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

Tesis de Maestría

Luís Carlos Vargas Cogollo

Universidad Autónoma de Bucaramanga
Universitat Oberta de Catalunya
Maestría en Software Libre
Bucaramanga
2014

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

Tesis de Maestría

Luís Carlos Vargas Cogollo

Directora:
Dra. Liliana Calderón Benavides

Universidad Autónoma de Bucaramanga
Universitat Oberta de Catalunya
Maestría en Software Libre
Bucaramanga
2014

Nota de Aceptación

_____

_____

_____

_____

_____

Jurado

_____

Jurado

Bucaramanga, julio de 2014

A Dios, quien nos obsequió el don de la vida; a la memoria de mi abuela, gracias por tus incansables cuidados; a Valentina, el motorcito que mueve mi mundo; a mi esposa, quien constantemente me ayuda a continuar en empresas como ésta; a mis padres, quienes con su incondicional ayuda han hecho posible que hoy alcance esta meta; a mi familia, con cuyo apoyo siempre puedo contar.

Agradecimientos:

A la Doctora Liliana Calderón Benavides, mi directora de tesis, quien me brindó su valioso y constante apoyo reflejado en su tiempo disponible para compartir conmigo sus conocimientos y experiencias en pro de la realización de esta tesis de Maestría.

Al equipo de traducción y revisión de la base de conocimientos en español, especialmente a mi esposa Yésica y a mi hermana Luisa Fernanda: su aporte fue prodigioso.

Al Ingeniero Jonatan De la Ossa, por su valioso aporte al mejorar la presentación de la captura de opiniones de Twitter, al desentrañar el formato JSON.

A todos los Docentes de quienes recibí formación en la Universitat Oberta de Catalunya, por haberme proporcionado muchas de las bases necesarias para alcanzar esta meta.

A los Docentes de la Universidad Autónoma de Bucaramanga de quienes recibí formación en investigación, por sus valiosas orientaciones.

A todas aquellas personas que de una u otra forma contribuyeron a la realización de este trabajo: Gracias.

Resumen

Palabras clave: Microblogging, Opinion Mining, Sentiment Analysis, SentiStrength, Tweet, Web 2.0

Este trabajo de investigación trata acerca de la minería de opiniones (opinion mining) basada en la recolección de textos a través de la Web, con el objetivo de realizar un análisis de los sentimientos de los colombianos con base en comentarios expresados en sitios web 2.0, para lo cual se tomó como tema central las elecciones presidenciales de Colombia en el año 2014.

Para la captura de opiniones se desarrolló una aplicación web capaz de recolectar tweets basados en una entidad específica; mientras que para la recolección de otros comentarios contenidos en sitios web de noticias, se llevó a cabo un proceso de web crawling. Gracias a estos instrumentos de captura de opiniones fueron recolectados, entre el 9 de mayo y el 15 de junio de 2014, 33570 comentarios relacionados con las elecciones presidenciales.

Los comentarios recolectados fueron sometidos a opinion mining por medio de la solución especializada de software SentiStrength, pero antes de esto fue necesario construir una base de conocimientos en español, ajustada al contexto colombiano.

La base de conocimientos resultante contiene 2878 términos, y es un aporte al estado del arte porque plantea los cimientos para la creación de bases de conocimientos en castellano con el objetivo de realizar opinion mining frente a procesos electorales, pues a la fecha no existía ninguna base de conocimientos disponible en español.

Mediante opinion mining se conoció la imagen que para los ciudadanos reflejó cada uno de los candidatos en las dos vueltas presidenciales, y se construyó un análisis general que permitió comprender mejor las tendencias de los colombianos durante las elecciones presidenciales en Colombia en 2014, y evidenció que para los ciudadanos hubo una perspectiva mayoritariamente negativa acerca de dicho proceso, soportada principalmente en la inconformidad con los candidatos y sus propuestas. Dicha inconformidad se pudo reflejar en los resultados de la minería de opiniones, los cuales indicaron que los candidatos Juan Manuel Santos y Oscar Iván Zuluaga obtuvieron una polaridad o imagen negativa con respecto al resto de candidatos durante la primera vuelta.

El análisis general realizado fue contrastado con resultados arrojados por varias encuestas acerca del proceso de elecciones, realizadas por firmas encuestadoras muy reconocidas, hallándose una importante relación entre los resultados de dichas encuestas y los resultados arrojados por la minería de opiniones realizada en este trabajo.

Se concluye que la minería de opiniones es un área capaz de identificar tendencias en la población, por lo cual puede ser utilizada para predecir fenómenos o situaciones, incluida la política, y por ende facilitar la toma de decisiones.

Tabla de Contenido

Índice de Figuras

Índice de Tablas

Índice de Algoritmos

Capítulo 1

Introducción

El presente trabajo de investigación está basado en el estudio y análisis de información recopilada desde diferentes sitios web 2.0[1], mediante la captura de comentarios y un proceso organizado de opinion mining (minería de opiniones) que permitan construir un análisis del contexto colombiano, de acuerdo a las opiniones, sentimientos y emociones detectados en información expresada por los ciudadanos.

Desde la aparición de las redes sociales de Internet, los servicios de microblogging y el auge de la Web 2.0[2], el número de colombianos que utilizan a diario esas herramientas para expresar sentimientos, emociones y opiniones a través de Internet mediante comentarios (texto), es cada vez mayor. Dichos comentarios pueden constituir una inmensa fuente de información esencial para tratar de comprender el porqué de algunas conductas o tendencias en la población colombiana frente a un tema en particular, mediante la identificación de fenómenos subjetivos en los comentarios gracias a las tareas de opinion mining[3].

El campo de opinion mining es de reciente aparición, y se considera que solo hasta hace pocos años está siendo explotado como una convergencia de técnicas y herramientas de software que permiten obtener respuestas acerca de diferentes acontecimientos que pueden ser rastreados desde la Web. El tema seleccionado en esta investigación corresponde a las Elecciones Presidenciales en Colombia llevadas

[1] http://www.ite.educacion.es/formacion/materiales/155/cd/modulo_1_Iniciacionblog/concepto_de_web_20.html

[2] Ibídem

[3] http://www.lct-master.org/files/MullenSentimentCourseSlides.pdf

a cabo los días 25 de mayo y 15 de junio de 2014, debido a la gran importancia que tuvo este suceso para todo el país, y a la gran cantidad de información expresada por los ciudadanos que este tema generó diariamente en la Web, específicamente en *Twitter*[4]*, elespectador.com*[5] *y semana.com*[6], durante todo el proceso electoral.

La recopilación de comentarios para las posteriores tareas de opinion mining se llevó a cabo, por una parte, mediante el desarrollo de una aplicación web, la cual permite capturar y almacenar comentarios de Twitter (tweets) en una base de datos. Los comentarios expresados por los colombianos en los sitios de noticias elespectador.com y semana.com, fueron capturados mediante un proceso de web crawling. Los comentarios de Twitter fueron capturados y almacenados diariamente, desde el 9 de mayo hasta el 15 de junio; mientras que los comentarios de elespectador.com y semana.com solamente fueron almacenados cuando se generaron noticias exclusivamente relacionadas con los candidatos y su papel durante las elecciones.

La minería de opiniones se realizó utilizando la herramienta SentiStrength[7], analizando cada uno de los comentarios recopilados y asignándoles una puntuación, con el objetivo de calcular posteriormente la polaridad global de todo el conjunto de comentarios recopilados y almacenados.

Con los resultados obtenidos mediante opinion mining, se realizó un análisis general que muestra cuáles fueron mayoritariamente las tendencias, sentimientos, opiniones y emociones de los colombianos frente al tema de las elecciones presidenciales de 2014 y sus candidatos. Este análisis general indicó que los colombianos tuvieron una posición bastante negativa durante todo el proceso de las elecciones

[4] https://twitter.com/

[5] http://www.elespectador.com/noticias

[6] http://www.semana.com/Home

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

presidenciales, y que una gran mayoría de los ciudadanos no se sintió identificada con alguno de los candidatos presidenciales en especial, lo cual se acentuó más en la segunda vuelta de las elecciones. Sin embargo también se pudo evidenciar que ciertos candidatos gozaron de cierta favorabilidad, según los comentarios de los colombianos, durante algunas etapas del proceso electoral.

También se estableció una relación entre dicho análisis general y otras informaciones disponibles públicamente acerca de las elecciones presidenciales de Colombia en 2014, como lo fueron varias encuestas realizadas por prestigiosas firmas como Gallup Colombia Ltda., y Datexco Company S.A., acerca de las elecciones y sus candidatos. Por medio de esa relación se observó que algunos de los resultados de dichas encuestas coincidieron en gran medida con el análisis general realizado, porque se pudo observar que las tendencias marcadas por el análisis realizado (basadas en las opiniones de los ciudadanos) sobre la imagen favorable o desfavorable de algún candidato, fueron similares a algunos de los resultados individuales que indicaron las encuestas para ese mismo candidato.

La minería de opiniones basada en la recopilación de comentarios expresados mediante sitios web 2.0 es un proceso muy atractivo para la comunidad científica y el sector empresarial, que permite obtener conocimiento relevante acerca del valor que los cibernautas le dan a muchos temas, productos y servicios; y que también facilita la comprensión de ciertas tendencias de la población frente a determinados acontecimientos relacionados con el contexto en donde se lleve a cabo la investigación. En este trabajo se ha tomado como tema central las elecciones presidenciales de Colombia en 2014, pero también es posible realizar el análisis con cualquier otro tema de actualidad acerca del cual se deseen conocer los sentimientos, y en general, las tendencias de los cibernautas.

1.1 Planteamiento del Problema y Motivación

Los vertiginosos avances de la tecnología, reflejados mayoritariamente mediante la conectividad global a través de modernos equipos móviles, han dado lugar al desarrollo y expansión de una nueva sociedad en la que las personas han dejado de ser consumidoras de contenido virtual y han pasado a ser generadoras de los mismos, construyendo de esta forma espacios de debate abiertos en donde se discuten diferentes temas.

Lo anterior ha traído como consecuencia la existencia de un enorme volumen de datos que se genera día a día en los diferentes sitios web 2.0, que se encuentra disponible y es de gran utilidad para detectar tendencias, obtener conclusiones y tomar decisiones en diferentes sectores del mercado.

El principal problema que existe para la obtención y tratamiento de ese enorme volumen de datos, con el objetivo de realizar una estimación de la percepción global que tienen los usuarios acerca de un tema en particular, es que se hace necesario contar con un equipo multidisciplinario que se encargue de buscar de manera rutinaria y repetitiva comentarios realizados en los diferentes sitios web 2.0 desde donde se desea recuperar información, lo cual no solamente resulta excesivamente tedioso sino costoso.

Una solución a este problema consiste en realizar minería de opiniones con base en información obtenida de manera automática, por medio de herramientas de software encargadas de recuperar información desde los principales servicios de microblogging (e.g., Twitter) y desde otros sitios web 2.0 en donde se publiquen noticias diariamente, con la finalidad de conocer la percepción general de los usuarios acerca de un tema, producto o servicio determinado.

En Colombia se han llevado a cabo estudios y publicaciones relacionadas con la minería de opiniones (sentiment analysis), entre las cuales cabe destacar el artículo de Pérez et al., denominado: *Sentiment Analysis in Colombian Online Newspaper Comments* [20], en el cual se presentan los resultados y conclusiones de los autores acerca de que los comentarios realizados en diarios on-line constituyen una información muy valiosa para la minería de opiniones. En el artículo se presenta el desarrollo y los resultados de lo que ellos llamaron NOA (News Opinion Analyzer), y que definen como una aplicación web para detectar las tendencias de opinión empleando sentiment analysis y tecnología NoSQL. Los autores explican que la aplicación NOA puede diagnosticar la calidad de los comentarios de una fuente web, y realizaron las pruebas en tres diarios on-line, recopilando de los mismos la información necesaria para llevar a cabo la minería de opiniones.

Otra investigación basada en minería de opiniones es la llevada a cabo por los estudiantes Víctor Cadena y José Romero, de la Universidad de los Andes (Colombia) en el año 2012, titulada: *From Big Data to Content Analysis*[8], en la cual se tuvo como prioridad el desarrollo de una aplicación funcional que proporcionara 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 realizar el sentiment analysis utilizaron *SentiStrength*, la 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.

A nivel internacional existen diversos trabajos relacionados con el sentiment analysis y con los comentarios expresados por los usuarios en sitios de microblogging tales como Twitter[9]. En el artículo de Mendoza, Poblete y Castillo, denominado *"Twitter Under Crisis: Can we trust what we RT?"* [15], los autores toman como eje central el

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

[9] https://twitter.com/

terremoto ocurrido en Chile en el año 2010, analizando los comentarios publicados en Twitter 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. Estas tareas se relacionan directamente con la clasificación de textos, en este caso para analizar la confiabilidad de los tweets en situaciones de emergencia (fenómenos naturales como terremotos, inundaciones, etc., situaciones de orden público, incendios, entre otros). Para realizar el mencionado análisis, los investigadores caracterizaron datos de Twitter relacionados con el terremoto, estudiando Twitter como una plataforma de propagación de mensajes reales y ficticios, discutiendo finalmente el comportamiento de las personas frente a este tipo de información obtenida mediante redes sociales de Internet.

Los mismos autores, Castillo, Mendoza y Poblete en otro artículo: *Information Credibility on Twitter* [8], describen su estudio basado en 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 sin intención. Los métodos que utilizaron para evaluar la veracidad de los tweets se basan en características relevantes de dichos tweets, como el texto de los mensajes, la citación de fuentes externas, el comportamiento de los re-tweets (RT) y las características de publicación de los usuarios. 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., denominado:

Sentiment Analysis of Twitter Data [1], de la Universidad de Columbia (New York), 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: el modelo unigrama, el modelo basado en características, y el modelo basado en núcleo de árbol, los cuales son 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].

Este es el primer trabajo que se realiza en Colombia para realizar análisis de los sentimientos de los colombianos frente a un proceso electoral. Como principales aportes en esta tesis se tiene, en primer lugar, la traducción, revisión y configuración de una base de conocimientos que contiene 2878 términos ajustados al idioma español, y especialmente al contexto colombiano, la cual facilita la minería de opiniones sobre temas electorales en Colombia por medio del uso de la herramienta de software SentiStrength[10]. Es necesario resaltar que sin la mencionada traducción de la base de conocimientos, no hubiese sido posible efectuar la minería de opiniones de los comentarios relacionados con las elecciones presidenciales, debido a que no fue posible hallar una base de conocimientos disponible en español. Esta base de conocimientos es un aporte muy importante al estado del arte, y plantea los cimientos para la creación de bases de conocimientos en castellano con el objetivo de realizar opinion mining frente a procesos electorales en Colombia.

También se construyó una herramienta de recuperación de información, basada en la captura de tweets, la cual reduce el tiempo en la obtención de comentarios muy valiosos para realizar minería de opiniones.

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

De igual manera, se creó una base de datos para almacenar todos los comentarios capturados, ordenados de forma descendente: desde el más reciente hasta el más antiguo, la cual puede servir de apoyo a otros investigadores interesados en realizar estudios basados en opinion mining relacionados con temas electorales.

1.2 Hipótesis de Investigación

"Las opiniones expresadas por los usuarios de servicios de microblogging y otros sitios web 2.0, contienen información necesaria y suficiente para realizar un análisis de los sentimientos de los colombianos".

1.3 Objetivos

1.3.1 Objetivo General

Desarrollar un estudio que permita realizar un análisis de los sentimientos de los colombianos, con base en el tratamiento computacional de las opiniones y demás información subjetiva expresada por internautas colombianos en Twitter y otros medios digitales reconocidos en el país.

1.3.2 Objetivos Específicos

- ✓ Recopilar comentarios publicados en Twitter, elespectador.com y semana.com, que contengan opiniones y demás información subjetiva relacionada con el proceso de elecciones presidenciales en Colombia en el año 2014.

- ✓ Utilizar técnicas de sentiment analysis, a la vez basadas en métodos de machine learning, para extraer información subjetiva contenida en comentarios previamente recopilados y relacionados con el contexto virtual colombiano.

- ✓ Realizar un análisis general que permita comprender los sentimientos, opiniones, emociones y tendencias de los colombianos frente a las elecciones presidenciales de Colombia en 2014, con base en la interpretación de los resultados obtenidos mediante sentiment analysis.

- ✓ Relacionar el análisis de las elecciones presidenciales de 2014 con información pública acerca del proceso de elecciones (encuestas) en Colombia, demostrando la creciente importancia que posee el sentiment analysis para los diferentes sectores de la población nacional.

1.4 Estructura de la Tesis

Este trabajo se encuentra organizado en seis capítulos, en los cuales se abordan diferentes aspectos que permiten alcanzar los objetivos planteados.

A continuación se presenta el capítulo 2, correspondiente al Marco Teórico, en el cual se presentan las bases conceptuales que dan cimiento a esta investigación. Se comienza definiendo el concepto de sentiment analysis, y luego se explican sus principales aplicaciones, disciplinas, métodos y algoritmos asociados, así como los principales factores que dificultan la minería de opiniones. Posteriormente se toca el tema de las redes sociales de Internet, los servicios de microblogging, y otros medios digitales de expresión muy populares en Colombia. Se prosigue con el concepto de recuperación de la información, y se hace énfasis en los web crawlers. Finalmente se presentan los conceptos de subjetividad, sentimientos, emociones y opiniones, los cuales están íntimamente relacionados con el campo de la minería de opiniones.

En el capítulo 3, titulado Desarrollo de la Propuesta, se detalla minuciosamente el proceso llevado a cabo para realizar la investigación. Se mencionan las técnicas e instrumentos utilizados para la recolección de datos. También se explican las tareas y experimentos realizados para efectuar el proceso de minería de opiniones y calcular la polaridad global de los comentarios analizados con SentiStrength. En este capítulo se desarrolla el primer y segundo objetivo de la investigación.

A lo largo del capítulo 4: Resultados, se describen todos los resultados obtenidos durante el proceso de recopilación de la información, la asignación de puntuaciones asignadas por SentiStrength, y los resultados del cálculo de la polaridad global de varios ficheros que contienen comentarios expresados por los colombianos. En este capítulo también se presenta un análisis general construido a partir de estos resultados mencionados. Finalmente se presenta un contraste entre el análisis general obtenido y los resultados de varias encuestas aplicadas por prestigiosas firmas como Gallup Colombia y Datexco Company. En este capítulo es desarrollado el tercer y cuarto objetivo de la investigación.

En el capítulo 5: Conclusiones, se establecen las conclusiones en relación con los objetivos planteados y los resultados obtenidos con el desarrollo de este trabajo.

Finalmente, en el capítulo 6, denominado Recomendaciones y Trabajo Futuro, se dan a conocer las contribuciones de este trabajo, algunas recomendaciones generales, y una propuesta de trabajo futuro.

Capítulo 2

Marco Teórico

Los modernos sitios web brindan a los usuarios la posibilidad de generar contenido mediante comentarios (texto), lo cual es una invaluable fuente de información para diversos sectores del mercado debido a que es posible aprovechar ese gran cúmulo de información que diariamente se genera y circula por la Web para realizar lo que se ha venido a llamar: minería de opiniones o análisis de sentimientos.

A continuación se presenta la base conceptual que sustenta el análisis del contexto colombiano a través de opinion mining propuesto en el presente trabajo de investigación. Se comienza con la definición formal de sentiment analysis, sus principales aplicaciones y tareas asociadas, principales obstáculos y algunas herramientas disponibles para realizar minería de opiniones. También se abordan las redes sociales de Internet y los servicios de microblogging como Twitter; así como los diarios y revistas on-line que favorecen las tareas de sentiment analysis. Posteriormente se definen los conceptos de subjetividad, sentimientos, opiniones y emociones, los cuales se relacionan con la minería de opiniones. Finalmente se mencionan ciertos aspectos de gran relevancia en cuanto al contexto político de las elecciones presidenciales en Colombia.

