Emotions and Opinions: An Analysis of the Colombian Context through Opinion Mining

Emociones y Opiniones: Un Análisis del Contexto Colombiano a través de Opinion Mining

Liliana Calderón, Luis C. Vargas

Facultad de Ingeniería de Sistemas, Universidad Autónoma de Bucaramanga, Colombia.

mcalderon@unab.edu.co
lvargas487@unab.edu.co

Abstract

This paper discusses the opinion mining based on the capture of opinions through the Web, in order to perform an analysis of the feelings of Colombians based on comments made, taking as its central theme the presidential elections in Colombia in 2014.

To catch opinions was developed a web application to collect tweets based on a specific entity; while for the collection of other comments in news websites, was carried out a process of web crawling. With these capture instruments reviews were collected 33570 comments between 9 May and 15 June 2014.

Through opinion mining were able to confirm the image they had citizens about each of the candidates in the two presidential rounds, and a general analysis which allowed evidence that there was a generally negative outlook on the presidential elections in Colombia was built.

The results from opinion mining conducted indicated that during the first presidential round, the candidates with the highest ratio of positive comments were: Clara Lopez, Enrique Peñalosa and Marta Lucia Ramirez, respectively; Clara Lopez however was the candidate who best image transmitted to most voters. In the second round the candidate who scored the highest ratio of positive comments was Juan Manuel Santos.

These results largely agreed with reality, especially in the second round, which was obtained as winner of the presidential election candidate Juan Manuel Santos.

***Keywords**:* Microblogging, Opinion Mining, Sentiment Analysis, SentiStrength, Tweet, Web 2.0

1. Introducción

La minería de opiniones [3] (Opinion Mining) es un campo de estudio de reciente aparición que está siendo cada vez más utilizado en el desarrollo de investigaciones que, con base en el tratamiento de textos por medio de herramientas de software, intenten explicar comportamientos o tendencias de los usuarios, e inclusive realizar predicciones acerca del éxito de un producto o servicio. Este campo de estudio también se conoce con el nombre de Sentiment Analysis.

El objetivo principal del sentiment analysis es identificar y extraer fenómenos subjetivos (i.e., sentimientos, emociones y opiniones) en determinados textos. Una de sus principales tareas es la clasificación de la polaridad de un determinado texto, para lo cual se utiliza mayoritariamente una clasificación basada en tres categorías: [positivo / neutral / negativo], en donde neutral indica que no hay una tendencia clara en el texto acerca de si es positivo o negativo, por lo que podría tratarse de una expresión no subjetiva.

Debido a la aparición de las redes sociales de Internet, los servicios de microblogging y

el auge de la Web 2.0, el número de personas que utilizan a diario estas herramientas para expresar sentimientos, emociones y opiniones por medio de comentarios (textos), es cada vez mayor. Toda esa información constituye una inmensa fuente de información, la cual puede ser esencial para tratar de comprender el porqué de algunas conductas o tendencias en la población frente a un tema en particular, por medio de la identificación de fenómenos subjetivos en los comentarios, gracias a la minería de opiniones.

En Colombia diariamente son utilizados diversos sitios web 2.0 para generar controversia acerca de temas de mucha importancia para la población nacional, lo cual deja como secuela un enorme volumen de comentarios disponibles públicamente, pero que no están siendo aprovechados por la mayoría de las empresas para realizar opinion mining con el objetivo de detectar tendencias, obtener conclusiones y realizar predicciones. Este tipo de estudios puede facilitar la toma de decisiones en diferentes sectores el mercado.

Este artículo propone una estrategia para recopilar comentarios automáticamente desde el servicio de microblogging *Twitter*, y desde los sitios web 2.0 de noticias en Colombia: *elespectador.com* y *semana.com*, y posteriormente realizar opinion mining sobre los comentarios recolectados con el objetivo de realizar un análisis de los sentimientos de los colombianos. El tema central que se seleccionó fueron las elecciones presidenciales de Colombia llevadas a cabo los días 25 de mayo y 15 de junio de 2014, debido a la gran importancia que tuvo dicho suceso para el país, y a la gran cantidad de comentarios que se generaron diariamente en los sitios mencionados.

El proceso de opinion mining sobre los comentarios recolectados se llevó a cabo por medio del uso del software SentiStrength[1], la traducción de 2878 términos del inglés al español para la creación y configuración de la base de conocimientos, y el desarrollo de un algoritmo para calcular la polaridad global (positiva, neutral o negativa) de cada uno de los grupos de comentarios recolectados, almacenados en ficheros planos.

El artículo tiene la siguiente estructura: en la sección 2 se presenta el estado del arte. La sección 3 describe el marco metodológico. Los resultados obtenidos son presentados en la sección 4. Finalmente, en la sección 5 se describen las conclusiones y se presenta una propuesta de trabajo futuro.

2. Estado del Arte

Diferentes propuestas han sido realizadas por diferentes autores para detectar tendencias en la Web por medio de opinion mining. Pérez et al. [5] presentaron el desarrollo y los resultados de un analizador de opiniones denominado NOA (News Opinion Analyzer), el cual definen como una aplicación web para detectar las tendencias de opinión utilizando sentiment analysis y tecnología NoSQL.

Otro estudio desarrollado en Colombia con base en opinion mining es el realizado por Víctor Cadena y José Romero, de la Universidad de los Andes, titulado *From Big Data to Content Analysis* [6], mediante el cual se desarrolló una aplicación funcional para proporcionar valor agregado teniendo en cuenta un análisis de contenidos. Dichos contenidos fueron recopilados desde la red social Facebook, y también desde otros sitios web de noticias de actualidad, por medio de un proceso de web crawling. Para la realización de minería de opiniones utilizaron el software SentiStrength.

