UNA REVISIÓN A LA MINERIA DE OPINIONES Y LOS RETOS DEL PNL

A review of opinion mining and NLP challenges

HECTOR NIGRO[1]

Recibido: 11 de noviembre de 2019. Aceptado: 15 de diciembre de 2019
DOI: http://dx.doi.org/10.21017/rimci.2020.v7.n13.a80

Resumen

En los últimos años se ha generado un crecimiento en el análisis de las redes sociales para tener una idea de lo que la gente piensa sobre los temas de interés actuales, sin embargo, los sistemas de minería de texto originalmente diseñados para tipos de textos más regulares, como los artículos de noticias, pueden necesitar adaptarse para tratar publicaciones de redes sociales como Facebook, tweets, etc. En este artículo, se presenta una reflexión sobre temas relacionados con la minería de opinión de las redes sociales y los desafíos que imponen en un sistema de procesamiento de lenguaje natural (PNL).

Palabras clave. Mineria de opiniones; Procesamiento de Lenguaje Natural; Redes Sociales.

Abstract

In recent years, there has been a growth in the analysis of social networks to get an idea of what people think about current topics of interest, however, text mining systems originally designed for more regular types of texts Like news articles, they may need to be adapted to deal with social media posts like Facebook, tweets, etc. In this article, a reflection is presented on issues related to mining opinion from social networks and the challenges they impose on a natural language processing system (NLP).

Key words. Opinion mining; Natural Language Processing; Social networks.

I. Introducción

En esta nueva era de la sociedad de la información, donde los pensamientos y las opiniones se comparten a ritmo acelerado a través de las redes sociales, la información cobra valor y las herramientas que pueden dar sentido al contenido de estas redes son primordiales. Para hacer un mejor uso de esta información, se requiere distinguir lo que es importante e interesante y poder hacer una clasificación al respecto. Al realizar esta segmentación existen multiples beneficios para las empresas, los gobiernos, etc., al comprender lo que el público piensa acerca de sus productos y servicios, pero también le interesa a las grandes instituciones de conocimiento público poder recopilar, recuperar y preservar toda la información relacionada con ciertos eventos y su desarrollo. Con el tiempo, la difusión de información a través de las redes sociales también puede desencadenar una cadena de reacciones a tales situaciones y eventos que finalmente conducen a cambios administrativos, políticos y sociales.

El contenido y el lenguaje que generan los usuarios es muy dinámico y cambia rápidamente para reflejar las fluctuaciones sociales y sentimentales de las personas. Tambien existe diversidad en las opiniones y acciones del público, desde simples botones «Me gusta» hasta artículos completos. Las actividades de los usuarios en sitios de redes sociales a menudo se desencadenan por eventos es-

1 Ingeniero de Sistemas (Unicen), Magister en Ciencias Políticas y Sociales (Flacso), Candidato a Doctor en Matemática Computacional e Industrial Aplicada (Unicen). Docente Investigador. Correo electrónico: oscarnigro@unicer.com.ar

pecíficos y entidades relacionadas (por ejemplo, eventos deportivos, celebraciones, crisis, artículos nuevos, personas, ubicaciones) y temas (por ejemplo, politicos, calentamiento global, crisis financiera, gripe porcina) .

En ese sentido, la explotación de Web 2.0 y la sabiduría de las multitudes pueden hacer que el archivo web sea un proceso más selectivo y basado en el significado. El análisis de los medios sociales puede ayudar a seleccionar material para su inclusión, proporcionando una evaluación del contenido a través de la web social, mientras que la minería de los medios sociales en sí puede enriquecer los archivos, avanzando hacia la preservación estructurada en torno a categorías semánticas.

En este artículo, nos centramos en los desafíos en el desarrollo de herramientas de minería de opinión que, junto con el reconocimiento de entidades, temas y eventos, forman la piedra angular para el análisis de redes sociales teniendo como base la adaptación de las herramientas de minería de opinión a los medios sociales, y los desafíos que imponen en un sistema de procesamiento de lenguaje natural (PNL).