2.1 Sentiment Analysis

También conocido como Opinion Mining [9], es un campo de estudio de reciente aparición, que se encarga de la identificación y extracción de sentimientos, emociones y opiniones en determinados textos. Actualmente esta identificación se

realiza principalmente con base en los contenidos generados en redes sociales de Internet, servicios de microblogging, y otros medios digitales de expresión[11] tales como diarios y revistas on-line, en los que se observa un proceso de retroalimentación entre la entidad que genera los contenidos iniciales y los usuarios interesados en los temas publicados.

Según Bing Liu [14], el término *sentiment analysis* apareció quizá por primera vez en el artículo de Nasukawa & Yi [16] en el año 2003; mientras que el término *opinion mining* apareció por primera vez en el artículo de Dave, Lawrence & Pennock [9] también en el año 2003. En la industria es más frecuentemente utilizado el término sentiment analysis, pero en la academia ambos (opinion mining y sentiment analysis) son ampliamente utilizados [14].

El sentiment analysis está relacionado con áreas tales como el procesamiento de lenguajes naturales, la minería de textos y la lingüística computacional. Pang & Lee [19] sugieren que el sentiment analysis puede ser considerado como una sub-área del "subjectivity analysis", el cual consiste en clasificar contenidos de acuerdo a si son objetivos o subjetivos [4].

2.1.1 Aplicaciones del sentiment analysis

Una de las principales aplicaciones del sentiment analysis es conocer qué piensan los lectores (usuarios) de un determinado texto (e.g., comentarios en micro-blogs y otros sitios web), esto con el fin de saber cuáles son las opiniones, sentimientos, tendencias, etc. que experimentan las personas, en este caso en el contexto colombiano, acerca de diferentes temas, personajes públicos, productos, servicios, etc., para realizar comparaciones y/o evaluaciones que generalmente se pueden

[11] http://www.banrepcultural.org/blaavirtual/ayudadetareas/periodismo/losmediosdecomunicacion.htm

llevar a cabo mediante la clasificación de la polaridad de los comentarios de los usuarios.

Otras aplicaciones se despliegan de la anterior, entre las cuales se pueden mencionar:

- “Detección de lenguaje ofensivo
- Sistemas de recomendación
- Análisis de citas bibliográficas
- Resumen y recopilación automática
- Seguimiento histórico de las tendencias de los clientes
- Seguimiento de la popularidad de candidatos políticos
- Aplicación a otras disciplinas o ciencias” (Cruz, 2008)

La aplicación relacionada con el seguimiento de la popularidad de candidatos políticos es la que mejor se ajusta al caso específico tomado para realizar sentiment analysis mediante esta investigación, contextualizándose en el seguimiento, procesamiento y posterior análisis de las elecciones presidenciales de Colombia en el año 2014.

2.1.2 Clasificación de la polaridad de un texto

Entre las principales tareas del sentiment analysis se encuentra la clasificación de la polaridad de un texto dado, para lo cual se utiliza por lo general una clasificación en dos o tres categorías: [positivo / negativo] o [positivo / neutral / negativo]. Neutral indica que no hay una tendencia clara en la expresión (texto) acerca de si es positiva o negativa, por lo tanto podría tratarse de una expresión con contenido objetivo, es decir que no expresa subjetividad.

Otro tipo de clasificación es un sistema de escala asociada a números desde -10 hasta +10, desde el más negativo hasta el más positivo, en donde los números representan las opiniones que se tienen sobre un tema, persona, producto, servicio, etc.

También existe la clasificación de polaridad basada en la identificación de subjetividad u objetividad: lo cual se logra clasificando un texto determinado en dos estados posibles: objetivo o subjetivo.

Un tipo de clasificación de polaridad aún más avanzada tiene que ver con la clasificación de textos de acuerdo a tres estados emocionales: “enojado, triste y feliz”.

El tipo de clasificación de polaridad de textos que se tuvo en cuenta para este trabajo es la clasificación en tres (3) categorías: [positivo / neutral / negativo]. Los textos analizados corresponden a comentarios expresados por cibernautas colombianos acerca de las elecciones presidenciales en 2014. La escala que se manejó está conformada por números desde -5 (comentarios extremadamente negativos) hasta +5 (comentarios extremadamente positivos).

2.1.3 Herramientas disponibles para realizar opinion mining

Existen varias herramientas de software que permiten realizar tareas de opinion mining sobre palabras, frases o textos más grandes. La mayoría de estas herramientas ofrecen una interfaz (generalmente clasificación de textos) mediante su sitio web oficial, con lo cual es posible introducir palabras o frases cortas para su evaluación en cuanto a la detección de subjetividades y a la asignación de una polaridad. Por lo general estas interfaces web son limitadas y no permiten el procesamiento de grandes cantidades de textos, pero algunas de esas herramientas

ofrecen la descarga de una versión instalable (fichero binario) que permite analizar varias frases, e incluso ficheros planos completos.

Algunas herramientas de software existentes para facilitar la minería de opiniones son:

- *SentiWordNet*[12]: recurso léxico para la minería de opiniones, el cual permite analizar palabras en inglés en busca de una polaridad positiva o negativa, o determinar si esa palabra es objetiva (neutral, que no expresa subjetividad). No está disponible en español y tampoco permite cargar ficheros planos para su análisis.

- *SenticNet*[13]: se trata de un recurso semántico disponible públicamente para efectuar sentiment analysis a nivel de conceptos. Ofrece una sección "demo" en su sitio web, mediante la cual es posible conocer la polaridad de palabras y frases cortas, pero no permite el análisis de ficheros con líneas de texto en su interior. Esta herramienta tampoco ofrece compatibilidad para realizar sentiment analysis con textos escritos en el idioma español.

- *WordNet-Affect*[14]: esta herramienta permite representar conceptos afectivos correlacionados con palabras afectivas. Las palabras o conceptos afectivos que representan un estado emocional se individualizan con emociones marcadas por etiquetas. El idioma español no es soportado por esta herramienta.

Para la realización de opinion mining en este trabajo de investigación se utilizó la herramienta de software SentiStrength, por las razones que se describen a continuación.

[12] http://sentiwordnet.isti.cnr.it/
[13] http://sentic.net/
[14] http://wndomains.fbk.eu/wnaffect.html

2.1.3.1 SentiStrength

Es una herramienta de software para realizar opinion mining, cuyo lenguaje por defecto es el inglés, pero permite su configuración a otros 13 idiomas, incluyendo el español. El objetivo principal de SentiStrength es estimar la *fuerza* positiva o negativa de los sentimientos (subjetividades) detectados en palabras, frases y documentos de texto plano (ficheros .txt).

SentiStrength clasifica los textos sometidos a análisis, mediante la asignación de puntuaciones (scores) a cada palabra de un texto, representadas por medio de dos categorías de escalas de emociones: positiva y negativa. Esta característica, sumada a la posibilidad de importar y analizar ficheros planos para realizar sentiment analysis, y la compatibilidad con el español, fueron determinantes para seleccionar a SentiStrength como el software de minería de opiniones en esta investigación.

Las escalas de emociones mencionadas consisten en asignar scores negativos: desde *-1* hasta *-5* (siendo -1 neutral y -5 muy negativo), y scores positivos: desde *+1* hasta *+5* (siendo 1 neutral y 5 muy positivo).

El proceso de configuración de los ficheros necesarios para personalizar el idioma de SentiStrength (al español) se describe detalladamente en el capítulo 3 de este trabajo. Estos ficheros conforman la base de conocimientos sobre la cual SentiStrength realiza el proceso de clasificación de cada palabra.

2878 términos (conformados por palabras y otras expresiones) fueron traducidos al español para lograr conformar la base de conocimientos en español, la cual es el resultado de un esfuerzo realizado por un equipo interdisciplinario para este trabajo novedoso en el contexto colombiano.

La versión de SentiStrength utilizada en este trabajo fue la 2.2 (ver figura 1), la cual se puede descargar desde el sitio web oficial, especialmente para realizar investigaciones de tipo académico.

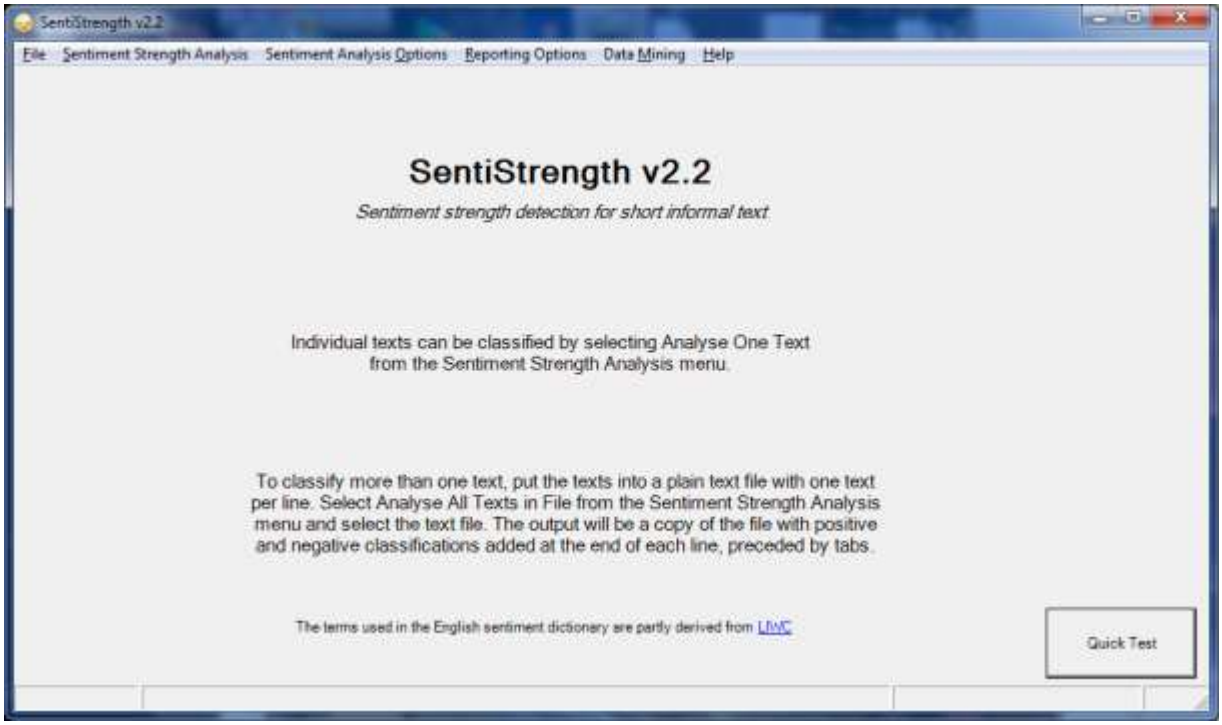


Figura 1. Menú principal de SentiStrength v2.2

2.1.4 Factores que dificultan las tareas de opinion mining

Existen diversos factores que hacen que las tareas de opinion mining, por ejemplo la clasificación de polaridad de un texto, se tornen especialmente difíciles. Entre dichos factores se pueden destacar:

- Presencia de ambigüedad en las oraciones (textos)
- Se debe tener una gran visión (conocimiento de aspectos universales)
- Existe una dependencia muy fuerte entre las oraciones y el contexto.

La dependencia del contexto llama mucho la atención, sobre todo en países como Colombia, en donde se suelen utilizar varias palabras para hacer referencia a un mismo objeto, situación o lugar.

A continuación se citan algunos ejemplos de textos (traducidos al español) mostrados por Pang & Lee [19], los cuales explican un poco mejor la existencia de las dificultades que pueden presentarse en las tareas de sentiment analysis:

a) *"Ella recorre toda la gama de emociones de A a B."*
   No hay palabras ostensiblemente negativas.

b) *"Si usted está leyendo esto porque es su fragancia querida, por favor usar en casa exclusivamente, y cierre las ventanas con cinta."*
   No hay palabras ostensiblemente negativas.

c) *Esta película debe ser brillante. Suena como un gran argumento, los actores son de primer grado, y el reparto es bueno también, y Stallone está intentando entregar un buen rendimiento. Sin embargo, no se puede sostener.*
   La mayoría de las palabras son positivas, pero es una expresión negativa.

2.1.5 Sentiment Analysis y la Web 2.0

La Web 2.0 ha traído consigo el desarrollo de sitios web centrados en los usuarios, quienes ahora están en capacidad de generar contenidos y han pasado de ser usuarios pasivos a ser usuarios muy activos de la Internet.

Tal es el caso de los sitios web que permiten a los usuarios realizar comentarios y subir contenidos digitales sobre algún tema común para una comunidad virtual; este tipo de sitios, generalmente micro-blogs, redes sociales de Internet, o portales de noticias, tiene una concurrida comunidad de usuarios que aportan su grano de arena

para expresar sus opiniones (sentimientos). Con esta proliferación de opiniones las empresas y centros de investigación han visto una gran oportunidad para efectuar clasificaciones y realizar análisis basados en esas opiniones, realizar valoraciones, brindar recomendaciones, entre otras muchas tareas que se resumen en un nuevo campo de estudio: el sentiment analysis.

Gracias a la Web 2.0 y al sentiment analysis, se están generando y tratando computacionalmente nuevas fuentes de información que antes parecía imposibles de obtener; una de las tareas de mayor interés es conocer la opinión general (o sentimientos) de una ciudad, región o país – como es el caso de Colombia – acerca de diferentes hechos que se viven cotidianamente, y que para la mayoría de la población representan ya sean alegrías o tristezas, satisfacción o frustración, odio o amor, sólo por mencionar algunos de los sentimientos más comunes.

El tema de las elecciones presidenciales escogido en esta investigación abarca todo ese conjunto de sentimientos que la población nacional siente y expresa a través de los sitios web 2.0, por lo cual este tema fue seleccionado para realizar el presente análisis del contexto colombiano a través de minería de opiniones.

2.2 Redes Sociales de Internet, y otros Medios Digitales de Expresión

El término “red social” (en inglés: social network) puede comprenderse como la representación de una estructura social que indica la relación que existe entre un conjunto de entidades (personas, organizaciones) conectados mediante un criterio específico, ya sea de tipo familiar, académico, laboral, de amistad, entre otros. A menudo éste término se utiliza de manera indistinta para referirse a red social de Internet, también conocido como “servicio de red social”.

Una red social de Internet se puede definir como un medio de comunicación social que presta servicios a los usuarios para encontrar personas y relacionarse en línea; las personas que conforman las redes sociales de Internet generalmente comparten algún tipo de relación, entre las cuales se destaca la amistad, y están interesadas en compartir contenidos o ver los de otros usuarios, los cuales son generados mediante tecnologías pertenecientes a la Web 2.0.

Según cifras oficiales arrojadas por un estudio[15] del Ministerio de tecnologías de la información y las comunicaciones (MinTIC[16]) en Colombia, indican que a inicios del año 2013 cuatro de cada diez colombianos usaban a diario redes sociales de Internet, y además, para septiembre de 2013 Colombia ocupó el noveno lugar a nivel mundial en cuanto al uso de las redes sociales de Internet[17]. Estas cifras demuestran la gran importancia y fuerza que cobran cada vez más las redes sociales de Internet en el contexto colombiano, especialmente para el planteamiento y discusión de temas de mucho interés para la población nacional, como los que se relacionan con el sector político.

Las redes sociales de Internet se dividen en: redes sociales internas, y redes sociales externas (e.g., Twitter[18], Facebook[19], Tumblr[20]).

En Colombia se utiliza mayoritariamente Facebook y Twitter, respectivamente, sin embargo se ha observado que en los últimos 3 años Twitter se ha convertido en la red social preferida por los colombianos para publicar contenidos de interés y generar controversia mediante la Red.

[15] http://www.portafolio.co/portafolio-plus/uso-las-redes-sociales-colombia
[16] http://www.mintic.gov.co
[17] http://www.portafolio.co/portafolio-plus/social-media-week-bogota
[18] https://twitter.com/
[19] https://www.facebook.com/
[20] https://www.tumblr.com/

2.2.1 Microblogging

El término microblogging (también conocido como micro-blog) hace referencia a un servicio que permite la publicación de mensajes cortos (generalmente sólo de texto) por medio de sitios tales como Twitter.

El contenido de los mensajes publicados mediante este servicio varía dependiendo del contexto, pero por lo general estos mensajes contienen conversaciones, noticias, debates, avances sobre novedades de interés, autopromociones, entre otros.

Por medio de los servicios de microblogging los usuarios pueden expresar su estado de ánimo a otros usuarios, sus ideas u opiniones acerca de un tema determinado, lo cual ubica a este tipo de servicios como uno de los más importantes actualmente en Internet por ser una inmensa fuente de datos de gran relevancia para diversos sectores del mercado.

2.2.1.1 Twitter

Es un servicio de microblogging muy popular creado por Jack Doersey en marzo de 2006, el cual para el año 2014 supera los 220 millones de usuarios activos en todo el mundo[21]. Permite a usuarios registrados enviar mensajes cortos, de 140 caracteres como máximo, los cuales reciben el nombre de tweets (tuits, en español).

Para comunicarse mediante Twitter los usuarios registrados pueden utilizar varias alternativas:

[21] http://www.pulzo.com/tecnologia/144756-twitter-ya-tiene-227-millones-de-usuarios-en-todo-el-mundo-y-contando

- Utilizando *hashtags*, por medio del uso del símbolo # seguido del nombre del grupo o personaje relacionado con el mensaje (e.g., #Elecciones2014).

- Utilizando el símbolo @ y seguidamente el nombre del usuario (e.g., @JuanManSantos), para dirigirse hacia un usuario de Twitter específico.

- Retuiteando (re-tweet o RT) mensajes de otros usuarios, es decir citándolos mediante la copia de su mensaje y nombre de usuario en Twitter.

Twitter coloca a disposición de los programadores, o personas con conocimientos en informática, la versión 1.1 de su API, mediante la cual es posible establecer comunicación entre aplicaciones desarrolladas por el usuario (e.g., utilizando PHP) y entre aplicaciones creadas en Twitter (https://apps.twitter.com/), con el objetivo de compartir (recibir y enviar) información relacionada con diversos tweets.

Durante el desarrollo de esta investigación se construyó una aplicación web en el lenguaje de programación PHP, la cual se conecta con la API de Twitter por medio de unas credenciales de autenticación generadas por una aplicación creada en Twitter. El objetivo principal es extraer tweets relacionados con las elecciones presidenciales en Colombia en el año 2014 y los cinco candidatos que aspiraron a la presidencia, por medio del carácter # (haciendo uso de hashtags).

Esta facilidad que ofrece Twitter para obtener información sobre tweets relacionados con cualquier tema o usuario, sumado a la facilidad que brinda para la adaptabilidad con aplicaciones desarrolladas en lenguajes de programación web, fue el principal motivo para escoger a este servicio de microblogging como una de las fuentes de información desde la cual se realizaron capturas de comentarios (texto) relacionados con las elecciones presidenciales de Colombia en el año 2014.

2.2.2 Medios digitales de expresión en Colombia

Debido a la masificación de las NTIC, en Colombia se han construido diferentes plataformas web informativas conformadas por elementos audiovisuales y con formatos de interacción, las cuales permiten que el público esté constantemente enterado de los últimos acontecimientos del país, y también brindan la posibilidad de retroalimentación mediante la expresión de opiniones (comentarios) de los ciudadanos en las diferentes noticias publicadas.

