

Caso de estudio de análisis de sentimientos en Twitter: Tratado de libre comercio de América del Norte

Diego Aguilar, Grigori Sidorov, Ildar Batyrshin

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

diego.aguilar.m15@gmail.com, sidorov@cic.ipn.mx, batyr1@gmail.com

Resumen. En este artículo se implementan técnicas de minería de texto, procesamiento del lenguaje natural y aprendizaje automático para realizar un análisis de sentimientos en Twitter; utilizando para ello el tema central del Tratado de libre comercio de América del Norte. Se muestran los resultados de aplicar diferentes algoritmos de aprendizaje automático; donde se concluye que al usar en conjunto los algoritmos se obtiene el mejor rendimiento con el 87.87 % de exactitud. Se implementó una aplicación en tiempo real para el análisis de sentimientos en Twitter.

Palabras clave: Análisis de sentimientos, minería de texto, aprendizaje automático, Twitter, Tlcan.

Case Study of Sentiment Analysis in Twitter: North American Free Trade Agreement

Abstract. This paper text mining techniques, natural language processing and automatic learning are implemented in order to develop a sentiment analysis on tweets with the topic of North American Free Trade Agreement (NAFTA). The results of applying different machine learning algorithms are shown; where it is concluded that the conjunction of all algorithms obtains the best performance with the 87.87 % of precision. A real time application was implemented for the analysis of feelings on Twitter.

Keywords: Sentiment analysis, text mining, machine learning, Twitter, Nafta.

1. Introducción

Recientemente se ha incrementado el interés por la extracción de conocimiento de manera automática en grandes bancos de datos en los campos de economía, ciencias, salud, etc.

Además actualmente se cuenta con un gran flujo de información en redes sociales y esta es una gran oportunidad para desarrollar algoritmos que en tiempo real generen información útil del mundo que nos rodea.

1.1. Tratado de Libre Comercio de América del Norte

El tratado de libre comercio de América del Norte (TLCAN) es un acuerdo entre Canadá, Estados Unidos y México para crear una zona en donde se reduzcan los costos de comercio e intercambio de bienes entre los países involucrados.

Originalmente el pacto lo comprendían Canadá y Estados Unidos; sin embargo a principios de los noventa se anexó a México al TLCAN iniciando así una nueva era de acuerdos diplomáticos entre los tres países. El TLCAN tiene como principales objetivos:

- Facilitar el comercio de bienes y servicios,
- Promover competencia económica,
- Impulsar la inversión entre los tres países,
- Proteger los derechos de propiedad intelectual,
- Establecer una cooperación trilateral para mejorar los beneficios del tratado.

El TLCAN ha sido motivo de discusiones en torno al beneficio que ofrece a sus partes, en donde hay opiniones encontradas las cuales lo apoyan [3] o bien indican que ha traído dificultades políticas y económicas a los países involucrados [2].

TLCAN en la actualidad Durante la campaña presidencial de Donald Trump se anunció que se buscaría una renegociación del TLCAN y en caso de no lograr un acuerdo satisfactorio su gobierno abandonaría el tratado. Esto provocó discusiones y desacuerdos entre los países cuando Trump se convirtió en presidente de Estados Unidos.

Lo anterior hace que las negociaciones del TLCAN y su estudio en redes sociales generan información valiosa respecto al sentimiento general que se tiene sobre el tema.

El resto del artículo esta comprendido como sigue: La Sección 2 muestra el estado del arte de los trabajos relacionados al análisis de sentimientos en Twitter. La descripción de los métodos de minería de texto, procesamiento de lenguaje natural y algoritmos de aprendizaje automático son descritos en la Sección 3. Los resultados obtenidos se muestran en la Sección 4 respectivamente. Finalmente, en la Sección 5 son presentadas las conclusiones y el trabajo a futuro.

2. Trabajos relacionados

Como se menciona en la Sección 1 el interés por parte de la comunidad científica para el análisis de textos ha aumentado. No obstante la mayor parte de las investigaciones realizadas se enfocan en el idioma inglés, para los cuales se han alcanzado altos porcentajes de clasificación correcta y se han generado corpus de calidad [1].

2.1. Análisis de sentimientos en Twitter

Específicamente la plataforma de Twitter se ha utilizado ampliamente para tareas de agrupamiento, clasificación de textos, detección de tópico, etc.