Existen muchos desafíos inherentes a la aplicación de técnicas típicas de minería y análisis de sentimientos en los medios sociales. Los microposts como los tweets son, en cierto sentido, el tipo de texto más desafiante para las herramientas de minería de textos, y en particular para la minería de opinión, ya que no contienen mucha información contextual y suponen mucho conocimiento implícito. La ambigüedad es un problema particular, ya que no se puede emplear fácilmente la información de referencia central: a diferencia de las publicaciones de blog y los comentarios, los tweets generalmente no siguen un hilo de conversación, y aparecen mucho más aislados de otros tweets. También exhiben mucha más variación en el lenguaje, tienden a ser menos gramaticales que las publicaciones más largas, contienen mayúsculas poco ortodoxas y hacen uso frecuente de emotividades, abreviaturas y hashtags, que pueden formar una parte importante del significado. Por lo general, también contienen un uso extenso de ironía y sarcasmo, que son particularmente difíciles de detectar por una máquina.

La mayoría de las técnicas de minería de opinión utilizan el aprendizaje automático, pero esto no ayuda en aplicaciones en las que intervienen varios dominios, idiomas y tipos de texto diferentes, porque los modelos tienen que ser entrenados para cada uno y se requieren grandes cantidades de datos de entrenamiento para obtener buenos resultados. Por lo general, los clasificadores creados utilizando métodos supervisados [1], se desempeñan bien en las tareas de detección de polaridad, pero cuando se utilizan en nuevos dominios, la precisión se reduce desastrosamente [2]. Mientras que algunos trabajos se han centrado el uso de diferentes palabras clave en tipos de texto similares [3].

II. Estado del arte

Diversos textos presentan una revisión amplia y detallada de las técnicas tradicionales de detección automática de sentimientos [4]. En general, las técnicas de detección de sentimientos estan basados en métodos de aprendizaje automático[5-7]; por medio de una colección de términos de sentimientos conocidos y precompilados. Los enfoques de aprendizaje automático hacen uso de características sintácticas y / o lingüísticas [8].

Sin embargo, estas técnicas relativamente exitosas a menudo fallan cuando se mueven a nuevos dominios o tipos de texto, porque son inflexibles con respecto a la ambigüedad de los términos de sentimiento. El contexto en el que se usa un término puede cambiar su significado, particularmente para adjetivos [9]. Varias evaluaciones han demostrado la utilidad de la información contextual [10], y han identificado palabras de contexto con un alto impacto en la polaridad de los términos ambiguos [11].

Las técnicas de minería de opinión han comenzado a centrarse en las redes sociales, combinadas con una tendencia hacia su aplicación como un mecanismo proactivo en lugar de reactivo. La comprensión de la opinión pública puede tener importantes consecuencias para la predicción de eventos futuros. Una de las aplicaciones más obvias de esto es para las predicciones del mercado de valores: descubrieron que, contrariamente a la expectativa de que si los mercados de valores caían, el estado de ánimo público también se volvería más negati-

vo, de hecho, una caída del público como precursor de una caída en el mercado de valores [12]. Casi todo el trabajo sobre minería de opinión de Twitter ha utilizado técnicas de aprendizaje automático [8]. Tenían como objetivo clasificar los tweets arbitrarios sobre la base de sentimientos positivos, negativos y neutrales, construyendo un clasificador binario simple que utilizara características de n-gramas y POS, y entrenados en instancias que habían sido anotadas de acuerdo con la existencia de positivos y emociones negativas.

También existe una gran cantidad de herramientas comerciales basadas en búsquedas para realizar análisis de sentimientos de tweets. En general, el usuario ingresa un término de búsqueda y recupera todos los tweets positivos y negativos (y a veces neutros) que contienen el término, junto con algunos gráficos, como gráficos circulares o gráficos.