Entre los medios digitales informativos más utilizados en Colombia se encuentran los blogs, las emisoras radiales virtuales, y las versiones digitales de los medios impresos: diarios y revistas.

2.2.2.1 Diarios on-line

Internet ha revolucionado la manera cómo las personas se informan acerca de los diferentes acontecimientos locales, nacionales e internacionales, dando paso a la creación de sitios web dedicados a informar constantemente a los internautas sobre diferentes noticias clasificadas en varias secciones, manteniendo el estilo de la prensa tradicional. A estos sitios web se les conoce como diarios on-line.

Una característica fundamental de los diarios on-line actuales, es que permiten a los usuarios realizar comentarios sobre las diferentes noticias publicadas, generando de esta manera contenidos al estilo de la Web 2.0, lo cual constituye una excelente fuente de información para lograr identificar y extraer emociones y subjetividades a partir de dichos comentarios.

El Espectador[22] es uno de los diarios más antiguos de América, y el de mayor antigüedad y trayectoria en Colombia. Fue fundado en Medellín en el año 1887, y su versión on-line (www.elespectador.com) se inauguró en el año 1996, permitiendo de esa manera el acceso global a las noticias a través de la Web, y permitiendo a los usuarios (lectores) suscribirse y comentar las noticias.

Debido a estas razones, *elespectador.com* fue seleccionado en esta investigación como una de las fuentes de información a partir de las cuales se recolectaron comentarios expresados por los usuarios, con relación a las elecciones presidenciales de 2014.

2.2.2.2 Revistas on-line

De la misma forma en que la prensa tradicional ha llevado sus contenidos a la Web, también lo han hecho varias revistas colombianas que hablan de política, actualidad, entretenimiento, entre otros. Una de ellas es Semana[23], la cual fue fundada en 1946, y cuya versión on-line fue inaugurada en el año 1998.

Esta revista es una de las más leídas en Colombia, publica temas de actualidad en diferentes secciones, relacionados con situaciones que despiertan el interés de los colombianos, y permite la realimentación mediante los comentarios que los usuarios expresan sobre cada noticia o publicación.

Por tales motivos *semana.com* también fue escogida como fuente de información para recolectar comentarios relacionados con las elecciones presidenciales en Colombia.

[22] http://www.elespectador.com/noticias

[23] http://www.semana.com/Home

2.3 Recuperación de la Información

La recuperación de la información (IR) es una importante área de la informática que se ocupa de ofrecer a las personas un fácil acceso a información de su interés, contenida en documentos electrónicos y cualquier tipo de colección documental digital.

De acuerdo a la definición presentada por Baeza et al. [3], "IR trata acerca de la representación, almacenamiento, organización y acceso a los elementos de información, como documentos, páginas web, catálogos en línea, registros estructurados y semi-estructurados, objetos multimedia. La representación y la organización de los elementos de información deben ser tales que proporcionan a los usuarios un fácil acceso a la información de su interés".

En este trabajo de investigación la información fue recuperada a partir de la Web, específicamente a través de comentarios generados por los usuarios en documentos o páginas presentes en los sitios de noticias *elespectador.com* y *semana.com*, y también en comentarios generados por los usuarios en el servicio de microblogging Twitter.

2.3.1 Recuperación de documentos desde la Web

La recuperación de la información desde la Web es posible mediante el diseño de soluciones que permitan recorrerla en busca de documentos con ciertas características, y que además almacenen dichos documentos de forma local. Existen múltiples soluciones que cumplen con esas funciones, y todas están basadas en el concepto de *Web Crawling*.

2.3.1.1 Web Crawling

Un *Web Crawler* (también conocido como *araña Web)* es un software encargado de recorrer el entramado de páginas web (documentos HTML) de manera automática y sistemática. Este tipo de software especializado se conoce como 'robots de Internet' o 'robots de la Web' (webbot), y actúa como un visitante que se dedica a recorrer la Web para recuperar documentos, aprovechando la estructura de grafo que posee la misma, de tal manera que es capaz de descargar de forma local y procesar las páginas web recuperadas.

Para su funcionamiento generalmente los web crawlers disponen de un conjunto inicial de URLs, las cuales se conocen como *semillas*. A medida que el web crawler trabaja, va descargando las páginas Web asociadas con las semillas, y también busca otras URLs dentro de dichas semillas, de tal forma que cada nueva URL hallada se agrega a la lista de URLs que el web crawler debe visitar, lo cual se conoce como recolección de URLs.

Dependiendo de la información que se desee recuperar de la Web, la manera en que ésta será recorrida, así como otras restricciones relacionadas con el nivel de profundidad de las búsquedas, los web crawlers pueden clasificarse en:

- *Crawlers de propósito general:* utilizan una fuente (información) como punto de partida para el proceso de recuperación de la información, y la búsqueda que efectúan es en profundidad, intentando recorrer la mayor cantidad posible de nodos a través de la Web. Un ejemplo de este tipo de web crawlers son los motores de búsqueda.
- *Crawlers distribuidos:* son varios web crawlers que se encuentran distribuidos y conectados mediante una red de computadoras, buscando aumentar la capacidad de procesamiento, con el objetivo de optimizar el flujo de información y evitar la obtención de información incorrecta o no deseada.

- *Crawlers focalizados o temáticos:* se refieren a una especialización de los crawlers de propósito general, cuyo objetivo general es recuperar información a través de la Web que únicamente pertenezca a un tema en particular o a un conjunto de sitios web definidos previamente.

Para la recuperación de información relacionada con las elecciones presidenciales de 2014 desde los sitios *elespectador.com* y *semana.com*, se utilizó un web crawler con características de focalización, llamado "Win Web Crawler" (ver figura 2).

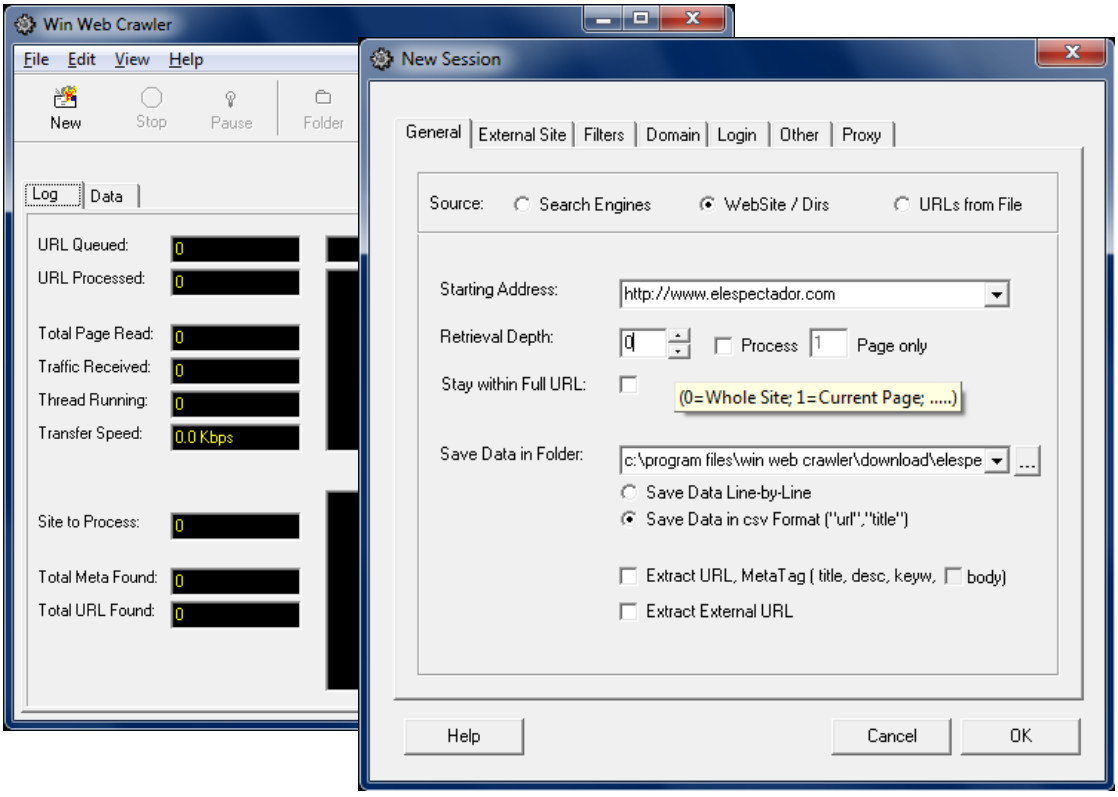


Figura 2. Interfaz de Win Web Crawler y ajustes generales para iniciar recuperación de la información (IR)

Como se muestra en la figura 2, Win Web Crawler permite especificar el sitio web (URL) a partir del cual iniciar la búsqueda de información, así como establecer el nivel de profundidad (por defecto se deja en cero [0] para buscar en el sitio entero),

utilizar un filtro de palabras o frases sobre las cuales buscar información, y seleccionar el directorio en el cual se guardarán los resultados de la búsqueda.

En la figura 3 se puede apreciar la inclusión de los nombres de los candidatos y del tema de las elecciones presidenciales en general, para recuperar páginas web desde el sitio especificado previamente (elespectador.com), y desde una fecha específica:

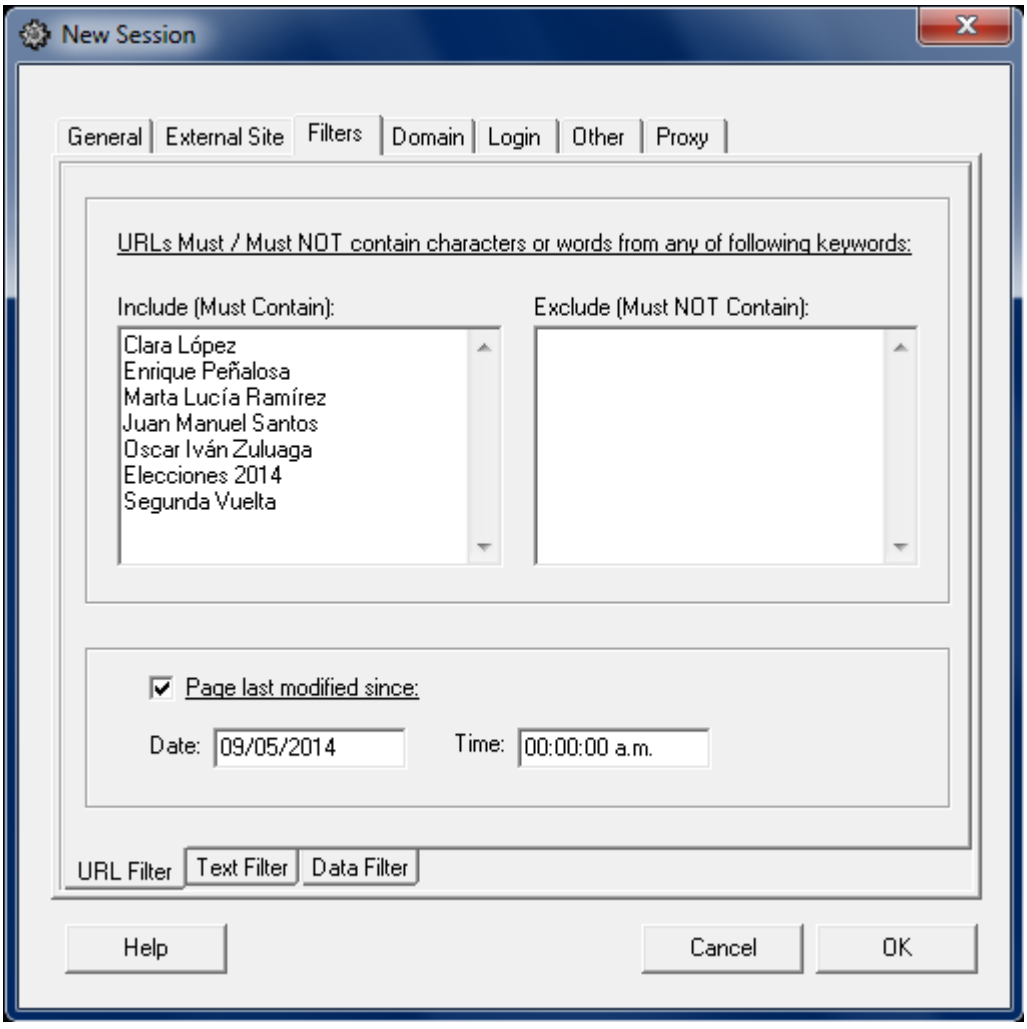


Figura 3. Configuración de filtro de búsqueda en Win Web Crawler

2.4 Subjetividad, Sentimientos, Opiniones y Emociones en los Seres Humanos

Gracias a las expresiones faciales es posible exteriorizar emociones en determinados momentos. Sin embargo muchas investigaciones en el campo del sentiment analysis son llevadas a cabo para intentar conocer la circunstancia personal, estado de ánimo o comportamiento de una persona, con base en sus comentarios expresados mediante la escritura (texto).

Cuando una persona escribe una frase o párrafo, seguramente quiere transmitir a los lectores diferentes sensaciones y enfoques con base en el tema tratado; precisamente con el sentiment analysis se intenta conocer cuáles son las tendencias de las de personas con respecto a un tema en particular, por lo general determinando la polaridad (positiva – neutral – negativa) de las opiniones acerca de un tema, por medio de la detección de subjetividades con base en el análisis de textos, sin que sea necesario un análisis de las expresiones musculares (estados faciales) de las personas.

La terminología utilizada en la definición de las diferentes tareas que se ocupan de la subjetividad, y los conceptos involucrados en ellas, aún no es uniforme en la comunidad investigativa, debido a la gran cantidad de investigaciones relacionadas llevadas a cabo durante los últimos años [4]. No obstante, a continuación se presentan varias definiciones relacionadas con el campo del sentiment analysis y con las tareas que con él se relacionan, para comprender mejor la relación entre este campo de reciente aparición y algunos aspectos inherentes a la conducta humana.

2.4.1 Subjetividad

La subjetividad puede definirse como la “expresión lingüística de las opiniones, los sentimientos, las emociones, las evaluaciones, las creencias y las especulaciones de alguien” [24].

Según la teoría del conocimiento, “subjetividad es la propiedad de las percepciones, argumentos y lenguaje basados en el punto de vista del sujeto, y por tanto influidos por los intereses y deseos particulares del mismo”.

Una tarea común cuando se compara la subjetividad con la objetividad, es determinar si un texto únicamente contiene información acerca de los hechos (objetividad), o si contiene contenido dogmático (subjetividad). Puede darse el caso de que el texto contenga ambos tipos de expresiones (objetivas y subjetivas).

Para Wiebe et al. [25] la subjetividad es un sentido pragmático equivalente al término *“estado privado”* [6], el cual es definido como un término general que abarca opiniones, sentimientos, creencias, emociones, pensamientos, objetivos, evaluaciones y juicios.

Un ejemplo de frase subjetiva, teniendo en cuenta la definición propuesta por Wiebe [24], es: “¡El partido fue increíble!”. Por otra parte, un ejemplo de frase objetiva es: “La selección colombiana de fútbol derrotó cinco a cero a su oponente”.

2.4.2 Sentimientos

Los sentimientos están fuertemente ligados a las emociones: puede decirse que los sentimientos son el resultado de las mismas. Van den Bos [23] define sentimiento como “la experiencia subjetiva consciente de la emoción”. Por otra parte, Scherer

[22] también afirma que los sentimientos son sólo una pequeña parte de una emoción.

Alexandra Balahur [4] describe el sentimiento como un tipo de opinión que hace referencia a la emoción, en el sentido de que es “la expresión de una evaluación basada en la emoción que el escritor siente”. Las tareas de sentiment analysis intentan extraer atributos y componentes de un documento de texto que contenga opiniones (o sentimientos), buscando determinar la polaridad de los comentarios expresados.

2.4.3 Opiniones

La opinión suele asociarse a los juicios subjetivos, y se puede definir como el concepto que podemos tener acerca de algo o alguien. En otras palabras, una opinión es un juicio que se forma a partir de algo que es cuestionable.

Kim & Hovy [13] definen la opinión desde el punto de vista computacional “como una relación cuádruple [Tema, Titular, Afirmación, Sentimiento] en la que el titular considera una afirmación sobre el tema, y en muchos casos asociada un sentimiento, como bueno o malo, con la creencia”.

Un ejemplo de opinión subjetiva puede ser: “Esta computadora es la mejor de la actualidad”. Puede verse que en la frase se puede conocer lo que el emisor piensa acerca de la computadora en cuestión, dependiendo de sus conocimientos informáticos y otra serie de aspectos muy personales como sus gustos, ideas, etc.

En cambio, una opinión objetiva, como: “Los síntomas indican que se trata de apendicitis”, puede ser la emitida por un experto, en la cual se puede observar que no hay presencia de sentimientos o juicios subjetivos.

No todas las opiniones son subjetivas, así como no todas las opiniones tienen sentimientos asociados a ellas. Kim & Hovy [13] ofrecen los siguientes ejemplos para ilustrar la diferencia entre opiniones con sentimientos y opiniones sin sentimientos:

- "Creo que el mundo es plano."
- "Es probable que Gap vaya a la quiebra." (Gap se trata de una empresa).

Ambas oraciones expresan opiniones, pero ninguna contiene sentimientos.

2.4.4 Emociones

La emoción es un fenómeno muy complejo al que aún no se le ha dado una definición aceptada universalmente. Sin embargo, se puede definir como una "variación profunda pero efímera del estado de ánimo de la persona"[24], como respuesta a un estímulo externo (encuentro con personas, lugares, cosas) o interno (recuerdos importantes).

Para Ekman [10] la función principal de la emoción es "movilizar a los organismos para reaccionar rápidamente con los encuentros entre las personas, o entre personas y otras cosas o hechos". Las emociones son responsables de organizar rápidamente las respuestas de algunos sistemas biológicos como los músculos y expresiones faciales, la voz, el sistema endocrino y el sistema nervioso autónomo.

Las emociones permiten a toda persona establecer una posición con respecto al ambiente que le rodea. En el artículo *Autonomic nervous system activity distinguishes among emotions* Ekman [11] propone un modelo de seis emociones

[24] http://definicion.de/emocion/

universales (también llamadas familias de emociones básicas), las cuales son: ira, asco, miedo, felicidad, tristeza y sorpresa.

La política es una de las áreas o sectores que más genera reacciones en las personas; durante sus campañas, los candidatos aspirantes a la presidencia de Colombia trataron activamente de generar empatía con los ciudadanos (potenciales votantes) y de provocar emociones positivas para obtener su voto.

Lo anterior deja en claro que realizar minería de opiniones sobre los comentarios de los cibernautas, relacionados con procesos electorales, es una poderosa herramienta para analizar las tendencias de los electores, modificar aspectos de una campaña política [17, 21], o analizar los resultados obtenidos al finalizar un proceso de elecciones como el llevado a cabo en Colombia para escoger al presidente de la república durante el período 2014 – 2018.

2.5 Aspectos Relevantes en el Contexto Político de las Elecciones

Según el artículo 188 de la Constitución Política de Colombia[25], el Presidente "simboliza la unidad nacional y al jurar el cumplimiento de la Constitución y de las leyes, se obliga a garantizar los derechos y libertades de todos los colombianos". Las elecciones presidenciales constituyen un hecho de vital importancia para el país, debido a que mediante ellas se elige a un ciudadano cualificado para representar al país durante un período de 4 años.

A continuación se describe el proceso que, según la Constitución Política de Colombia, debe seguirse para elegir al Presidente, y también se efectúa una breve reseña de los candidatos a la Presidencia de Colombia para el período 2014 – 2018.