[1] http://sentistrength.wlv.ac.uk/

Mendoza et al. [4] tomaron como eje central el terremoto ocurrido en Chile en el año 2010, analizando los comentarios publicados en *Twitter*[2] desde el día del terremoto hasta los días inmediatamente siguientes al mismo. Los autores describen cómo es posible detectar rumores en los tweets usando análisis agregado a los mismos, y concluyeron en su estudio que los rumores en Twitter difieren de las noticias reales, debido a que dichos rumores son mucho más cuestionados que las noticias veraces.

En otro estudio realizado en Chile, Castillo et al. [2] describen métodos automáticos para evaluar la credibilidad de ciertos conjuntos de tweets capturados, la mayoría de los cuales consideraron que son veraces, pero también consideraron que existen muchos tweets que contienen mensajes falsos y rumores, muchas veces emitidos sin intención. Los autores concluyeron que existen diferencias considerables entre este tipo de mensajes propagados, las cuales pueden ser utilizadas para clasificar de manera automática como "creíbles" o "no creíbles" los tweets, con precisión dentro de un rango del 70% al 80%.

Otras importantes contribuciones para la minería de opiniones de comentarios de Twitter pueden ser halladas en el artículo publicado por Agarwal et al. [1], en cuya publicación los autores dan a conocer sus experimentos con base en el análisis de grandes conjuntos de datos basados en tweets para conocer la polaridad de los comentarios utilizando tres tipos de modelos basados en algoritmos para realizar tareas de sentiment analysis, y cuya finalidad es realizar clasificaciones de textos en dos tipos: [positivos o negativos], y en tres tipos: [positivos, negativos o neutrales].

3. Marco Metodológico

En esta sección se explica detalladamente cómo se realizó la captura de comentarios, qué herramienta se utilizó para realizar opinion mining, los pasos que fue necesario efectuar para que los comentarios fuesen procesados en español, así como el proceso de calcular la polaridad global de varios comentarios capturados.

3.1 Instrumentos de captura de opiniones

A continuación se describen los instrumentos utilizados para la recolección de opiniones desde el servicio de microblogging Twitter[3], y desde los sitios web de noticias on-line de El Espectador[4] y Semana[5].

3.1.1 Instrumento para recolectar opiniones desde Twitter

Para la captura de opiniones desde Twitter se desarrolló una aplicación web en PHP con conexión a base de datos MySQL. Con el objetivo de lograr que la aplicación web extrajera tweets relacionados con las elecciones presidenciales y sus candidatos, se seleccionaron siete hashtag, buscando que los resultados del sentiment analysis aplicado a los tweets no fuera parcializado.

Los hashtag seleccionados fueron los más usados por los colombianos durante los meses de mayo y junio de 2014 para referirse a los candidatos presidenciales y al proceso de elecciones en general. Los hashtag fueron los siguientes:

[2] https://twitter.com/
[3] https://twitter.com/
[4] http://www.elespectador.com/noticias
[5] http://www.semana.com/Home

Hashtag	Descripción
#ClaraLopez	Para la captura de tweets relacionados con Clara López Obregón, candidata por la coalición de los partidos polo democrático alternativo y unión patriótica.
#EnriquePeñalosa	Para capturar tweets relacionados con Enrique Peñalosa Londoño, candidato por el partido alianza verde.
#MartaLuciaRamirez	Tweets relacionados con Marta Lucía Ramírez, candidata por el partido conservador colombiano.
#JuanManuelSantos	Captura de tweets relacionados con Juan Manuel Santos, candidato por el partido social de unidad nacional.
#OscarIvanZuluaga	Para la captura de tweets relacionados con Oscar Iván Zuluaga, candidato por el partido centro democrático.
#Elecciones2014	Hashtag muy utilizado para referirse a las elecciones presidenciales, especialmente a partir de la segunda vuelta.
#SegundaVuelta	Captura de tweets también relacionados con el proceso de elecciones a partir de la segunda vuelta.

Tabla 1. Descripción de los hashtag utilizados para la recolección de opiniones desde Twitter

Para acceder a los tweets relacionados con los hashtag seleccionados, fue necesario acceder a la API de Twitter mediante la creación de una aplicación de Twitter, cuyas credenciales de autenticación son utilizadas por la aplicación desarrollada en PHP.

3.1.2 Instrumento para recolectar opiniones desde elespectador.com y semana.com

La captura de comentarios expresados en *elespectador.com* y en *semana.com* se realizó mediante un proceso de web crawling, gracias a lo cual se descargaron y almacenaron noticias relacionadas con las elecciones presidenciales entre el 9 de mayo y el 15 de junio de 2014, específicamente las noticias directamente relacionadas con los cinco candidatos aspirantes a la Presidencia de Colombia.

3.2 Clasificación de comentarios

Para realizar la clasificación de los comentarios se utilizó el programa SentiStrength[6], el cual es una herramienta especializada en minería de opiniones que arroja resultados con una alta precisión, y funciona clasificando textos mediante la asignación de puntajes a cada palabra.

Esta herramienta fue creada como parte del proyecto *CyberEmotions*[7], el cual fue fundado por la Unión Europea bajo el Seventh Framework Programme (FP7). SentiStrength se encuentra disponible a través del link: *http://sentistrength.wlv.ac.uk/#Download*. Ha sido aplicada en varios proyectos de investigación, entre los cuales se pueden mencionar: *Sentiment in Twitter events*[8]; *A large-scale sentiment analysis for Yahoo! Answers*[9]; *Answers, not links: extracting tips from yahoo! answers to address how-to web queries*[10]; y *Sentiment-focused web crawling*[11].

[6] http://sentistrength.wlv.ac.uk/
[7] http://www.cyberemotions.eu/
[8] http://www.scit.wlv.ac.uk/~cm1993/papers/SentimentInTwitterEvents_preprint.doc
[9] http://research.yahoo.com/node/3699
[10] http://dl.acm.org/citation.cfm?id=2124369

SentiStrength permite analizar palabras, frases cortas, o cargar ficheros planos desde el equipo local, con el objetivo de asignar una puntuación (score) global, la cual se divide en dos categorías:

- Escala de emociones Positivas: desde 1 (neutral) hasta 5 (muy positivo).
- Escala de emociones Negativas: desde -1 (neutral) hasta -5 (muy negativo).