A. Minería de opinión

Existe una serie de aplicaciones iniciales para la minería de opinión desde las redes sociales utilizando GATE [13], un conjunto de herramientas disponible gratuitamente para el procesamiento del lenguaje. Basados en la identificación en tweets de sentimientos sobre partidos políticos, existen aplicaciones que se extienden a un análisis genérico del sentimiento sobre cualquier tipo de entidad o evento mencionado, dentro de dos dominios específicos: la actual crisis financiera y la música. En ambos casos, se realiza un primer análisis de sentimiento básico al asociar un sentimiento positivo, negativo o neutral a cada objetivo de opinión relevante, junto con una polaridad Puntuación. En los escenarios actuales, esto podría ser cualquier entidad o evento que sea pertinente para el dominio y el caso de uso. En la musica, este podría ser el evento general, una banda o el desempeño particular de una banda en el concierto, o algún sub-evento como un espectáculo de luces que ocurrió durante la presentación. En la parte politica, esto podría ser una persona político como tal, una organización, un evento como una huelga general o una reunión relevante que tuvo lugar.

B. Extracción de la entidad

La aplicación de minería de opinión primero requiere que el cuerpo se anote con entidades y eventos. Para esto también existen aplicaciones de reconocimiento predeterminado de Entidad, para encontrar menciones de Persona, Ubicación, Organización, Fecha, Hora, Dinero y Porcentaje, que brindan valores de características, en los patrones lingüísticos para la minería de opinión..

C. Reconocimiento de eventos

Además de las entidades, también se identifican eventos para ser utilizados como posibles objetivos para las opiniones, y como entrada para otros procesos como la extracción de temas. Los eventos pueden expresarse mediante elementos de texto como predicados verbales y sus argumentos, frases nominales encabezadas por nominalizaciones («crecimiento económico»), combinaciones adjetivo-sustantivo («medida gubernamental»; «dinero público») y sustantivos referentes a eventos ("crisis", "inyección de efectivo"). El método basado en patrones implica el reconocimiento de entidades y las relaciones entre ellas para encontrar eventos y situaciones específicos del dominio.

D. Análisis de los sentimientos

El enfoque que se toma para el análisis de sentimientos está basado en reglas centrándose en la creación de una serie de subcomponentes que tienen un efecto en la puntuación de un sentimiento. El cuerpo principal de la aplicación de minería de opinión involucra un conjunto de gramáticas que crean anotaciones en segmentos de texto. Las reglas gramaticales utilizan información de diccionarios geográficos combinada con características lingüísticas e información contextual para crear un conjunto de anotaciones y características, que pueden modificarse en cualquier momento por otras reglas. El conjunto de listas de diccionarios geográficos contiene pistas útiles y palabras de contexto.

Una vez que las palabras que llevan el sentimiento han sido emparejadas, se hace un intento de encontrar una relación lingüística entre una entidad o evento en la oración o frase, y una o palabras que llevan el sentimiento, como un adjetivo que lleva el sentimiento modificando una entidad o en aposición con ella, o un verbo portador de asentimiento cuyo sujeto u objeto directo es una entidad. Si se encuentra dicha relación, se crea una anotación de Sentimiento para esa entidad o even-

