Análisis de opinión en Twitter para la valoración de películas

Universidad Nacional del Centro de la Provincia de Buenos Aires

Propuesta de Tesis de Grado

Integrante: Francisco Boato

Director: Dr. Marcelo Armentano

Fecha: 12/06/2013

1. Introducción

Twitter es, hoy en día, el servicio de microblogging más utilizado en el mundo. Microblogging permite escribir, publicar y compartir mensajes de texto cortos (en el caso de Twitter 140 caracteres como máximo y son llamados "tweets"). En la actualidad los usuarios escriben y comparten con otros usuarios aspectos de su vida personal, intereses, opiniones de distintos tópicos y hasta se desarrollan discusiones de temas de actualidad. En los últimos años el microblogging (y Twitter en particular) ha desplazado a otros medios de comunicación on-line (como por ejemplo blogs tradicionales y lista de emails). Actualmente se estima que existen más de 500 millones de usuarios en Twitter[1]

El uso masivo y globalizado de twitter ha generado una fuente de información extremadamente grande, constantemente actualizada y que en su mayoría es de público acceso a todos los usuarios. Es por este motivo que se necesitan herramientas que permitan el análisis de esta información. Como se dijo anteriormente los usarios utilizan Twitter para, entre otras cosas, expresar sus opiniones acerca de una gran variedad de aspectos (política, tecnología, libros, films, religión, alimentación, etc.). El aspecto en común que comparten todas estas opiniones es la subjetividad ya que el usuario expresa su propio pensamiento o sentimiento en relación a alguien o algo a través de un texto de no más de 140 caracteres. Es evidente entonces la utilidad que tendría una aplicación capaz de extraer la "polaridad" o el "valor" del sentimiento expresado por el usuario. Por ejemplo, las empresas de producción podrían estar interesadas en saber cuál es la opinión de la comunidad en relación a un nuevo producto; las empresas de servicios podrían tener una manera de evaluar la calidad del servicio brindado; y los partidos políticos podrían saber cuál es la imagen de determinado personaje político o la aceptación o no de determinada ley o proyecto. A una persona podría interesarle la opinión general de un libro, o de una película y así decidir si vale la pena o no comprar el libro o ver el film.

Debido al inmenso volumen de información que existe en Twitter, la aplicación de técnicas de Data Mining para analizar el sentimiento de un tweet resulta una alternativa ideal. En primer lugar se pueden diferenciar dos tipos de tweets: los tweets que expresan un hecho o situación objetiva o neutra, sin incluir ningún juicio de valor (por ejemplo.... "Llueve" o "Hoy elecciones presidenciales") de aquellos tweets subjetivos o que contienen un "sentimiento" (por ejemplo "Otra vez lluvia.... que tristeza" o "Hoy elecciones presidenciales.... todos con Mauricio Kirchner!!!"). Tomando en consideración los tweets que contienen sentimiento, la primera diferenciación obvia que cabe mencionar es si son sentimientos positivos o negativos (por ejemplo "Iron Man 3 está buenísimo" o bien "No me gustó el 3D del último IRON man"). Sin embargo existen otros tweets donde se expresan sentimientos positivos y negativos al mismo tiempo (por ejemplo "el argumento de Iron Man 3 es muy interesante, sin embargo la versión 3D no vale la pena") en estos casos se considera un tweet con sentimiento Mixto. Resumiendo podemos considerar el valor o polaridad de un tweet como: Neutro (no contiene sentimiento), Positivo, Negativo o Mixto.

Una alternativa para poder determinar el valor (o polaridad) de un tweet utilizando técnicas de Data Mining, es utilizar modelos de clasificación. Un modelo de clasificación se puede definir como la construcción de un modelo que mapea (clasifica) una instancia dentro de una clase previamente definida *(FAYYAD, Gregory, & Padhraic, 1996)*. El modelo se deriva a partir de un conjunto de datos de entrenamiento (es decir un dataset de tweets ya clasificados). La clasificación de los tweets se puede hacer automáticamente, a partir de reglas previas (por ejemplo a partir de la existencia de emoticones que reflejen el humor del usuario), o bien se puede clasificar manualmente los tweets a

[1] "Twitter reaches half a billion accounts More than 140 millions in the U.S" – 30-07-2012.
http://semiocast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US

partir del criterio de la persona encargada de realizar esta clasificación (se puede contar con una guía que ayude a disminuir la subjetividad).