En el taller TASS [7] del año 2012 se propuso un corpus de Tweets en español para realizar las tareas de análisis de sentimientos y clasificación de tópicos. El corpus contiene un total de 70000 tweets distribuidos en 10 temas y escritos por 200 conocidas personalidades de habla hispana. Durante el taller el porcentaje más alto para análisis de sentimientos obtuvo un 71.12% de precisión utilizando tres clases (*pos*, *neg*, *neu*).

Antonio Fernández et al. [1] en su investigación realizaron una comparación de diversa técnicas de procesamiento de lenguaje natural (PLN) y aprendizaje automático (ML por sus siglas en inglés *machine learning*) para un corpus de Tweets en español con la finalidad de realizar tareas de identificación de tópicos y análisis de sentimientos.

Concluyeron que aunque la mayoría de los trabajos se centra en el idioma inglés, recientemente ha crecido la investigación en otros idiomas. Para la selección e tópicos la precisión más alta alcanzada fue 58%, mientras un 42% para clasificación de sentimientos. En su estudio utilizaron los algoritmos Ibk, Complement Naive Bayes, Naive Bayes Multinomial, Random Committe y SMO.

En el trabajo de Sidorov et al. [6] se realizó un estudio exhaustivo para análisis de sentimientos en Twitter, utilizando la herramienta WEKA ¹ se comprobaron los resultados de clasificación de tweets en español para los algoritmos Naive Bayes, J48 y SVM. Como parte del preprocesamiento de datos utilizaron módulos para corrección de errores, POS-tagging y negación de sentimientos. Los autores concluyen que los parámetros ideales de configuración son: (1) utilizar unigramas como características, (2) usar solamente las clases *positivo* y *negativo*, (3) un corpus de al menos 3,000 tweets para entrenamiento, (4) un corpus balanceado arrojó resultados peores que uno des-balanceado y (5) las SVM presentan la mejor clasificación.

3. Materiales y métodos

El panorama general para extraer información de sentimientos en tweets puede agruparse en cuatro pasos. Primero se debe reunir un corpus dedicado al tema a tratar (en este caso tweets relacionados al TLCAN); después se realizan métodos de PLN para limpiar obtener las características de los tweets. Finalmente, se entrenan los algoritmos de ML y se evalúan con criterios de medición estándar.

3.1. Corpus de tweets

Twitter es una plataforma de red social en donde los usuarios pueden expresar pensamiento o noticias en un párrafo breve (actualmente el límite es de 280

¹ Disponible en línea: <https://www.cs.waikato.ac.nz/ml/weka/>

caracteres). En el año 2018 Twitter cuenta con 330 millones de usuarios activos al mes con soporte para 40 idiomas [4].

Con el API de Twitter para desarrolladores obtuvimos un corpus con 5054 tweets relacionados al tema TLCAN, es decir que en su texto incluyeran “#TLCAN” para el periodo de Abril a Mayo del 2018. Los tweets fueron clasificados de forma manual (*Gold standard*) en una división de tres clases:

- POS Ejemplo: “EU optimista ante TLCAN.”
NEG Ejemplo: “No ha sido fácil negociar el TLCAN con México y Canadá: Trump.”
NEU Los tweets neutrales son de carácter informativo o bien contienen una combinación de sentimientos positivos y negativos. Ejemplo: “Pese avances en TLCAN, no hubo luz en puntos críticos.”

La Tabla 1 muestra la distribución de clase en el corpus; como se observa, las clases POS y NEG se encuentran más balanceadas en comparación con la clase NEU que predomina en el corpus.

Tabla 1. Distribución de clases en el corpus.

Clase	No. tweets	Porcentaje
POS	1,342	24.1 %
NEG	1,618	28.9 %
NEU	2,625	47 %
Total	5,585	100 %

3.2. Preprocesamiento

Cabe mencionar que la tarea de clasificación de sentimientos es difícil incluso para los humanos, en donde a veces no posible llegar a un acuerdo respecto a la polaridad del tweet [7].

Además los textos en tweets están sujetos a errores de ortografía, abreviaciones, modismos y humor o sarcasmo [6].

De forma similar al proceso seguido en [6]; utilizamos los siguientes métodos para preprocesar los tweets:

Corrección de errores: Para corregir algunos de los errores en el corpus se utilizó un diccionario en donde manualmente se introducen las correcciones ortográficas a los errores más comunes. Por ejemplo, los acentos en las preguntas “Qué, Cómo,Cuál, Dónde”.