[25] http://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=4125

2.5.1 Proceso de elección del Presidente de la República

Los requisitos para ser Presidente la República, según el artículo 191 de la Constitución Política de Colombia[26], son: ser colombiano por nacimiento, ciudadano en ejercicio y mayor de treinta años.

Con respecto a la elección del Presidente, la Constitución en su artículo 190, plantea: "El Presidente de la República será elegido para un período de cuatro años, por la mitad más uno de los votos que, de manera secreta y directa, depositen los ciudadanos en la fecha y con las formalidades que determine la Ley".

En caso de uno de los candidatos presidenciales obtenga una votación igual o superior a la mitad más uno del total de los votos, ganará las elecciones, y en este caso se dice que ganó en Primera Vuelta.

La Constitución, en el mismo artículo 190, agrega: "Si ningún candidato obtiene dicha mayoría, se celebrará una nueva votación que tendrá lugar tres semanas más tarde, en la que sólo participarán los dos candidatos que hubieren obtenido las más altas votaciones. Será declarado Presidente quien obtenga el mayor".

Este hecho es más conocido como Segunda Vuelta, y se celebra una nueva instancia de las votaciones para elegir al Presidente de los colombianos, buscando lograr la transparencia de las elecciones presidenciales.

En las elecciones para la elección de Presidente, llevadas a cabo en Colombia en el año 2014, hubo segunda vuelta debido a que ninguno de los cinco candidatos iniciales aspirantes a la Presidencia logró obtener una votación mayor o igual a la mitad más uno de los votos. A la segunda vuelta solo pasaron los dos candidatos

[26] http://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=4125

que obtuvieron las más altas votaciones: Oscar Iván Zuluaga y Juan Manuel Santos, respectivamente.

2.5.2 Reseña de los candidatos a la Presidencia de Colombia en el año 2014

Las elecciones presidenciales de Colombia, llevadas a cabo en el año 2014 con el objetivo de elegir al primer mandatario de los colombianos para el período 2014 – 2018, contó con la inscripción y participación de los siguientes ciudadanos, en calidad de candidatos presidenciales:

- **Clara Eugenia López Obregón:** "economista, dirigente y política colombiana, graduada en economía de la Universidad de Harvard y como doctora en derecho financiero y tributario por la Universidad de Salamanca, adicional a lo anterior, en el ámbito académico ha sido profesora de la Universidad del Rosario y Universidad de los Andes"[27].

  Candidata por la colación de los partidos políticos Polo Democrático Alternativo y Unión Patriótica. Su candidatura se definió dos años antes de las elecciones, y se ratificó el 13 de marzo de 2014, día en que se selló la alianza entre el Polo Democrático Alternativo y la Unión Patriótica para las elecciones presidenciales, inscribiendo como candidata a la Vicepresidencia a la ciudadana Aída Avella.

  En la hoja de vida de Clara López cabe destacar que ha sido concejal, secretaria de gobierno, contralora y alcaldesa mayor (encargada) de Bogotá.

- **Enrique Peñalosa Londoño:** "se graduó en Economía e Historia en la Universidad de Duke. Realizó sus estudios de maestría en Administración en el Institut International d'Administration Publique y un doctorado en

[27] http://es.wikipedia.org/wiki/Clara_L%C3%B3pez_Obreg%C3%B3n

Administración Pública en la Universidad de París II (Universidad Panthéon-Assas)"[28].

Candidato por el partido político Alianza Verde. Su candidatura se definió el 9 de marzo de 2014, día de las elecciones parlamentarias. Su fórmula vicepresidencial fue la ciudadana Isabel Segovia.

Enrique Peñalosa ha sido representante a la cámara por Cundinamarca, y alcalde mayor de Bogotá.

- **Marta Lucía Ramírez de Rincón:** "abogada y política colombiana. Fue ministra de comercio exterior, embajadora en Francia y la única mujer en ocupar el cargo de ministro de defensa de Colombia. En el año 2006 resultó elegida como senadora de la república por el Partido de la U"[29].

  Candidata por el partido Conservador Colombiano. Su candidatura fue definida el 26 de enero de 2014, en una convención realizada al interior del partico conservador. Como fórmula vicepresidencial fue inscrito el ciudadano Camilo Gómez.

  Marta Lucía Ramírez ha desempeñado importantes cargos, tales como: viceministra de comercio exterior, embajadora de Colombia en Francia, senadora de la República, ministra de comercio exterior, y ministra de defensa.

- **Juan Manuel Santos Calderón:** "político, periodista y economista colombiano, actual presidente de Colombia desde el 7 de agosto de 2010. Tras desempeñarse como periodista, incursionó a la política haciéndose miembro del Partido Liberal Colombiano, fue ministro de Comercio Exterior durante el gobierno de César Gaviria, más tarde hizo parte del gobierno del

[28] http://es.wikipedia.org/wiki/Enrique_Pe%C3%B1alosa

[29] http://es.wikipedia.org/wiki/Marta_Luc%C3%ADa_Ram%C3%ADrez

conservador Andrés Pastrana Arango en el que se desempeñó como ministro de Hacienda"[30].

Candidato por el partido Social de Unidad Nacional. El 20 de noviembre de 2013 anunció públicamente su intención de aspirar a la reelección presidencial. Su fórmula vicepresidencial fue el ciudadano Germán Vargas Lleras.

Juan Manuel Santos ha desempeñado importantes cargos, entre los cuales cabe destacar: Presidente de Colombia para el período 2010 – 2014, ministro de defensa, ministro de comercio exterior, y ministro de hacienda y crédito público.

- **Oscar Iván Zuluaga Escobar:** "economista, empresario y político colombiano. Se desempeñó como Ministro de Hacienda en el segundo período de gobierno de Álvaro Uribe Vélez. Ha sido concejal y alcalde de su municipio, Pensilvania, dirigente empresarial y gremial, y Senador de la República. Como Senador fue cofundador del Partido de la U".[31]

  Candidato por el partido Centro Democrático. Su candidatura fue inscrita durante los días 25 y 26 de octubre de 2013, en la convención nacional de ese partido, llevada a cabo en la ciudad de Bogotá. Su fórmula vicepresidencial fue el ciudadano Carlos Holmes Trujillo.

  Entre los cargos que ha desempeñado Oscar Iván Zuluaga se pueden destacar: concejal y alcalde del municipio de Pensilvania (Caldas), senador de la República, y ministro de hacienda y crédito público.

[30] http://es.wikipedia.org/wiki/Juan_Manuel_Santos

[31] http://es.wikipedia.org/wiki/%C3%93scar_Iv%C3%A1n_Zuluaga

Capítulo 3

Desarrollo de la Propuesta

En este capítulo se describe el tipo de investigación, las técnicas e instrumentos de recolección de datos, así como todo el proceso llevado a cabo para capturar, analizar y procesar los datos. El análisis de datos utilizando en este trabajo es de tipo cuantitativo.

3.1 Tipo de Investigación

Debido a que este trabajo trata acerca de situaciones asociadas a un tema de reciente aparición y poco explotado en Colombia (opinion mining), la investigación llevada a cabo es de tipo exploratoria. "Los estudios exploratorios nos sirven para aumentar el grado de familiaridad con fenómenos relativamente desconocidos. Por lo general determinan tendencias e identifican relaciones potenciales entre variables".[32]

3.2 Definición de Población y Muestra

La población estuvo conformada por todos los colombianos que expresaron comentarios relacionados con las elecciones presidenciales de 2014, a través de Twitter, elespectador.com y semana.com, durante los días cercanos a ese proceso.

[32] http://metodologadelainvestigacinsiis.blogspot.com/2011/10/tipos-de-investigacion-exploratoria.html

La muestra del estudio la conformaron los internautas colombianos de quienes se capturaron comentarios expresados en Twitter, elespectador.com y semana.com, en el período comprendido entre el 9 de mayo y el 15 de junio de 2014, con relación a las elecciones presidenciales.

3.3 Técnicas de Recolección de Datos

Debido a que los datos necesarios para esta investigación se encontraban ya disponibles a través de los diferentes sitios web 2.0 de donde se realizaron capturas de información, dichos datos se recolectaron mediante herramientas de software diseñadas para este estudio específico.

Las encuestas que conformaron la información pública disponible en Colombia hasta antes de la segunda vuelta, fueron tomadas de diferentes medios de comunicación electrónica, y principalmente fueron tenidas en cuenta las encuestas realizadas por las prestigiosas firmas: Gallup Colombia, y Datexco Company.

Es necesario aclarar que los datos recolectados para esta investigación se encuentran públicamente disponibles a través de la Web, por lo tanto no se realizó manejo de información confidencial que pueda dar lugar a acciones legales por parte de los ciudadanos de quienes se capturó información expresada y que guarde relación con las elecciones presidenciales de 2014. En Colombia la protección de datos personales se reglamenta de acuerdo a la Ley 1581 de 2012 (Habeas Data), en la cual se define como base de datos al “conjunto organizado de datos personales que sea objeto de tratamiento”.

En este trabajo no se realizó recolección de datos personales tales como nombres, apellidos, identificación, teléfonos, entre otros, que den cuenta de información vital de los ciudadanos; sólo se capturaron los comentarios expresados por los ciudadanos,

el nombre de usuario, y la fecha de publicación de cada comentario. Por todo lo anterior se reitera que no se abusó de la información personal de los ciudadanos debido a que los pocos datos tomados fueron expresados en sitios web 2.0 de dominio público.

3.4 Procedimiento Experimental

En esta sección se describirán todas las acciones que se llevaron a cabo para poder efectuar la captura, minería y cálculo de la polaridad de las opiniones relacionadas con las elecciones presidenciales.

3.4.1 Instrumentos de captura de opiniones

Para la captura de comentarios de Twitter en esta investigación, se desarrolló una aplicación en PHP con conexión a base de datos MySQL, en la cual se almacenan los tweets expresados por los usuarios.

La base de datos diseñada cuenta con una estructura en la cual se pueden almacenar los siguientes datos:

- ID del tweet
- Usuario de Twitter (e.g., @colombiana123)
- Comentario (Contenido del tweet)
- Fecha y hora del tweet

Para lograr que la aplicación desarrollada extrajera los tweets relacionados con las elecciones presidenciales y sus candidatos, se seleccionaron cuidadosamente siete (7) hashtag, uno para candidato presidencial y dos más para el proceso de elecciones en general, buscando que los resultados del sentiment analysis aplicado a

los tweets no fuera parcializado. Por ejemplo, no se obtiene el mismo resultado al usar el hashtag *#VotoPorZuluaga* que al usar el hashtag *#OscarIvanZuluaga*, pues al realizar la prueba, con el primer hashtag se obtuvieron en su mayoría comentarios positivos para ese candidato, emitidos por sus simpatizantes. En cambio cuando se utilizó el segundo hashtag los comentarios obtenidos fueron positivos, neutrales y negativos, permitiendo una verdadera objetividad en esta investigación.

El algoritmo que se encarga de la captura y almacenamiento de tweets en la base de datos, se describe en las siguientes líneas:

```
1. <?php
2. // Llamado a la librería 'twitteroauth', requerida para este ejercicio:
3. require_once("lib/twitteroauth.php");
4. // Valor almacenado del hashtag a partir del cual capturar tweets:
5. $search = $_POST['txt_hashtag'];
6. // Especificar número máximo de tweets a capturar (por ejemplo 1000):
7. $num_tweets = 1000;
8. // Credenciales de acceso a la aplicación desarrollada en Twitter:
9. $consumerkey = "u6ETmM3K1WPs73DmB8x***";
10. $consumersecret = "zO8yJ5WW9b86ZB0xQTGQQl5vuBwkiwyT5yivHaBv***";
11. $accesstoken = "283853004-ByMRVlQfizb8IssMhNIQtOOs697k9K5flusdY***";
12. $accesstokensecret = "bxV2lrhkszoNE5X5UeCQMEk0FSpSkEVN1aaVBRzf6D***";
13. // Conexión a la aplicación de Twitter, con las credenciales suministradas:
14. function getConnectionWithAccessToken($cons_key, $cons_secret, $oauth_token, $oauth_token_secret) {
15. $connection = new TwitterOAuth($cons_key, $cons_secret, $oauth_token, $oauth_token_secret);
16. return $connection;
17. }
18. $connection = getConnectionWithAccessToken($consumerkey, $consumersecret, $accesstoken, $accesstokensecret);
19. // Reemplazar el signo caracter (o almohadilla) por su equivalente en ASCII (23):
20. $search = str_replace("#", "%23", $search);
21. // Establecer conexión para obtener los tweets relacionados con el hashtag indicado:
22. $tweets=$connection->get
("https://api.twitter.com/1.1/search/tweets.json?q=".$search."&count=".$num_tweets);
23. // Almacenar el nombre de la tabla en la que se deben guardar los tweets, según el hashtag seleccionado:
24. switch ($_POST['txt_hashtag']) {
25. case '#ClaraLopez':
26. $tabla = 'tweets_lopez';
27. break;
28. case '#EnriquePeñalosa':
```

```
29. $tabla = 'tweets_penalosa';
30. break;
31. case '#MartaLuciaRamirez':
32. $tabla = 'tweets_ramirez';
33. break;
34. case '#JuanManuelSantos':
35. $tabla = 'tweets_santos';
36. break;
37. case '#OscarIvanZuluaga':
38. $tabla = 'tweets_zuluaga';
39. break;
40. case '#Elecciones2014':
41. $tabla = 'tweets_elecciones2014';
42. break;
43. case '#SegundaVuelta':
44. $tabla = 'tweets_2a_vuelta';
45. break;
46. }
47. // Establecer conexión con MySQL (Usuario root y el respectivo password):
48. $conexion = mysql_connect('localhost', 'root', '12345');
49. if (!$conexion) {
50. die('No es posible establecer la conexión: '.mysql_error());
51. }
52. //
53. foreach ($tweets['statuses'] as $key) {
54. $id_tweet = $key['id_str'];
55. echo '<b>'."ID del tweet: ".'</b>'.$id_tweet.'<br/>';
56. $usuario = $key['user']['screen_name'];
57. echo '<b>'."Usuario: @".'</b>'.$usuario.'<br/>';
58. $comentario = $key['text'];
59. echo '<b>'."Tweet: ".'</b>'.$comentario.'<br/>';
60. $fecha_hora = date('d-m-Y H:i:s', strtotime($key['created_at']));
61. echo '<b>'."Fecha y hora del tweet: ".'</b>'.$fecha_hora.'<br/>';
62. // Establecer codificación en UTF8 para que se reconozcan caracteres como las tildes, la Ñ, y otros.
63. mysql_set_charset('utf8');
64. // Indicar el nombre de la base de datos en la que se guardarán los tweets capturados:
65. $db = 'db_elecciones';
66. mysql_select_db($db, $conexion);
67. // Grabar registros en la base de datos:
68. $sql = "INSERT INTO ".$tabla."(id_tweet, usuario, comentario, fecha_hora) VALUES('$id_tweet', '$usuario', '$comentario', '$fecha_hora')";
69. $result = mysql_query($sql, $conexion);
70. }
71. // Cerrar conexión MySQL:
72. mysql_close($conexion);
73. ?>
```

Algoritmo 1. Script para la captura y almacenamiento de tweets en MySQL

El funcionamiento del algoritmo 1 se puede apreciar al acceder al primer módulo de la aplicación web, denominado 'Herramienta de captura de tweets acerca de las elecciones presidenciales', el cual se encuentra visible desde el menú principal (índex) de la aplicación, tal como se puede observar en la figura 4:

Figura 4. Aplicación web desarrollada para la captura de tweets y análisis de resultados obtenidos mediante SentiStrength

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. A continuación se muestran los hashtag seleccionados para capturar los tweets, ordenados alfabéticamente de acuerdo al primer apellido de los candidatos a la presidencia de Colombia en el año 2014:

- *#ClaraLopez*: hashtag 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 la captura de 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*: para la captura de tweets relacionados con Juan Manuel Santos, candidato por el partido social de unidad nacional.
- *#OscarIvanZuluaga*: captura de tweets relacionados con Oscar Iván Zuluaga, candidato por el partido centro democrático.
- *#Elecciones2014*: hashtag muy utilizado por los colombianos para referirse a las elecciones presidenciales en general, especialmente a partir de la segunda vuelta.
- *#SegundaVuelta*: para la captura de tweets expresados por los colombianos sobre el proceso de elecciones presidenciales a partir de la segunda vuelta.

El resultado obtenido al utilizar los hashtag en mayúsculas, minúsculas, o combinaciones de éstas, es el mismo. Por ejemplo, los tweets obtenidos tras usar el hashtag *#MartaLuciaRamirez* son los mismos que los obtenidos al usar el hashtag *#martaluciaramirez*.

Para poder acceder a los tweets relacionados con los hashtag seleccionados para este estudio fue necesario acceder a la API de Twitter, lo cual se logró por medio de la creación de una aplicación de Twitter, cuyas credenciales de autenticación (consumer key, consumer secret, access token, y access token secret) son utilizadas por la aplicación desarrollada en PHP para lograr extraer los tweets.

La captura de comentarios expresados en elespectador.com y en semana.com se realizó utilizando una araña Web (crawler) llamada Win Web Crawler, mediante la 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 a la presidencia, y en cuyas noticias se expresaron comentarios por parte de los colombianos, en relación con su posición (polaridad) con un(a) candidato(a) en particular.

Como hasta la fecha el programa escogido para realizar la minería de opiniones (SentiStrength v2.2) solamente soporta la importación de ficheros planos (.txt), se ofrece por medio del módulo de captura de tweets de la aplicación web desarrollada una opción para exportar los tweets capturados (únicamente el campo “comentario”) a un fichero plano, para que el contenido de dicho fichero pueda ser leído posteriormente y sin problemas por SentiStrength.

3.4.2 Clasificación de comentarios

Para realizar la clasificación de los comentarios recolectados con relación a las elecciones presidenciales se utilizó el programa SentiStrength[33], el cual es una herramienta desarrollada para realizar sentiment analysis (opinion mining) mediante la estimación de la *fuerza* de sentimientos positivos y negativos detectados en textos.

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

SentiStrength puede ser descargado desde el sitio web del proyecto (http://sentistrength.wlv.ac.uk/download.html). Esta herramienta 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).

Se tiene entonces por ejemplo que a la palabra *ahogar* se le ha asignado un score de -4 (por tratarse de una palabra bastante negativa); mientras que a la palabra *felicidad* se le ha asignado un score de 5 (extremadamente positiva).

SentiStrength tiene varias abreviaturas registradas en su idioma original, las cuales son muy comunes en Twitter, por ejemplo: *LOL* (Laughing Out Loud). En este trabajo dichas abreviaturas se han traducido a su equivalente en español, y si no lo tienen han sido reemplazadas por otras similares del contexto colombiano.

El idioma predeterminado de SentiStrength es el inglés, lo cual implica un obstáculo inicial para realizar minería de opiniones en el contexto colombiano, pero es posible configurar SentiStrength para procesar frases en español realizando modificaciones en varios de sus ficheros de clasificación; de tal manera que en este trabajo ese obstáculo se superó mediante la traducción del contenido de dichos ficheros, los cuales constituyen la base de conocimientos sobre la cual se clasificaron los comentarios.

La traducción que en este trabajo se realizó a la base de conocimientos es un aporte muy importante al estado del arte que desde este trabajo se está haciendo al área del sentiment analysis, pues a la fecha no existía una base de conocimientos en español para realizar procesos de opinion mining sobre procesos electorales.

