Índice general

Índice de figuras

Índice de cuadros

Lista de Acrónimos

GSV	*Global Sentiment Value*
IR	*Information Retrieval*
ME	*Maximum Entropy*
NB	*Naïve Bayes*
NLP	*Natural Language Processing*
SVM	*Support Vector Machine*
SW	*Sentiment Words*

Lista de Símbolos

Simbolo	*Descripcion*
Simbolo	*Descripcion*
Simbolo	*Descripcion*

Capítulo 1

Capítulo Introductorio

En este capítulo van:

Introducción

Problema, motivación

Objetivos generales y específicos

Escritos que podrían ser reutilizados:

Las herramientas mediadoras de redes sociales permiten a sus usuarios compartir publicaciones entre sí. Los autores de dichos mensajes escriben sobre sus vidas, comparten opiniones sobre diferentes tópicos y discuten problemas actuales. A medida que más y más usuarios publican mensajes acerca de los productos y servicios que utilizan, estos sitios de mediación se convierten en una fuente valiosa de descripción de las opiniones y sentimientos de las personas. Tales datos pueden ser eficientemente utilizados para estudios sociales (Pak y Paroubek, 2010). Estos datos pueden ser convertidos en información que permita estimar la sensación general acerca de un determinado tópico o entidad.

El análisis de sentimientos fue objeto de estudio a diferentes niveles de granularidad. Inicialmente se ocuparon del análisis de críticas acerca de productos específicos, con el objetivo de clasificarlos como "recomendables" o "no recomendables", a partir de asociaciones semánticas de las oraciones (Turney, 2002; Pang y

Lee, 2004). Otros estudios clasificaron los sentimientos al nivel de oraciones, definiéndolas como "positivas" o "negativas" (Hu y Liu, 2004; Kim y Hovy, 2004). Más recientemente, atraídos por la variedad y diversidad de los datos generados, fueron objeto de análisis los mensajes compartidos en mediadores de redes sociales, los cuales tienen una longitud limitada y son generados en tiempo real. Algunos trabajos que abordaron estas fuentes son los de Pak y Paroubek (2010), que clasificaron los mensajes de Twitter como "positivos", "negativos" o "neutrales"; Go y otros (2009), que utilizaron los tuits con emoticones para clasificarlos como "positivos" o "negativos"; y Saralegi Urizar y San Vicente Roncal (2012), que clasificaron las publicaciones de Twitter en español como "muy positivas", "positivas", "neutrales", "negativas", "muy negativas" o "ninguna de las anteriores".

Capítulo 2

Minería y análisis computacional de sentimientos

Las opiniones son fuertemente influyentes sobre el comportamiento de las personas en la mayoría de sus actividades. Nuestras creencias, decisiones y percepción de la realidad están condicionadas por la visión que perciben las demás personas en nuestro entorno. El área de estudio denominada "análisis de sentimientos", conocida también como "minería de opiniones" se ocupa de conceptos tales como opinión, sentimiento, actitud y evaluación (Liu, 2012).

En este capítulo definimos el análisis computacional como un Proceso de Descubrimiento de Conocimiento (KDD por sus siglas en inglés, *Knowledge Discovery in Databases*) y como un área de estudio del Procesamiento de Lenguaje Natural (NLP por sus siglas en inglés, *Natural Language Processing*). Luego, se presentan las estrategias más utilizadas en el campo, además de una breve descripción de los trabajos similares más recientes que utilizan dichas estrategias para la clasificación de sentimientos.

2.1. Proceso de Descubrimiento de Conocimiento

Los datos son un conjunto discreto de hechos sobre ciertos eventos; los datos en sí no describen las condiciones de un evento, o posibles interpretaciones sobre el mismo, solamente sumarizan características de dichos eventos. Su importancia radica en que sirven de materia prima en el proceso de generación de información.

Los datos se convierten en información cuando el receptor de dichos datos da un sentido o interpretación a los mismos, entonces podemos definir a la información como un conjunto de datos con sentido, relevancia y propósito para el receptor de los datos.

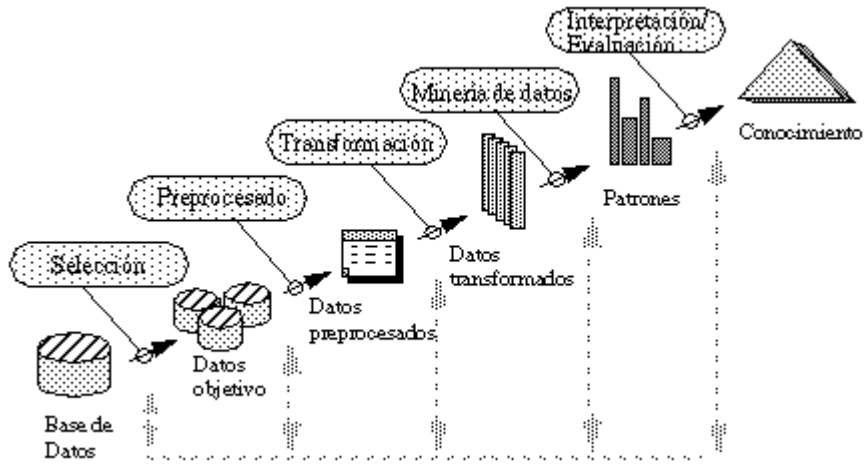

Figura 2.1: Etapas en el proceso de KDD

A partir de informaciones obtenidas, podemos definir KDD como el proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y comprensibles en un conjunto de datos (Fayyad y otros, 1996). Este proceso puede ser aplicado al análisis de sentimientos, puesto que buscamos identificar patrones en un lenguaje de entrada, que describan significativamente las opiniones del tal manera que podamos clasificarlas, por ejemplo, como "positivas" o "negativas" y presentarlas como información generada.