Una alternativa a los modelos de clasificación para extraer la polaridad subjetiva de un texto es aplicar algoritmos basados en léxico. Estos algoritmos se basan en diccionarios donde se le da una polaridad a las palabras, y en base a la polaridad de las palabras que componen un texto se puede extraer el sentimiento del texto como un único concepto semántico.

El objetivo de este trabajo es evaluar diferentes modelos que permitan determinar la polaridad de un texto corto, tomando como fuente de información la red social Twitter. Se tomarán en cuenta solo los tweets que hagan referencia a algún film (o película). Se evaluarán y compararán entre sí, tanto diversos modelos de clasificación, como algoritmos basados en léxico, explorando diferentes técnicas de pre-procesamiento de texto que permitan llegar a una mejor performance de los modelos evaluados.

Los modelos planteados se integrarán a una aplicación que tome como entrada el nombre de un film y, a partir de la búsqueda y extracción de polaridad de los tweets que lo referencian, se emitirá una valoración de dicho film.

2. Trabajos Relacionados

Desde hace varios años que se realizan muchos estudios en relación a la detección de sentimiento u opiniones en textos. En *(TURNEY, 2002)* se realizó un algoritmo de aprendizaje no supervisado que permitía clasificar reviews como una recomendación positiva ("Thumbs up") o una recomendación negativa ("Thumbs down"), a partir de la extracción de la orientación semántica de las frases que componen cada review. En el mismo año *(PANG, LEE, & VITYANATHAN, 2002)* realizaron un estudio donde evaluaron distintos algoritmos de Data Mining (Naives Bayes, maximum entropy classification, y support vector machines o SVM), sobre una base de datos de opiniones de films, para determinar si la opinión era positiva o negativa; logrando la mejor performance con SVM.

En *(PANG & LEE, 2008)* se realizó un estudio dentro del campo de Opinion Mining y Sentiment Analysis, que tenía como objetivo el análisis y el diseño de un sistema de procesado y búsqueda de información subjetiva (opiniones o revisiones), una de cuyas tareas más importantes es la determinación del sentimiento de una opinión. Hasta ese momento no existían muchos estudios sobre sentimientos contenidos en blog y mucho menos acerca de microblogging.

En *(READ, 2005)* se realizó un estudio aplicando técnicas de Data Mining (SVM y Bayes) para predecir el sentimiento (positivo o negativo) de textos con diferentes tópicos, contextos y fechas de edición; a partir de la inclusión o no de emoticones (tipo ☺ o ☹); los datos fueron extraídos de grupos de noticias de Usenet. En *(GO, HUANG, & BHAYANI, 2009)* se aplicaron técnicas de aprendizaje supervisado (SVM y Maximum Entropy) para determinar el sentimiento (negativo o positivo) de tweets, que contenían emoticones, a partir de una previa clasificación manual de los mismos. El problema surgió cuando se trató de utilizar 3 clases (positivo, negativo y neutro) ya que no se logró una buena precisión. En *(PAK & PAROUBEK, 2010)* se continuó en la línea de trabajo de *(GO, HUANG, & BHAYANI, 2009)* ; centrándose en la recolección de los datos y en el entrenamiento de los clasificadores; y se logró una mejora de la eficiencia de los clasificadores usando 3 clases (positivo, negativo y neutro).

La inclusión de sentimientos "mixtos" se consideró por primera vez en *(OUNIS, MACDONALD, & SOBOROFF, 2008)*. Este estudio es un resumen de los "Blog Track" que formaron parte de las TREC (Text Retrieval Conference) de 2006 y 2007. En este resumen se habla de la importancia de establecer si un blog es subjetivo (o si se incluye opiniones), y en particular se estableces 3 niveles de subjetividad: NEGATIVO (cuando se incluye expresiones de opinion negativas), POSITIVO (cuando la opinion es positiva) y MIXTO (cuando incluye sea opiniones positivas y negativas).

En *(WILSON, WIEBE, & HOFFMANN, 2005)* se realiza un estudio de Sentiment analysis a nivel de oración determinando primero si una oración es neutral o polar (subjetiva). En caso de ser polar se procede a determinar si la polaridad es: Positiva (sean emociones, evaluaciones o determinaciones positivas), Negativa (sean emociones, evaluaciones o determinaciones), Doble (cuando la oración contiene polaridad positiva y negativa) o bien Neutra (cuando se trata de una oración subjetiva que no es ni positiva ni negativa, x ejemplo una especulación).