3.4.2.1 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, negativos o neutrales. En este caso dicho entrenamiento se realizó por medio de la traducción de palabras al español y la revisión de los puntajes asignados a cada palabra, los cuales pueden encontrarse en las escalas de emociones positivas o negativas mencionadas anteriormente.

A continuación se detallarán todos los pasos necesarios para lograr que SentiStrength reconozca palabras, frases y documentos de texto plano en español, para lo cual es necesario traducir al español el contenido de siete (7) ficheros, 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).

La traducción de esta base de conocimientos (ficheros) fue realizada por un equipo conformado por tres colaboradores: una docente de Lengua Castellana, un docente de Inglés, y una Psicóloga. La colaboración de estas personas fue fundamental debido a la gran cantidad de términos que hubo que traducir del inglés al español, específicamente al contexto colombiano.

El número de términos contenidos por la base de conocimientos original (en inglés) era de 2815, pero durante el proceso de traducción se encontraron varios casos en los que una palabra en inglés, por ejemplo *free*, en español traduce dos cosas bien diferentes: *gratis* y *libre*, lo cual obligó a ensanchar el número de términos resultantes para la base de conocimientos en español, quedando exactamente 2878 términos conformados por palabras, combinaciones de palabras, expresiones interrogativas, expresiones de negación, y emoticonos.

Al final de la traducción de los siete ficheros, la Psicóloga y el autor de esta tesis se encargaron de verificar la calidad del trabajo realizado mediante la revisión de los términos traducidos y la prueba de algunas palabras y frases en español introducidas en SentiStrength, para observar los resultados arrojados y observar la clasificación asignada (positiva, negativa o neutral) por este programa a cada grupo de palabras y frases.

Los ficheros cuyo contenido se tradujo al español fueron los siguientes:

1. '*EmotionLookupTable.txt*': este fichero contiene palabras que tienden a indicar sentimientos en el ser humano. Se tradujeron al español las palabras que contenía el fichero original, pero se conservó la puntuación (score) previamente asignada a cada palabra. Por ejemplo, una de las palabras que contiene este fichero (traducido al español) es: *amada* (en inglés: *loved*), cuyo score de 4 (bastante positiva) se mantuvo al momento de la traducción. También está incluida en el fichero la palabra *decapitar* (*behead*), cuyo score es de -5 (es una palabra muy negativa). Adicionalmente, en este fichero también fueron incluidas nuevas palabras relacionadas con el contexto colombiano y con el proceso electoral, entre las cuales se pueden mencionar: *mermelada (-3), paz (4), chuzadas (-4), carrusel (-3), verraquera (3), narcotráfico (-4), chiviado (-2), parapolítica (-4), lagarto (-3),* entre otras.

   A las palabras consideradas como neutrales en este fichero se les ha asignado un score de *-1* ó *1*, sin embargo el -1 tiende más a lo negativo que a lo positivo, mientras que el 1 tiende más a lo positivo.

   Para comprender mejor el significado de los puntajes asignados a cada palabra en este fichero, puede observarse la siguiente tabla:

Score	Descripción
-5	Sentimiento negativo muy fuerte.
-4	Sentimiento negativo fuerte.
-3	Sentimiento negativo moderado.
-2	Sentimiento negativo leve.
-1	No expresa sentimiento (neutral con tendencia a -)
1	No expresa sentimiento (neutral con tendencia a +)
2	Sentimiento positivo leve.
3	Sentimiento positivo moderado.
4	Sentimiento positivo fuerte.
5	Sentimiento positivo muy fuerte.

Tabla 1. Escala de sentimientos negativos y positivos usada por SentiStrength

2. '*QuestionWords.txt*': fichero que contiene un listado de palabras que casi siempre indican que una frase es de tipo interrogativo. En este fichero se pueden encontrar palabras tales como: "qué", "cuándo", "por qué", "cómo", entre otras. SentiStrength permite agregar nuevas palabras a este fichero, las cuales en este caso indiquen preguntas comúnmente realizadas en el contexto colombiano.

3. '*NegatingWordList.txt*': contiene una lista de palabras que casi siempre indican la negación de una palabra o una frase, por medio de los términos (traducidos al español): "*no*", "*nunca*", "*no puedo*", entre otros. Es posible agregar nuevos términos a este fichero, en este caso dependiendo de palabras y expresiones utilizadas en diferentes regiones de Colombia para expresar negación.

4. '*BoosterWordList.txt*': en este fichero se encuentra una lista de palabras que pueden aumentar o disminuir el sentimiento de una palabra que siga a continuación, por ejemplo: "*algunos*", "*muy*", "*totalmente*". A cada palabra le es asignado un puntaje de refuerzo que indica el aumento o disminución en el sentimiento dado por la palabra. La puntuación (score) que se maneja en este fichero puede ser: -2, -1, 1, 2, de tal manera que -2 indica una gran

disminución en el sentimiento, mientras que 2 indica un gran aumento en el sentimiento.

5. '*SlangLookupTable.txt*': se trata de un listado de palabras que se consideran de la jerga común. En realidad las palabras contenidas en este fichero son abreviaturas utilizadas para ahorrar tiempo y caracteres. Las siguientes líneas muestran ejemplos de las abreviaturas contenidas en este fichero:

   m/f male or female
   lulab love you like a brother

   En español estos términos (abreviaturas) se tradujeron por los que se usan diariamente en el contexto, por ejemplo para las anteriores abreviaturas sería:

   m/f masculino o femenino
   tqm te quiero mucho

6. '*IdiomLookupTable.txt*': este fichero contiene frases o expresiones comunes, las cuales denotan sentimientos diferentes del sentimiento de las palabras originales que las conforman. Puede tomarse como ejemplo la frase o expresión "*shock horror*", la cual se tradujo como "*algo sorprendido*", y el score que se le ha asignado es ligeramente negativo (-2), sin embargo las palabras que la conforman son moderada y fuertemente negativas (-3 y -4). Se pueden agregar más frases o expresiones comunes a este fichero, dependiendo del idioma y del contexto. Para este trabajo fueron agregadas muchas expresiones comunes en el ámbito político, especialmente en lo que a elecciones se refiere. Entre esas expresiones se pueden mencionar: *conflicto armado (-2), libertad de expresión (3), falsos positivos (-5), muy feliz (4), carro bomba (-5), mente abierta (3),* entre otros.

7. '*EmoticonLookupTable.txt*': en este fichero se encuentran registrados muchos íconos gestuales o emoticonos (secuencias de caracteres ASCII que expresan emociones). Cada emoticono registrado (e.g., "*:-)*", "*:'(*", "*XD*", "^_^", "*:-S*", "<3",

“*:-*”, “*:p*”) tiene un score asignado para indicar la polaridad expresada. Se pueden añadir al fichero tantos emoticonos como se desee, pero se recomienda que sean real y frecuentemente utilizados en el contexto nacional mediante los sitios web 2.0 y redes sociales de Internet en general.

Todos los ficheros de configuración que acaban de describirse se almacenan por defecto en la ruta *C:\SentStrength_Data* (en una instalación de Windows), pero SentiStrength ofrece una opción para registrar una nueva ubicación (directorio) en donde colocar los ficheros de configuración que para esta investigación fueron traducidos al español.

Esta opción se encuentra en el menú *File* de SentiStrength, y se puede observar en la figura 5:

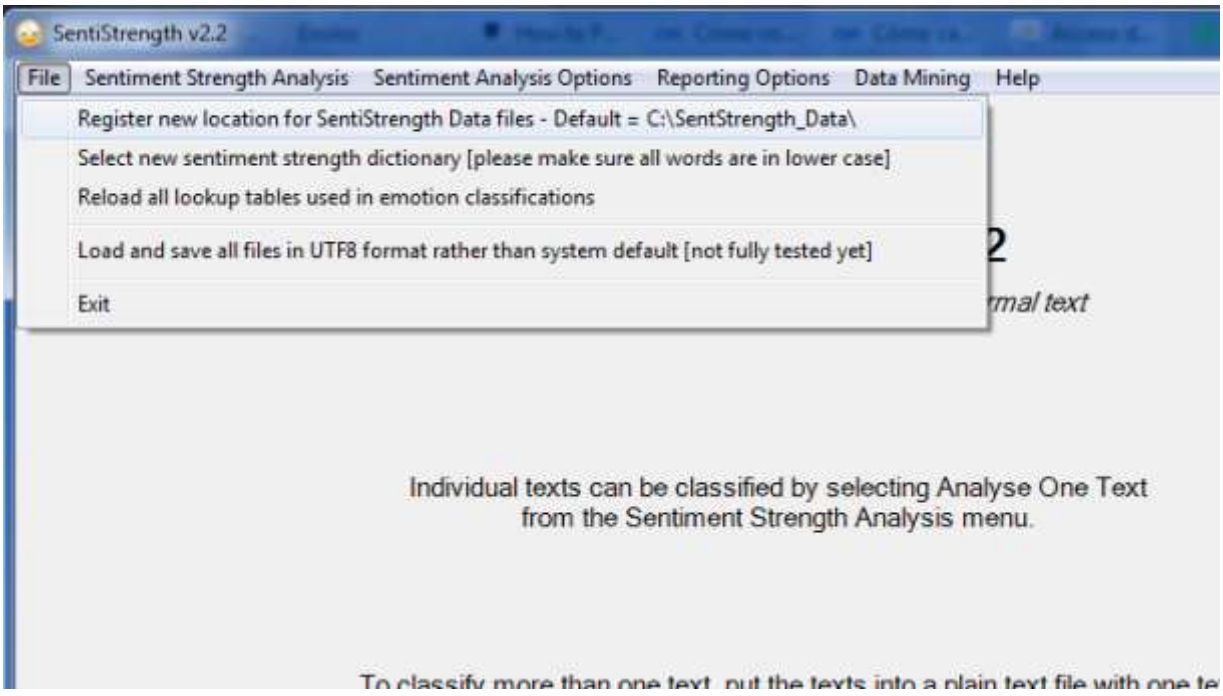


Figura 5. Registrar nueva ubicación de ficheros de configuración de SentiStrength

3.4.2.2 Realizando opinion mining con SentiStrength

Después de realizado el proceso de entrenamiento (traducción y asignación de scores), para que SentiStrength pueda detectar y procesar palabras en español, se realizó la minería de opiniones (sentiment analysis) con esta herramienta, importando los ficheros planos en donde se encuentran almacenados los comentarios recolectados acerca de las elecciones presidenciales de Colombia llevadas a cabo en el año 2014.

Estos ficheros planos se encuentran disponibles por medio de la aplicación web producto de esta investigación, en el módulo de base de datos completa de comentarios capturados. En total son siete ficheros .txt, cada uno de los cuales contiene comentarios relacionados con un candidato presidencial en particular, y con las elecciones presidenciales en general, ordenados de manera descendente: desde el más reciente hasta el más antiguo.

La importación de los ficheros que contienen los comentarios se realizó a través de la opción *Analyse ALL Texts in File*, la cual se encuentra en el menú *Sentiment Strength Analysis* de SentiStrength (ver figura 6).

Un aspecto de suma importancia es que los ficheros .txt que vayan a ser importados para su análisis en SentiStrength, deben estar codificados con el formato UTF-8, para que se reconozcan y posteriormente se muestren correctamente las tildes, la letra *Ñ* y otros caracteres propios del idioma español. Las líneas de texto de los ficheros .txt deben estar dispuestas una debajo de otra, es decir que no debe haber un comentario seguido de otro, sino debajo del mismo.

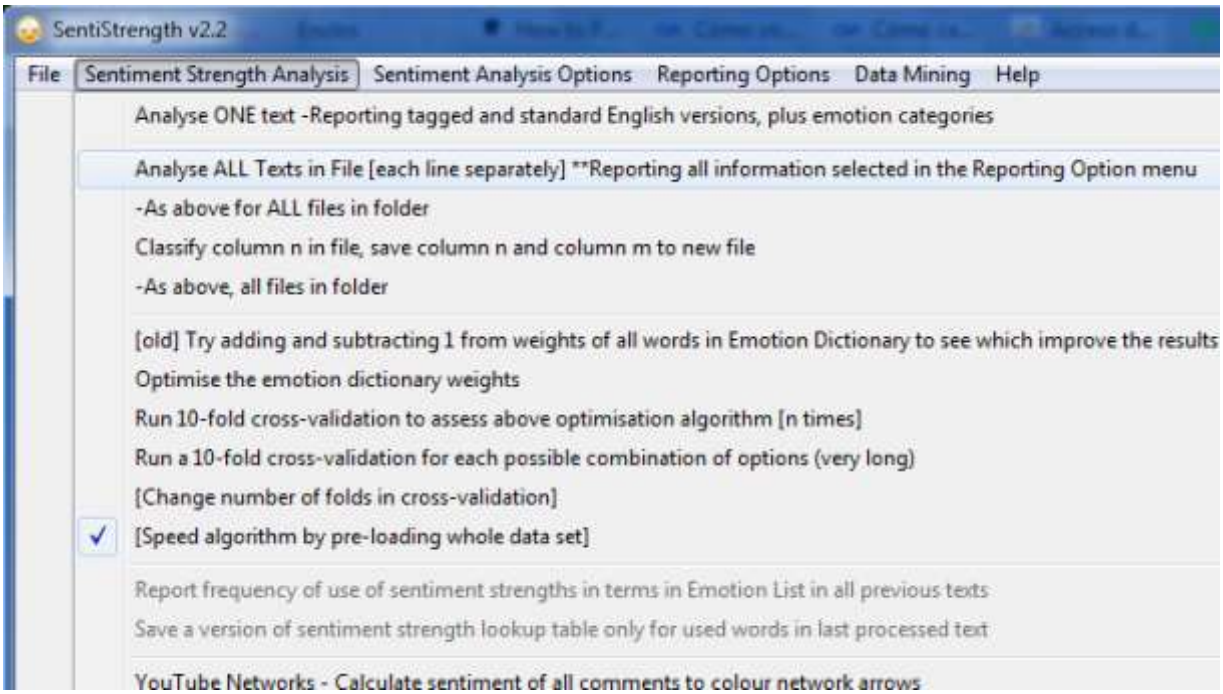


Figura 6. Opción para importar y analizar ficheros planos (.txt) en SentiStrength

También es importante mencionar que SentiStrength no tiene en cuenta la primera línea (fila) de un fichero .txt importado, debido a que el programa entiende que en esa primera línea se encuentra solamente el encabezado o la información de las columnas (campos).

Antes de comenzar con el análisis de un fichero .txt importado, SentiStrength pide que se especifique qué columna contiene el texto a procesar. En esta investigación los ficheros planos exportados desde la base de datos contienen solamente una columna: el comentario, por lo cual se debe introducir el número *1* en la caja de texto donde el programa pide que se especifique ese valor (ver figura 7) para que se pueda dar inicio a la minería de opiniones.

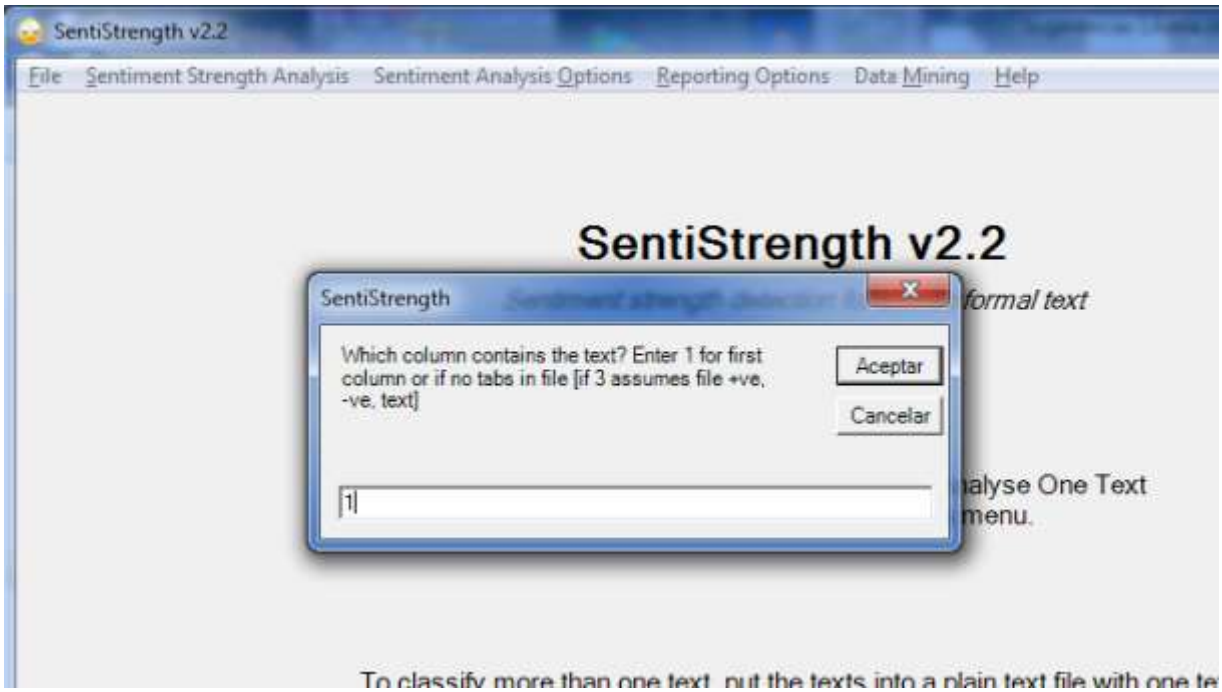


Figura 7. Establecer columna que contiene el texto a analizar en un fichero plano

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

La figura 8 representa la estructura de resultados común para todas las líneas de un fichero plano generado por SentiStrength tras haber realizado minería de opiniones:

Figura 8. Estructura de resultados de un fichero plano generado por SentiStrength

Un ejemplo de línea que contiene la estructura mostrada en la figura 6 es el siguiente:

#ClaraLopez la mejor propuesta de país, los demás, qué manada de payasos. *3* *-2*

El valor *3* corresponde a las palabras "*la mejor*", debido a que en el fichero *IdiomLookupTable.txt* esta combinación de palabras tiene registrado un score de +3 (es positiva). Por otra parte, el valor *-2* corresponde a la palabra *payasos*, pues esta palabra en el contexto político colombiano se usa para denigrar, y en el fichero *EmotionLookupTable.txt* esta palabra fue catalogada como negativa (score de -2).

Lo anterior indica que el comentario (frase) capturado y analizado es positivo, por un estrecho margen, debido a que el valor que se encuentra en la escala positiva (3) se trata de un sentimiento positivo moderado, mientras que el valor ubicado en la escala negativa (-2) se trata de un sentimiento negativo leve. Por lo tanto la polaridad positiva se impuso en esta ocasión. Sin embargo no siempre es así, existen líneas en los diferentes ficheros planos analizados que dieron como resultado polaridades negativas (e.g., *2 -5*), y otras líneas de texto en las que SentiStrength arrojó resultados neutrales (e.g., *3 -3*).

Cuando en una frase existe más de una palabra o término positivo, y/o más de uno negativo, SentiStrength toma el score más significativo de cada uno de ellos para asignar la puntuación al final de la línea que contiene la frase en el fichero plano.

Las palabras cortas que indican conjunciones, pronombres, entre otros, generalmente tienen un score de *1*, pues se les cataloga como neutrales y no inciden de gran forma en la puntuación total que se le asigna a cada frase o comentario expresado por los ciudadanos.

3.4.2.2.1 Cálculo de la polaridad global de un fichero generado por SentiStrength después de haber realizado 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 (ubicados en la columna 3), y en una segunda variable acumular la sumatoria de los scores negativos (ubicados en la columna 4), para comparar si esas dos sumatorias son iguales o determinar cuál de las dos es mayor.