Además se eliminaron los símbolos extras como “(”, “), ”, “!”, “.”, “:”, “;”, etc.

Etiquetas especiales: Una característica de los tweets es la presencia de *hashtags* (#) los cuales generalmente son utilizados para etiquetar tweets en un tema en particular. Además, frecuentemente los tweets contienen URLs con direcciones a páginas de noticias o bien utilizan *emojis*.

Para lidiar con lo anterior se realiza una limpieza del texto. en un primer acercamiento se cambiaron los URL y emojis por etiquetas especiales (\$URL\$ y \$SMILEY\$); no obstante, en un segundo enfoque se eliminaron por completo estas etiquetas de los tweets, ya que así se mejoraron los resultados de clasificación.

Por otra parte, los *hashtags* generalmente son escritos utilizando un modelo de *CamelCase* si son varias palabras juntas, e.g “#FelizJueves”. Por lo cual se utilizó un método para eliminar el símbolo (#) y separarlos en palabras, de tal forma que el ejemplo anterior sería: “Feliz Jueves”.

POS-tagging: Después de realizar una limpieza a los textos utilizamos el paquete *Natural Language Toolkit* (NLTK) de Python para el etiquetado de *Part of Speech*; una vez realizado este procedimiento obtenemos la clasificación de las palabras en: verbos, adjetivos, sustantivos, etc. De este conjunto de etiquetas realizamos un filtro para conservar solamente los adjetivos, adverbios y verbos.

Lematización: En un primer enfoque se utilizó la lematización de las palabras para reducir el número de características en el vector de entrada; no obstante, se obtuvieron mejores resultados al no utilizar este método por lo que se optó por no utilizarlo en la versión de tiempo real.

Negación de polaridad: Para el análisis de sentimientos es muy importante considerar la negación del sentido de la oración, es decir que el uso de ciertas palabras pueden cambiar el significado del texto. Particularmente, en este trabajo cuando se encuentra la palabra “no”, por ejemplo en “no estoy feliz”; se utiliza el método descrito en [6] y se anexa a la palabra inmediata siguiente. Quedando así la oración anterior como: “no_estoy feliz”.

Es importante destacar que existen otros métodos para combatir la negación, en [1] al encontrar la palabra “no” cambian la polaridad de las siguientes tres palabras. Además, existen adjetivos como “muy” o “poco” que acentúan o disminuyen el valor del sentimiento [5]. En este trabajo son omitidas este tipo de palabras.

3.3. Clasificadores

Usamos clasificadores implementados en las paqueteras NLTK y Sklearn de Python, dichos algoritmos son: Naive Bayes, gradiente descendente estocástico (SGDC), máquina de vectores de soporte (LinearSVC).

Los algoritmos utilizan de entrada un vector de características formadas por uni-gramas y filtradas usando POS-tagging.

Del corpus de tweets utilizamos el 80 % para entrenamiento y el 20 % para pruebas. Para estabilizar los resultados obtenidos por los clasificadores utilizamos el método *k-cross validation*.

3.4. Medidas de evaluación

Calculamos la exactitud de cada uno de los clasificadores en el set de pruebas con la Ecuación 1:

$$exactitud = \frac{\text{respuestas correctas}}{\text{total de respuestas}}. \quad (1)$$

Utilizamos la Ecuación 2 para calcular la exhaustividad de los clasificadores:

$$exhaustividad = \frac{\text{respuestas correctas}}{\text{total de posibles respuestas correctas}}. \quad (2)$$

Finalmente, la medida F1 utiliza los valores anteriores para mostrar un rendimiento general (Ecuación 3):

$$F1 = \frac{2 \times exactitud \times exhaustividad}{exactitud + exhaustividad}. \quad (3)$$

3.5. Votación de clase

Para la clasificación de tweets en tiempo real utilizamos un votador, el cual clasifica el texto con las mejores versiones de los algoritmos de ML y calcula la moda de la clase pronosticada. Si existe un acuerdo entre los votadores con un umbral mayor al 70 % entonces se clasifica al tweet con el sentimiento acordado, en caso contrario, no existe un acuerdo entre los votantes y el tweet se deshecha.