to, con características que denotan la polaridad (positiva o negativa) y el puntaje de polaridad. La puntuación inicial asignada se basa en la escucha del diccionario geográfico de las palabras de opinión relevantes. El concepto detrás de la calificación (y la decisión final sobre la polaridad del sentimiento) es que la calificación predeterminada de una palabra puede ser alterada por varias pistas contextuales. Por ejemplo, típicamente una palabra negativa encontrada en una asociación lingüística invertirá la polaridad de positiva a negativa y viceversa. De manera similar, si el sarcasmo se detecta en la declaración, la polaridad se invierte (en la mayoría de los casos, el sarcasmo se usa junto con una declaración aparentemente positiva, para reflejar una negativa, aunque esto puede no ser necesariamente cierto para otros idiomas además del inglés). Los adverbios que modifican un adjetivo de sentimiento generalmente tienen el efecto de aumentar su intensidad, lo que se refleja multiplicando el factor de intensidad del adverbio (definido en una lista de diccionario geográfico) por la puntuación existente del adjetivo. Por ejemplo, si «brillante» tuvo un puntaje de 0.4, y «absolutamente» tuvo un factor de intensidad de 2, entonces el puntaje de «brillante» aumentaría a 0.8 cuando se encuentra en la frase «absolutamente brillante». Actualmente, los factores de intensidad se definen manualmente, pero algunos de estos también podrían generarse automáticamente cuando se derivan morfológicamente de un adjetivo (por ejemplo, podríamos usar la puntuación de sentimiento del adjetivo «brillante» definido en nuestra lista de adjetivos para generar un factor de intensidad para el verbo ad «brillantemente»). Las palabras juradas, por otro lado, tienen un papel un poco más complejo y en temas como la política y la religión, donde las personas tienden a tener opiniones muy fuertes.

Los emoticones se procesan como otras palabras que transmiten sentimientos, de acuerdo con otra lista de diccionarios geográficos, si ocurren en combinación con una entidad o evento. Por ejemplo, el tuit «Todos votaron por pepe :-(« sería anulado como negativo con respecto al objetivo «pepe». De lo contrario, en cuanto a las malas palabras, si una oración contiene una carita sonriente pero ninguna otra entidad o evento, la oración se anota anotada con el asentimiento, con el valor de la carita sonriente de la lista del diccionario geográfico. Una vez que todos los subcomponentes se han pasado por encima del texto, se produce una salida final para cada segmento que lleva el sentimiento, con un polaridad (positiva o negativa) y una puntuación.

E. Problemas multilingües

Otro artefacto de las redes sociales que consisten en blogs, foros, páginas de Facebook, colecciones de Twitter, etc., a menudo son multilingües. Los comentarios y tweets pueden estar en cualquier idioma, por lo tanto, se emplea una herramienta de identificación de idioma para determinar el idioma de cada oración. La identificación del idioma en los tweets es un problema particular, debido a su corta longitud (140 caracteres como máximo) y la ubicuidad de los tokens independientes del idioma (RT (retweet) , hashtags, @menciones, números, URL, emoticones). A menudo, una vez que se eliminan, un tweet contendría menos de 4 o 5 palabras, algunos incluso no tendrían palabras «adecuadas».

F. Desafíos impuestos por las redes sociales

Además de los factores ya discutidos, las redes sociales imponen una serie de desafíos adicionales en un sistema de minería de opinión. Incluso cuando un rastreador está restringido a temas específicos e identifica correctamente las páginas relevantes, esto no significa que todos los comentarios sobre dichas páginas también sean relevantes. Este es un problema particular para las redes sociales, donde las discusiones y los hilos de comentarios pueden divergir rápidamente en temas no relacionados, a diferencia de las revisiones de productos que rara vez se desvían del tema en cuestión.

Hay varias formas en que se puede tratar de lidiar con el problema de relevancia. Primero, se tiene el clasificador para tweets o comentarios que sean relevantes, por ejemplo, ignorar los tweets si contienen ciertos términos. En segundo lugar, utilizar la agrupación para encontrar oraciones o segmentos de opinión relacionados con ciertos temas, y descartar aquellos que quedan fuera de estos temas. Este es probablemente el enfoque más prometedor, especialmente dado que se utiliza algoritmos de agrupación de temas, aunque corre el riesgo de que se omitan algunos comentarios relevantes.