En el proceso de KDD podemos distinguir una serie de etapas, como se puede observar en la Figura 2.1, las cuales describimos a continuación:

2.1.1. Selección

En esta etapa, el objetivo es escoger una fuente o conjunto de fuentes de datos como entrada. La selección de las fuentes se da a partir del análisis previo del problema, de tal manera que los datos escogidos sean representativos del problema y resulten útiles

para descubrir conocimiento. Además, se pueden generar nuevos repositorios de datos a partir del conjunto de datos seleccionado. La recolección, descripción y exploración de datos son tareas comunes en esta fase del proceso.

2.1.2. Preprocesamiento

Esta etapa involucra operaciones de limpieza y preprocesamiento de los datos, así como también de integración de diferentes fuentes de datos. Algunas tareas comunes que efectúan en esta etapa son la selección de columnas para el análisis de datos, la limpieza de los datos con "ruido", eliminación de registros repetidos y la definición de acciones sobre los campos con registros nulos.

2.1.3. Transformación

La transformación de datos se efectúa con el objetivo de conseguir una caracterización o representación de los datos, acorde a los objetivos planteados y, si es posible, reducir la cantidad de variables o atributos considerados, buscando un buen desempeño en las siguientes fases del proceso. Algunas tareas comunes en esta etapa son la discretización de valores, agrupaciones por tipos y la reducción de variables.

2.1.4. Extracción de Patrones

En esta etapa se llevan a cabo trabajos de decisión sobre los modelos y parámetros adecuados para obtener resultados coherentes y esperados sobre los datos, mapeando dichos modelos con las tareas adecuadas de minería para cada criterio u objetivo definido en el modelado. Las tareas comunes en esta fase del proceso son la clasificación, agrupación, asociación, visualización y sumarización de elementos, además de la detección de desvíos, estimación de valores, análisis de relacionamiento, entre otras.

2.1.5. Interpretación y Evaluación

De acuerdo a los datos obtenidos, este paso primariamente implica una retroalimentación de las etapas previas con el fin de corregir errores o mejorar ciertos procedimientos. Además se realiza una visualización sobre los patrones y modelos obtenidos de manera a

comprender dichos resultados y extraer conocimientos. Durante las acciones de comprensión y descubrimiento de conocimiento, se pueden presentar casos en los cuales los resultados generen conflictos con algunos conocimientos previos, por tanto la resolución de dichos inconvenientes puede implicar la modificación o actualización de ciertos parámetros aplicados en pasos anteriores.

2.2. Procesamiento de Lenguaje Natural

Para extraer las opiniones o sentimientos (información) a partir de un texto, el análisis del mismo debe ser tratado como un problema de NLP. Como tal, debe considerar todos los aspectos de ella, como por ejemplo la desambiguación de palabras, las negaciones, intensificaciones, decrementaciones, entre otros.

El problema específico de extracción de opiniones, es un problema de NLP restringido puesto que no es necesario entender completamente la semántica de cada oración o documento, mas debe interpretar únicamente ciertos aspectos de ellos, como los sentimientos positivos o negativos y las entidades o tópicos a los que se refieren. De todas maneras, existen también otros problemas aún no resueltos de NLP que agregan ciertas dificultades al proceso de análisis.

Para clasificar como positivos o negativos los sentimientos expresados en un documento o mensaje, existen dos principales enfoques comúnmente utilizados: la estrategia basada en léxicos, y la estrategia de aprendizaje de máquina (Zhang y otros, 2011), los cuales son abordados en las siguientes secciones.

En esta sección abordamos tanto los aspectos básicos como las dificultades mencionadas de NLP, además de algunas técnicas de preprocesamiento comúnmente utilizadas previamente a la aplicación de estrategias de clasificación.

2.2.1. Aspectos básicos de NLP

El estudio de NLP conlleva el conocimiento de ciertos conceptos de lingüística y su implicancia en la transmisión de las ideas contenidas en una oración o documento. En esta sección tratamos algunos

de ellos, como la desambiguación de palabras, la negación, la intensificación y la atenuación.

La desambiguación de palabras es un aspecto a llevar en cuenta, puesto que en el lenguaje de entrada pueden existir varias palabras que presentan polisemia, es decir cuando una misma palabra representa diversos significados o acepciones. Esto conduce a que parte del problema considerado es identificar el sentido de la palabra dentro de una determinada oración o documento. Por ejemplo, la palabra "vaya" podría representar un verbo conjugado en presente subjuntivo, en modo imperativo, o incluso una interjección que denota admiración o asombro. La diferenciación de dichos significados es una capacidad humana sencilla, pero debe poder reproducirse además en los algoritmos de interpretación que sean desarrollados.

La negación es un elemento lingüístico que, como lo dice su nombre, se utiliza para negar una oración entera o alguna idea contenida en ella, mediante la utilización de adverbios de negación. Tales adverbios pueden ser utilizados en distintos órdenes para referirse a la idea negada, es decir pueden aparecer antes, en medio o después de ellas. Algunos de estos adverbios de negación son "no", "nunca", "jamás", "tampoco", "nada" y "negativamente".

La intensificación y la atenuación consiste en la utilización de ciertos adverbios como recursos lingüísticos para aumentar o disminuir la intensidad de alguna idea transmitida. Algunos intensificadores de ejemplo son "muy", "mucho", "demasiado" y "sobremanera", mientras que adverbios como "poco", "menos", "solamente", "apenas", "justo" son atenuadores.

