de transformación del espacio vectorial de entrada, a uno de dimensión $m \gg n$, con el fin de hacer que las clases sean linealmente separables por un hiperplano (ver figura 2.2), en una dimensión distinta al del espacio vectorial original, puesto que no lo son en el espacio vectorial de origen. De la misma forma, w^t representa el hiperplano de dimensión m que divide ambas clases, y b la separación con respecto al eje de coordenadas.

Las nuevas muestras son clasificadas de acuerdo al resultado de la función discriminante $g(x)$, según la regla de clasificación:

$$g(X) = \begin{cases} > 0 & \text{para } y = 1 \\ < 0 & \text{para } y = -1 \end{cases} \tag{2.11}$$

Para la búsqueda del hiperplano óptimo, SVM intenta maximizar el margen entre las muestras más cercanas de las dos clases, minimizando la cantidad de muestras que no están dentro del margen, resolviéndolo como el siguiente problema de minimización:

$$\text{mín}\, Q(w, b, E_i) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} E_i \tag{2.12}$$

sujeto a $y_1(w^t\phi(X) + b) \geq 1 - E_1$ y con $E_i \geq 0, i = 1, .., n$, dónde $x_i \in \mathbb{R}^n$ corresponde a las muestras etiquetadas con $y_1 \in \{-1, 1\}$, E_i mide el error de entrenamiento de cada muestra y C es un factor de ponderación. Cuando el número de variables (n+1) a etiquetar es pequeño, el problema puede ser resuelto utilizando una técnica de programación cuadrática. Sin embargo en algunos casos dónde el número de variables es alto, es necesario convertir el problema en su dual equivalente dónde el número de variables corresponde a la cantidad de datos de entrenamiento N, enfocando el problema como una búsqueda de multiplicadores de Lagrange α_i, los que se obtienen maximizando la siguiente función:

$$\text{máx}\, Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j \phi(x_i)^t\phi(x_j) \tag{2.13}$$

sujeto a $\sum_{i=1}^{N} \alpha_i y_i = 0, 0 <= \alpha_i <= C$. Con lo que es posible calcular los parámetros que definen al hiperplano óptimo:

$$w = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i) \tag{2.14}$$

Ahora el problema consiste en evaluar la función $\phi(x_i)$, ya que no se conoce de forma explícita, o es dificil de evaluar. Sin embargo, una posibilidad es utilizar una técnica denominada *Truco de Kernel*, con la cuál es posible evaluar los vectores de entrada en una función de kernel $K(x_i, x_j)$, con lo que es posible reescribir la función de discriminación como:

$$g(X) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x_j) + b \tag{2.15}$$

donde $K(x_i, x_j)$ puede evaluarse utilizando un kernel de tipo lineal $K_l(x_i, x_j) = (x_i^t * x_j)$, o uno gaussiano $K_{RBF}(x_i, x_j) = exp(-\gamma\|x_i - x_j\|^2)$.

Para obtener buenos resultados en la clasificación se recomienda escalar los datos, normalizando los valores presentes en los vectores a un rango de $[0, 1]$ [31], cuya principal ventaja es evitar que los atributos con un mayor rango numéricos tenga una mayor preponderancia a los de menores rangos numéricos. Otra ventaja es evitar dificultades numéricas durante el

cálculo del hiperplano de separación.

2.4 Sistemas de micro-blogging

El micro-blogging es un tipo de servicio basado en el blogging que permite a sus usuarios enviar y recibir mensajes cortos (140 caracteres), que generalmente incluyen texto, imágenes o vínculos a otros sitios webs. Éstos mensajes están asociados a una línea de tiempo del usuario emisor y pueden ser vistos por cualquier usuario de Internet, o bien un grupo restringido seleccionados por el dueño de la cuenta[1].

El contenido de un micro-blog difiere de los sistemas tradicionales de blogging pues posee un tamaño de mensaje más reducido. En los sistemas de micro-blogging también es posible encontrar una red social, conformada por los usuarios que *siguen* las actualizaciones de otros (que análogamente forman un grafo dirigido cuyos aristas son la función "sigue a"). Un sistema de micro-blogging popular es Twitter, el cual posee las siguientes características:

Largo de mensaje Se pueden enviar 140 caracteres como máximo, donde se pueden incluir direcciones de sitios webs, referencia a otros usuarios y hashtags.

Followers Son aquellos usuarios que están suscritos a las actualizaciones de otro usuario.

ReTwitt Reenvío de un mensaje escrito por un usuario dentro de la red de amigos.

Como anteriormente se comentó, dentro de los mensajes es posible hacer referencia a usuarios, para esto se utiliza el símbolo "**@**" , por ej.: *"creo que* ***@rodrigozuniga*** *está perdiendo el tiempo"*. Los usuarios también han creado un sistema para etiquetar conversaciones llamada **hashtags**, identificados con el símbolo "#", con lo que básicamente se consigue identificar los mensajes de una conversación en torno a algún tema en particular. Por ejemplo, utilizando el buscador de mensajes de Twitter, es posible identificar la siguiente conversación en torno al hashtag **#chiledebate**[2] (mensajes ordenados desde el más reciente al más antiguo):

- @Eacuna88 dijo: ***#chiledebate*** *Le esta puro ayudando el publico po, que hagan preguntas buenas!! mal elejidos hoy*

[1]Información obtenida desde la sección de ayuda de Twitter.com `http://help.twitter.com/forums/10711/entries` (Marzo, 2010)

[2]Chile Debate es un programa de televisión dónde se entrevistan candidatos a la Presidencia de Chile

- @bastiangb dijo: *Viendo a @sebastianpinera en* ***#ChileDebate***
- @elenkos dijo: ***#ChileDebate*** *@tele13online: Cuales son sus Estudios y preparaciones academicas?? Opina que ser presidente deberia exigir ser profesional??*
- @jcrucesv dijo: *las preguntas pencas del público* ***#chiledebate***
- @jretamal dijo: *Viendo a tatan (@sebastianpinera) en* ***#ChileDebate*** *en #Canal13*
- @maenava dijo: ***#chiledebate*** *@tele13online alguien entendió su propuesta de protección social mezcló todo y no fue claro en su respuesta.*
- @gonzagz dijo: *y dele con su millón de empleos, eso es una verdadera UTOPIA* ***#chiledebate***

Una de las características más importantes de los sistemas de micro-blogging es que están diseñados para ser un medio de comunicación informal que permite de manera fácil y rápida dar a conocer lo que "está sucediendo" en un instante de tiempo. Es así como se han desarrollado aplicaciones para todos los sistemas operativos de computador y para diferentes dispositivos móviles, con el fin de estar disponibles cuando el usuario necesite hacerlo. Es importante destacar que la espontaneidad que proveen al usuario los sistemas de micro-blogging permite que los mensajes no se ajusten necesariamente a un tópico en particular, permitiendo usos para los cuales no fueron diseñados[3]. Dentro de los usos más conocidos de Twitter se tiene el seguimiento de eventos en directo donde poca gente tiene acceso, dando la oportunidad a la gente que no asistió a enterarse mucho antes de que aparezcan los pormenores del evento en prensa. También se ha utilizando en el intercambio de opiniones durante la transmisión de películas o debates en televisión (ej.: el primer y segundo debate para las elecciones presidenciales de Chile). Twitter también se ha usado en otros casos, donde por ejemplo los habitantes de Edmon, Oklahoma (EE. UU.) fueron capaces de ubicar el tornado que azotó dicha localidad, o para informar el estado de las carreteras luego del terremoto que azotó Chile el pasado 27 de febrero de 2010[4]. En resumen, los mensajes escritos tienen una ventana de tiempo dónde son válidos, y es sólo en ese contexto donde logran un real valor para el resto de los usuarios.

En el área de análisis de sentimientos originalmente se ha trabajado con documentos extensos en comparación a un mensaje de micro-blogging. Por ejemplo los estudios realizados

[3]Los Retwitts fueron creados por usuarios que necesitaban reenviar mensajes, característica que finalmente fue soportada oficialmente en Twitter

[4]Información extraída desde el artículo Twitter en Wikipedia español `http://es.wikipedia.org/wiki/Twitter` (Marzo, 2010)

por Pang [2] utilizan reviews de películas o el trabajo de Turney [7] que se basa en reviews de automóviles, utilizan un tipo de documento que resume la opinión de los usuarios en párrafos que claramente superan el límite de 140 caracteres, lo que determina la cantidad de información posible de transmitir por mensaje. El reducido tamaño de un mensaje de micro-blogging obliga al usuario a sintetizar lo que desea decir, obviando técnicas del discurso tales como repetir la idea utilizando sinónimos o dar ejemplos para clarificar la opinión al lector, restrigiendo la estructura del mensaje a la expresión de una idea utilizando una o más frases para aquello.