El idioma predeterminado de la base de conocimientos de SentiStrength es el inglés, lo cual implica un obstáculo inicial para realizar opinion mining en el contexto colombiano.

Este obstáculo fue superado mediante la traducción y ensanchamiento de la base de conocimientos al español, la cual contiene 2878 términos revisados y ajustados al contexto colombiano; esto constituye un valioso aporte al estado del arte porque no existía una base de conocimientos en español disponible para realizar opinion mining.

3.2.1 Personalización del idioma de SentiStrength

SentiStrength es una herramienta de clasificación que requiere de un entrenamiento que permita determinar si los comentarios (conjuntos de palabras) suministrados son positivos, neutrales o negativos. En este estudio dicho entrenamiento se realizó por medio de la traducción de los términos de la base de conocimientos al español, y la revisión de los puntajes asignados a cada palabra.

Para lograr que SentiStrength reconociera términos en español, fue necesario traducir el contenido de siete de sus ficheros de configuración, los cuales constituyen la base de conocimientos mediante la cual se realiza la evaluación de la polaridad de cada grupo de palabras (frases o comentarios). Exactamente se tradujeron y revisaron 2878 términos conformados por palabras, combinaciones de palabras, expresiones interrogativas, expresiones de negación, y emoticonos.

3.2.2 Realizando opinion mining con SentiStrength

Para realizar la minería de opiniones con SentiStrength, fueron importados los ficheros planos en donde se encuentran almacenados los comentarios recolectados.

Cuando SentiStrength termina de analizar todo el texto de un fichero plano, genera un nuevo fichero .txt con los resultados de la minería de opiniones; ese nuevo fichero contendrá los comentarios originales acompañados de una puntación al final de cada uno de ellos. La puntuación de cada comentario dependerá de las palabras que lo conforman, y consiste en dos valores separados por una tabulación, correspondientes a la fuerza positiva (e.g., 4) y a la fuerza negativa (e.g., -3) que tiene el comentario.

Fig. 1: Estructura de resultados de un fichero plano generado por SentiStrength

La figura 1 representa la estructura de resultados común para todas las líneas de un

[11] http://dl.acm.org/citation.cfm?id=2398564

fichero plano generado por SentiStrength tras haber realizado opinion mining.

Las palabras cortas que indican conjunciones, pronombres, entre otros, generalmente tienen un score de 1, pues se les cataloga como neutrales.

3.2.2.1 Cálculo de la polaridad global de un fichero generado por SentiStrength después de realizar opinion mining

Después de obtenidos los nuevos ficheros planos que contienen la puntuación al final de cada comentario, se realizó un proceso que consiste en analizar cada uno de esos nuevos ficheros .txt por medio de un script desarrollado en PHP.

El script permite leer línea por línea cada uno de los nuevos ficheros .txt, y acumular en una primera variable la sumatoria de los scores positivos, y en una segunda variable acumular la sumatoria de los scores negativos, para comparar si esas dos sumatorias son iguales o determinar cuál de las dos es mayor.

Cuando el script termina de recorrer el fichero se comparan los valores acumulados en ambas variables, de tal manera que pueden existir 3 casos:

a) El valor acumulado en la primera variable es igual al valor acumulado en la segunda, por lo cual se obtendría una clasificación global de ese fichero como “neutral”.
b) El valor acumulado en la primera variable es mayor que el acumulado en la segunda, por lo cual se obtendría una clasificación global de ese fichero como “positivo”.
c) El valor acumulado en la primera variable es menor que el acumulado en la segunda, obteniendo entonces una clasificación global de ese fichero como “negativo”.

4. Resultados Obtenidos

En esta sección se presentan los resultados que se obtuvieron al utilizar los instrumentos diseñados para la captura de opiniones desde Twitter, elespectador.com y semana.com. También se describen los resultados obtenidos al realizar opinion mining por medio del software SentiStrength, con base en el procesamiento de comentarios almacenados en ficheros planos.

Posteriormente se construye un análisis general de la información obtenida mediante el proceso de opinion mining, y finalmente se realiza un contraste entre dicho análisis general obtenido e información pública recopilada acerca de las elecciones presidenciales: encuestas de Gallup Colombia Ltda.[12], y Datexco Company S.A.[13]

4.1 Descripción de los resultados obtenidos mediante la aplicación de instrumentos

En la tabla 2 se da a conocer el número de comentarios recolectados para cada uno de los candidatos y para el proceso de elecciones en general:

Candidato presidencial o hashtag utilizado	Comentarios de 1ª vuelta	Comentarios de 2ª vuelta	Total comentarios de 1ª y 2ª vuelta
Clara López	2469	457	2926
Enrique Peñalosa	2750	54	2804

[12] http://www.gallup.com/region/es-xm/americas.aspx
[13] http://datexco.com/int/index.php?lang=es

Candidato presidencial o hashtag utilizado	Comentarios de 1ª vuelta	Comentarios de 2ª vuelta	Total comentarios de 1ª y 2ª vuelta
Marta Lucía Ramírez	388	210	598
Juan Manuel Santos	1095	1852	2947
Oscar Iván Zuluaga	1262	862	2124
#Elecciones2014	0	15918	15918
#SegundaVuelta	0	6253	6253
Total comentarios recolectados para esta investigación:			**33570**

Tabla 2. Totalización de comentarios recolectados desde Twitter, elespectador.com y semana.com

Como se puede observar en la tabla 2, también se capturaron tweets mediante los hashtag *#Elecciones2014* y *#SegundaVuelta*, lo cual se hizo porque durante el transcurso de la segunda vuelta los usuarios de Twitter comenzaron a utilizar muchísimo esos dos hashtag para referirse, tanto a los dos candidatos que pasaron a la segunda vuelta, como al proceso de elecciones en su recta final.