2.2.2. Dificultades de NLP

A pesar de ser objeto de estudio hace varias décadas y de ser llevada a diversas aplicaciones, existen problemas de NLP que aún no fueron completamente resueltos, por tanto deben ser llevadas en cuenta pues pueden resultar influyentes sobre los resultados. Las más comunes de ellas son las palabras dependientes del contexto, las múltiples orientaciones semánticas en un mismo texto y el sarcasmo.

Las palabras dependientes del contexto son aquellas que no pueden ser definidas como orientadas positivamente o negativamente, sin conocer el dominio del tema que se esté abordando. Una palabra puede ser utilizada para referirse positivamente hacia una entidad en un determinado dominio; pero a la vez dicha palabra puede tener una implicancia negativa en un dominio diferente. Por ejemplo, el adjetivo "pequeño" puede ser visto como positivo refiriéndose al tamaño de un teléfono celular; o bien como negativo si es referido al tamaño de una porción de comida servida en un comedor.

Las múltiples orientaciones semánticas se presentan cuando se efectúa el análisis de un texto que puede estar compuesto por varias oraciones, o bien varias ideas y éstas pueden tener polaridades diferentes. La entidad a la que se está refiriendo el texto puede ser la misma, pero distintas oraciones pueden referirse a diferentes características de ella. Por ejemplo, en la siguiente referencia a una cámara fotográfica: "La calidad de las imágenes es muy buena, pero la duración de la batería es muy corta.", la opinión acerca de la característica de calidad de las imágenes es positiva, mientras que la duración de la batería es vista negativamente. Ambas aparecen en una misma oración y además se refieren a la misma entidad, la cámara fotográfica.

El sarcasmo es una figura que genera dificultades pues cuando se encuentra presente en una oración, pretende dar a entender lo contrario a lo que las palabras expresan. Es decir, se pueden encontrar varias palabras con orientación positiva, con la intención de manifestar negativismo de manera evidente, o viceversa.

2.2.3. Preprocesamiento de entradas

Este paso se efectúa previamente con el objetivo de limitar y refinar los espacios de búsqueda. En esta etapa del proceso se define cuáles entradas son relevantes y la disposición más eficiente en la que pueden ser introducidas al proceso de clasificación. Algunas técnicas de preprocesamiento son las siguientes:

- **Utilización de n-gramas:** Los datos pueden ser introducidos en forma de n-gramas. Un n-grama consiste en una subsecuencia de n-elementos de una secuencia dada. En el estudio del lenguaje natural, se pueden introducir elementos de distinta granularidad, como letras, sílabas o

palabras. Sin embargo, en el análisis de sentimientos, generalmente se utilizan las palabras. El caso base para la disposición de los datos es la utilización de las palabras como unigramas (subsecuencias de un elemento). Otra variante muy utilizada y que es objeto de varios experimentos con resultados positivos, es la introducción de bigramas (subsecuencias de dos palabras) con el objetivo de aumentar la precisión en la determinación de la polaridad de los textos, partiendo de la idea que con ellas se puede lograr una mejor evaluación de intensificadores, negadores y otros fenómenos gramaticales; por ejemplo, "muy bueno", "no apto" son bigramas significativos.

- **Eliminación de stopwords:** Las *stopwords* son un conjunto de palabras de uso muy frecuente y que son omitidas en el proceso de búsqueda, pues no representan entradas significativas (ruído) para el análisis pertinente; en este caso, cargas de polaridad. Varias de ellas son propias de la gramática, como artículos, nombres, preposiciones y otras categorías léxicas; algunos ejemplos son: "acá", "ahí", "algún", "del", "en", entre otras. Otras palabras no significativas para el análisis de sentimientos pueden ser aquellas relativas al dominio, pues son utilizadas de manera muy frecuente y tampoco implican orientación a alguna polaridad determinada; por ejemplo, en una crítica de cine, estas palabras podrían ser "actor", "actuación", "sonido", entre otras.

- **Stemming:** El método de *stemming* consiste en obtener la raíz semántica de cada palabra con el objetivo de asociar las palabras con dicha raíz. Con esto se logra reducir las formas derivadas de las palabras a su forma base. Por ejemplo, la raíz del verbo "aprender" es "aprend", con lo que se logra asociar las conjugaciones de dicho verbo como "aprende", "aprendo", "aprendiendo", etc.

- **Lematización** La lematización es similar al método de *stemming*, a diferencia de que éste toma como raíz de las palabras un subconjunto de letras de la misma; mientras que en el proceso de lematización, se busca hallar el lema de una palabra, es decir su forma genérica o como la encontraríamos en un diccionario. Por ejemplo, mediante lematización se puede lograr reducir la forma "hice" a su lema "hacer", mientras que esto no sería posible mediante *stemming*.

- **Análisis de estructuras gramaticales:** *Está pendiente el desarrollo de este párrafo.*

2.3. Estrategia basada en léxicos

En cualquier lenguaje, existen palabras comúnmente utilizadas para expresar opinión o "sentimiento" con una orientación determinada respecto a una entidad en particular. Dicha orientación puede ser positiva o negativa; por ejemplo: "bueno", "genial" y "espectacular" son entradas positivas; mientras que "malo", "pésimo" y "horrible" son entradas negativas. Estas palabras, denominadas *sentiment words* (SW, por sus siglas en inglés) en conjunto conforman un léxico de sentimientos u opiniones. Para conformar dicho léxico, pueden ser tomados como fuentes distintos textos escritos, cuyos contenidos traten de evaluar o expresar opinión sobre alguna entidad, como por ejemplo críticas, libros, recetas, páginas Web, entre otros.