El reducido tamaño de los mensajes de micro-blogging impacta de dos maneras en las técnicas utilizadas para el análisis de sentimiento. En primer lugar, las representaciones asociadas al mensaje contienen menos información en comparación a un documento completo (discurso, cartas, reviews, etc...), lo que podría influir en la calidad de los resultados obtenidos al utilizar clasificadores automáticos de texto. En segundo lugar, una de las limitantes para ocupar técnicas que hacen uso de la estructura del lenguaje (etiquetas léxicas, pequeñas gramáticas o shallow parsing) es que los usuarios generalmente utilizan términos tales como modismos o expresiones que reflejan opinión, que no están incluídos en bases de conocimiento. Normalmente estas expresiones son comunes a un tipo de usuario y sólo pueden ser capturadas creando corpus con estos nuevas expresiones.

Capítulo 3

Análisis automático de opiniones basado en clasificación de texto

En este capítulo se describe y desarrolla un modelo basado en clasificación automática de texto para el análisis de opinión en sistemas de micro-blogging. También se describe un prototipo de análisis automático de opiniones para Twitter, donde se incluye el modelo propuesto.

3.1 Introducción

En esta memoria se construyó un prototipo computacional capaz de analizar mensajes de micro-blogging, con el objetivo de extraer la polaridad acerca del estado de satisfacción sobre un tópico en particular. El enfoque utilizado se basa en la extracción de opinión a nivel de documento, asumiendo que un mensaje de micro-blogging posee sólo un **opinion holder** que sólo expresa la opinión sobre un **objeto** en particular.

El prototipo realizado utiliza técnicas de clasificación automáticas como Naive Bayes y Support Vector Machines, dividiendo la tarea de análisis de opinión en dos sub-actividades: detección de subjetividad y detección de polaridad, con total independencia de recursos externos tipo Wordnet o bases de conocimiento. Los datos de entrada del prototipo son mensajes de Twitter obtenidos a través de una búsquedas por palabras claves que están relacionadas con un concepto y el periodo de tiempo en el que se debe realizar la búsqueda. Como salida, el prototipo entrega la cantidad de personas que opinan positivamente y negativamente dentro del periodo de tiempo, además de unas métricas que indican la tendencia que tienen los usuarios en relación al concepto.

Para la construcción del prototipo se adaptaron algunas técnicas utilizadas en el análisis de opinión, con el propósito de ajustarlas a las características del sistema de micro-blogging Twitter. Algunas de las características que se tomaron en cuenta incluyen:

1. **Longitud de los mensajes de micro-blogging** (140 caracteres): normalmente las técnicas de análisis de sentimiento utilizan un vector de features basados en frecuencia inversa de términos, pues hay términos que se repiten con frecuencia en documentos, sin embargo eso no sucede en los mensajes de micro-blogging. Para la representación de mensajes se propuso la utilización de una representación basada en un vector binario que sólo indica si un término está presente dentro del mensaje, asignando la misma importancia a cada uno de los términos presentes en el mensaje y además, simplifica el proceso de entrenamiento-clasificación pues no se necesita reescalar los datos para obtener buenos resultados en las SVM. En el procesamiento del corpus también se hicieron ajustes, incluyendo términos que se repiten al menos dos veces por cada clase (no es opinion, positiva o negativa) dentro del corpus. Además, en la etapa de clasificación con las SVM se utilizarán los kernels RBF y Lineal pues presentan un buen desempeño cuando el tamaño del corpus posee una menor cantidad de muestras en comparación a la cantidad de features por mensaje.

2. **Utilización de hashtags**: para capturar el sentido que los usuarios le dan al mensaje con los hashtags se extrajeron todas las etiquetas y se agregaron dentro del vector de *features*. También se agregó un término espacial llamado *hashtag* que indica la existencia de un hashtag dentro del mensaje.

3. **Asociación a un periodo de tiempo**: dado que los mensajes están muy ligados al contexto temporal donde fueron escritos, el prototipo provee unas métricas para conocer la cantidad de usuarios que mantienen y cambian su opinión durante un periodo de tiempo.

4. **Recuperación de mensajes**: dado que no es posible obtener todos los mensajes que son enviados a Twitter, el prototipo incluye un mecanismo de extracción de mensajes mediante la *API* . Esta realiza una búsqueda por palabras claves dentro de los mensajes enviados la última semana, o bien desde una fecha posterior.

3.2 Arquitectura del sistema

La arquitectura para el prototipo de detección de subjetividad y de polaridad se basa un sistema *backend - frontend*, que consiste a grandes rasgos en dos procesos. El *backend* se encarga de realizar el análisis de opinión de los mensajes de micro-blogging, y es un proceso que está constantemente ejecutando, mientras que el proceso *frontend* presenta un resumen que contiene la información de los mensajes procesados cada vez que un usuario lo requiera.

El *backend* consiste en una serie de tareas, dividas básicamente en recuperación, limpieza de mensajes, y la clasificación de los mismos (ver figura 3.1), estructurada de la siguiente forma:

1. **Recuperación de mensajes y sus metadatos:** Este módulo se encarga de recuperar los mensajes asociados a palabras claves correspondientes al dominio de trabajo. Aquí se extraen los metadatos asociados a los mensajes, por ejemplo el usuario emisor, la fecha cuando se envió y el concepto asociado.

2. **Limpieza de texto del mensaje:** Se encarga de decodificar entidades HTML presentes en los mensajes, extraer tags HTML, y limpiar URLs.

3. **Procesamiento y selección de tokens:** Se seleccionan los tokens que representan palabras, *hashtags* y nombres de usuarios, mediante la utilización de expresiones regulares

(ver tabla 3.1).

4. **Selección de términos:** Se obtiene la etiqueta léxica (POS) de las palabras seleccionadas en el paso anterior, y filtra aquellos tokens que son irrelevantes para el análisis de opinión. Posteriormente estos términos se agrupan en unigramas o bigramas.

5. **Creación de modelo vectorial:** Crea el modelo vectorial asociado a los features del mensaje procesado. Para esto, se utilizan de tres modelos: frecuencia de términos (tf), frecuencia y presencia de términos.

6. **Detección de subjetividad:** Se encarga de seleccionar aquellos mensajes que contienen una opinión.

7. **Detección de polaridad:** Se encarga de obtener la polaridad de aquellos mensajes que en la etapa de subjetividad se detectaron como subjetivos, es decir, los que poseen una opinión.

Token	Expresión regular
Hashtag	#[A-Za-z0-9]+
Nombre de usuario	@[A-Za-z0-9]+

Tabla 3.1: Expresiones regulares para capturar nombres de usuario y hashtags

Los componentes de *detección de opinión* y de *detección de polaridad* se realizan en base a dos tareas:

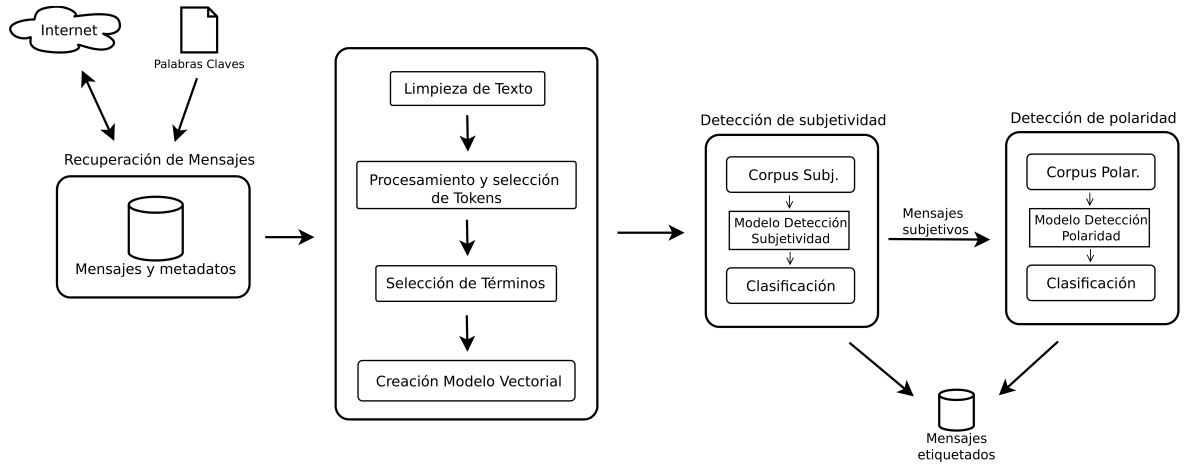

Figura 3.1: Arquitectura del proceso *backend* del prototipo de análisis de opinión

1. **Entrenamiento:** Se genera un modelo de detección de subjetividad y detección de polaridad, con mensajes previamente etiquetados, con el fin de ser utilizado por técnicas de clasificación. Para la detección de subjetividad, el clasificador se entrena utilizando las etiquetas **not** y **yes**, que representan los mensajes que no son una opinión y los que sí lo son, respectivamente. Para la detección de polaridad (u orientación de la opinión), se utilizaron las clases **pos** (positivo) y **neg** (negativo), cada una representando el tipo de sentimiento asociado a un mensaje.