Como el objetivo es medir la fuerza global de las polaridades seleccionando la de mayor valor (scores acumulados), se tiene un inconveniente inicial porque los valores de la columna 4 están precedidos por el signo menos (-). Esto se solucionó agregando una instrucción en el script que multiplica por -1 la sumatoria obtenida de scores negativos, con lo cual se convierte ese valor a positivo.

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 (e.g., 5000) es igual al valor acumulado en la segunda variable, por lo cual se obtendría una clasificación global de ese fichero como "neutral". Para el tema de las elecciones presidenciales, esta clasificación indicaría una posición neutral de los colombianos frente al candidato en particular del cual se analizaron los comentarios recolectados (fichero .txt).

b) El valor acumulado en la primera variable (e.g., 4500) es mayor que el valor acumulado en la segunda variable (e.g., 3950), por lo cual se obtendría una clasificación global de ese fichero como polaridad "positiva". Esta clasificación indicaría una posición positiva, es decir muy favorable, de los colombianos frente al candidato que se esté analizando.

c) El valor acumulado en la primera variable (e.g., 4830) es menor que el valor acumulado en la segunda variable (e.g., 5700), obteniendo entonces una clasificación global de ese fichero como polaridad "negativa". Esto indicaría una posición negativa, o sea poco favorable, de los colombianos frente a ese candidato.

El algoritmo que realiza este proceso se describe de manera completa en las siguientes líneas:

```
1. <?php
2. // Fichero .txt a cargar (e.g., puntuaciones de los comentarios sobre Enrique Peñalosa)
3. $fichero=file('Resultados_Tweets_EnriquePenalosa.txt');
4. // Iniciar en cero el contador de fila y los acumuladores de puntuaciones
5. $i = 0;
6. $CantPos = 0;
7. $CantNeg = 0;
8. // Ciclo: recorrer el fichero mientras exista una fila en el mismo
```

```
9.  while ($fichero[$i] != NULL) {
10. // Almacenar en la variable $row todo el contenido de la fila
11. $row = $fichero[$i];
12. // Generar array que almacene los datos de la fila, separándolos por tabulación
13. $sql = explode("\t", $row);
14. // Incrementar contador de fila en 1
15. $i++;
16. $CantPos = $CantPos+$sql[2];
17. $CantNeg = $CantNeg + $sql[3];
18. }
19. // Convertir a positivo el valor de los scores negativos acumulados
20. $CantNeg = $CantNeg*-1;
21. // Imprimir acumulado de scores positivos y negativos
22. echo 'Acumulador de puntuaciones positivas: '.$CantPos.'<br/>';
23. echo 'Acumulador de puntuaciones negativas: '.$CantNeg.'<br/><hr/>';
24. // Imprimir polaridad global del fichero analizado
25. if ($CantPos == $CantNeg) {
26. echo 'El fichero tiene polaridad global NEUTRAL.'; }
27. else
28. if ($CantPos > $CantNeg) {
29. echo 'El fichero tiene polaridad global POSITIVA.'; }
30. else
31. if ($CantPos < $CantNeg) {
32. echo 'El fichero tiene polaridad global NEGATIVA.'; }
33. ?>
```

Algoritmo 2. Script para calcular la polaridad global de un fichero plano generado por SentiStrength

En la tabla 2 se puede observar un ejemplo que explica detalladamente cómo se puede determinar la polaridad global de un fichero plano generado por SentiStrength, tras haber realizado opinion mining. En dicha tabla se muestran dos columnas iniciales que indican los scores positivos y negativos asignados a determinados comentarios contenidos en un fichero plano. En la tercera columna de la tabla se va detallando la polaridad global (acumulada) del fichero tras comparar los scores positivo y negativo en cada comentario.

En este ejemplo los scores iniciales son: *1* y *-2*, por lo que debe recordarse que *1* corresponde a un sentimiento neutral, mientras que *-2* corresponde a un sentimiento negativo leve. Por tal razón comienza imponiéndose la polaridad negativa: *-1* (-2 menos 1 = -1).

Score Pos.	Score Neg.	Detalle de la polaridad global (acumulada)
1	-2	*Comienza imponiéndose la polaridad Negativa.* Global: -1 (-2 menos 1 = -1)
1	-2	Global: -2 (-2 menos 1 = -1 más -1 acumulado)
3	-5	Global: -4 (-5 menos 3 = -2 más -2 acumulado)
1	-1	Global: -4 (-1 menos 1 = 0 más -4 acumulado)
2	-3	Global: -5 (-3 menos 2 = -1 más -4 acumulado)
1	-1	Global: -5 (-1 menos 1 = 0 más -5 acumulado)
2	-1	Global: -4 (-1 menos 2 = 1 más -5 acumulado)
1	-2	Global: -5 (-2 menos 1 = -1 más -4 acumulado)
1	-1	Global: -5 (-1 menos 1 = 0 más -5 acumulado)
2	-1	Global: -4 (-1 menos 2 = 1 más -5 acumulado)
1	-3	Global: -6 (-3 menos 1 = -2 más -4 acumulado)
4	-1	Global: -3 (-1 menos 4 = 3 más -6 acumulado)
2	-1	Global: -2 (-1 menos 2 = 1 más -3 acumulado)
1	-1	Global: -2 (-1 menos 1 = 0 más -2 acumulado)
5	-2	Global: +1 (-2 menos 5 = 3 más -2 acumulado)
3	-1	Global: +3 (-1 menos 3 = 2 más 1 acumulado) *Termina imponiéndose la polaridad Positiva.*
Total: |31|	**Total: |28|**	**Gana el sentimiento Positivo |31| por una diferencia de +3 sobre el sentimiento Negativo |28|**

Tabla 2. Determinación paso a paso de la polaridad global de un fichero plano con varios comentarios a los que SentiStrength les ha asignado puntuaciones

Nótese que en la última fila de la tabla 2 los totales se han expresado con valor absoluto (sin tener en cuenta su signo) porque lo que interesa es acumular las puntuaciones en cada columna para luego comparar cuál polaridad tiene mayor fuerza.

El bloque conformado por las líneas 25 – 32 del algoritmo 2, muestra que al final del recorrido de todas las líneas de un fichero plano generado por SentiStrength, se imprime un mensaje que indica si el fichero tiene polaridad global neutral, positiva o negativa.

3.5 Confiabilidad del Procedimiento

La recopilación de texto para realizar opinion mining estuvo sujeta a la publicación de comentarios por parte de ciudadanos colombianos durante el período comprendido entre el 9 de mayo y el 15 de junio de 2014, con relación a las elecciones presidenciales en Colombia. Por tal motivo solamente se seleccionaron tres fuentes de información: Twitter, elespectador.com, y semana.com, porque estos medios se encuentran entre los más preferidos por los colombianos para expresar sus sentimientos, emociones y opiniones constantemente.

La aplicación desarrollada en PHP, en conjunto con la base de datos y los ficheros de texto plano que se pueden exportar, permitieron organizar la información a ser tratada mediante el programa SentiStrength, el cual es una solución probada en muchos entornos a nivel internacional y con soporte para varios idiomas, incluido el español.

Los hashtag seleccionados para capturar los comentarios de Twitter fueron los más utilizados por los colombianos para referirse al proceso de elecciones presidenciales en general, y a los candidatos presidenciales de manera individual, lo cual se probó por medio de la comparación de la cantidad de tweets capturados al utilizar dos o más hashtag similares para un candidato específico. Por ejemplo, el día 16-05-2014 a las 13:44 horas, se encontró la siguiente cantidad de tweets para estos hashtag asociados con algunos de los candidatos presidenciales:

- Oscar Iván Zuluaga: 51 tweets relacionados con el hashtag *#OIZuluaga*, y 80 tweets relacionados con el hashtag *#OscarIvanZuluaga*
- Clara López: 43 tweets relacionados con el hashtag *#ClaraLopez*, y 11 tweets relacionados con el hashtag *#ClaraLopezObregon*
- Enrique Peñalosa: 15 tweets relacionados con el hashtag *#Peñalosa*, y 38 tweets relacionados con el hashtag *#EnriquePeñalosa*

La recolección de comentarios desde los sitios web de noticias elespectador.com y semana.com se inició por medio de la descarga y almacenamiento de varias páginas web que contienen noticias relacionadas con los candidatos a la presidencia de Colombia en el año 2014. Este proceso se efectuó gracias al software de araña web Win Web Crawler. Posteriormente los comentarios contenidos en las páginas descargadas fueron copiados y pegados al final de los ficheros planos en los que se almacenaron los tweets capturados mediante la aplicación web. La selección de dichos comentarios se realizó de manera aleatoria, sin detenerse a leer los comentarios completos y por ende conocer su polaridad, buscando garantizar la imparcialidad en la investigación y no parcializar la información seleccionada a favor o en contra de algún(a) candidato(a).

La traducción al español de las palabras contenidas en los siete ficheros de configuración de SentiStrength se realizó con mucha cautela, cuidando mantener la puntuación asignada a cada palabra con el fin de obtener resultados concordantes con la realidad, y por ende lograr que SentiStrength efectuara una correcta asignación de scores positivos y negativos al final de cada línea con base en la fuerza total de los sentimientos que conforman las frases (comentarios).

El margen de error que puede existir en la asignación de los scores positivos y negativos al final de cada línea, está determinado por la cantidad de palabras mal escritas por parte de los cibernautas en algunos comentarios (por ejemplo debido a la omisión de caracteres). Esta dificultad se supera fácilmente porque SentiStrength a las palabras mal escritas o no reconocidas, les asigna un score neutral, buscando no afectar la polaridad global del texto analizado.

El cálculo de la polaridad global de los ficheros generados por SentiStrength (tras haber realizado opinion mining) estuvo a cargo del script (algoritmo 2) desarrollado en PHP para acumular la sumatoria de las puntuaciones positivas y negativas asignadas por SentiStrength. Este script fue probado varias veces con ficheros

planos más pequeños (con menos líneas) que los originales, con el objetivo de comprobar si los resultados arrojados por el algoritmo eran correctos, es decir, verificando los scores positivos y negativos acumulados, así como el juicio final emitido por el algoritmo con respecto a la polaridad global (positiva, neutral, o negativa) de cada fichero plano.

3.6 Plan de Presentación de Resultados

Los resultados obtenidos durante toda esta fase de opinion mining y asignación de polaridad global, se analizan con detalle en el próximo capítulo, mediante la construcción de un análisis general que permite comprender de mejor manera los sentimientos, emociones, opiniones y tendencias de los ciudadanos colombianos con relación al proceso de elecciones presidenciales que se llevó a cabo los días 25 de mayo y 15 de junio de 2014.

Además, en el próximo capítulo se establece una relación entre dicho análisis general y entre información pública correspondiente a resultados de encuestas realizadas por las firmas Gallup Colombia y Datexco Company, las cuales estuvieron muy pendientes de las opiniones de los ciudadanos colombianos acerca del proceso de elecciones presidenciales en el año 2014.

Capítulo 4

Resultados

Todos los resultados obtenidos al realizar este trabajo investigativo son presentados en este capítulo. Se presentan los resultados que se obtuvieron al utilizar los instrumentos diseñados para la recopilación de la información: instrumento de recolección de tweets, e instrumentos de recolección de comentarios de elespectador.com y semana.com, todos relacionados con el tema de las elecciones presidenciales de 2014. 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., y Datexco Company S.A.

4.1 Descripción de los resultados obtenidos mediante la captura de opiniones

Por medio de la aplicación web desarrollada para la captura y almacenamiento de tweets, se logró recolectar una gran cantidad de comentarios que posteriormente fueron exportados a ficheros planos (.txt). El mecanismo de exportación utilizado por la aplicación organiza los comentarios en orden descendente (i.e., desde el más reciente hasta el más antiguo), y solamente exporta a los ficheros planos el campo ‘comentario’, para poder analizar individualmente cada comentario con SentiStrength.

Posteriormente se agregaron al final de esos ficheros planos, varios comentarios contenidos en las páginas web recolectadas mediante web crawling desde elespectador.com y semana.com; estos últimos comentarios fueron seleccionados al azar, y copiados y pegados a los ficheros planos que contienen tweets relacionados con los cinco candidatos presidenciales.

Finalizado el proceso de inclusión de todos los tipos de comentarios en los ficheros planos relacionados con los cinco candidatos presidenciales, se realizó una división del contenido de cada uno de estos ficheros, con el fin de contar con tres (3) ficheros por cada candidato, de tal manera que en el primero de ellos se almacenaron los comentarios capturados que se generaron durante la primera vuelta (del 9 al 25 de mayo), en el segundo los comentarios capturados que se generaron durante la segunda vuelta (del 26 mayo al 15 de junio), y en el tercero la totalidad de los comentarios. El objetivo de esta división consistió en observar y posteriormente analizar los cambios, con base en opiniones, que tuvieron los electores con respecto a los candidatos durante todo el transcurso de las elecciones presidenciales.

A continuación se presenta la cantidad de comentarios que fueron capturados desde Twitter por medio del uso de los hashtag seleccionados para la recopilación de opiniones relacionadas con cada candidato presidencial, iniciando el día 9 de mayo y finalizando el día 15 de junio de 2014:

Fecha de captura	#Clara Lopez	#Enrique Peñalosa	#MartaLucia Ramirez	#JuanManuel Santos	#OscarIvan Zuluaga
9 de mayo	9	1	4	31	29
10 de mayo	4	0	2	30	9
11 de mayo	3	1	0	16	1
12 de mayo	22	2	5	54	5
13 de mayo	25	4	2	41	4
14 de mayo	29	7	6	43	6
15 de mayo	73	2	10	94	5
16 de mayo	17	4	1	35	12

Fecha de captura	#Clara Lopez	#Enrique Peñalosa	#MartaLucia Ramirez	#JuanManuel Santos	#OscarIvan Zuluaga
17 de mayo	24	0	5	44	32
18 de mayo	30	3	10	110	159
19 de mayo	41	572	17	49	166
20 de mayo	102	12	12	29	59
21 de mayo	248	1478	78	21	104
22 de mayo	96	311	5	49	38
23 de mayo	470	85	46	79	270
24 de mayo	515	61	21	74	105
25 de mayo	716	147	130	215	198
26 de mayo	105	15	15	147	234
27 de mayo	25	1	12	67	37
28 de mayo	13	0	4	48	36
29 de mayo	11	1	83	18	29
30 de mayo	19	0	1	24	17
31 de mayo	8	0	2	42	9
1° de junio	3	0	0	48	5
2 de junio	1	0	0	31	7
3 de junio	1	0	1	22	5
4 de junio	23	1	3	36	45
5 de junio	30	0	2	33	21
6 de junio	34	3	4	44	55
7 de junio	13	8	1	122	38
8 de junio	9	0	0	17	8
9 de junio	6	0	0	38	29
10 de junio	18	0	0	100	69
11 de junio	8	0	0	35	17
12 de junio	6	0	0	120	21
13 de junio	1	0	0	173	18
14 de junio	0	0	0	37	12
15 de junio	11	0	9	594	78
Total:	**2769**	**2719**	**491**	**2810**	**1992**

Tabla 3. Información sobre capturas individuales de comentarios recolectados desde Twitter, para cada candidato presidencial

Como se indica en la tabla 3, el candidato Juan Manuel Santos fue quien generó mayor número de tweets en el servicio de microblogging Twitter, durante el período comprendido del 9 de mayo al 15 de junio de 2014; mientras que la candidata Marta Lucía Ramírez fue quien menos tweets generó para el mismo período.

Mediante la tabla 4 se da a conocer la totalidad de comentarios recopilados desde Twitter, elespectador.com y semana.com, 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
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 4. Totalización de comentarios recolectados desde Twitter, elespectador.com y semana.com

En total, gracias a los instrumentos de captura de opiniones, se capturaron 33570 comentarios expresados por los colombianos en elespectador.com, semana.com y Twitter. Sobre estos comentarios se realizó la minería de opiniones por medio de SentiStrength.

Como se puede observar en la tabla 4, también se capturaron tweets mediante los hashtag *#Elecciones2014* y *#SegundaVuelta*, lo cual se hizo porque durante el transcurso de la segunda vuelta presidencial 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, por lo

cual los hashtag mencionados marcaron tendencia en Twitter desde el 26 de mayo de 2014 hasta finalizadas las elecciones presidenciales.

4.2 Descripción de los resultados obtenidos mediante Opinion Mining

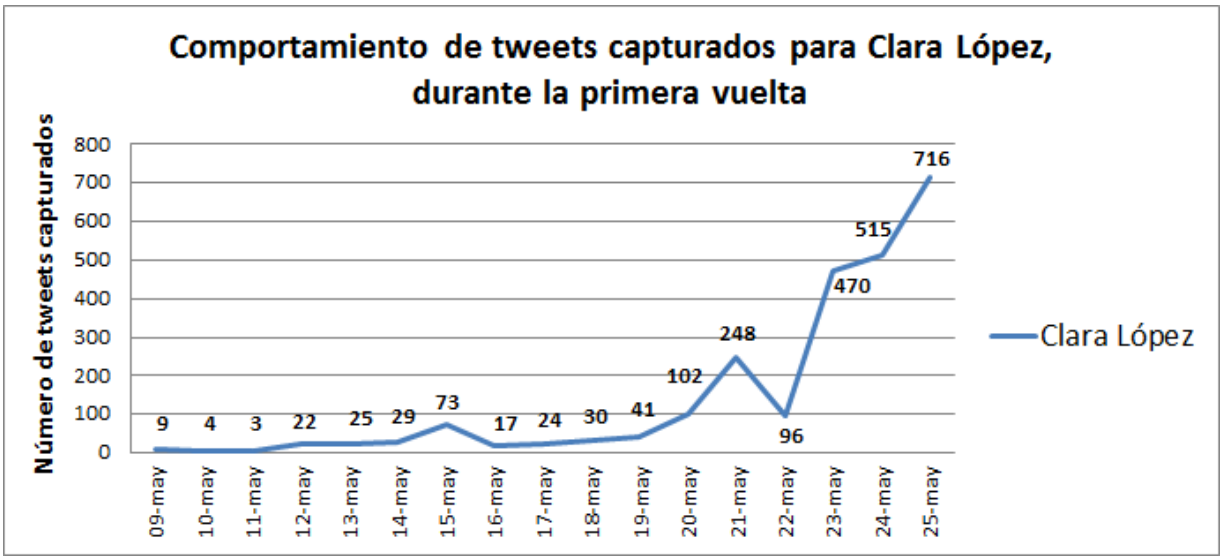
Todos los comentarios capturados fueron sometidos a opinion mining mediante su procesamiento con el software SentiStrength, para lo cual se analizaron uno a uno los ficheros planos en los que se agruparon los comentarios. Los comentarios directamente relacionados con los candidatos presidenciales, fueron separados en tres (3) ficheros, correspondientes a primera vuelta, segunda vuelta y proceso global.

A continuación se detallan todos los resultados arrojados por SentiStrength, para todos los ficheros analizados y posteriormente valorados mediante la asignación de puntuaciones. Estos resultados se muestran organizados por el número de puntuaciones positivas, negativas, y la polaridad global de cada fichero, lo cual se logró gracias al cálculo de la polaridad global que se obtiene por medio de la ejecución del algoritmo # 2, el cual se ejecutó mediante el script *'leer_fichero.php'*, como se puede apreciar en cada una de las figuras que muestran los resultados.