Las SW podemos clasificar en palabras descriptivas y comparativas. Las palabras descriptivas expresan calificación sobre una entidad, tales como las citadas en el párrafo anterior; mientras que las palabras comparativas establecen confrontación entre dos entidades, generando opiniones superlativas. Por ejemplo, en la oración "El producto A es mejor que el producto B" no se establece calificación positiva ni negativa para alguno de los productos, sino que solamente en comparación al producto B, el A es mejor. Esto no implica que "mejor" sea exclusivamente una SW comparativa para cualquier oración, sino que cumple tal función en el contexto dado.

En algunos léxicos, la orientación semántica de las palabras es cuantificada generalmente por valores numéricos, que luego son introducidos en una función de puntuación, para determinar la polaridad de una opinión acerca de una entidad determinada. Dicha función varía según el análisis de distintos autores; así como los criterios de cuantificación de las orientaciones semánticas. En algunos trabajos, se limitan a asignar signos positivos y negativos (+1 y -1) a las palabras (Ding y otros, 2008); mientras que en otros, se emplean un intervalo más amplio para obtener un mejor balance de la suma final de polaridades, como (Taboada y otros, 2011) que utilizan un intervalo de puntuación de -5 a +5. Estas puntuaciones, así como las funciones definidas para el cálculo de polaridad, pueden influir en la precisión de resultados logrados.

En (Liu, 2012) se identifican tres tipos de estrategia para compilar SW que conformen un léxico: la generación manual, la generación basada en un diccionario y la generación basada en un corpus.

2.3.1. Léxico generado manualmente

Seguir esta estrategia implica un costo muy alto de trabajo y de tiempo, por tanto generalmente no es utilizada de manera exclusiva. Tiene su utilidad en combinación con estrategias automatizadas mediante tareas tales como una verificación final de palabras, puesto que con los métodos automatizados se pueden generar errores.

2.3.2. Léxico basado en un diccionario

La construcción de un diccionario implica un listado de sinónimos y antónimos para cada palabra. Un método sencillo de generación es utilizar como semillas algunas SW, a partir de las cuales se pueden encontrar las demás palabras del diccionario a partir de los sinónimos y antónimos de las semillas. A medida que dichos sinónimos son agregados como nuevas palabras semilla, el diccionario va creciendo sucesivamente. Algunos trabajos que se ocupan de la generación de un diccionario son (Hu y Liu, 2004) y (Kim y Hovy, 2004).

La ventaja de utilizar un diccionario es que se puede encontrar fácil y rápidamente un gran número de SW con sus orientaciones semánticas. A pesar de que la lista resultante puede incluir varios errores, es posible efectuar una verificación manual de limpieza para refinar el diccionario. Por otro lado, la principal desventaja es que la lista de palabras generadas en un diccionario son de propósito muy general (no orientadas al dominio de aplicación) y dependientes del contexto. Esta dificultad puede tratada mediante la generación de un corpus.

2.3.3. Léxico basado en un corpus

Esta estrategia fue mayormente aplicada frente a dos escenarios:

1. Dado un listado base de SW conocidas, de propósito general, se descubren nuevas SW con sus orientaciones para un determinado dominio.

2. Adaptar un léxico de propósito general a uno nuevo utilizando un corpus para aplicaciones de análisis de sentimientos en el dominio.

Construir un léxico de sentimientos para un dominio específico puede no ser suficiente puesto que en el mismo dominio, una misma palabra puede ser positiva en un contexto, pero negativo en otro. Esta estrategia también puede ser utilizada para construir un léxico de propósito general en el caso que se encuentre disponible un corpus muy grande y diverso, pero para tal propósito es más efectiva la elaboración de un diccionario, puesto que el mismo ya está compuesto por todas las palabras. Algunos trabajos que tratan la generación de un corpus son (Turney, 2002) y (Ding y otros, 2008).

2.4. Estrategia basada en aprendizaje de máquina

La estrategia basada en aprendizaje de máquina, también referida como estadística o clasificadora de textos, consiste en entrenar un clasificador para determinar sentimientos positivos, negativos o neutrales. El objetivo es generalizar comportamientos a través de información previamente recolectada, denominada conjunto de entrenamiento, en forma de ejemplos, que luego puedan ser utilizados para evaluar un texto en función a sus similaridades.

En términos generales, un conjunto de entrenamiento, llamada en inglés *training set* es un conjunto de datos que se utiliza para realizar pruebas iniciales sobre un modelo conceptual esperando conseguir una descripción comprensible de dicho modelo y a partir del mismo, la generación de conocimiento.

Podemos identificar dos tipos de aprendizaje en esta estrategia: supervisado y no supervisado.

En un método de aprendizaje supervisado, se produce una función de correspondencia entre las entradas y salidas deseadas del sistema; como por ejemplo clasificar las palabras según su categoría

léxica (verbos, sustantivos, adjetivos, etc) o categoría funcional (determinantes, cuantificadores, pronombres, etc), por lo que se necesita una base de conocimiento previamente elaborada, compuesta por ejemplos de etiquetados anteriores.

Los métodos de aprendizaje no supervisado se llevan a cabo sobre conjuntos de ejemplos conformados únicamente por entradas al sistema; sin etiquetado de categorías, a diferencia de los métodos supervisados; por lo que el sistema debe tener capacidad de reconocimiento de patrones para etiquetar nuevas entradas.