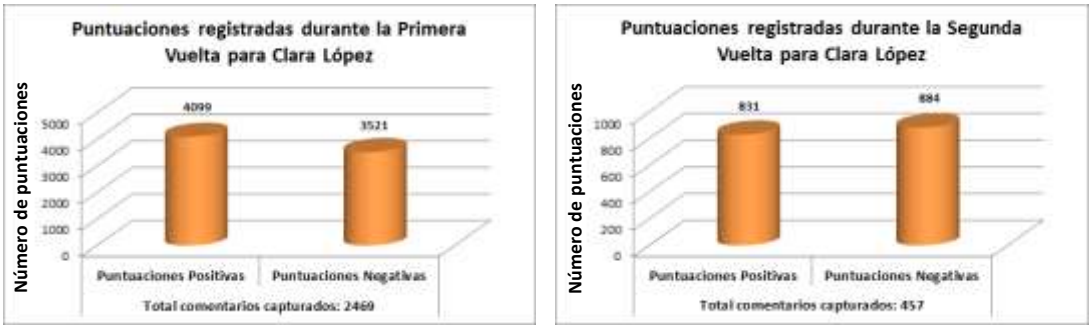
4.2 Descripción de los resultados obtenidos mediante Opinion Mining

A continuación se describen los resultados obtenidos para cada uno de los candidatos presidenciales. Estos resultados se muestran organizados por el número de puntuaciones positivas, negativas, y la polaridad global de los comentarios.

4.2.1 Resultados de opinion mining obtenidos para la candidata Clara López

El cálculo de la polaridad de los comentarios analizados con relación a esta candidata, para la primera vuelta, determinó que se obtuvo una polaridad *Positiva*, mientras que para la segunda vuelta la polaridad fue *Negativa* (figuras 2.a y 2.b).

Para todo el proceso (primera y segunda vuelta), el cálculo de la polaridad determinó que hubo un acumulado de 4930 puntuaciones positivas, frente a 4403 negativas, obteniéndose una polaridad global *Positiva* para esta candidata.



a) Primera vuelta b) Segunda vuelta

Fig. 2: Comportamiento de puntuaciones registradas para la candidata Clara López

4.2.2 Resultados de opinion mining obtenidos para el candidato Enrique Peñalosa

Con relación a este candidato, el cálculo de la polaridad determinó que tanto para la primera como para la segunda vuelta se obtuvo una polaridad *Positiva* (figuras 3.a y 3.b). Para todo el proceso, el cálculo arrojó un acumulado de 3951 puntuaciones positivas, frente

a 3813 negativas, obteniéndose una polaridad global *Positiva*.

a) Primera vuelta

b) Segunda vuelta

Fig. 3: Comportamiento de puntuaciones registradas para el candidato Enrique Peñalosa

4.2.3 Resultados de opinion mining obtenidos para la candidata Marta Lucía Ramírez

Para esta candidata el cálculo de la polaridad, para la primera vuelta, arrojó una polaridad *Positiva*, mientras que para la segunda vuelta se obtuvo una polaridad *Negativa* (figuras 4.a y 4.b).

Para todo el proceso, el cálculo de la polaridad determinó que hubo un acumulado de 1021 puntuaciones positivas, frente a 1021 negativas, de tal manera que se obtuvo una polaridad global *Neutral* para esta candidata.

a) Primera vuelta

b) Segunda vuelta

Fig. 4: Comportamiento de puntuaciones registradas para la candidata Marta Lucía Ramírez

4.2.4 Resultados de opinion mining obtenidos para el candidato Juan Manuel Santos

Con respecto a este candidato el cálculo de la polaridad, tanto para la primera como para la segunda vuelta, indicó una polaridad *Negativa* (figuras 5.a y 5.b).

Para todo el proceso, el cálculo arrojó un acumulado de 4791 puntuaciones positivas, frente a 5504 negativas, obteniéndose una polaridad global *Negativa*.

a) Primera vuelta b) Segunda vuelta

Fig. 5: Comportamiento de puntuaciones registradas para el candidato Juan Manuel Santos

4.2.5 Resultados de opinion mining obtenidos para el candidato Oscar Iván Zuluaga

Con relación a este candidato, el cálculo de la polaridad también determinó que se obtuvo una polaridad *Negativa* tanto para la primera como para la segunda vuelta (figuras 6.a y 6.b).

Para todo el proceso, el cálculo de la polaridad determinó que hubo un acumulado de 3345 puntuaciones positivas, frente a 3689 negativas, de tal manera que se obtuvo una polaridad global *Negativa* para este candidato.

a) Primera vuelta b) Segunda vuelta

Fig. 6: Comportamiento de puntuaciones registradas para el candidato Oscar Iván Zuluaga

4.2.6 Resultados de opinion mining obtenidos para el hashtag *#Elecciones2014*

El cálculo de la polaridad del fichero en el cual fueron almacenados los tweets capturados mediante el hashtag *#Elecciones2014*, arrojó un acumulado de 22057 puntuaciones positivas, frente a 22112 negativas, indicando una polaridad *Negativa*.

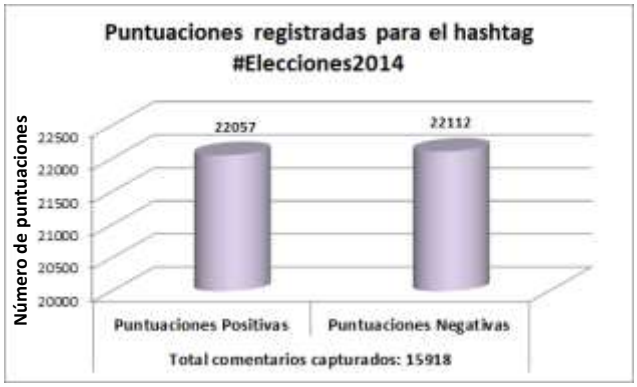


Fig. 7: Cálculo de la polaridad global del fichero contenedor de tweets capturados por medio del hashtag #Elecciones2014

4.2.7 Resultados de opinion mining obtenidos para el hashtag *#SegundaVuelta*

Con respecto a los tweets capturados por medio del hashtag *#SegundaVuelta*, la polaridad obtenida también fue *Negativa* (figura 8), tras obtener un acumulado de 9279 puntuaciones positivas, frente a 9569 negativas.