Existen también métodos basados en léxico para extraer la polaridad de palabras, frases o textos. En *(ESULI & SEBASTIANI, 2006)* se presenta SentiWordNet, un léxico basado en los synset de WordNet *(MILLER, 1995)*, donde a cada synset(s) se le asocian 3 valores Obj(s), Pos(s) y Neg(s); que representarían el grado de objetividad, positividad y negatividad de cada synset (la suma de estos 3 valores da siempre 1). Posteriormente en *(BACCIANELLA, ESULI, & SEBASTIANI, 2010)* se presentó SentiWordNet 3.0, que pasó a utilizar una versión más actualizada de WordNet y que utiliza una variante en el algoritmo que da los valores a los synset.

En este proyecto se plantea la aplicación y evaluación de diversos algoritmos de Data Mining para predecir la polaridad de tweets relacionados a films. Debido a la particularidad de estos textos (son cortos, contienen abreviaciones, utilizan palabras y expresiones de Slang, contienen de errores ortográficos, referencian a páginas web y a otros usuarios, etc.), se probarán y evaluarán distintos

métodos de pre-procesamiento (algunos clásicos y otros específicos a este dominio) con el objetivo de encontrar el pre-procesamiento que maximice la performance, para el presente caso de estudio, de los modelos de clasificación planteados.

Se trabajará con 4 posibles polaridades de un texto (negativo, neutro, positivo y mixto), aprovechando que no hay muchos trabajos con este universo de valores. La razón por lo que se hace esto es hacer los clasificadores "más flexibles" y de esta manera acercarlos más a la realidad, ya que es muy común el hecho de expresar sentimientos positivos y negativos en una misma frase referenciando un mismo objeto (en este caso film).

Por último, se buscará dar una utilidad práctica orientada a los usuarios finales, desarrollando un aplicativo que facilite la decisión de ver o no ver un film en base a la búsqueda y extracción de sentimiento de las opiniones expresadas por otras personas a través de Twitter.

3. Trabajo Propuesto

Como se mencionó en la introducción, el objetivo de este trabajo es evaluar distintos métodos que permitan clasificar un tweet en base a su polaridad, usando como base de entrenamiento las opiniones vertidas por los usuarios de Twitter. En esta red social, los usuarios expresan opiniones en una gran diversidad de dominios, desde electrodomésticos, teléfonos y libros, hasta religión y política. Sin embargo, una palabra puede tener una connotación positiva en un dominio, pero negativa en otro *(LIU, 2010)* (por ejemplo "This film was too long" puede ser algo negativo, sin embargo "The life of the batery was long" es sin duda positivo). Es por esto que se decidió especializar los modelos propuestos en el dominio de las películas o films.

El sistema propuesto se divide en dos aplicaciones principales. La primera (en adelante llamada "*Constructor de Clasificadores*") será una aplicación capaz de crear clasificadores que podrán **inferir** la polaridad de un tweet en particular y evaluar la efectividad del mismo. La segunda aplicación (en adelante llamada "Valuador de film") sería la que, usando los clasificadores construidos, emitirá un informe del valor que los usuarios de Twitter dan a un film determinado, a partir de la búsqueda y extracción de la polaridad de los tweets que mencionan dicho film.

El objetivo del *Constructor de Clasificadores* es dar al usuario la posibilidad de crear y evaluar distintos tipos de clasificadores.

El primer paso para construir un clasificador es tener un conjunto de datos de entrenamiento (en adelanto llamado Dataset). Este dataset se puede construir en la primera etapa del *Constructor de Clasificadores*, llama **Recolección de Datos**. El dataset básicamente es un conjunto de tweets cada uno asociado a su valor de opinión. En primer lugar se descargarán un número importante de tweets que hagan referencia un conjunto de films utilizados como "semilla". Luego se filtrarán aquellos tweets no relevantes o que puedan dificultar el aprendizaje. Por último se determinará manualmente el sentimiento de cada tweet.

En las secciones previas se discutió el valor de opinión que puede tomar un texto, una opinión y en nuestro caso un tweet. En la mayoría de los trabajos existentes en la literatura se trabaja con 3 valores de opinión (Neutro, Positivo y Negativo). En este proyecto se decidió utilizar 4 posibles clases (se agrega el nivel Mixto). Esto se debe a que si la idea es dar valor de opinión a un tweet como una unidad indivisible y, como se mencionó previamente, a pesar de la limitación de 140 caracteres, existen tweets que expresan opiniones distintas y a veces opuestas en relación a un mismo film.