Algunos clasificadores de aprendizaje de máquina son *Naïve Bayes* (NB), *Maximum Entropy* (ME) y *Support Vector Machines* (SVM). Las formulaciones de cada algoritmo en este documento están basadas en la propuesta de (Pang y otros, 2002), en el que se utiliza el siguiente entorno de variables:

$\{f_1, f_2, \ldots, f_m\}$**, un conjunto predefinido de características que pueden aparecer en el documento.**

$n_i(d)$**, la cantidad de ocurrencias de** f_i **en el texto** d**.**

$\vec{d} := (n_1(d), n_2(d), \ldots, n_m(d))$**, es el vector que representa al texto** d**.**

2.4.1. Naïve Bayes

Una técnica para la clasificación de texto es asignar a un determinado texto d **la clase** $c^* = arg\,max_c\, P(c|d)$**.**

El clasificador *Naïve Bayes* deriva de la regla de Bayes,

$$P(c \,|\, d) \;=\; \frac{P(c)P(d \,|\, c)}{P(d)},$$

donde $P(d)$ **no influye en la selección de** c^***. Para estimar el término** $P(d \,|\, c)$**, en esta técnica se lo descompone asumiendo que las** f_i **son independientes entre sí, resultando:**

$$P_{NB}(c \,|\, d) \;:=\; \frac{P(c)\,(\prod_{i=1}^{m}\, P(f_i \,|\, c)^{n_i(d)})}{P(d)}.$$

El algoritmo basado en el teorema de Bayes, aplicado al análisis de textos, consiste en clasificar cada palabra de acuerdo a la probabilidad de que pertenezca a un determinado grupo. Tales agrupaciones pueden estar determinadas por categorías gramaticales o por polaridad. La clasificación por etiquetas gramaticales (***POS tagging***) permite, por ejemplo, establecer la probabilidad de que la palabra siguiente a un sustantivo sea un verbo; con el objetivo de determinar qué influencia tiene una palabra en la polaridad de un texto. Un adjetivo es más descriptivo que un sustantivo para la inferencia del sentimiento transmitido en un texto. La asociación por polaridad relaciona las palabras que componen una oración con el sentimiento que ellas en conjunto transmiten. Es decir, las palabras utilizadas en una oración considerada positiva son, con mayor probabilidad, palabras con asociación positiva para cualquier oración.

La dificultad en la aplicación de este algoritmo es que está cimentado en la hipótesis de que las variables evaluadas son independientes entre sí; y como la gramática es dependiente del contexto, muchas entradas son clasificadas erróneamente, originando los fenómenos conocidos como "falsos positivos" o "falsos negativos". Sin embargo, los resultados pueden ser aún muy positivos con la determinación de un buen conjunto de entrenamiento.

2.4.2. Maximum Entropy

El clasificador por entropía máxima trata de generar la información reduciendo el sesgo al mínimo posible. A diferencia del algoritmo de ***Naïve Bayes***, no asume la independencia entre variables. El principio fundamental del algoritmo es que son preferidas las distribuciones más uniformes que satisfacen además todas las restricciones del problema (Nigam y otros, 1999).

En la aplicación del método se trata de identificar las características que definen una categoría, luego para cada una de ellas estimar su valor esperado sobre el conjunto de entrenamiento para tratarlas como restricción en el modelo de distribución (Lee y Renganathan, 2011).

Formulando el algoritmo, el término $P(c \mid d)$ toma la siguiente forma:

$$P_{ME}(c \mid d) := \frac{1}{Z(d)} exp\left(\sum_i \lambda_{i,c} F_{i,c}(d,c)\right),$$

donde $Z(d)$ es una función de normalización y $F_{i,c}$ es una función que relaciona una característica con la clase, definida como:

$$F_{i,c}(d,c') := \begin{cases} 1, & n_i(d) > 0 \text{ y } c' = c \\ 0 & \text{otros casos} \end{cases}$$

Los parámetros $\lambda_{i,c}$ son indicadores de peso; es decir, un valor mayor indica que f_i es un indicador más relevante para la clase c.

2.4.3. Support Vector Machines

La técnica de clasificación por *Support Vector Machines* efectúa la separación de puntos un espacio N-dimensional utilizando un hiperplano $(N-1)$-dimensional, a diferencia de las técnicas de Bayes y entropía máxima, que utilizan medidas probabilísticas para clasificar los puntos. Dado un conjunto de entrenamiento, el clasificador por SVM busca encontrar un hiperplano con el mayor margen posible, de manera que cada punto del conjunto de entrenamiento sea clasificado correctamente y que el hiperplano se encuentre a la máxima distancia posible de sus puntos más cercanos. La clasificación de las instancias de entrenamiento consiste en determinar a qué lado del hiperplano caen ellas. En la práctica, es posible que no se encuentre un hiperplano que separe los grupos perfectamente, por lo que los puntos pueden hallarse dentro del margen o del lado incorrecto del hiperplano.

Si representamos el hiperplano buscado por el vector $\vec{w}$, la búsqueda corresponde a un problema de optimización con restricciones, siendo $c_j \in \{1, -1\}$ (positiva y negativa) la clase correcta del texto d_j, la solución puede ser escrita como:

$$\vec{w} := \sum_j \alpha_j c_j \vec{d_j}, \quad \alpha_j \geq 0,$$

donde las α_j son obtenidas por resolución de un problema de optimización. Los vectores $\vec{d_j}$, tal que α_j es mayor que cero, son llamados vectores de soporte, dado que son los únicos que contribuyen al hiperplano $\vec{w}$.