Fig. 8: Cálculo de la polaridad global del fichero contenedor de tweets capturados por medio del hashtag #SegundaVuelta

4.3 Análisis general, con base en los resultados de Opinion Mining

Este análisis general, se divide principalmente en dos secciones relacionadas con las dos vueltas que tuvo el proceso de elecciones presidenciales, analizando el papel que jugó cada candidato involucrado en cada una de las dos vueltas.

También se analizas la fluctuación o comportamiento de tweets capturados para cada candidato presidencial, desde el 9 de mayo hasta el 15 de junio de 2014.

En la figura 9 puede observarse que antes de las votaciones de primera vuelta, para los candidatos Clara López y Enrique Peñalosa se generó un alto número de tweets. Sin embargo estos dos candidatos, al no pasar a segunda vuelta, disminuyeron su participación en los comentarios capturados a partir de la segunda vuelta.

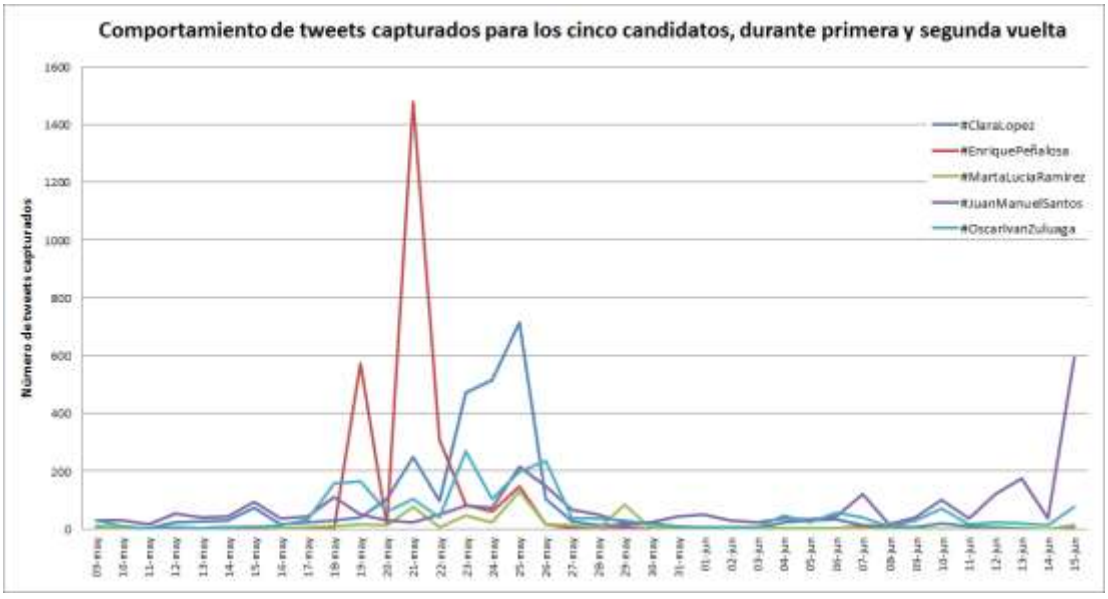


Fig. 9: Comportamiento en el tiempo del número de tweets expresados para cada candidato presidencial, durante primera y segunda vuelta

En la figura 9 puede observarse que los candidatos Juan Manuel Santos y Oscar Iván Zuluaga fueron quienes se mantuvieron más constantes en la generación de tweets para la primera y segunda vuelta.

Puede observarse también que la candidata Marta Lucía Ramírez fue quien menos tweets generó a lo largo de las elecciones presidenciales.

4.3.1 Análisis basado en la primera vuelta presidencial y sus candidatos

Los resultados obtenidos mediante minería de opiniones han sido utilizados en este trabajo para comprender la imagen que tenían los ciudadanos acerca de cada candidato presidencial, y del proceso de elecciones presidenciales de 2014 en general. Obsérvese en la figura 10 las polaridades globales calculadas con base en los cinco candidatos para la primera vuelta, de acuerdo a las opiniones recolectadas.

En la figura 10 se puede observar que los candidatos: Clara López, Enrique Peñalosa y Marta Lucía Ramírez, tuvieron una imagen positiva durante la primera vuelta.

Todo lo contrario ocurrió con los candidatos: Juan Manuel Santos y Oscar Iván Zuluaga, quienes según la minería de opiniones tuvieron poca favorabilidad durante la primera vuelta.

Para determinar cuál de los candidatos tuvo la imagen más positiva en la primera vuelta, según la minería de opiniones efectuada, bastó con obtener la mayor diferencia (resta) entre las puntuaciones positivas y negativas para cada uno de los candidatos que obtuvieron una polaridad global positiva (según la figura 10).