4. Resultados

En esta sección se describen los resultados obtenidos del entrenamiento. La Tabla 2 muestra los mejores porcentajes de exactitud por algoritmo; también se muestra el resultado de combinar todos los clasificadores en un votador, cuya exactitud alcanzó un 87.8 % en el conjunto de pruebas.

Tabla 2. Resultados para algoritmos de entrenamiento.

Algoritmo	Exactitud	Exhaustividad	F1
Naive Bayes	65.86 %	65.92 %	65.76 %
Multinomial NB	66.40 %	64.87 %	65.44 %
C4.5	65.02 %	64.44 %	64.66 %
Bernoulli NB	66.81 %	64.41 %	62.90 %
Regresión logística	70.79 %	68.00 %	69.03 %
SGDC	68.22 %	67.12 %	67.22 %
LinearSVC	69.84 %	69.19 %	69.42 %
Votador	88.60 %	87.35 %	88.44 %

Utilizando el API de Twitter para realizar un streaming de datos es posible obtener en tiempo real todos los tweets relacionados al tema TLCAN. Después, son clasificados con el votador y si supera el umbral mencionado en la Sección 3.5 se guarda el resultado para finalmente visualizarlo en dos gráficas.

Análisis de sentimientos en el tiempo: La Figura 1 muestra la evolución del sentimiento general sobre el tema TLCAN. En el eje horizontal se muestra el tiempo en donde un tweet se publicó; mientras que en el eje vertical se muestra el valor del sentimiento, cuyo rango se encuentra en $[-1, 1]$

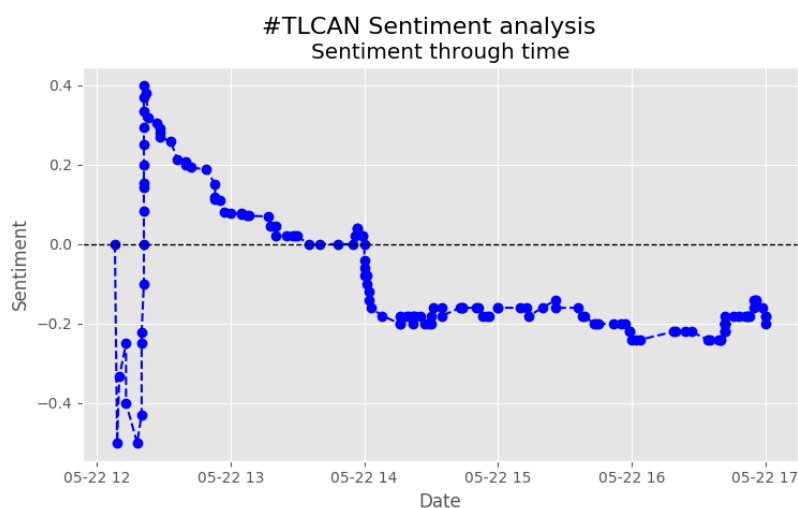


Fig. 1. Análisis de sentimientos a través del tiempo.

El valor de asociación es calculado a través de un promedio móvil simple. Si tenemos un vector $X = [x_1, x_2, \dots, x_n]$ de tamaño n , donde cada x_i representa el valor de una clase de sentimiento, entonces su promedio móvil es calculado como se muestra en la Ecuación 4. Para la Figura 1 se utiliza $k = 50$:

$$sma_i = \frac{\sum_j^{i-1} x_j}{k}, \text{ donde: } i > j. \quad (4)$$

Distribución del sentimiento: La Figura 2 muestra la distribución porcentual del sentimiento en tweets respecto al tema TLCAN. Como se observa, para este periodo de tiempo la mayoría de los tweet tenían cargado un sentimiento negativo; esto posiblemente se deba a que ocurrió un evento externo que afectó a la negociación del tratado (por ejemplo algún anuncio político del presidente de USA).