Ha sido demostrado varias veces que este método es altamente efectivo para la categorización de textos, obteniendo generalmente mejores resultados que la técnica de Bayes (Joachims, 1998; Pang y otros, 2002; Kennedy e Inkpen, 2006).

2.5. Métricas de evaluación

La evaluación de la clasificación se efectúa primariamente con el objetivo de comparar el desempeño de distintos sistemas o métodos entre sí. La relevancia y utilidad de los resultados obtenidos prevalecen sobre los tiempos y espacios de respuesta cuando evaluamos una estrategia de análisis de textos. No existe un único criterio que permita determinar que un método es mejor que otro; por ello, existen diversas métricas para medir la efectividad de la información obtenida del análisis.

Las métricas que presentamos en esta sección están basadas en la matriz de confusión o de clasificación. Dicha matriz contrasta la estimación del modelo clasificador para un elemento dado, con la clase real asociada a dicho elemento.

	Clase$_1$	**Clase$_2$**	**Clase$_n$**
Clase$_1$	$VClase_1$	$FClase_2$	$FClase_n$
Clase$_2$	$FClase_2$	$VClase_2$	$FClase_n$
Clase$_n$	$FClase_n$	$FClase_2$	$VClase_n$

Cuadro 2.1: Matriz de confusión o de clasificación

Dado un conjunto de datos E, donde cada elemento e puede ser representado como un vector $< e_1, e_2, \ldots, e_n >$ de n atributos, y un conjunto de clases C; cada vector $e \in E$ tiene una clase $c \in C$ asociada. La clasificación consiste en asignar una clase $c \in C$ para cada elemento $e \in E$, y el objetivo es cuantificar la efectividad de dichas asignaciones. Para dicho efecto, las filas de la matriz de

confusión representan las clases reales a la que pertenecen cada uno de los elementos y las columnas representan a la predicción del modelo clasificador, como se observa en el Cuadro 2.1.

Referenciando los subíndices de filas y columnas como $[i, j]$, una entrada Verdadera representa un acierto, es decir si el clasificador ha determinado que un elemento e pertenece a la $Clase_j$ y dicho elemento efectivamente está asociado a la $Clase_j$; mientras que una entrada Falsa representa un error de clasificación, es decir que el elemento e no pertenecía efectivamente a la $Clase_j$.

Definimos a continuación algunas métricas de evaluación, en cuyas definiciones referenciamos a los aciertos y errores con el subíndice de la fila i a la que pertenecen en la matriz de confusión, como Verdadero de la $Clase_i$ o Falso de la $Clase_i$:

2.5.1. Precisión

La precisión respecto a la $Clase_i$ indica el número de predicciones correctas sobre el total de predicciones de dicha clase. Es decir, indica la probabilidad de que un elemento e pertenezca realmente a la $Clase_i$ dado que fue identificado como tal por el clasificador. Está dado por la proporción de verdaderos de la $Clase_i$ en relación a los totales identificados como de dicha clase:

$$Precisión = \frac{VClase_i}{VClase_i + FClase_i}$$

2.5.2. Recall

El *recall* respecto a la $Clase_i$ indica el número de predicciones acertadas como $Clase_i$ sobre el total de elementos que realmente pertenecen a dicha clase. Es la probabilidad de que un elemento e que pertenece realmente a la $Clase_i$ sea clasificado como tal. Está dado por la razón entre los verdaderos de la $Clase_i$, sobre los elementos reales de la $Clase_i$, en la matriz de confusión dados por la suma de los verdaderos de la $Clase_i$ más los falsos de las demás clases.

$$Recall = \frac{VClase_i}{VClase_i + FClase_j} \quad donde \quad j \in \{1, 2, \ldots, n\} \quad con \quad j \neq i$$

2.5.3. Medida-F

Relaciona las medidas de precisión y recall mediante una media con el mismo peso para ambas. Es conocida también como F_1 por la uniformidad de los pesos.

$$Medida - F = \frac{2 \times Precisión \times Recall}{Precisión + Recall}$$

2.6. Trabajos similares más recientes

Analizamos algunos de los trabajos publicados más recientemente con el fin de conocer las últimas propuestas de solución que fueron utilizadas efectivamente para análisis de sentimientos.

Esta sección aún debe ser reorganizada.

(Moreno-Ortiz, Pérez Hernández, 2013)

La información proporcionada por Moreno-Ortiz y Pérez Hernández es el *Global Sentiment Value* (GSV), el cual es un valor numérico (en una escala entre 0 - 10) asignado a sentimientos de acuerdo a una entrada de texto. Para obtener este valor numérico, se realizan cómputos sobre varios valores previos que influyen en el resultado del GSV. Uno de los más influyentes es el *Affect Intensity*, el cual modula el GSV para reflejar el porcentaje de palabras con cierta carga de sentimientos que se encuentran en un texto de entrada.

El *Affect Intensity* no es un contador de palabras (positivas, negativas o neutras), sino que cuenta segmentos de texto que correspondan a unidades léxicas. El mismo se define como el porcentaje de segmentos o unidades léxicas con cierta carga de sentimientos. Esta medida no es utilizada de manera directa para computar el

valor global de sentimientos de un texto, sin embargo es utilizado como un paso intermedio para ajustar los límites superior e inferior de las valencias relacionadas con el grado de sentimientos.