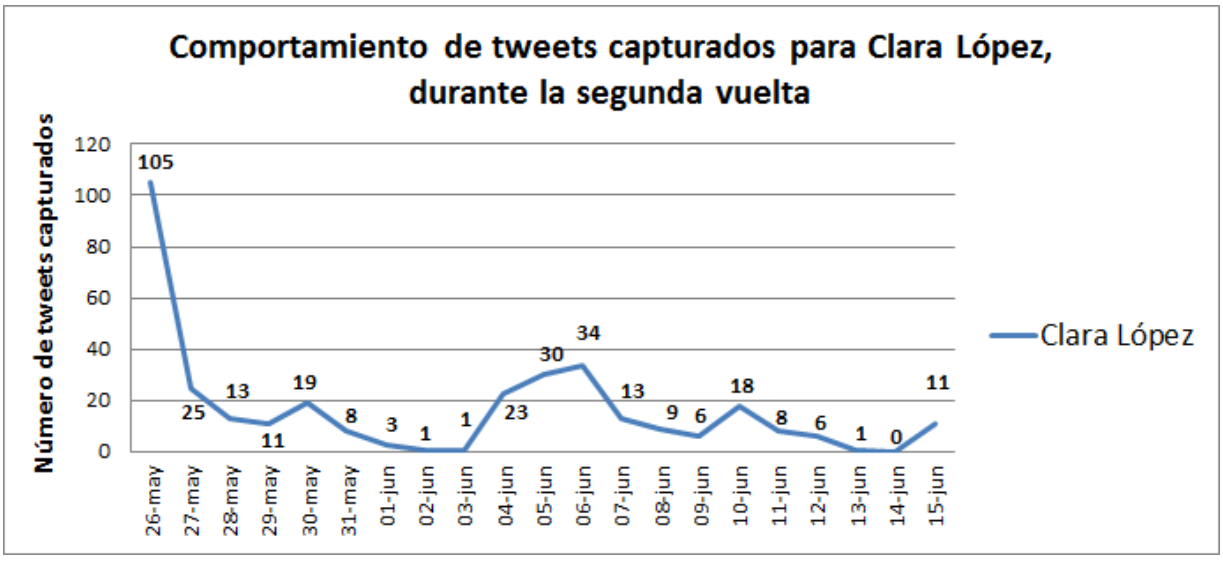
4.2.1 Resultados de opinion mining para la candidata Clara López

La fluctuación de tweets descargados por fecha para la candidata Clara López puede ser de mucha utilidad para comprender los resultados arrojados por la minería de opiniones y el cálculo de la polaridad de los comentarios.

La figura 9 describe la fluctuación de tweets capturados con relación a la candidata presidencial Clara López. La captura de tweets se realizó diariamente, iniciando el día 9 de mayo de 2014, y finalizando el día 15 de junio del mismo año.



a) Primera vuelta



b) Segunda vuelta

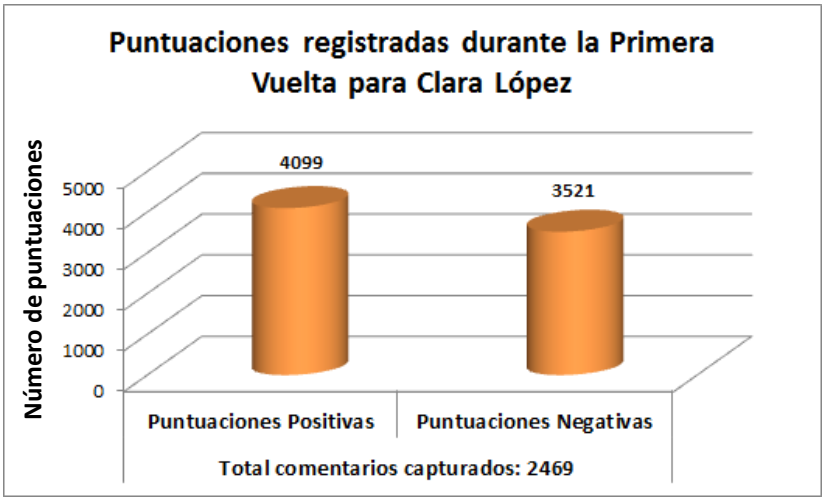
Figura 9. Comportamiento en el tiempo del número de tweets capturados para Clara López, durante primera y segunda vuelta

En la figura 9.a se puede observar un incremento del número de tweets capturados para la candidata Clara López, desde el 9 de mayo hasta el día de las votaciones en primera vuelta, alcanzando para ese día el máximo número de tweets capturados en

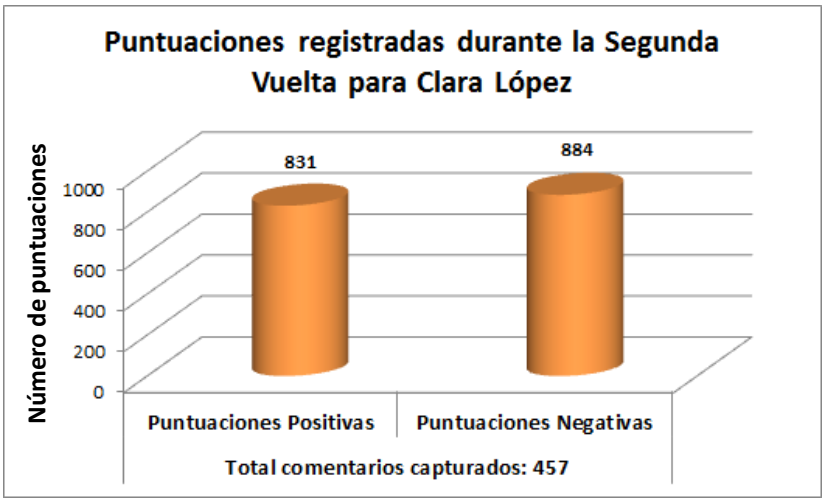
la primera vuelta. Esto indica que la popularidad y aceptación de Clara López fue tomando mucha fuerza durante el transcurso de la primera vuelta presidencial.

La figura 9.b muestra un decremento en el número de tweets capturados para Clara López después de las votaciones en primera vuelta: fue disminuyendo el número de tweets generados con relación a esta candidata durante la segunda vuelta. Esto obedece a que la candidata Clara López no pasó a la segunda vuelta presidencial, y las menciones en Twitter en torno a ella fueron disminuyendo porque se concentraron en los dos candidatos que pasaron a la segunda vuelta.

El cálculo de la polaridad de los comentarios analizados con relación a esta candidata, para la primera vuelta, determinó que hubo un acumulado de 4099 puntuaciones positivas, frente a 3521 puntuaciones negativas, por lo cual se obtuvo una polaridad *Positiva* (figura 10.a), determinada por lo que expresaron los colombianos acerca de Clara López durante la primera vuelta presidencial. Para la segunda vuelta, el cálculo de la polaridad arrojó un acumulado de 831 puntuaciones positivas, frente a 884 negativas, obteniendo esta vez una polaridad global *Negativa* (figura 10.b) durante la segunda vuelta.



a) Primera vuelta



b) Segunda vuelta

Figura 10. Comportamiento de puntuaciones registradas para la candidata Clara López, durante la primera y la segunda vuelta

Finalmente, el cálculo de la polaridad para la primera y segunda vuelta, arrojó un acumulado de 4930 puntuaciones positivas, frente a 4405 negativas, de tal manera que se obtuvo una polaridad global **Positiva** para la candidata Clara López, teniendo en cuenta un total de **2926** comentarios capturados.

La tabla 5 muestra, según los comentarios capturados, cuáles fueron los cinco términos positivos que más utilizaron los colombianos para referirse a Clara López durante las elecciones presidenciales. La palabra "Felicidad" fue la palabra más positiva utilizada, pero la palabra positiva que más se repitió fue "Paz".

Término o palabra utilizada	Puntuación (score)	Frecuencia absoluta
Felicidad	+5	2
Paz	+4	126
Inteligente	+4	10
Felicitaciones	+4	7
Esperanza	+3	20

Tabla 5. Términos positivos más utilizados para referirse a Clara López

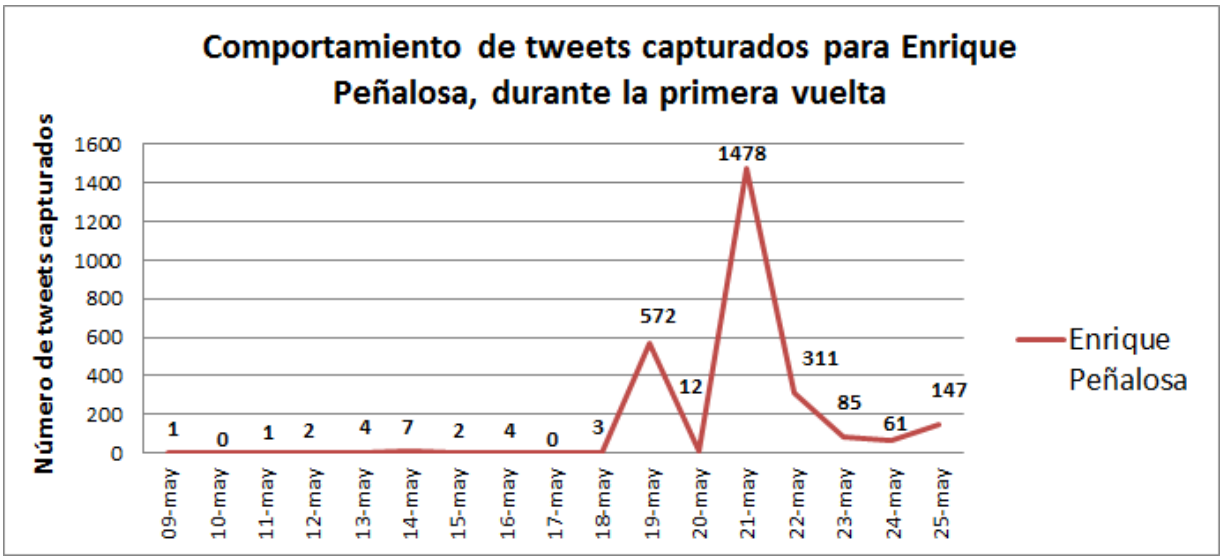
En la tabla 6 se observa cuáles fueron los cinco términos negativos que más se utilizaron para referirse a Clara López durante las elecciones. Las palabras "Terroristas" y "Asesinas" fueron las palabras más negativas utilizadas, y de estas dos la que más se repitió fue "Terroristas".

Término o palabra utilizada	Puntuación (score)	Frecuencia absoluta
Terroristas	-5	7
Asesinas	-5	2
FARCSantos	-4	2
Violencia	-4	4
Secuestro	-4	2

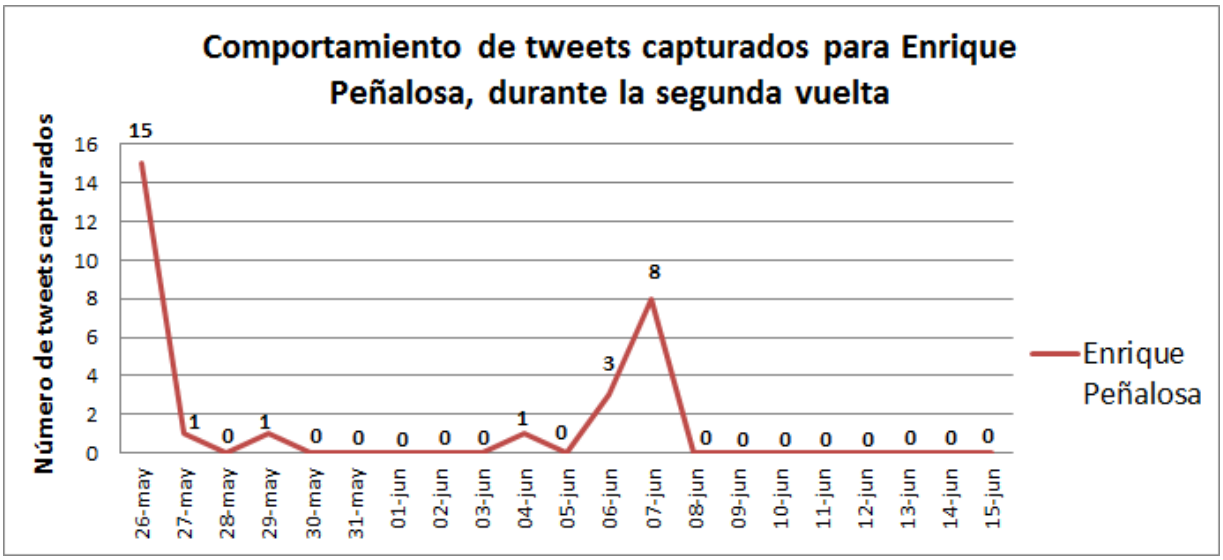
Tabla 6. Términos negativos más utilizados para referirse a Clara López

4.2.2 Resultados de opinion mining para el candidato Enrique Peñalosa

La fluctuación de tweets descargados por fecha para el candidato Enrique Peñalosa, descrita mediante la figura 11, permite comprender mejor los resultados arrojados por la minería de opiniones y el cálculo de la polaridad de los comentarios.



a) Primera vuelta



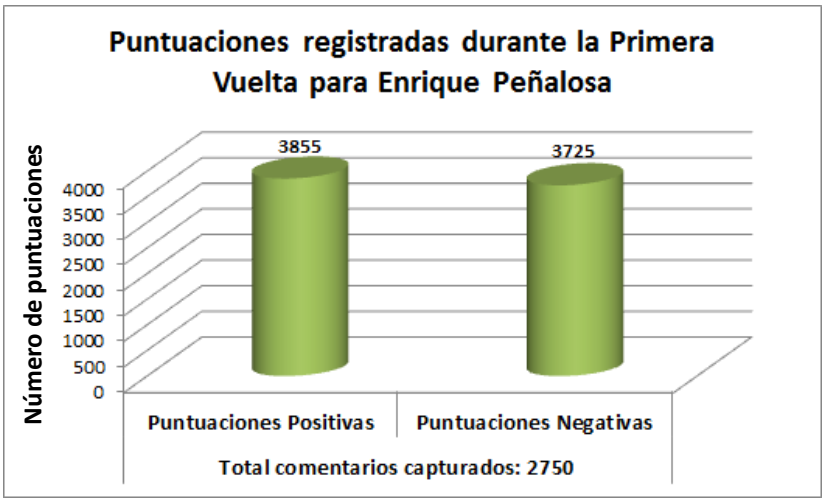
b) Segunda vuelta

Figura 11. Comportamiento en el tiempo del número de tweets capturados para Enrique Peñalosa, durante primera y segunda vuelta

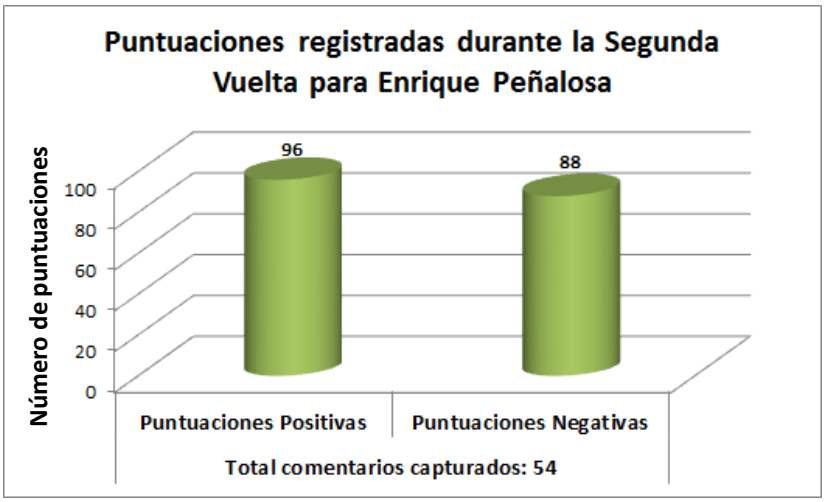
La figura 11.a indica que el número de tweets capturados para el candidato Enrique Peñalosa fue muy escaso durante los primeros días, y sólo tuvo un aumento muy significativo durante los días 19 y 21 de mayo; a partir de entonces su popularidad en Twitter fue decayendo, para tener un leve aumento el día 25 de mayo.

La figura 11.b muestra que el número de tweets capturados para Enrique Peñalosa después de las votaciones en primera vuelta fue disminuyendo a tal punto que en la mayoría de los días previos a la segunda vuelta, dicho número de tweets fue igual a cero. Esto se debe que el candidato Enrique Peñalosa tampoco pasó a la segunda vuelta presidencial, y fue desapareciendo de las menciones realizadas en Twitter por los electores debido a su escasa popularidad a partir de la segunda vuelta.

Para este candidato, con relación a la primera vuelta, el cálculo de la polaridad arrojó un acumulado de 3855 puntuaciones positivas, frente a 3725 negativas, obteniéndose una polaridad *Positiva* (figura 12.a). En cuanto a la segunda vuelta, el cálculo de la polaridad arrojó un acumulado de 96 puntuaciones positivas, frente a 88 negativas, obteniendo una polaridad *Positiva* (figura 12.b) durante la segunda vuelta.



a) Primera vuelta



b) Segunda vuelta

Figura 12. Comportamiento de puntuaciones registradas para el candidato Enrique Peñalosa, durante la primera y la segunda vuelta

La polaridad para todo el proceso (primera y segunda vuelta) arrojó un acumulado de 3951 puntuaciones positivas, frente a 3813 negativas, obteniéndose finalmente una polaridad global **Positiva** para el candidato Enrique Peñalosa, teniendo en cuenta un total de **2804** comentarios capturados.

La tabla 7 muestra cuáles fueron los cinco términos positivos que más utilizaron los colombianos para referirse a Enrique Peñalosa durante las elecciones presidenciales. Las palabras "Paz" y "Excelente" fueron las más positivas utilizadas, y la palabra "Paz" fue la que más se repitió.

Término o palabra utilizada	Puntuación (score)	Frecuencia absoluta
Paz	+4	104
Excelente	+4	3
Respeto	+3	7
Dignidad	+3	23
Ganar	+3	4

Tabla 7. Términos positivos más utilizados para referirse a Enrique Peñalosa

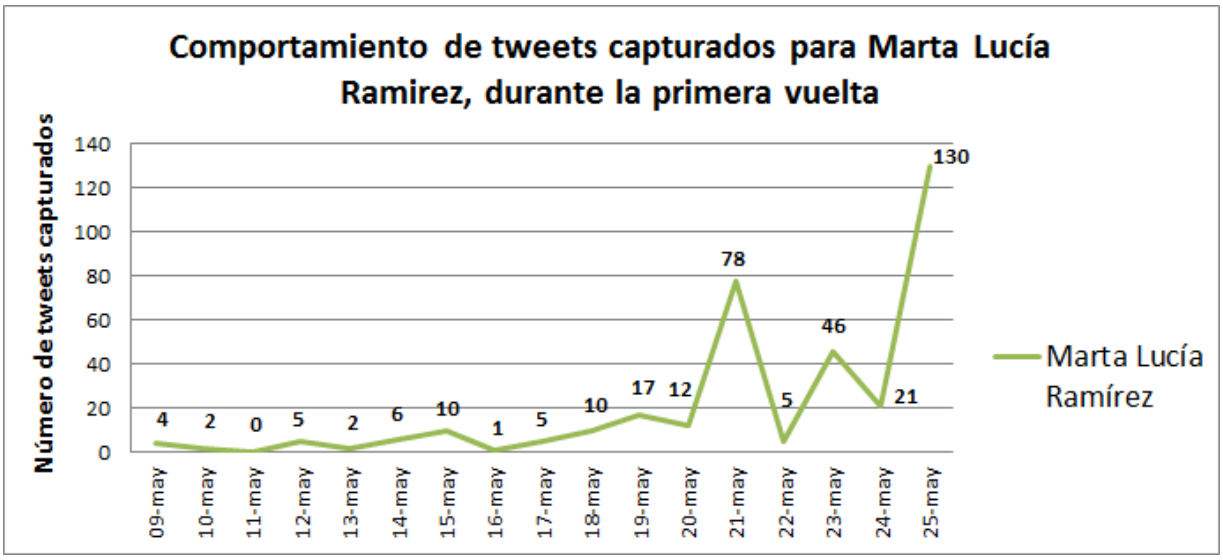
En la tabla 8 se observa cuáles fueron los cinco términos negativos que más se utilizaron para referirse a Enrique Peñalosa durante las elecciones. Las palabras "Terrorismo" y "Narcoterroristas" fueron las palabras más negativas utilizadas, pero la palabra que más se repitió fue "Guerrilla".