G. Identificación del objetivo

Un problema que enfrentan muchos enfoques basados en la búsqueda para el análisis del sentimiento es que el tema del documento recuperado no es necesariamente el objeto del sentimiento contenido allí. De modo que, si bien la polaridad de la opinión puede ser correcta, el tema o el objetivo de la opinión puede ser algo totalmente diferente. Una forma en la que se trata este problema es mediante el uso de un enfoque centrado en la entidad, mediante el cual primero se identificam la entidad relevante y luego se busca opiniones semánticamente relacionadas con esta entidad, en lugar de tratar de decidir cuál es el sentimiento sin referencia. Hay varias formas en que las oraciones que contienen sentimientos pero que no tienen un enlace obvio de opinión pueden ser anotadas. En la actualidad, simplemente se identifica estas intenciones como «que contienen sentimientos», pero no se hacen suposiciones sobre el objetivo.

H. Negación

Los clasificadores de sentimientos de palabras más simples tienen la debilidad de que no manejan bien la negación; la diferencia entre las frases «no bueno» y «bueno» es algo que se ignora en un modelo de unigrama, aunque tienen significados completamente diferentes. Una posible solución es incorporar características de mayor alcance, como estructuras de dependencia de orden superior, lo que ayudaría a capturar patrones más completos y sutiles, como en la oración "Sorprendentemente, la calidad de construcción está muy por encima del par, considerando el resto de las características « en donde el término «sorprendentemente» debería negar parcialmente el sentimiento general positivo [4]. Otra forma de lidiar con la negación, evitando la necesidad de analizar la dependencia, es capturar simples patrones como «no es útil» o «no emocionante» insertando unigramas como «NO útil» y «NO emocionante» respectivamente.

I. Información contextual

Los medios sociales, y en particular los tweets, generalmente asumen un nivel mucho más alto de conocimiento contextual y mundial por parte del lector que los textos más formales. Esta información puede ser muy difícil de obtener automáticamente.

Una ventaja de los tweets, en particular, es que tienen una cantidad mínima de metadatos asociados con ellos que pueden ser útiles, no solo para el resumen y la agregación de opiniones sobre una gran cantidad de tweets, sino también para la desambiguación. Los ejemplos de estos metadatos incluyen la fecha y la hora, el número de seguidores de la persona que tuitea, la ubicación de la persona e incluso su perfil. Por ejemplo, se puede tener información sobre la afiliación política de esa persona mencionada en su perfil, que podemos usar para ayudar a decidir si su tweet es sarcástico cuando parecen ser positivos sobre una figura política en particular. Debido a que cada persona registrada en Twitter tiene un ID único, se puede desambiguar entre diferentes personas con el mismo nombre, algo que puede ser problemático en otros tipos de texto.

J. Volatilidad en el tiempo

Los medios sociales, especialmente Twitter, exhiben una dinámica temporal muy fuerte. Más específicamente, las opiniones pueden cambiar radicalmente con el tiempo, de positivas a negativas y viceversa. Dado que también existe una correlación entre los dos dominios, también deben explorarse modelos conjuntos de opiniones políticas y opiniones del mercado financiero. Para abordar este problema, los diferentes tipos de opiniones posibles se asocian como propiedades ontológicas con las clases que describen entidades, hechos y eventos, descubiertas a través de técnicas de extracción de información y técnicas de anotación semántica.

Las opiniones y los sentimientos extraídos tienen una marca de tiempo y se almacenan en una base de conocimiento, que se enriquece continuamente a medida que entran nuevos contenidos y opiniones. Una pregunta particularmente desafiante es cómo detectar nuevas opiniones emergentes, en lugar de agregar la nueva información a una opinión existente para la entidad dada. Las contradicciones y los cambios también fueron capturados y utilizados para rastrear las tendencias a lo largo del tiempo, en particular a través de la fusión de opiniones.

K. Evaluación

La evaluación de la minería de opinión puede ser complicada, por varias razones. Primero, las

opiniones son a menudo subjetivas, y no siempre está claro lo que pretendía el autor.