2. **Clasificación:** Se asigna una etiqueta a un mensaje, ya sea para determinar si contiene o no un opinión, y si es que contiene una opinión, de qué tipo es. Para asignar una etiqueta, el clasificador utiliza los modelos creados para la detección de subjetividad y detección de polaridad.

El *frontend* consiste en la creación y visualización de un infome resumen que contiene las métricas que describen el comportamiento de los usuarios en un periodo de tiempo (ver sección 3.4). El filtrado de mensajes se realiza a través de una búsqueda por palabras claves dentro de la base de datos de mensajes etiquetados, restringiendo la selección a un período de tiempo en particular. El *frontend* (ver figura 3.2) está conformado por las siguientes etapas:

1. **Filtro de mensajes:** Obtiene un subconjunto de mensajes etiquetados por el *backend*, cuya selección está restringida a un periodo de tiempo y un conjunto de palabras claves que debe estar presente dentro de los mensajes.

2. **Cálculo de Métricas:** Realiza cálculos para obtener la tendencia de las opiniones de usuarios en un periodo de tiempo determinado.

3. **Creación de informe resumen:** Muestra mediante un gráfico de tiempo la variación de las métricas

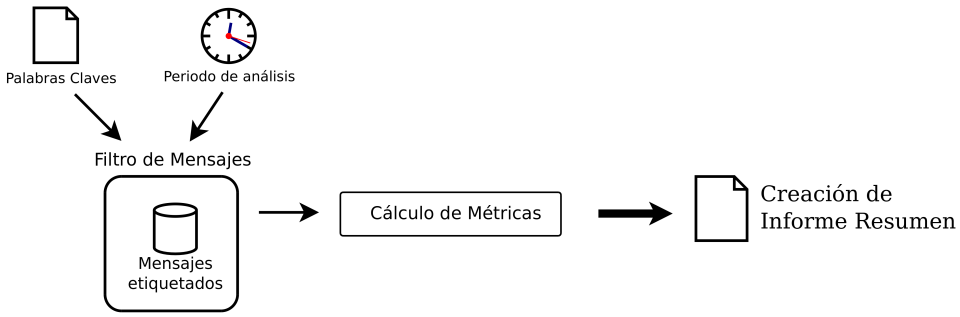

Figura 3.2: Arquitectura del proceso *frontend* del prototipo de análisis de opinión

La interfaz de usuario para acceder al *frontend* consiste en un formulario web para ingresar las palabras claves y una página web para visualizar las métricas del informe resumen.

3.3 Representación de mensajes

Existen diferentes mecanismos para representar textos, que incluyen: frecuencia de términos, presencia de términos o frecuencia inversa de términos, donde términos se refiere a una palabra o un grupo de ellas (expresada, por ejemplo, en n-gramas). Para esta memoria se utilizó el modelo de presencia de términos, el cual contiene vectores que sólo indican la presencia de un término $t_{i,j}$ en un mensaje m_j:

$$a_{(i,j)} = \begin{cases} 0 & \text{, para } t_{(i,j)} = 0 \\ 1 & \text{, para } t_{(i,j)} >= 1 \end{cases} \tag{3.1}$$

donde, $a_{(i,j)}$ representa la existencia de un término, $t_{(i,j)}$ es el número de veces que la palabra t_i aparece en el mensaje m_j. Esta codificación permite asignarle la misma importancia a cada palabra presente dentro del mensaje, pues sólo indica si un término está presente dentro del mensaje a clasificar, y además permite realizar un paso menos en la etapa de entrenamiento-clasificación, pues no es necesario normalizar los vectores cuando se utilizan las SVM [31].

Para el desarrollo de los experimentos se probaron otras formas de codificar los mensajes, utilizando la medida *TF* (Term Frequency), la cual cuantifica la cantidad de veces que se repite un término en un mensaje, y la medida *TF-IDF* (Term Frequency - Inverse Document Frequency) que penaliza aquellos términos con mayor frecuencia dentro de un corpus (pues entregan menor información por mensajes), definida cómo:

$$a_{(i,j)} = t_{(i,j)} * log(\frac{N}{n_j}) \tag{3.2}$$

donde, $a_{(i,j)}$ representa el valor *TF-IDF* de un término, $t_{(i,j)}$ es el número de veces que la palabra t_i aparece en el mensaje m_j , n_j es el número de mensajes dónde aparece la palabra t_j y N es el número total de textos.

3.4 Métricas de visualización de cambios

Una de las principales características de los servicios de micro-blogging es que los mensajes están altamente asociados a una línea de tiempo, permitiendo generar entre los usuarios discusiones mientras se realiza otra actividad. Por ejemplo, enviar mensajes a servicios de micro-blogging mientras se mira un programa de televisión, con el objetivo de comentar lo que sucede en el programa. Las aplicaciones desarrolladas hasta ahora se han limitado a proveer una línea de tiempo con la cantidad de opiniones positivas y negativas en un periodo de tiempo[1].

En el desarrollo de esta memoria también se ha definido una métrica para verificar el cambio en las tendencias de polaridad en los usuarios de micro-blogging, con el fin de visualizar un cambio en la tendencia de los usuarios. Esto permitiría comprobar los cambios de opinión que generan intervenciones realizadas en los medios de comunicación (por ejemplo, una nueva campaña de marketing) en los usuarios del sistema.

La métrica muestra un análisis realizado en un periodo de tiempo (un día, una hora, un mes, etc...), bajo los siguientes supuestos:

1. Si un usuario (*opinion holder*) opina de forma contradictoria, es decir, envía un mensaje positivo y otro negativo en el mismo periodo de tiempo, los mensajes enviados por el usuario en ese período se descartan del análisis.

2. Sólo se considera la polaridad a nivel de usuario hacia un *objeto*, y no a nivel de mensaje.

Sea P_i el conjunto de usuarios que opinan positivamente sobre un *objeto* en el periodo i de tiempo, sea N_i el conjunto de usuarios que opinan negativamente sobre un *objeto* en el periodo de tiempo i. Ahora, asumiendo que i puede tomar valores discretos consecutivos de t unidades de tiempo (un minuto, una hora, un día, etc.), se pueden definir las siguientes métricas:

Usuarios que mantuvieron una opinión positiva: $M_{pos} = |P_i \cap P_{i+1}|$

Usuarios que mantuvieron una opinión negativa: $M_{neg} = |N_i \cap N_{i+1}|$

Usuarios nuevos con opinión positiva: $N_{pos} = |P_{i+1}| - |P_i \cap P_{i+1}|$

[1]Para ver una lista de aplicaciones de análisis de sentimiento en Twitter visite `http://spreadsheets.google.com/ccc?key=tVfLhBxao70Fk9bAtebp_xQ&zx=1272397799411`

Usuarios nuevos con opinión negativa: $N_{neg} = |N_{i+1}| - |N_i \cap N_{i+1}|$

Usuarios que cambiaron de opinión positiva a negativa: $C_{pos2neg} = |P_i \cap N_{i+1}|$

Usuarios que cambiaron de opinión negativa a positiva: $C_{neg2pos} = |N_i \cap P_{i+1}|$

Con estas medidas es posible construir un gráfico sobre el eje tiempo que muestra la evolución de las métricas, reflejando los distintos tipos de cambio de opinión que pueden darse en el transcurso de dos periodos de tiempo.