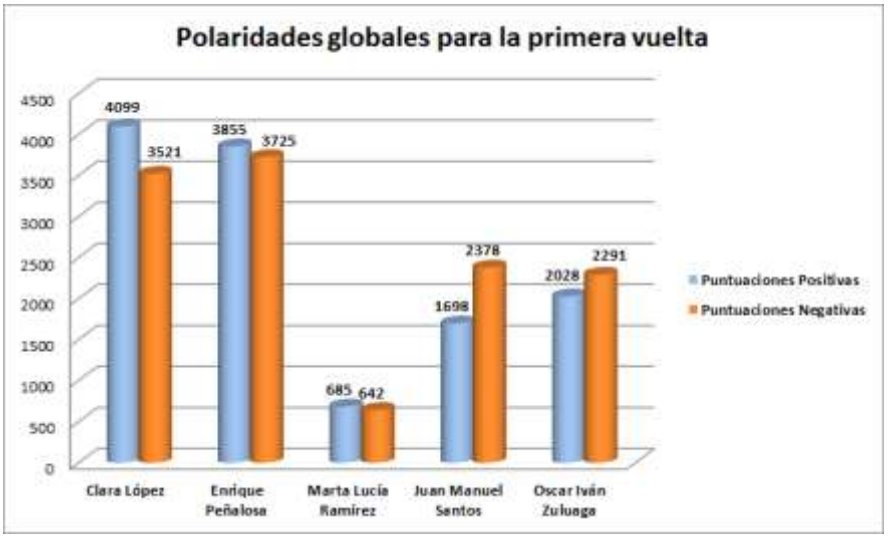


Fig. 10: Resumen de polaridades globales calculadas con base en los cinco candidatos para la primera vuelta

La candidata presidencial que obtuvo la mayor diferencia entre las puntuaciones positivas y negativas fue Clara López. Esto indica que, según los comentarios recolectados, Clara López tuvo la imagen más positiva durante la primera vuelta presidencial.

La firma Gallup Colombia Ltda. realizó la encuesta llamada "La gran encuesta de los medios #6"[14], en cuyos resultados indicaron que el candidato Oscar Iván Zuluaga obtendría la mayor votación en la primera vuelta, con una intención de voto del 29,3%, mientras que el candidato Juan Manuel Santos ocuparía el segundo lugar, con una intención de voto del 29.0%, por lo cual ambos pasarían a la segunda vuelta..

En cuanto a la imagen de los candidatos para la primera vuelta, obtenida mediante opinion mining, los resultados indicaron que la candidata Clara López obtendría la votación más alta en la primera vuelta (figura 10), seguida por el candidato Enrique Peñalosa. Estos resultados no coincidieron con los resultados descritos en la encuesta de Gallup para la primera vuelta.

La firma Datexco Company S.A. realizó una encuesta denominada "Tercera gran encuesta para las elecciones presidenciales 2014"[15], cuyos resultados indicaron que el candidato Juan Manuel santos, con una intención de voto del 25,5% obtendría la mayor votación en la primera vuelta, mientras que el candidato Enrique Peñalosa, con una intención de voto del 17.1%, ocuparía el segundo lugar. Los resultados de esa encuesta no se compaginaron con la realidad.

Comparando los resultados de esta encuesta de Datexco con la imagen de los candidatos obtenida mediante opinion mining, puede observarse que ambos resultados indicaron que el candidato Enrique Peñalosa obtendría la segunda mayor votación en la primera vuelta, por lo cual se demuestra que los resultados obtenidos mediante minería de opiniones estuvieron muy relacionados con la mencionada encuesta realizada por la firma Datexco.

4.3.2 Análisis basado en la segunda vuelta presidencial y sus candidatos

Los candidatos que obtuvieron las dos votaciones más altas en las elecciones de primera vuelta fueron: Oscar Iván Zuluaga y Juan Manuel Santos.

[14] http://media.caracoltv.com/LA-GRAN-ENCUESTA-DE-LOS-MEDIOS-ELECCIONES-2014--6.pdf
[15] http://www.eltiempo.com/Multimedia/infografia/encuestapresidencial/

La imagen que los ciudadanos tuvieron acerca de estos dos candidatos, según los resultados de opinion mining, se puede analizar observando la figura 11. Según los resultados arrojados por la minería de opiniones, ambos candidatos tuvieron una imagen negativa para los ciudadanos durante la segunda vuelta (figura 11).

Como en este caso ambas polaridades fueron negativas, se procedió a determinar cuál de los dos candidatos tuvo la imagen menos negativa en la segunda vuelta, de acuerdo a la minería de opiniones efectuada, obteniendo esta vez la menor diferencia (resta) entre las puntuaciones negativas y positivas para los dos candidatos.

El candidato Juan Manuel Santos fue quien obtuvo la menor diferencia entre las puntuaciones negativas y positivas. Esto quiere decir que con base en los comentarios recolectados, Juan Manuel Santos tuvo la imagen menos negativa durante la segunda vuelta presidencial. Estos resultados generados a partir de la minería de opiniones coincidieron con la realidad, pues el candidato Juan Manuel Santos derrotó en las elecciones del 15 de junio a su contendor Oscar Iván Zuluaga.

Una explicación a lo anterior podría consistir en analizar las alianzas que se formaron después de las elecciones de primera vuelta. Dichas alianzas quedaron así:

- *Juan Manuel Santos*: formó alianza con Clara López.
- *Oscar Iván Zuluaga:* formó alianza con Marta Lucía Ramírez.

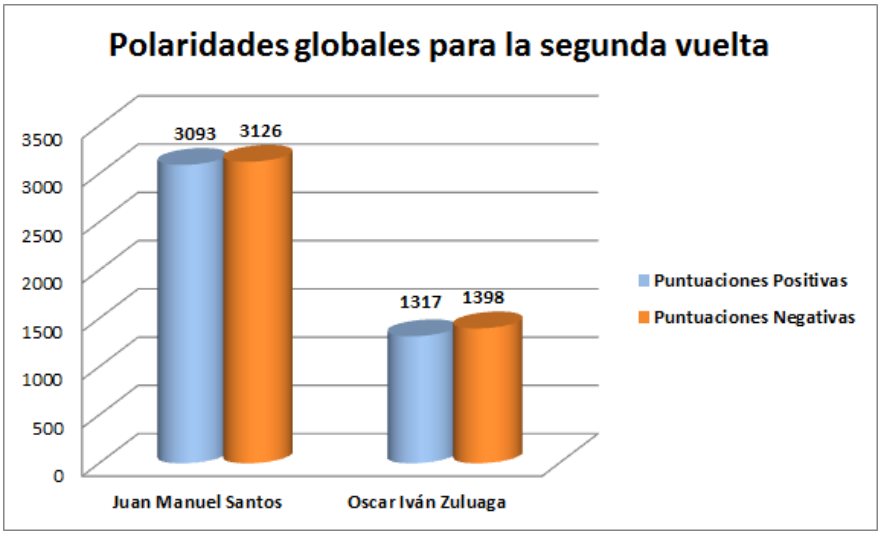


Fig. 11: Resumen de polaridades globales calculadas con base en los dos candidatos para la segunda vuelta