El procedimiento del cómputo para el valor del GSV localiza los segmentos léxicos positivos y negativos, asignando a cada uno valores iguales dentro de la escala definida por el *Affect Intensity*, sin embargo se va incrementando dicho valor a cada lado de la escala definida. El método resultante para obtener el GSV a partir de un texto de entrada está determinado por una formulación matemática que involucra la adición de las sumatorias de las sentencias negativas y positivas encontradas, y multiplicadas por la escala acotada del *Affect Intensity*; todo esto dividido la resta entre la cantidad total de unidades léxicas menos la cantidad de unidades léxicas consideradas neutras.

(Vilares, Alonso y Gómez-Rodríguez, 2013)

En contraste con las propuestas léxicas, esta propuesta se basa en obtener la estructura sintáctica del texto para tratar las construcciones lingüísticas e identificar los elementos de la frase que están implicados en ellas.

El pre procesamiento del texto implica la unificación de expresiones compuestas que actúan como una sola unidad de significado, la normalización de signos de puntuación y la segmentación del texto en oraciones y a la vez dividir cada oración en tokens para palabras y signos de puntuación.

El análisis sintáctico posterior se inicia creando un árbol de dependencias sintácticas mediante el cual se identifican relaciones binarias entre los elementos de una oración. Cada vínculo binario constituye una dependencia, que se anota con la función sintáctica que relaciona los dos términos. Para el análisis semántico, esta propuesta se respalda en el diccionario de polaridad para adjetivos, sustantivos, verbos, adverbios e intensificadores propuesto por (Brooke, Tofiloski, y Taboada, 2009). Es importante señalar que los valores numéricos asociados a los intensificadores tienen un significado distinto, ya que representan el porcentaje por el que modifican el sentimiento de la expresión que afectan; por ejemplo, a

un amplificador "muy" se le asocia el valor 0.25 y el decrementador "en absoluto" el valor -1.

Una diferencia notoria con respecto al trabajo de (Brooke, Tofiloski, y Taboada, 2009) es la de utilizar el árbol de dependencias para determinar la parte de la frase que se ve afectada por la modificación de los intensificadores, negación y oraciones adversativas (determinando las oraciones subordinadas y la subordinante). Ciertos factores que afectan a la clasificación de la polaridad no se han considerado, por ejemplo, el problema de la polaridad cambiante para determinados términos según el dominio en el que aparezcan (Pang y Lee, 2008), la ironía y el sarcasmo como figuras literarias para expresar opinión.

Capítulo 3

Formulación del problema propuesto

En este capítulo van:

Introducción (Presentación del caso de análisis)

Modelo conceptual (selección de atributos, etc)

Objetivos y criterios de éxito

Caracterización de Twitter

Extracción de datos, diseño de pruebas (Implementaciones, balance de carga, etc)

3.1. Caracterización de mensajes en Twitter

Los mensajes compartidos en Twitter poseen varias características únicas, que los diferencian de los demás corpus de datos citados previamente. (Go y otros, 2009) identifican cuatro de ellas:

- Longitud máxima: Cada publicación puede estar compuesta por una cantidad máxima de 140 caracteres, incluyendo signos de puntuación y símbolos.
- Alta disponibilidad: La API de Twitter permite obtener tuits libremente, a través de su función de búsqueda. Mediante esta función es

posible recolectar una cantidad de tuits considerable para conformar los conjuntos de entrenamiento y de evaluación.

- Modelo del lenguaje: Los mensajes de Twitter pueden generados desde diversos dispositivos, incluyendo los teléfonos celulares. Esto conlleva a que los errores ortográficos y expresiones informales se hagan mucho más frecuentes en comparación a otros dominios.
- Dominio: Los tuits se refieren a diversos temas, lo cual es relevante puesto que un gran porcentaje de estudios pasados fueron centrados en un dominio específico.

Existen además ciertas convenciones de lenguaje de los tuits:

- "RT" es un acrónimo de retuit, que se coloca delante de un mensaje para indicar que el mismo está siendo repetido o compartido.
- El caracter "#" es utilizado para marcar, organizar y clasificar tuits por temas o categorías.
- El caracter "@" es empleado para referirse a una cuenta por su nombre de usuario.

Es frecuente además que los tuits vayan acompañados de emoticones y vínculos de Web.

Capítulo 4

Despliegue y análisis de resultados

En este capítulo van:

Herramientas utilizadas

Selección de atributos, de las tareas, etc

Descripción paso a paso de las variantes utilizadas, tunning de algoritmos

Desplegar gráficos de resultados

Capítulo 5

Conclusiones y trabajos futuros

En este capítulo van las conclusiones y trabajos futuros.

Capítulo 6

Bibliografía

1. Agarwal, Apoorv; Xie, Boyi; Vovsha, Ilia; Rambow, Owen; Passonneau, Rebecca. Columbia University, 2011. Artículo, LSM '11 "Proceedings of the Workshop on Languages in Social Media", Páginas 30-38, Association for Computational Linguistics Stroudsburg, PA, USA. "Sentiment Analysis of Twitter Data".

2. Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier. "Modern Information Retrieval". USA: ACM Press, 1999.

3. Ding, Xiaowen; Liu, Bing; Yu, Philip S. University of Illinois at Chicago, Chicago, IL, 2008. Publicación, WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining, Páginas 231-240, ACM New York, NY, USA. "A Holistic Lexicon-Based Approach to Opinion Mining".

4. Fayyad, Usama; Piatetsky-Shapiro; Smyth, Padhraic. Publicación, Advances in knowledge discovery and data mining, P'aginas 1-34, American Association for Artificial Intelligence Menlo Park, CA, USA, 1996. "From Data Mining to Knowledge Discovery".

5. Giménez-Lugo, Gustavo. Universidade de São Paulo. "Um Modelo de Sistemas Multiagentes para Partilha de Conhecimento Utilizando Redes Sociais Comunitárias". Tesis de Doctorado en Ingeniería, 2004.