3.5 Descripción del corpus

Los datos de entrada para las tareas de entrenamiento y clasificación del prototipo provienen de un corpus creado especialmente para esta memoria, que consiste en mensajes etiquetados por usuarios de Twitter. Para esto se desarrolló una interfaz web dónde se muestran mensajes recuperados utilizando la API publicada por Twitter[2], que se utiliza para buscar mensajes que contienen ciertas palabras claves en un idioma en particular (en este caso, español).

El dominio escogido para recolectar el corpus fue las elecciones presidenciales de Chile 2009, por lo tanto se utilizaron términos asociados a cada uno de los candidatos y algunos *hashtags* utilizados por los usuarios de Twitter (ver Tabla 3.2).

#debate09	marco enriquez	frei	piñera	arrate
#chiledebate	enriquez ominami	@noticiasfrei	@sebastianpinera	@JorgeArrate
#debateanatel	@marco2010	#frei	#piñera	@arrate2009
chile candidato	#meo			#arrate

Tabla 3.2: Términos utilizados para buscar mensajes en Twitter

La recolección de mensajes utilizando los términos asociados al dominio se inició el día 8 de agosto de 2009 y finalizó el día 29 de septiembre de 2009, logrando capturar 56456 mensajes (el número real de mensajes extraídos es mayor al presentado, pues se realizaron reducciones de mensajes duplicados por *ReTwitts*). El conjunto de mensajes a etiquetar se redujo a una muestra aleatoria de un 10 % (cinco mil mensajes), asumiendo que cada mensaje tiene la misma probabilidad de ser escogido. Finalmente se logró etiquetar un total de **1112** mensajes, conformando un corpus de:

[2]Más información en el sitio oficial de la API de Twitter en `http://apiwiki.twitter.com/`

- **476** mensajes etiquetados como **objetivos** (que no poseen una opinión).
- **251** mensajes etiquetados como subjetivos con polaridad **positiva**.
- **385** mensajes etiquetados como subjetivos con polaridad **negativa**.

3.6 Clasificación de texto

Para clasificar mensajes subjetivos y detectar polaridad se utilizaron dos clasificadores con un buen desempeño en el área de análisis de polaridad a nivel de documento [2]. Básicamente estos clasificadores asignan una etiqueta C a una representación vectorial de un mensaje (features), en base a datos de entrenamiento que ya se poseen mediante un corpus. La clasificación se realiza mediante el conocimiento adquirido de las muestras de las etiquetas, haciendo uso en el caso de Naive Bayes de probabilidades, y en el caso de SVM de algoritmos de optimización y álgebra lineal. Se propone la utilización de estos clasificadores pues en estudios anteriores han logrado el mejor desempeño tanto para la detección de frases subjetivas y polaridad [2, 16].

Los ajustes realizados para el clasificador Naive Bayes consisten en definir la probabilidad a priori a partir de los porcentajes que posee el corpus para los mensajes de cada clase, tanto para el análisis de subjetividad como de polaridad. Para la utilización del clasificador basado se definieron los parámetros $C = 1$ y el error de clasificación mínimo fue definido a 0,001. Además, se utilizaron los kernels Lineal y RBF (parámetro $\gamma = \frac{1}{Num.Features}$), pues han demostrado buenos resultados cuando los datos de entrenamiento poseen una mayor dimensión en comparación al número de muestras disponibles [31]. En este mismo sentido, es importante se nalar que la codificación de mensajes a un vector de presencia de términos hace menos complejo el proceso de entrenamiento-testing para las SVM, pues como el vector de features es binario, no se necesita escalar los datos para obtener un buen desempeño [31].

Capítulo 4

Experimentos y Resultados

En este capítulo se describen los diversos experimentos realizados con el modelo desarrollado, los resultados obtenidos y el análisis de los mismos.

4.1 Metodología de pruebas

La evaluación del prototipo propuesto considera la realización de algunas pruebas para la detección de opinión y polaridad, con el fin de obtener la mejor configuración para cada una de las técnicas y features utilizados para representar y clasificar un mensaje. Para lograr estos objetivos se realizaron los siguientes experimentos:

1. **Evaluación individual de cada clasificador:** Se varían los parámetros de configuración de cada clasificador y se evalúa cuál combinación entrega los mejores resultados en la detección de opinión y clasificación de polaridad en un mensaje de micro-blogging.

2. **Evaluación con el uso de distintos *features*:** Se utilizan distintos tipos de combinaciones de *features*, por ej: sólo palabras, sólo POS tags o palabras + POS tags, combinando los términos en unigramas o bigramas, bajo un modelo de frecuencia de términos (TF), frecuencia inversa de términos (*TF-IDF*) o presencia de términos, tanto para la detección de opinión como para la detección de polaridad.

Todas las pruebas descritas fueron realizadas utilizando un procedimiento de validación cruzada del tipo k-fold, con $k = 5$, con el fin de dividir el conjunto de mensajes de cada clase en k subconjuntos de igual tamaño. Se utilizó $k = 5$ pues es el valor recomendado para realizar la validación cruzada en datos de alta dimensión donde se posee una baja cantidad de muestras [31].

Los resultados obtenidos por las distintas configuraciones de cada clasificador se analizan posteriormente con medidas estándares en problemas de **clasificación**, utilizando para ello una matriz de confusión. Una matriz de confusión es una tabla que contiene información acerca de las clases reales y las predicciones realizadas por un sistema de clasificación. Las distintas configuraciones para cada clasificador se evaluarán utilizando las medidas *Precision* y *Accuracy*, definidas en términos de la cantidad de total muestras N, cantidad de muestras de la clase positiva N_{pos}, cantidad de muestras de la clase negativa N_{neg}, cantidad de muestras etiquetadas como positiva C_{pos} y cantidad de muestras etiquetadas com negativas C_{neg}:

Accuracy: Promedio de resultados correctamente clasificados:

$$Accuracy = \frac{C_{pos} + C_{neg}}{N} \tag{4.1}$$

Precision: Proporción de muestras de clase k ($k \in \{pos, neg\}$) que fueron correctamente clasificadas:

$$Precision = \frac{C_k}{N_k} \tag{4.2}$$

En la evaluación de resultados para la detección de mensajes subjetivos se considera clase positiva a aquellos mensajes que pertenecen a la clase **yes** (mensajes subjetivos) y clase negativa a los de etiqueta **not** (mensajes objetivos). En la detección de polaridad se considera clase positiva a aquellos etiquetados con la etiqueta **pos** (polaridad positiva) y clase negativa a aquellos que son etiquetados con la clase **neg** (polaridad negativa).

En particular, la medida *Precision* será utilizada para evaluar los resultados de la detección de subjetividad, midiendo la tasa de éxito de clasificación para los mensajes correctamente etiquetados como subjetivos. Para la detección de subjetividad se utilizará la medida *Accuracy*, que medirá el promedio de mensajes correctamente clasificados con polaridad positiva o negativa.

Procesamiento del corpus

El procesamiento de los mensajes presentes en el corpus considera el etiquetado léxico (POS tags) utilizando la herramienta *Freeling*. Luego, se restringió la cantidad de unigramas y bigramas, seleccionando aquellos que aparecen al menos 2 veces dentro del corpus; esto con el fin de capturar aquellos términos que entregan información para al menos una clase.

4.2 Ajuste de parámetros en detección de subjetividad

Los clasificadores utilizados para la detección de subjetividad en mensajes de microblogging permiten realizar predicciones basándose en ejemplos de mensajes ya etiquetados como opiniones dentro de un corpus. Con el fin de obtener el mejor la mejor configuración para la detección de subjetividad, se realizaron pruebas variando el tipo de *feature* que mejor representa a los mensajes etiquetados como subjetivos dentro del corpus. La medida *Precision* fue escogida para evaluar el desempeño de cada configuración, pues la detección de subjetividad es un paso intermedio en el análisis de opinión, cuya salida sirve como entrada

a otro módulo, por lo tanto lo que más interesa es que tener la mayor certeza de que lo que se predice efectivamente corresponde a la realidad.

Las configuraciones de prueba se basan en la elección de términos basada en palabras y etiquetas POS, combinadas bajo un modelo de frecuencia de términos (TF), frecuencia inversa de términos (IDF) y presencia de términos (PRES).

Las pruebas para la detección de subjetividad se realizaron en un corpus que consta de **636** mensajes etiquetados como subjetivos (unión de los mensajes con polaridad positiva y negativa, clase **yes**) y **476** mensajes que no contienen opinión (mensajes objetivos, clase **not**).