El candidato Enrique Peñalosa decidió no apoyar a ninguno de los dos candidatos que pasaron a la segunda vuelta, y dejó a sus seguidores en libertad para elegir la alternativa que ellos consideraran mejor para la ciudadanía.

Para comprender mejor los resultados reales de las elecciones, se deben analizar las polaridades globales que tuvieron todos los candidatos durante la primera y la segunda vuelta. La figura 12 resume esa información. En dicha figura se puede observar que la candidata Clara López, quien hizo alianza con Juan Manuel Santos, obtuvo una imagen muy positiva, lo cual puedo haber favorecido enormemente, entre otros factores, la candidatura de Juan Manuel Santos en la segunda vuelta.

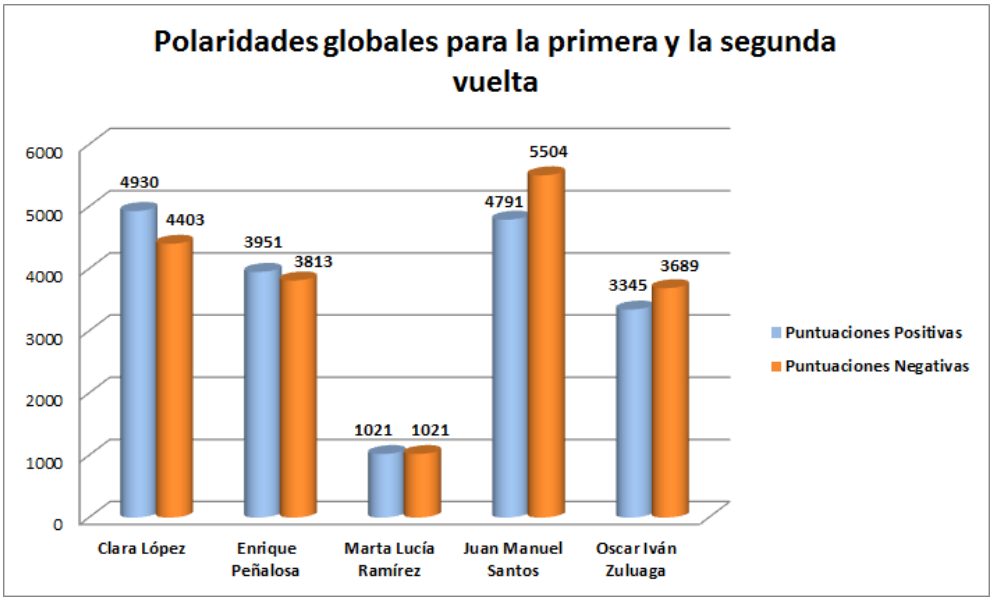


Fig. 12: Resumen de polaridades globales, calculadas con base en los cinco candidatos, para la primera vuelta y la segunda vuelta

La figura 12 también muestra que la candidata Marta Lucía Ramírez, quien para la segunda vuelta se vinculó a la campaña de Oscar Iván Zuluaga, obtuvo una imagen global neutral. Este importante factor, junto con otros más graves como los escándalos en los que se vio involucrado Oscar Iván Zuluaga por el tema de una supuesta interceptación ilegal de correos electrónicos relacionados con el proceso de paz que desde Cuba se adelantaba en ese momento entre el Gobierno de Colombia y la guerrilla de las FARC, pudo haber influido de manera muy negativa en la votación final de este candidato presidencial. Este suceso es más conocido como chuzadas al proceso de paz, las cuales se dice que fueron realizadas principalmente por un hacker colombiano, presuntamente contratado por el candidato Oscar Iván Zuluaga[16].

Se puede analizar que los dos candidatos que pasaron a la segunda vuelta tuvieron desde el principio una imagen negativa frente a los demás candidatos, lo cual se explica porque la mayoría de colombianos que expresaron comentarios mediante Twitter, elespectador.com y semana.com no querían que ganara Juan Manuel Santos ni Oscar Iván Zuluaga, sino que querían apostar por una nueva alternativa encarnada en uno de los otros tres candidatos.

Con relación a la segunda vuelta, Gallup Colombia Ltda. realizó una nueva gran encuesta de los medios[17], la cual indicó que el candidato Oscar Iván Zuluaga sería el ganador de las elecciones presidenciales de Colombia en 2014, con una intención de voto del 48,5%, sobre un 47,7% para el candidato Juan Manuel Santos. Como es sabido, estos resultados no se compaginaron con la realidad.

Con respecto a la imagen de los candidatos para la segunda vuelta, los resultados indicaron que el candidato Juan Manuel Santos tuvo la imagen menos negativa (figura 11), pues la diferencia entre puntuaciones negativas y positivas indicó que podría superar en las urnas al candidato Oscar Iván Zuluaga.

Datexco Company S.A. también realizó una nueva encuesta[18], la cual indicó que el

[16] http://www.elespectador.com/noticias/politica/acusan-zuluaga-y-uribe-de-chuzar-y-sabotear-el-proceso-articulo-490746
[17] http://www.elespectador.com/noticias/infografia/santos-y-zuluaga-diferencia-apenas-de-08-articulo-496690
[18] http://ow.ly/d/2equ

candidato Juan Manuel santos, con una intención de voto del 41,9% ganaría las elecciones presidenciales, mientras que el candidato Oscar Iván Zuluaga, con una intención de voto del 37.7%, sería derrotado. Los resultados de esta encuesta realizada por Datexco se cumplieron: el nuevo presidente elegido por los colombianos fue el candidato Juan Manuel Santos.