6. Go, Alec; Bhayani, Richa; Huang, Lei. Reporte técnico, Stanford University, 2009. "Twitter Sentiment Classification using Distant Supervision".

7. Hu, Minqing; Liu, Bing. University of Illinois at Chicago, Chicago, IL, 2004. Publicación, KDD '04 Proceedings of the tenth ACM SIGKDD in-

ternational conference on Knowledge discovery and data mining, Páginas 168-177, ACM New York, NY, USA. “Mining and summarizing customer reviews”.

8. Joachims, Torsten. Universität Dortmund, Germany, 1998. Publicación, ECML ’98 Proceedings of the 10th European Conference on Machine Learning, Páginas 137-142, Springer-Verlag London, UK. “Text Categorization with Suport Vector Machines: Learning with Many Relevant Features”.

9. Kennedy, Alistair; Inkpen, Diana. University of Ottawa, Ottawa, Canada, 2006. Artículo, Computational Intelligence, Volumen 22, N° 2, Mayo 2006, Páginas 110-125. “Sentiment Classification of Movie Reviews Using Contextual Valence Shifters”.

10. Kim, Soo-Min; Hovy, Eduard. University of Southern California, Marina del Rey, CA, 2004. Publicación, COLING ’04 Proceedings of the 20th international conference on Computational Linguistics, Artículo N° 1367, Association for Computational Linguistics Stroudsburg, PA, USA. “Determining the sentiment of opinions”.

11. Lee, Huey Yee; Renganathan, Hemnaath. Jalan Universiti, Ehsan, Malaysia, 2011. Publicación, SAAIP 2011 Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology, Páginas 89-93, Asian Federation of Natural Language Processing. “Chinese Sentiment Analysis Using Maximum Entropy”.

12. Liu, Bing. “Sentiment Analysis and Opinion Mining”. Morgan & Claypool Publishers, Mayo 2012.

13. Moreno-Ortiz, Antonio; Pérez-Hernández, Chantal. Universidad de Málaga, 2013. Artículo, Procesamiento del Lenguaje Natural, N° 50, Páginas 93-100, Sociedad Española para el Procesamiento del Lenguaje Natural. “Lexicon-Based Sentiment Analysis of Twitter Messages in Spanish”.

14. Nigam, Kamal. Carnegie Mellon University, Pittsburgh, PA; Lafferty, John. Carnegie Mellon University Pittsburgh, PA; McCallum, Andrew. Just Research, Pittsburgh, PA, 1999. Publicación, IJCAI-99 Workshop on Machine Learning for Information Filtering, Páginas 61-67. “Using maximum entropy for text classification”.

15. Pak, Alexander; Paroubek, Patrick. Université de Paris-Sud, 2010. Publicación. LREC ’10 Proceedings of the Seventh International Confe-

rence on Language Resources and Evaluation. “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”.

16. Pang, Bo. Cornell University, Ithaca, NY; Lee, Lillian. Cornell University, Ithaca, NY; Vaithyanathan, Shivakumar. IBM Almaden Research Center, San Jose, CA, 2002. Publicación, EMNLP ’02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volumen 10, Páginas 79-86, Association for Computational Linguistics Stroudsburg, PA, USA. “Thumbs up?: sentiment classification using machine learning techniques”.

17. Pang, Bo; Lee, Lilian. Cornell University, Ithaca, NY, 2004. Publicación, ACL ’04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Artículo N° 271, Association for Computational Linguistics Stroudsburg, PA, USA. “A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts”.

18. Prasad, Suhaas. Reporte técnico, Stanford University, 2010. “Microblogging Sentiment Analysis Using Bayesian Classification Methods”.

19. Saralegi Urizar, Xabier; San Vicente Roncal, Iñaki. Elhuyar Fundazioa, España, 2012. Artículo, TASS 2012 Workshop on Sentiment Analysis at SEPLN. “TASS: Detecting Sentiments in Spanish Tweets”.

20. Taboada, Maite. Simon Fraser University; Brooke, Julian. University of Toronto; Tofiloski, Milan. Simon Fraser University; Voll, Kimberly. University of British Columbia; Stede, Manfred. University of Postdam, 2011. Artículo, Journal Computational Linguistics Volume 37 Issue 2, Junio 2011, Páginas 267-307, MIT Press Cambridge, MA, USA. “Lexicon-based methods for sentiment analysis”.

21. Turney, Peter D. National Research Council of Canada, Ottawa, Ontario, Canada, 2002. Publicación, ACL ’02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Páginas 417-424, Association for Computational Linguistics Stroudsburg, PA, USA. “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews”.

22. Vilares, David; Alonso Miguel A.; Gómez-Rodríguez, Carlos. Universidade da Coruña, A Coruña, 2013. Artículo, Procesamiento del Lenguaje Natural, N° 50, Páginas 13-20, Sociedad Española para el Procesamiento del Lenguaje Natural. “Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias”.

23. Yang, Yiming; Liu, Xin. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1999. Publicación, SIGIR '99 Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Páginas 42-49, ACM New York, NY, USA. " A re-examination of text categorization methods".

24. Zhang, Lei; Ghosh, Riddhiman; Dekhil, Mohamed; Hsu, Meichun; Liu, Bing. HP Laboratories, 2011. Reporte Técnico, HPL-2011-89. "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis".

Apéndice A

Listado de palabras del dominio

Aquí podríamos incluir nuestro léxico de palabras positivas y negativas, y nuestra lista de stopwords.