Clasificador Naive Bayes

En las pruebas realizadas con el clasificador Naive Bayes se evaluaron las distintas configuraciones de features modificando la cantidad de features utilizados para la clasificación, considerando:

- Unigramas de palabras, etiquetas POS y la combinación de ambas.
- Bigramas de palabras y etiquetas POS.
- Combinación de unigramas y bigramas de palabras.

Los resultados de la figura 4.1 muestran que la configuración con la que se obtiene mejores resultados es la de combinación de unigramas de palabras y etiquetas POS, logrando una precisión de 0,72 para los modelos basados en frecuencia de términos, frecuencia inversa y presencia.

Clasificador SVM

El desempeño del clasificador SVM está directamente relacionado con el tipo de Kernel utilizado para evaluar los vectores de soporte. Para la selección de la mejor configuración de este clasificador se realizaron pruebas para dos tipos diferentes de Kernel: Lineal y Gaussiano (RBF), con un margen de error de entrenamiento de 0,001.

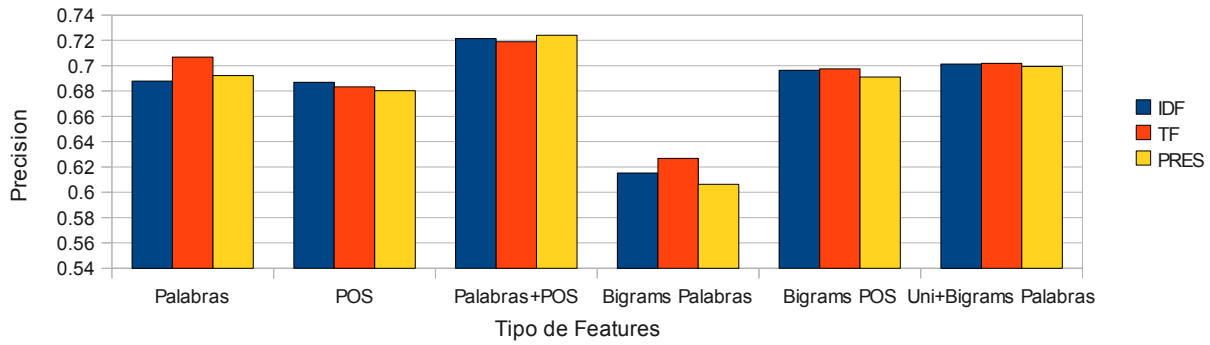

Figura 4.1: Precision para clasificador Naive Bayes en Detección de Subjetividad

El mejor resultado obtenido al utilizar un Kernel lineal (figura 4.2) logra una precisión de 0,71 utilizando palabras y etiquetas POS, tanto en frecuencia de términos como la frecuencia inversa, resultado muy similar al obtenido por el clasificador Naive Bayes en su mejor configuración. Los resultados obtenidos al utilizar Kernel Gaussiano (figura 4.3) arrojan que la mejor configuración logra 0,72 de precisión al utilizar palabras y etiquetas POS, logrando la mejor configuración para el clasificador SVM.

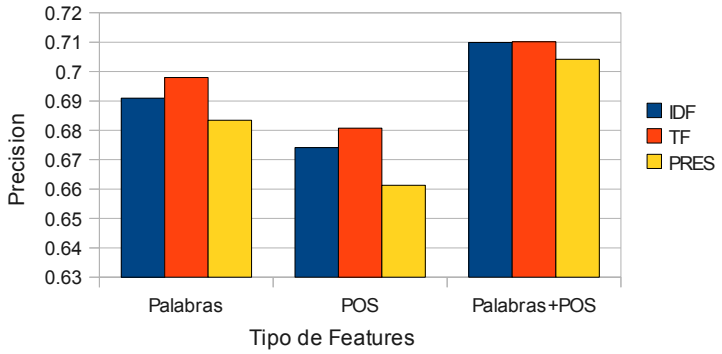

Figura 4.2: Precision para SVM con Kernel Lineal en Detección de Subjetividad

La evaluación del mejor clasificador para detectar mensajes subjetivos considera el valor medida *Accuracy* entre las configuraciones que lograron el desempeño más alto en el clasificador de Naive Bayes y SVM. El mejor resultado obtenido a través de la medida *Accuracy* muestra que la mejor configuración se obtiene mediante la utilización de SVM con un Kernel Gaussiano, utilizando frecuencia de términos de palabras y etiquetas POS, con un *Accuracy* de 0,68 y para Naive Bayes de 0,67.

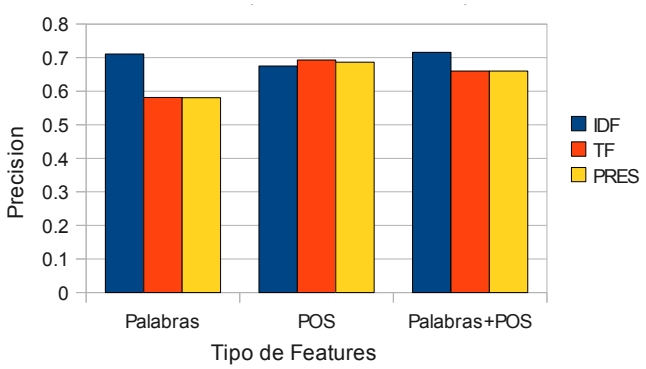

Figura 4.3: Precision para SVM con Kernel Gaussiano en Detección de Subjetividad

4.3 Ajuste de parámetros en detección de polaridad

El ajuste de parámetros para la detección de polaridad en mensajes subjetivos se basa en la evaluación de la medida *Accuracy* obtenida al utilizar los clasificadores Naive Bayes y SVM. Se utiliza la medida Accuracy para evaluar las distintas configuraciones pues esta medida incluye información global asociada a las dos clases que el clasificador debe predecir. En contraste a la detección de subjetividad, en la detección de polaridad interesa que las clases positivas y negativas obtengan en general un buen desempeño, pues las predicciones realizadas por el clasificador escogido serán utilizados directamente por las métricas de análisis de opinión propuestas en el capítulo 3.

Las de configuraciones de las *features* para la clasificación se basan en la elección de los siguientes términos:

- Unigramas de palabras
- Bigramas de palabras
- Unigramas + bigramas de palabras

Para las tres configuraciones de términos utiliza un modelo de frecuencia (*TF*), de frecuencia inversa de términos (*TF-IDF*) y de presencia de términos. Se utilizó bigramas tanto para el clasificador Naive Bayes cómo para SVM, pues con eso se espera capturar la información presente en algunas combinaciones gramaticales que incluyen negación (por ej.: "***no*** *quiero* que salga de presidente").

Las pruebas para la detección de polaridad se realizaron en un corpus que consta de **251**

mensajes etiquetados con opiniones positivas (clase positiva) y **385** mensajes con opiniones negativas (clase negativa).

Clasificador Naive Bayes

En las pruebas realizadas para el clasificador Naive Bayes se utilizó como probabilidad a priori, los porcentajes respectivos a la cardinalidad de mensajes que posee el corpus en cada clase.

Las pruebas obtuvieron resultados (figura 4.4) que indican la mejor configuración con 0,71 de Accuracy. Se logró el mismo desempeño al utilizar unigramas y unigramas+bigramas de palabras utilizando frecuencia de términos.

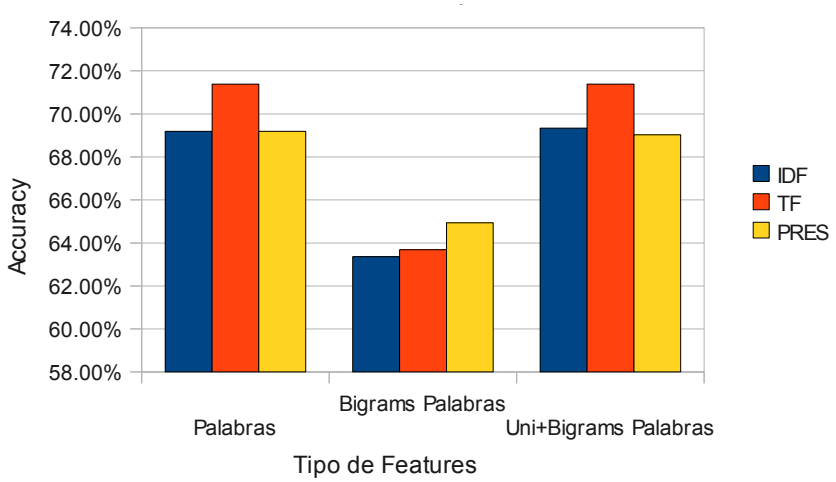

Figura 4.4: Accuracy para clasificador Naive Bayes en Detección de Polaridad

Clasificador SVM

El desempeño del clasificador SVM está directamente relacionado con el tipo de Kernel utilizado para evaluar los vectores de soporte. Para la selección de la mejor configuración de este clasificador se realizaron pruebas para dos tipos diferentes de Kernel: Lineal y Gaussiano (RBF), con un margen de error de entrenamiento de 0,001.