Los resultados de esta última encuesta realizada por Datexco, comparados con la imagen de los candidatos para la segunda vuelta obtenida mediante minería de opiniones (figura 11), tuvieron una estrecha relación, pues la minería de opiniones efectuada indicó que el candidato Juan Manuel Santos tuvo la imagen menos negativa para la segunda vuelta y por lo tanto ganaría las elecciones presidenciales.

4.3.2.1 Análisis basado en otros tweets capturados durante la segunda vuelta presidencial

Los resultados de opinion mining que fueron obtenidos al utilizar el hashtag *#Elecciones2014* indicaron que la mayoría de los tweets expresados por los colombianos fueron negativos: 22057 puntuaciones positivas, contra 22112 puntuaciones negativas. De igual manera, los resultados de opinion mining que se obtuvieron tras utilizar el hashtag *#SegundaVuelta* también indicaron que la mayoría de los tweets expresados por los colombianos fueron negativos: 9279 puntuaciones positivas, contra 9569 negativas.

Estos resultados demuestran que la mayoría de los ciudadanos que utilizaron Twitter para expresar opiniones acerca de las elecciones presidenciales de Colombia en 2014, tuvieron una perspectiva negativa acerca del proceso de las elecciones en general. Puede afirmarse que esa perspectiva negativa que tuvieron en su gran mayoría los colombianos acerca de las elecciones presidenciales se debió a la escasa credibilidad que para muchos ciudadanos tenían varios de los candidatos, principalmente por razones asociadas a la corrupción, escándalos políticos, entre otros.

5. Conclusiones y Trabajo Futuro

En general los resultados obtenidos en esta investigación demuestran que es posible utilizar la minería de opiniones para conocer tendencias de un grupo poblacional.

Los resultados obtenidos a partir de los instrumentos de captura de opiniones comprueban que es posible recuperar información desde la Web, especialmente desde sitios web 2.0, para su posterior análisis.

En cuanto a los resultados obtenidos mediante el proceso de opinion mining a través de SentiStrength, parece comprobarse que la veracidad de los resultados obtenidos está fuertemente ligada a la revisión y alimentación de la base de conocimientos sobre la cual se puede efectuar opinion mining.

Los resultados del análisis general realizado con base en la interpretación de los resultados obtenidos mediante opinion mining demuestran que la minería de opiniones es actualmente una herramienta clave para la predicción de situaciones o fenómenos, y que puede ayudar mucho en la toma de decisiones.

Por medio del contraste entre el análisis general realizado, y la información disponible sobre las elecciones por medio de las encuestas realizadas por Gallup y Datexco, parece demostrarse que la minería de opiniones puede ser utilizada en conjunto con las tradicionales encuestas para predecir resultados más cercanos a la realidad en determinadas

situaciones, y así facilitar la toma de decisiones.

Otras conclusiones obtenidas a través de esta investigación son las siguientes:

- El sentiment analysis es un campo cada vez más investigado, y la Web 2.0 ha contribuido en gran manera a este campo de estudio.
- El servicio de microblogging Twitter es tal vez la mayor fuente de información desde donde actualmente se puede recuperar información expresada por los usuarios sobre casi cualquier tema.
- La existencia de herramientas para realizar opinion mining con textos en español, así como la disponibilidad de bases de conocimiento configurables para estos procesos, son muy limitadas.

Como trabajo futuro se propone revisar y complementar la base de conocimientos, traducida del inglés al español y conformada por 2878 términos, con el objetivo de realizar nuevos experimentos basados en sentiment analysis en el contexto colombiano, y publicar los resultados obtenidos en pro de fortalecer las investigaciones en el campo del sentiment analysis.

Agradecimientos

Los autores agradecen a la Doctora Miriam Fernández, y al equipo creador de SentiStrength, por sus importantes contribuciones a este trabajo.

Referencias

[1] AGARWAL, A., XIE, B., VOVSHA, I., RAMBOW, O., & PASSONNEAU, R. (2011). Sentiment Analysis of Twitter Data. *En LSM 2011, actas del Workshop sobre lenguajes en medios de comunicación social*, pág. 30-38. http://aclweb.org/anthology//W/W11/W11-0705.pdf?CFID=262217785&CFTOKEN=33606415

[2] CASTILLO, C., MENDOZA, M., & POBLETE, B. (2011). *Information Credibility on Twitter. En WWW 2011, actas de la 20ª conferencia internacional en World Wide Web*, pág. 675-684. http://chato.cl/papers/castillo_mendoza_poblete_2010_twitter_credibility.pdf

[3] DAVE, K., LAWRENCE, S., & PENNOCK, D. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. En *WWW 2003, actas de la 12ª conferencia internacional en World Wide Web,* pág. 519-528. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.2424&rep=rep1&type=pdf

[4] MENDOZA, M., POBLETE, B., & CASTILLO, C. (2010). *Twitter Under Crisis: Can we trust what we RT? En SOMA 2010, actas del primer Workshop sobre análisis de medios sociales*, pág. 71-79. http://chato.cl/papers/mendoza_poblete_castillo_2010_twitter_terremoto.pdf

[5] PÉREZ, D., LOZANO, C., LÓPEZ, A., et al. (2012). Sentiment Analysis in Colombian Online Newspaper Comments. *Recent Progress in Data Engineering and Internet Technology*, vol. 157, pp. 113-119. http://link.springer.com/chapter/10.1007%2F978-3-642-28798-5_16#page-1

[6] http://sistemas.uniandes.edu.co/~isis4513/dokuwiki/doku.php?id=proyectogrupo03_2012-2