Es muy difícil evaluar los puntajes de polaridad, o cuán correcto es el puntaje. Sin embargo, si bien estos puntajes técnicamente representan fuerza de opinión, se pueden ver como un indicador de confianza. Por lo tanto, se espera que los sentimientos expresados con puntajes de alta polaridad tengan una alta precisión, y podemos adaptar nuestra evaluación en consecuencia, buscando tasas de precisión más altas a medida que aumenta el puntaje de polaridad.

Gran parte del éxito de la opinión centrada en la entidad herramienta minera depende de la calidad de las entidades y eventos extraídos. Debido a que se adopta una estrategia de alta precisión, a expensas potenciales del retiro, nuestro objetivo es minimizar este efecto.

Conclusiones

Si bien el desarrollo de las herramientas de minería de opinión esta en progreso, los resultados iniciales son prometedores y se confia en que las estrategias de retroceso inherentes a la metodología incremental permitirán un sistema exitoso. Abogamos por el uso de técnicas bastante poco profundas para gran parte del procesamiento lingüístico, utilizando, por ejemplo, la fragmentación en lugar del análisis completo. Por otro lado, los subcomponentes lingüísticos también se pueden usar como preprocesamiento inicial para proporcionar funciones para el aprendizaje automático, donde tales datos están disponibles, y actualmente estamos experimentando con tales técnicas. Con este trabajo de revisión se espera avanzar en la etapa de desarrollo se software e implementación de algoritmos y redes neuronales considerando la relevancia de la información que cobra valor en esta era de la sociedad de la información.

Referencias

[1] E. Boiy, Pieter Hens, Koen Deschacht, and Marie-Francine Moens. Automatic sentiment analysis of onlinetext. InProc. of the 11th International Conference on Electronic Publishing, Vienna, Austria. 2007.

[2] A. Aue and M. Gamon. Customizing sentiment classifiers to new domains: a case study. InProc. of the International Conference on Recent Advances in Natural Language Processing, Borovetz, Bulgaria. 2015.

[3] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In Annual Meeting-Association For Computational Linguistics, page 440. 2007.

[4] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis.Information Retrieval, 2(1). 2008.

[5] A. Popescu and O. Etzioni. Extracting product features and opinions from reviews. InProc. of the Confer-ence on Empirical Methods for Natural Language Pro-cessing (EMNLP), pages 339–346, Vancouver, Canada. 2005.

[6] A. Scharl and A. Weichselbraun. An automated ap-proach to investigating the online media coverage of USpresidential elections.Journal of Information Technol-ogy and Politics, 5(1):121–132. 2008.

[7] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon based methods for sentiment analysis. Computational Linguistics, 1(September 2010):1–41. 2011.

[8] A. Pak and P. Paroubek. Twitter as a corpus forsentiment analysis and opinion mining. Proceedings of LREC. 2010.

[9] A. C. Mullaly, C.L. Gagné, T.L. Spalding, and K.A.Marchak. Examining ambiguous adjectives inadjective noun phrases: Evidence for representation as ashared core meaning. The Mental Lexicon, 5(1):87–114. 2010.

[10] A. Weichselbraun, S. Gindl, and A. Scharl. Acontext-dependent supervised learning approach to sen-timent detection in large textual databases. Journal of Information and Data Management, 1(3):329–342. 2010.

[11] S. Gindl, A. Weichselbraun, and A. Scharl. Cross-domain contextualisation of sentiment lexicons. InProceedings of 19th European Conference on Artificial Intelligence (ECAI-2010), pages 771–776. 2010.

[12] Johan Bollen and Huina Mao. Twitter mood as astock market predictor. IEEE Computer, 44(10):91–94. 2011.

[13] H. Cunningham, D. Maynard, and V. Tablan. JAPE:a Java Annotation Patterns Engine (Second Edition). Re-search Memorandum CS–00–10, Department of Computer Science, University of Sheffield, November. 2000.