El mejor resultado obtenido al utilizar un Kernel lineal (figura 4.2) logra un Accuracy de 0,67 utilizando unigramas de palabras usando término de frecuencias. Al utilizar un Kernel Gaussiano (figura 4.3) se obtuvo un Accuracy de 0,69 al utilizar frecuencia inversa de

términos combinando unigramas y bigramas de palabras, obteniendo el mejor resultado para el clasificador SVM.

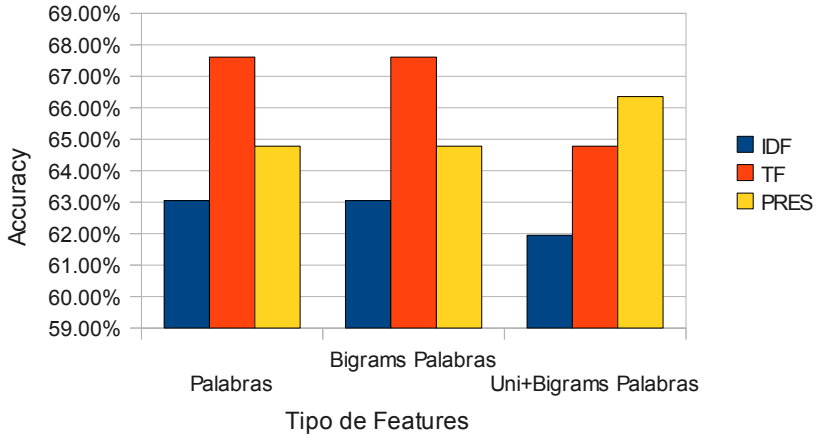

Figura 4.5: Accuracy para SVM con Kernel Lineal en Detección de Polaridad

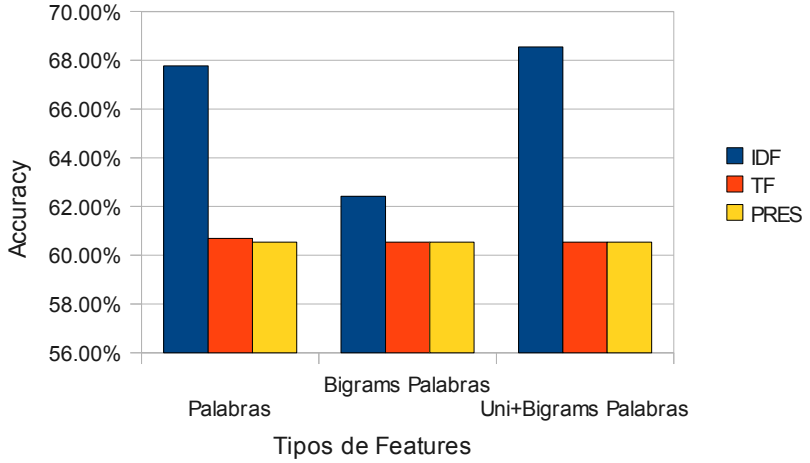

Figura 4.6: Accuracy para SVM con Kernel Gaussiano en Detección de Polaridad

En general, no se aprecian grandes diferencias entre usar unigramas de palabras y la combinación de unigramas+bigramas de palabras. Los resultados obtenidos indican que la mejor configuración se obtiene al utilizar Naive Bayes usando frecuencia de términos en unigramas+bigramas de palabras, logrando una Precision 0,65 para la clase positiva. En cuanto a la Presicion en la clase negativa, en todas las configuraciones se mantiene baja, alrededor de un 0,3. Esto puede deberse principalmente al método basado en bigramas para capturar las negaciones, que sumado a un corpus pequeño no es capaz de capturar todas las referencias negativas que combinan dos o más palabras. Una prueba con un corpus mayor podría dar otras hipótesis de lo que podría estar sucediendo.

Se esperaba que la combinación de las SVM junto a las features basadas en presencia de términos tuviese un mejor desempeño frente a las que utilizan frecuencia o frecuencia invertida de términos, pues a estas últimas no se les realizó la normalización de las features. El vector de presencia de términos podría no representar algunas frases donde se utiliza reiteradamente un término positivo junto a negativo, o viceversa (ver ejemplo en sección 2.2.3).

Los bajos resultados para la detección de subjetividad y polaridad podrían deberse a la poca cantidad de muestras que se posee y al desbalanceo de un 20 % que existe entre las clases que representan opinión, situación que podría no representar la diversidad de formas que se tienen para expresar opinión. Este problema se podría resolver creando un corpus más preciso, poniendo un especial énfasis en realizar una mejor distribución para las frases que representan opinión. Los bajos resultados también podrían deberse a la variación de términos que poseen los mensajes a clasificar. Para este prototipo no se realizaron correcciones a palabras que podrían estar mal escritas (omisión de tildes, dislexia, etc.), ni tampoco se realizó una unión de términos que bajo un contexto distinto significan los mismo, problema que podría resolverse integrando un red de sinonimia o una bases de conocimiento tipo WordNet.

En resumen, la combinación que entregó mejores resultados para detectar mensajes subjetivos consta de una SVM con Kernel Gaussiano que utiliza un vector de frecuencia de términos que incluye palabras y sus POS tags. El mejor clasificador para detectar polaridad es NaiveBayes, bajo un modelo de frecuencia de términos que incluyen las palabras del mensaje.

4.4 Prueba de correlación

Otra forma de evaluar el sistema consiste en medir la efectividad del prototipo propuesto comparando los resultados arrojados por el sistema versus datos reales tales como resultados de encuestas de opinión pública. Para esta prueba se analizaron mensajes relacionados con los candidatos presidenciales de Chile 2010. El análisis utiliza los mensajes escritos durante la realización la encuesta CEP, en Octubre de 2009. En la comparación, se utilizó el apartado "Evaluación de Personajes Políticos" de la encuesta CEP, que contiene un índice de evaluación positiva y negativa para cada uno de los candidatos.

La figura 4.7 muestra la comparación realizada entre las evaluaciones positivas y negativas de la encuesta nacional de opinión pública (CEP). La encuesta tiene un universo de 1505 personas entrevistadas y fue realizada durante el 8 y 30 de Octubre de 2009, en 141 comunas

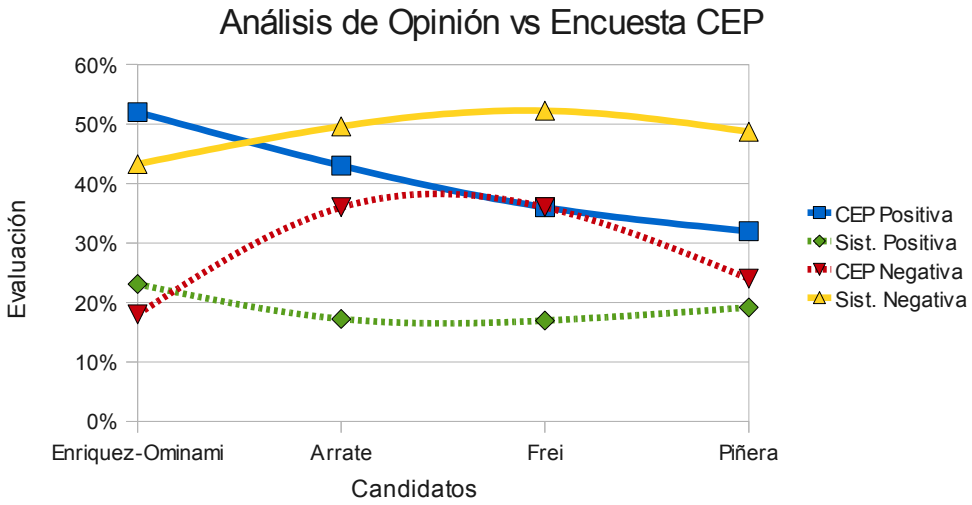

Figura 4.7: Análisis de Opinión vs Encuesta CEP

del país. El sistema procesó 41319 mensajes de 16245 usuarios de Twitter, cada uno de los cuales escribió palabras relacionadas con los candidatos a la presidencia. El sistema realiza el análisis en un tiempo aproximado de 16 minutos.