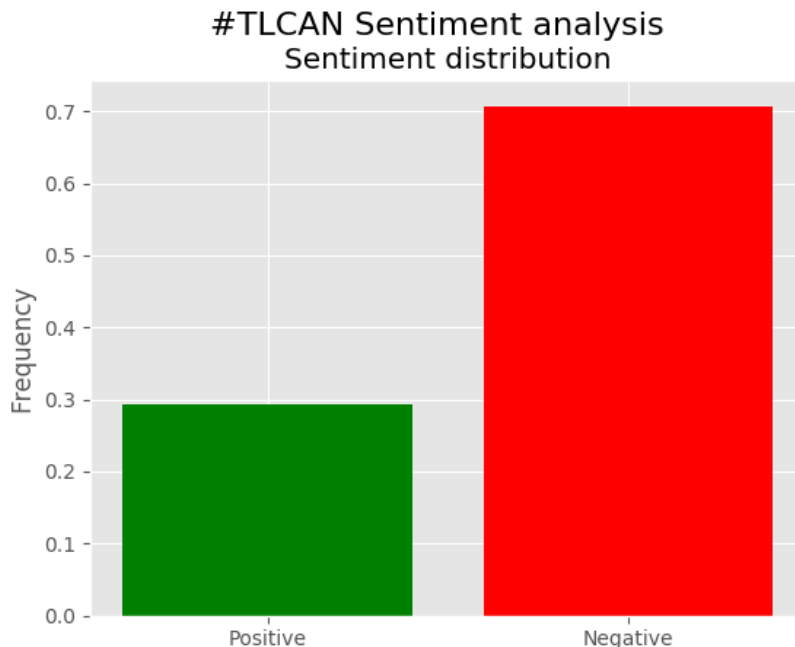


Fig. 2. Distribución porcentual de sentimientos, se oculta la etiqueta 'NEU' debido a que no aporta información para el sentimiento general sobre el tema.

5. Conclusiones y trabajo a futuro

En análisis de sentimientos en Twitter es un campo que promete mucho aún y es posible explotar las capacidades de los algoritmos de ML para obtener información valiosa del mundo que nos rodea; a través del análisis de grandes cantidades de datos generados por los usuarios diariamente.

En el presente trabajo se presentó una solución para el análisis de sentimientos de tweets en español con el tema TLCAN. Utilizando diversos algoritmos de ML encontramos que se obtiene un mejor resultado combinando los clasificadores en una capa de votación, con una exactitud del 88.6%. También creemos que la precisión de cada uno de los algoritmos puede mejorar si se utilizan características más sofisticadas y algunos métodos de preprocesamiento encontrados en el estado del arte.

También se presentó una aplicación para la clasificación y visualización de sentimientos en tiempo real, a través del streaming de tweets, usando el API proporcionado por Twitter. Estas aplicaciones pueden ser aprovechadas de diversas formas; por ejemplo, para analizar el efecto que tienen los eventos políticos o sociales sobre algún tema en específico.

Como trabajo a futuro proponemos utilizar otros temas políticos con su propio corpus de tweets, e.g opiniones de candidatos presidenciales en Twitter. Además, se puede diseñar un método que identifique eventos importante que afecten al sentimiento del tema tratado y a su vez poder visualizarlo en la gráfica. Finalmente, es posible utilizar diferentes métodos de visualización, así como calcular el valor general del sentimiento respecto al tema utilizando otras medidas.

Agradecimientos. Quisiera agradecer a mis compañeros Carolina Martín de Campo y Christian Maldonado por las sugerencias e ideas que aportaron al proyecto. Agradecemos el apoyo del proyecto CONACYT 240844 y el proyecto SIP 20181849.

Referencias

1. Anta, A.F., Chiroque, L.N., Morere, P., Santos, A.: Sentiment analysis and topic detection of Spanish tweets: A comparative study of of NLP techniques. *Procesamiento del lenguaje natural* 50, 45–52 (2013)
2. Buendía Rice, E.A.: Las promesas incumplidas del Tratado de Libre Comercio de América del Norte (TLCAN). *Análisis Económico* 29(72) (2014)
3. Moreno Brid, J.C., Rivas Valdivia, J.C., Ruiz Nápoles, P.: La economía mexicana después del TLCAN. *Revista Galega de Economía* 14(1-2) (2005)
4. Newberry, C.: 28 twitter statistics all marketers need to know in 2018. <https://blog.hootsuite.com/twitter-statistics/> (2018)
5. Polanyi, L., Zaenen, A.: Contextual valence shifters. In: *Computing attitude and affect in text: Theory and applications*, pp. 1–10. Springer (2006)
6. Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A., Gordon, J.: Empirical study of machine learning based approach for opinion mining in tweets. In: *Mexican international conference on Artificial intelligence*. pp. 1–14. Springer (2012)
7. Villena Román, J., Lana Serrano, S., Martínez Cámara, E., González Cristóbal, J.C.: TASS workshop on sentiment analysis at SEPLN (2013)