Para llevar a cabo la valoración manual de los tweets se definirán reglas orientativas que guíen el análisis de los tweets con el objetivo de limitar lo más posible la subjetividad de la evaluación.

Como resultado de la etapa de **Recolección de Datos** se obtiene un dataset compuesto de los siguientes campos: id-tweet, usuario, fecha, texto, valor.

La siguiente etapa del *Constructor de Clasificadores* es el **Pre-procesamiento**. Como ya fue mencionado previamente, los textos que formarán el dataset de estudio no son escritos en un lenguaje "correcto" o "bien formado" (abreviaciones, errores de ortografía, inclusión de links, etc.). Esto podría implicar una baja calidad de los datos que a su vez probablemente conduzcan a resultados de baja calidad, por este motivo se requiere llevar a cabo un pre-procesamiento que mejore la calidad de los datos. *(HAN, KAMBER, & PEI, 2006)*

En concreto las tareas que se evaluarán para mejorar la precisión del clasificador serán:

- Eliminación de campos innecesarios.
- Conversión del texto en minúsculas.
- Eliminación de links.
- Eliminación de referencias a otros usuarios.

- Eliminación de hashtag (#)[2].
- Reemplazo del nombre del film por el token [FILM].
- Corrector ortográfico básico.
- Traducción del Slang de Internet.
- Eliminación de Stopwords.
- Stemming.

Como resultado se obtiene un dataset semi-estructurado con 2 campos: texto (ya "limpio") y valor. Finalmente se debe transformar el formato del dataset para que sirva como entrenamiento de los clasificadores. Esta tarea se llama **tokenización**, que se puede definir como el proceso de segmentar un texto en unidades lingüísticas (como palabras, puntos, números, símbolos, etc.)[3] . De esta manera se obtiene como resultado el "dataset final", donde cada tupla consiste de un conjunto de "tokens" (o palabras) más el valor (neutro, positivo, negativo o mixto) dado por el usuario.

En esta etapa se podrán construir distintos dataset, aplicando distintos pre-procesamientos, que luego servirán de entrada a los distintos clasificadores y, a partir de los resultados obtenidos se determinará cuál es el mejor pre-procesamiento posible.

Finalmente, en la última etapa del *Constructor de Clasificadores* se posibilita al usuario crear, entrenar y evaluar los clasificadores. Inicialmente el usuario selecciona el algoritmo que desea utilizar. A los fines de evaluar la performance de distintos algoritmos de clasificación en el caso de estudio presentado, se brindara la posibilidad de elegir entre:

- Inducción de Árbol de decisión.
- Clasificador Bayesiano, un clasificador estadístico que se basa en el teorema de Bayes.
- Support Vector Machine (SVM), este clasificador realiza un mapeo de las instancias del dataset como puntos dentro de un espacio y luego busca crear las divisiones que separan las instancias que pertenece a distintas clases.

Una vez construido un nuevo clasificador, el usuario podrá entrenarlo (para luego aplicarlo para clasificar nuevos tweets) y tambien evaluar la performance del mismo (se utilizará un 10-fold cross validation).

Con el objetivo de comparar el desempeño de los modelos de clasificación frente a un algoritmo basado en léxico, se aplicará un algoritmo llamado SentiStrength *(Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010)* y *(Thelwall, Buckley, & Paltoglou, 2012)*. Este algoritmo lo que hace es asignar a un texto corto dos valores, que representan el sentimiento positivo (+ve) y negativo (-ve) contenido en dicho texto. Estos valores, además contienen la "magnitud" de dicho sentimiento, pudiendo tomar valores entre 1 (no contiene sentimiento) y 5 (mayor magnitud de sentimiento). Por ejemplo si un texto toma valore +ve=5 y –ve=1, significaría un texto con un sentimiento positivo muy fuerte y que no contiene sentimiento negativo; si +ve=3 y –ve=3 se trataría de un texto que contiene sentimientos (sean positivos que negativos) moderados. Este algoritmo utiliza métodos que analizan las gramáticas *de-facto* y los estilos ortográficos del ciber-espacio.