Para la prueba también se utilizaron las métricas descritas en la sección anterior (figuras 4.8, 4.9 y 4.10), con el objetivo de verificar que los usuarios mantienen una opinión similar durante el periodo de evaluación. El periodo utilizado para calcular las métricas es de tres días, capturando la opinión de los usuarios 3 veces por semana (suponiendo que los usuarios escriben al menos 3 mensajes en relación a los candidatos).

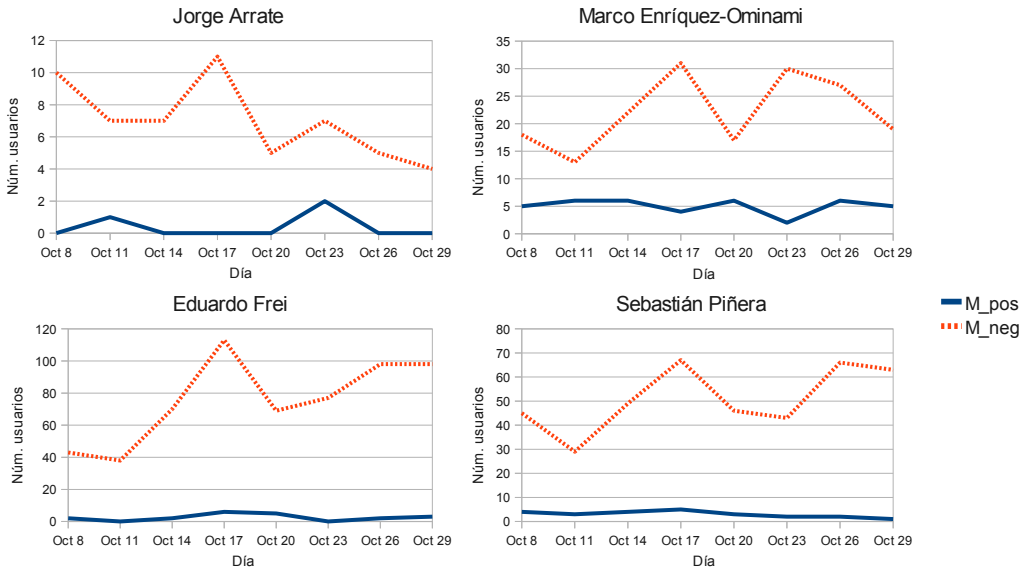

Figura 4.8: Variación de métricas M_{pos} y M_{neg} durante Encuesta CEP

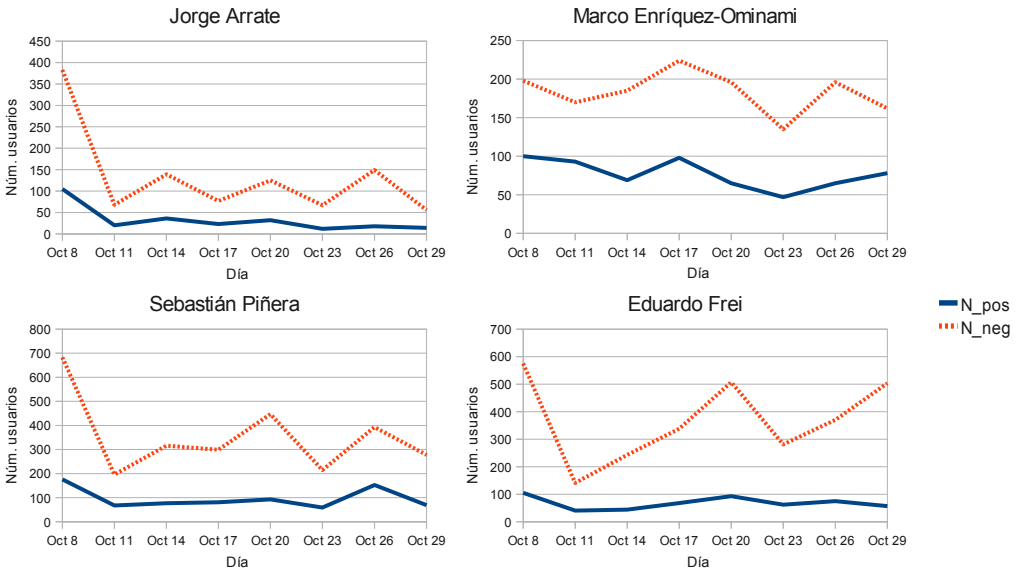

Figura 4.9: Variación de métricas N_{pos} y N_{neg} durante Encuesta CEP

Los resultados obtenidos para la polaridad positiva presentan una correlación lineal de $0,67$ y de 0,89 para la polaridad negativa, lo que demuestra una asociación fuerte entre los resultados del prototipo en polaridad negativa y la encuesta CEP. El resultado de las métricas demuestra que en general existe una marcada tendencia durante todo el mes, donde los usuarios mantienen su opinión durante el periodo de análisis. En particular, la opinión de los usuarios permanece negativa durante todo el mes, mostrando algunas variaciones durante el periodo, asimismo la polaridad se mantiene marginalmente positiva, aunque en

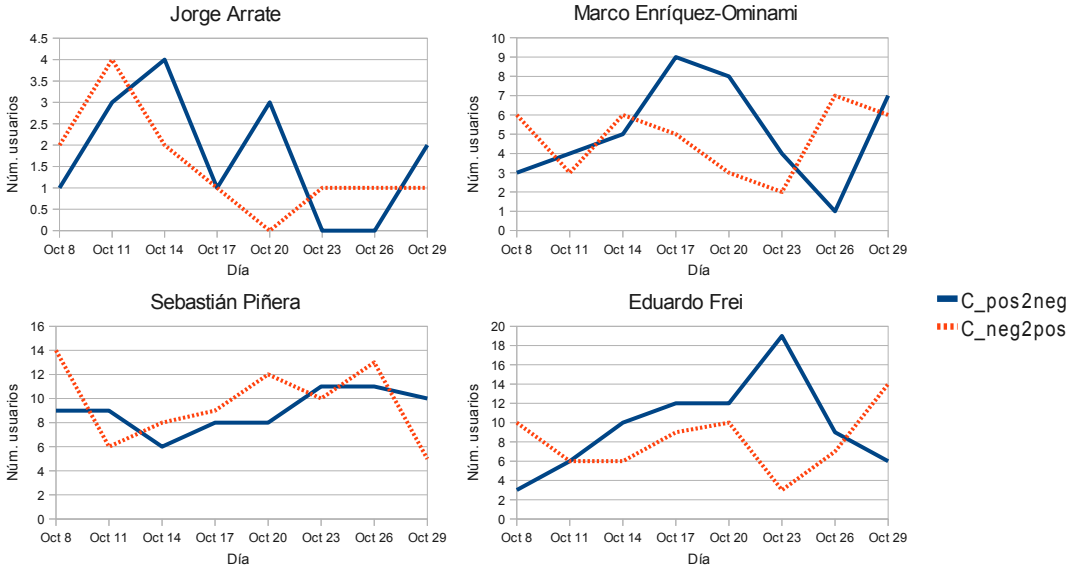

Figura 4.10: Variación de métricas $C_{pos2neg}$ y $C_{neg2pos}$ durante Encuesta CEP

algunos casos no hay usuarios que corroboren su opinión en el mismo periodo; esto se debe probablemnte a que no se envían mensajes con frecuencia. También es posible apreciar una pequeña cantidad de usuarios que cambia su opinión, probablemente debido a mensajes con comentarios por declaraciones en prensa o reacciones a mensajes de otros usuarios. En todo el tiempo del análisis, es posible ver una gran cantidad de usuarios nuevos que envían mensajes subjetivos, lo que problemente se debe a que éstos no realizan comentarios cada 3 días.

Resulta interesante apreciar una alta correlación para la detección de polaridad, pues los experimentos arrojaron resultados mediocres para la detección de polaridad negativa. Una de las razones a las que se podría deber este buen comportamiento es la distribución que posee el corpus: cuando se utiliza validación cruzada para verificar *Precision* o *Accuracy*, se entrena con una porción de los datos disponibles. Al entrenar con todos los datos se podría estar incluyendo información que no se procesa al realizar las pruebas con validación cruzada, lo que permitiría obtener mejores resultados en la clasificación final del prototipo.