---

[2] El símbolo # (llamado hashtag) es utilizado en Twitter para marcar palabras claves o tópicos de interés. Fue creado por los usuarios de Twitter para categorizar los mensajes. https://support.twitter.com/entries/49309-what-are-hashtags-symbols#

[3] "The Art of Tokenization", developerWorks, Jan 23, 2013 – https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en

Para poder comparar este algoritmo con los modelos de data-mining aplicados se necesita interpretar los resultados de SentiStrengh y "transformarlos" en uno de los 4 valores usados previamente (negativo, neutro, positivo o mixto). Las reglas de transformación utilizadas son las siguientes:

- Si +ve = 1 y –ve = -1 → valor = neutro
- Si +ve > 1 y –ve = -1 → valor = positivo
- Si +ve = 1 y –ve < -1 → valor = negativo
- Si +ve > 1 y –ve < -1 → valor = mixto

Por último se construirá una aplicación, destinada a un usuario final, que permita evaluar un film en particular a partir de las opiniones que los usuarios de Twitter emiten. El funcionamiento básico permitirá al usuario final de ingresar el nombre de un film y luego la aplicación buscará tweets que referencien este film. Una vez descargados los tweets se aplicará un algoritmo de clasificación para determinar la polaridad de cada texto. Finalmente se elaborará un informe final que se presentará al usuario, describiendo y resumiendo las opiniones relacionadas al film analizado, resultando en una sugerencia final recomendando o no de ir a ver dicho film.

4. Plan de Trabajo

- Relevamiento bibliográfico en el área de “Opinion Mining y Sentiment Analysis”. Tiempo estimado: 2 semanas
- Análisis y diseño Conceptual del *Constructor de Clasificadores*. Tiempo estimado: 1 semana
- Análisis, diseño e implementación de la etapa de Recolección de Datos. Tiempo Estimado: 3 semanas
- Construcción del dataset inicial; que comprende la descarga, filtrado inicial y valoración manual de los tweets que conformaran el dataset. Tiempo Estimado: 3 semanas
- Análisis, diseño e implementación de la etapa de pre-procesamiento. Tiempo estimado: 4 semanas
- Análisis, diseño e implementación de la etapa de Entrenamiento y Validación. Tiempo estimado: 3 semanas
- Realización de pruebas. Tiempo estimado: 2 semanas
- Análisis de los resultados obtenidos. Tiempo estimado: 1 semana.
- Análisis, diseño e implementación de la aplicación *Valuador de Film*. Tiempo estimado: 3 semanas.
- Redacción del informe final. Tiempo estimado: en forma paralela a las actividades anteriores y 2 semanas más luego de la finalización de las mismas.

Tiempo total: 24 semanas (6 meses)

5. Referencias

BACCIANELLA, S., ESULI, A., & SEBASTIANI, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10).* Valletta, Malta.

ESULI, A., & SEBASTIANI, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *Proceedings of LREC*, (págs. 417-422).

FAYYAD, U., Gregory, P.-S., & Padhraic, S. (1996). From Data Mining to Knowledge discovery in Databases. *AI Magazine Volume 17 Number 3*.

GO, A., HUANG, L., & BHAYANI, R. (2009). Twitter sentiment analysis. *Entropy vol 17*.

HAN, J., KAMBER, M., & PEI, J. (2006). *Data mining: concepts and techniques. Morgan kaufmann.*

LIU, B. (2010). Sentiment analysis and subjectivity. En N. INDURKHYA, & F. J. DAMERAU, *Handbook of natural language processing - second edition.* Chapman and Hall/CRC.

MILLER, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM - vol 38*, 39-41.

OUNIS, I., MACDONALD, C., & SOBOROFF, I. (2008). On the TREC blog track. *Proceedings of the International Conference on Weblogs and Social Media (ICWSM).*

PAK, A., & PAROUBEK, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of LREC.*

PANG, B., & LEE, L. (2008). *Opinion mining and sentiment analysis.* Foundations and trends in information retrieval.

PANG, B., LEE, L., & VITYANATHAN, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (págs. 79-86). Association for Computational Linguistics.

READ, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *En Proceedings of the ACL Student Research Workshop. Association for Computational Linguistics*, (págs. ent Research Workshop. Association for Computational Linguistics, 2005. p. 43-48).

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social Web. *Journal of the American Society for Information Science and Technology*.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal texT. *Journal of the American Society for Information Science and Technology*.

TURNEY, P. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th annual meeting on association for computational linguistics* (págs. 417-424). Association for Computational Linguistics.

WILSON, T., WIEBE, J., & HOFFMANN, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (págs. 347-354). Association for Computational Linguistics.