Capítulo 5

Conclusiones

En este capítulo se concluye acerca del modelo obtenido, haciendo énfasis en los resultados más interesantes. Como también el trabajo futuro a desarrollar.

5.1 Conclusiones

En este trabajo se desarrolló un prototipo de análisis automático de opinión en mensajes de micro-blogging. Las técnicas utilizadas para detectar los mensajes que poseen opinión y la polaridad presente en ellos usan herramientas para el análisis del lenguaje natural y de clasificación automática, mediante una caracterización vectorial de los mensajes.

Para obtener los mensajes, el sistema utiliza palabras claves (definidas manualmente) que representan una *objeto*, para posteriormente realizar una limpieza de símbolos y caracteres no legibles, para posteriormente seleccionar sólo aquellos mensajes que están escritos en español.

El sistema se evaluó para conseguir la mejor configuración entre varias representaciones de mensajes, así como la mejor técnica de clasificación. La evaluación consiste en calcular la efectividad del prototipo mediante las medidas *Precision* y *Accuracy*, tanto para la etapa de detección de mensajes subjetivos, como para la etapa de detección de polaridad. La mejor configuración para la detección de opinión se obtuvo utilizando SVM con kernel Gaussiano, utilizando frecuencia de términos de palabras y etiquetas POS. Para la detección de polaridad se obtuvo que el clasificador Naive Bayes usando frecuencia de términos en unigrams+bigrams de palabras logra los mejores resultados.

También se evaluó el sistema con datos reales, comparando los resultados de polaridad obtenidos por el prototipo frente a una encuesta de estudios públicos en el tópico de política (en particular, los candidatos a la presidencia 2010 de Chile). En esta prueba se logró una correlación de 0,69 para mensajes con polaridad positiva, y de 0,89 para mensajes clasificados negativos, logrando una correlación lineal con la realidad.

Un aspecto importante a destacar es el tamaño y cobertura del corpus. En el desarrollo de esta memoria se construyó el corpus desde cero con ayuda de voluntarios, sin embargo, posee una menor cantidad de muestras con polaridad positiva, lo que podría explicar la menor correlación obtenido para los mensajes negativos en la comparación con la encuesta.

Finalmente, este prototipo ha demostrado ser un buen punto de partida para analizar lo que la gente opina mediante la extracción de mensajes escritos en sistemas de micro-blogging, puesto que en el tópico de elecciones presidenciales la comparación con la encuesta de opinión pública está bastante acorde a los resultados obtenidos por el prototipo.

5.2 Trabajo futuro

Los resultados de los experimentos muestran que el modelo depende de la calidad del corpus y la información que se puede extraer de él, lo que en primer lugar sugiere la realización de un trabajo más amplio para la recolección de mensajes sobre algún tópico, considerando un balance adecuado entre mensajes que no son opinión y los que contienen polaridad positiva y negativa. También se debe probar otro tipo de técnicas de clasificación automática de texto, con el fin de evaluar el corpus actual. Asimismo resulta necesario investigar técnicas que permitan descubrir el tópico del concepto a buscar, con el fin de cargar un corpus por cada tópico.

En segundo lugar también es necesario que el sistema realice un análisis con más detalle dentro de los mensajes. En el prototipo se consideró que un mensaje contiene sólo una frase, suposición que ocurre con poca frecuencia, por lo tanto uno de los pasos a seguir es investigar sobre como detectar los cambios de contexto entre las diferentes frases presentes en los mensajes de micro-blogging.

En tercer lugar se propone la creación de un modelo híbrido que sea capaz de asociar palabras que significan lo mismo, con el fin de unificar conceptos asociados a cada polaridad. Se podría utilizar una red de palabras tipo WordNet u otra red semántica que contenga significados de palabras, para posteriormente asociar las palabras dentro de mensajes a conceptos descriptivos de polaridad dentro de la representación del mensaje.

En cuarto lugar es necesario utilizar el grafo social disponible por las redes micro-blogging con el objetivo de capturar el comportamiento de las subredes de amistad disponibles dentro del grafo. Por ejemplo, ser capaz de detectar si un grupo de usuarios que normalmente opinan de forma positiva hacia un concepto, están empezando a cambiar su opinión por comentarios realizados por usuarios con un alto nivel de influenza en la red.

Por último, el prototipo desarrollado en esta memoria requiere la automatización y unificación de algunos módulos. Por lo tanto es necesario implementar el código necesario que permita realizar las tareas de las diversas etapas de manera automática, incluyendo una interfaz web que permita a los usuarios suscribir palabras claves para que el sistema vigile aquellos conceptos y un módulo para reentrenar el sistema utilizando feedback de los usuarios (cuando no corresponde la etiqueta que asigna el clasificador). También se debe proveer de un mecanismo que permita acelerar el cálculo de las métricas, almacenando los resultados para un periodo de tiempo dado, con el fin de no realizar los cálculos cada vez que un usuario

ejecute el *frontend*.

Bibliografía

[1] K Dave, S Lawrence, and DM Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web*, pages 519–528, 2003.

[2] B Pang, L Lee, and S Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86, 2002.

[3] Bing Liu. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing*, chapter 26. CRC Press, second edition edition, 2010.

[4] N Kobayashi, K Inui, and Y Matsumoto. Opinion mining from web documents: Extraction and structurization. *Information and Media Technologies*, 2(1):326–337, 2007.

[5] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. *Proceedings of the 20th international conference on Computational Linguistics*, 2004.

[6] W Parrott. *Emotions in social psychology: Essential readings*. Psychology Press, 2000.

[7] Peter Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002.

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

[9] Soo-Min Kim and Eduard Hovy. Automatic identification of pro and con reasons in online reviews. *Proceedings of the COLING/ACL on Main conference poster sessions*, 2006.

[10] Sanjiv Das and Mike Chen. Yahoo! for amazon: Extracting market sentiment from stock message boards. *Proceedings of the 8th Asia Pacific Finance Association*, 2001.

[11] Vasileios Hatzivassiloglou and Janyce Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. *Proceedings of the 18th conference on Computational linguistics - Volume 1*, 2000.

[12] Janyce Wiebe. Learning subjective adjectives from corpora. *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, 2000.

[13] Janyce M Wiebe, Theresa Wilson, and Matthew Bell. Identifying collocations for recognizing opinions. *Proceedings ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, 2001.

[14] Nitin Jindal and Bing Liu. Identifying comparative sentences in text documents. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006.

[15] Andrea Esuli and Fabrizio Sebastiani. Determining term subjectivity and term orientation for opinion mining. *Proceedings EACL-06, the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006. 11th eacl.

[16] Jonnattan Gonzalez. Analisis de sentimientos sobre la web usando tecnicas de procesamiento automatico de textos. Master's thesis, Universidad de Concepción, Sep 2008.

[17] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004.

[18] H Yu and V Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003.

[19] V Hatzivassiloglou, JL Klavans, M Holcombe, R Barzilay, MY Kan, and KR McKeown. Simfinder: A flexible clustering tool for summarization. *Proceedings of the NAACL Workshop on Automatic Summarization*, pages 41–49, 2001.

[20] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003.

[21] Ellen Riloff, Janyce Wiebe, and Theresa Wilson. Learning subjective nouns using extraction pattern bootstrapping. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, 2003.

[22] Soo-Min Kim and Eduard Hovy. Automatic detection of opinion bearing words and sentences. *Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2005.

[23] Jaap Kamps, Maarten Marx, Robert J Mokken, and Maarten de Rijke. Using wordnet to measure semantic orientation of adjectives. *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004.

[24] C Osgood, G Suci, and P Tannenbaum. *The measurement of meaning*. University of Illinois Press, 1957.

[25] Vasileios Hatzivassiloglou and Kathleen McKeown. Predicting the semantic orientation of adjectives. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 1997.

[26] Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato, and V S Subrahmanian. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

[27] C Whitelaw, N Garg, and S Argamon. Using appraisal taxonomies for sentiment analysis. *Proceedings of the 2nd Midwest Computational*, 2005.

[28] Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. Structured models for fine-to-coarse sentiment analysis. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007. 45th acl.

[29] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Readings in speech recognition*, 1990.

[30] E Alpaydin. *Introduction to machine learning*. The MIT Press, 2004.

[31] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. Technical report, Technical report, Department of Computer Science National Taiwan University, 2009.