Término o palabra utilizada	Puntuación (score)	Frecuencia absoluta
Terrorismo	-5	1
Narcoterroristas	-5	2
FARCsantes	-4	2
Guerrilla	-4	55
Corrupción	-3	34

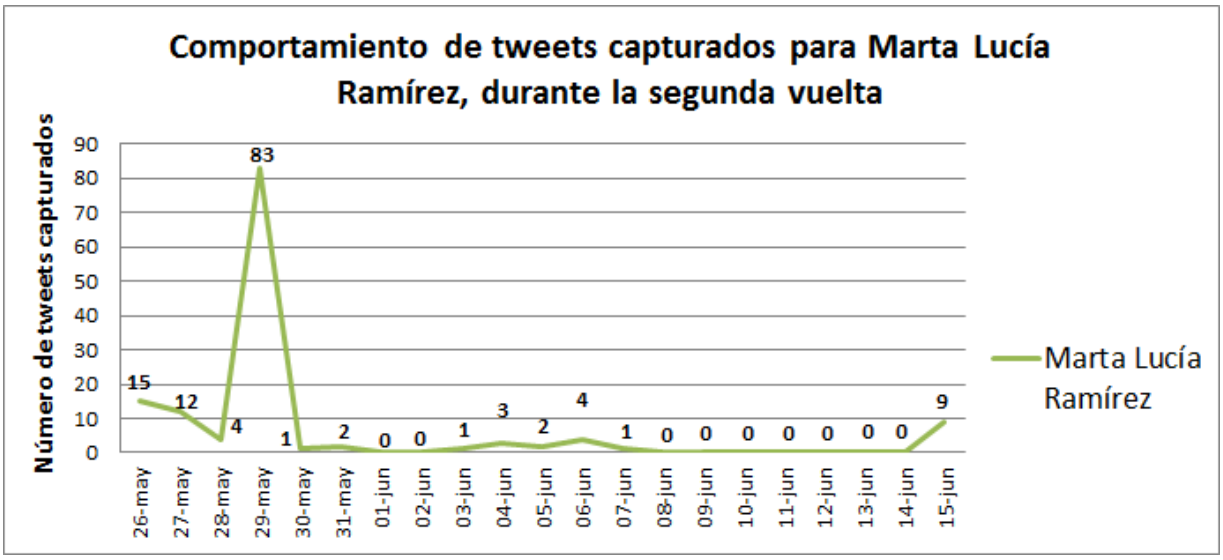
Tabla 8. Términos negativos más utilizados para referirse a Enrique Peñalosa

4.2.3 Resultados de opinion mining para la candidata Marta Lucía Ramírez

La figura 13 describe la fluctuación de tweets descargados por fecha para la candidata Marta Lucía Ramírez:



a) Primera vuelta



b) Segunda vuelta

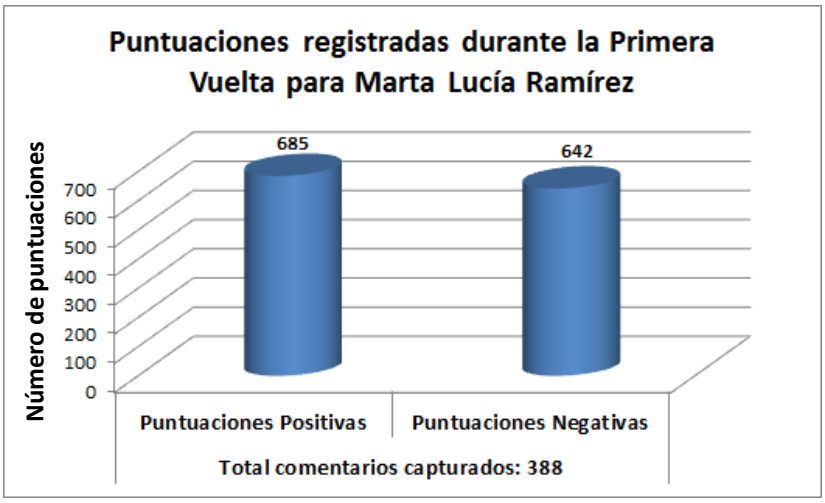
Figura 13. Comportamiento en el tiempo del número de tweets capturados para Marta Lucía Ramírez, durante primera y segunda vuelta

La figura 13.a muestra cómo la candidata Marta Lucía Ramírez generó muy pocos tweets durante los días previos a las votaciones de primera vuelta, notándose un leve incremento el día 21 de mayo, y posteriormente el día de las votaciones (25 de mayo). Lo anterior indica que esta candidata no gozaba de gran popularidad en Twitter en los primeros días de su campaña, sino que fue tomando fuerza a medida que se acercaban las votaciones de primera vuelta.

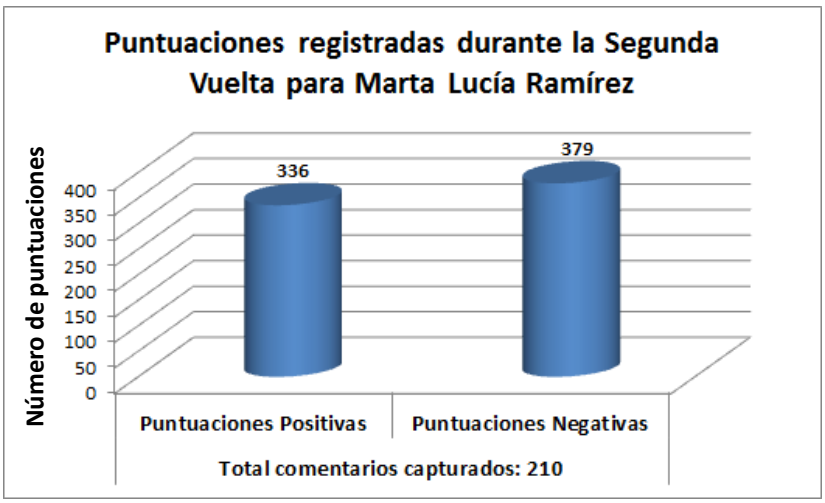
La figura 13.b indica que la popularidad en Twitter de Marta Lucía Ramírez fue decayendo constantemente a medida que se acercaban las votaciones de segunda vuelta, y a pesar de haber conformado alianza con el candidato Oscar Iván Zuluaga, a la candidata se le mencionó muy poco a partir del 29 de mayo.

Para la primera vuelta, el cálculo de la polaridad de los comentarios analizados con relación a la candidata Marta Lucía Ramírez, arrojó un acumulado de 685 puntuaciones positivas, frente a 642 negativas, obteniéndose una polaridad *Positiva* (figura 14.a) determinada por lo que expresaron los colombianos acerca de Marta Lucía Ramírez durante la primera vuelta presidencial.

Para la segunda vuelta, el cálculo de la polaridad arrojó un acumulado de 336 puntuaciones positivas, frente a 379 negativas, obteniendo de esta manera una polaridad *Negativa* (figura 14.b) con base en lo que expresaron los colombianos acerca de Marta Lucía Ramírez durante la segunda vuelta.



a) Primera vuelta



b) Segunda vuelta

Figura 14. Comportamiento de puntuaciones registradas para la candidata Marta Lucía Ramírez, durante la primera y la segunda vuelta

Finalmente, el cálculo de la polaridad para la primera y segunda vuelta, arrojó un acumulado de 1021 puntuaciones positivas, frente a 1021 negativas, obteniendo de esta manera una polaridad global **Neutral** para la candidata Marta Lucía Ramírez, teniendo en cuenta un total de **598** comentarios capturados.

En la tabla 9 se puede observar cuáles fueron los cinco términos positivos que más utilizaron los colombianos para referirse a Marta Lucía Ramírez durante las elecciones presidenciales. Las palabras “Paz” e “Inteligente” fueron las más positivas utilizadas, siendo “Paz” la palabra que más se repitió.

Término o palabra utilizada	Puntuación (score)	Frecuencia absoluta
Paz	+4	16
Inteligente	+4	4
Honesta	+3	5
Respeto	+3	3
Ganadora	+3	3

Tabla 9. Términos positivos más utilizados para referirse a Marta Lucía Ramírez

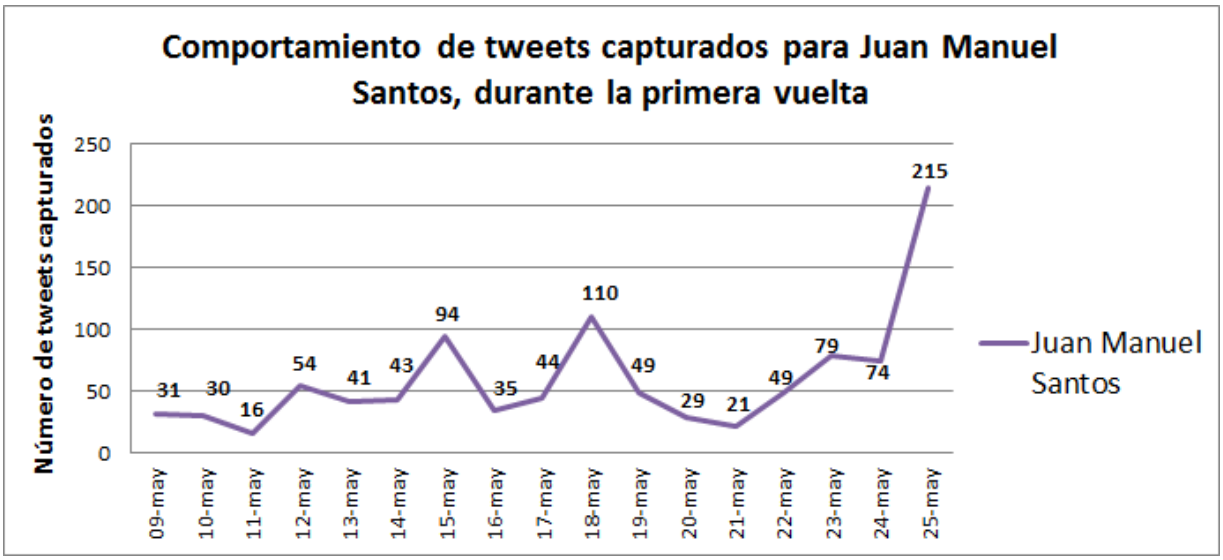
En la tabla 10 se indican los cinco términos negativos que más se utilizaron para referirse a Marta Lucía Ramírez durante las elecciones. La palabra “Terrorista” fue la más negativa utilizada, pero la palabra que más se repitió fue “Hp”.

Término o palabra utilizada	Puntuación (score)	Frecuencia absoluta
Terrorista	-5	1
Odio	-4	2
Narco	-4	4
Paraco	-4	5
Hp	-4	6

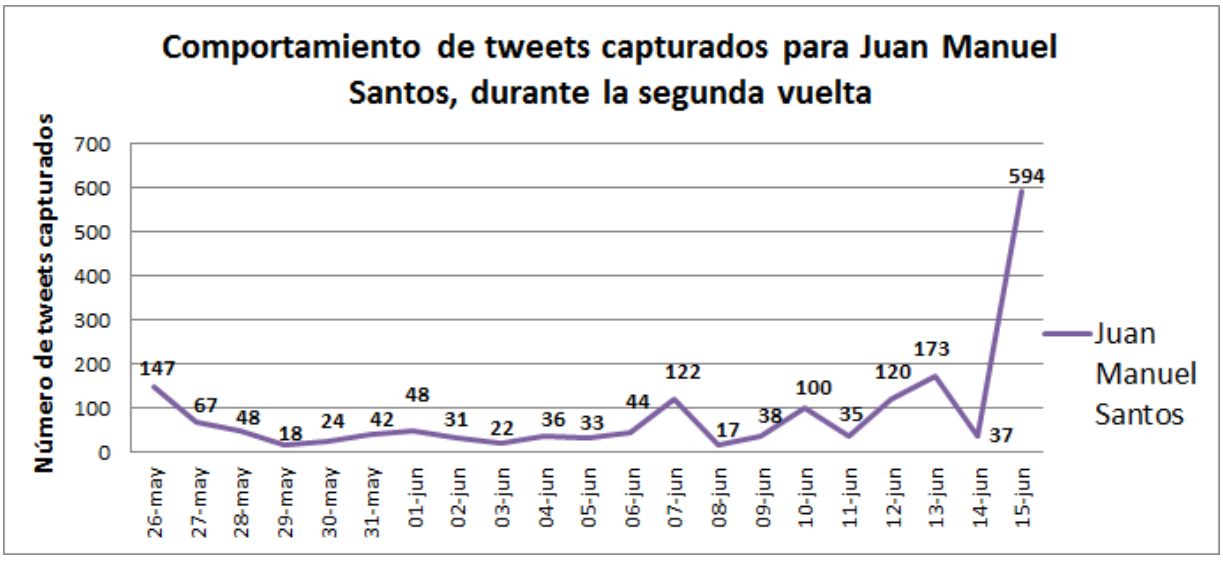
Tabla 10. Términos negativos más utilizados para referirse a Marta Lucía Ramírez

4.2.4 Resultados de opinion mining para el candidato Juan Manuel Santos

La fluctuación de tweets descargados por fecha para el candidato Juan Manuel Santos se describe mediante la figura 15, para comprender mejor los resultados arrojados por la minería de opiniones y el cálculo de la polaridad de los comentarios.



a) Primera vuelta



b) Segunda vuelta

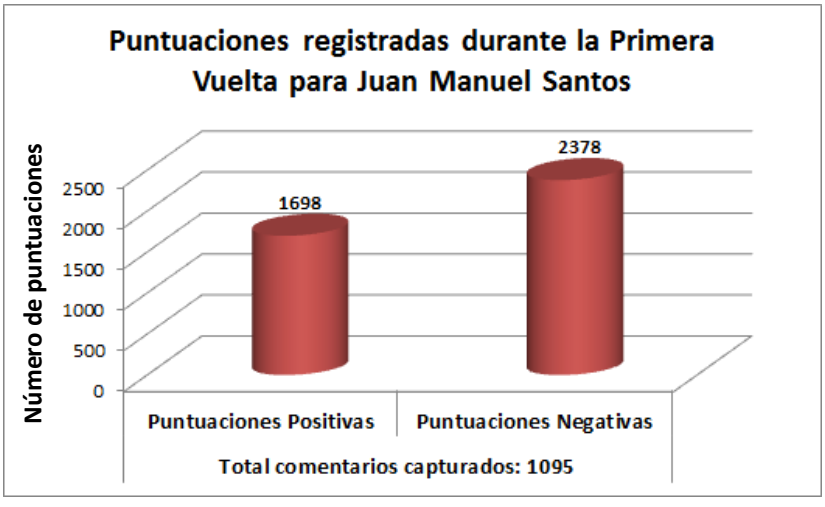
Figura 15. Comportamiento en el tiempo del número de tweets capturados para Juan Manuel Santos, durante primera y segunda vuelta

La figura 15.a muestra que el candidato Juan Manuel Santos tuvo varios altibajos en cuanto al número de opiniones expresadas acerca de él en Twitter durante la primera

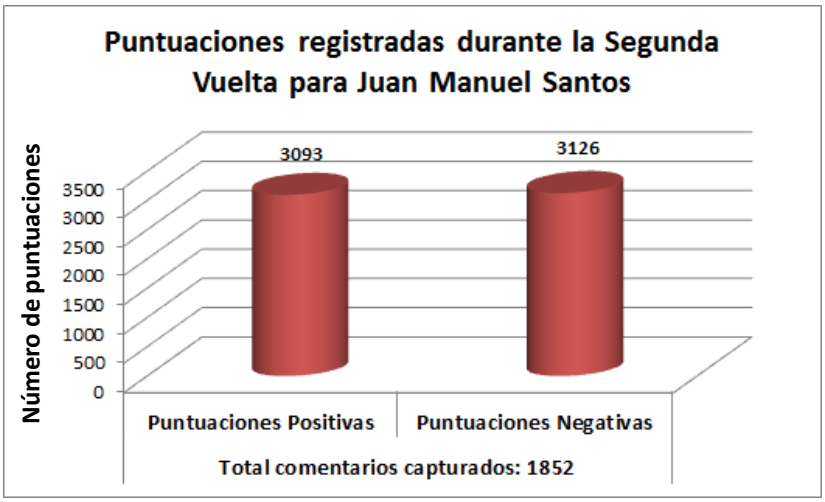
vuelta, y el número más alto de tweets que registró fue el día de las votaciones en primera vuelta, cuando obtuvo la segunda votación más alta.

En la figura 15.b se puede observar que después de la primera vuelta el número de tweets para Juan Manuel Santos fue disminuyendo y se mantuvo estable varios días antes de la votaciones de segunda vuelta, para luego recobrar un ligero incremento los días previos al 15 de junio, siendo este último el día en que más tweets se registraron acerca de este candidato, tras haber ganado las elecciones a la Presidencia de la República.

El cálculo de la polaridad de los comentarios analizados con relación al candidato Juan Manuel Santos, para la primera vuelta, arrojó un acumulado de 1698 puntuaciones positivas, frente a 2378 negativas, de tal manera que se obtuvo una polaridad *Negativa* (figura 16.a), basada en los comentarios expresados por los colombianos durante la primera vuelta. Con respecto a la segunda vuelta, el cálculo de la polaridad arrojó un acumulado de 3093 puntuaciones positivas, frente a 3126 negativas, obteniéndose nuevamente una polaridad *Negativa* (figura 16.b), esta vez basada en lo que expresaron los colombianos durante la segunda vuelta presidencial acerca del candidato Juan Manuel Santos.



a) Primera vuelta



b) Segunda vuelta

Figura 16. Comportamiento de puntuaciones registradas para el candidato Juan Manuel Santos, durante la primera y la segunda vuelta

Finalmente, el cálculo de la polaridad para la primera y segunda vuelta, arrojó un acumulado de 4791 puntuaciones positivas, frente a 5504 negativas, de tal manera que se obtuvo una polaridad global **Negativa** para el candidato Juan Manuel Santos, teniendo en cuenta un total de **2947** comentarios capturados.

La tabla 11 muestra cuáles fueron los cinco términos positivos que más utilizaron los colombianos para referirse a Juan Manuel Santos durante las elecciones presidenciales. Las palabras “Felicidad”, “Paz” y “Felicitaciones” fueron las más positivas utilizadas, y la palabra “Paz” fue la que más se repitió.

Término o palabra utilizada	Puntuación (score)	Frecuencia absoluta
Felicidad	+5	1
Paz	+4	440
Felicitaciones	+4	14
Respeto	+3	17
Esperanza	+3	7

Tabla 11. Términos positivos más utilizados para referirse a Juan Manuel Santos

Este gran número de veces que se repitió la palabra paz cuando los colombianos expresaron sus comentarios acerca de Juan Manuel Santos, puede que esté muy relacionado con el hecho de que este candidato ganara las elecciones en segunda vuelta, pues el tema de la paz es vital para todos los colombianos de bien.

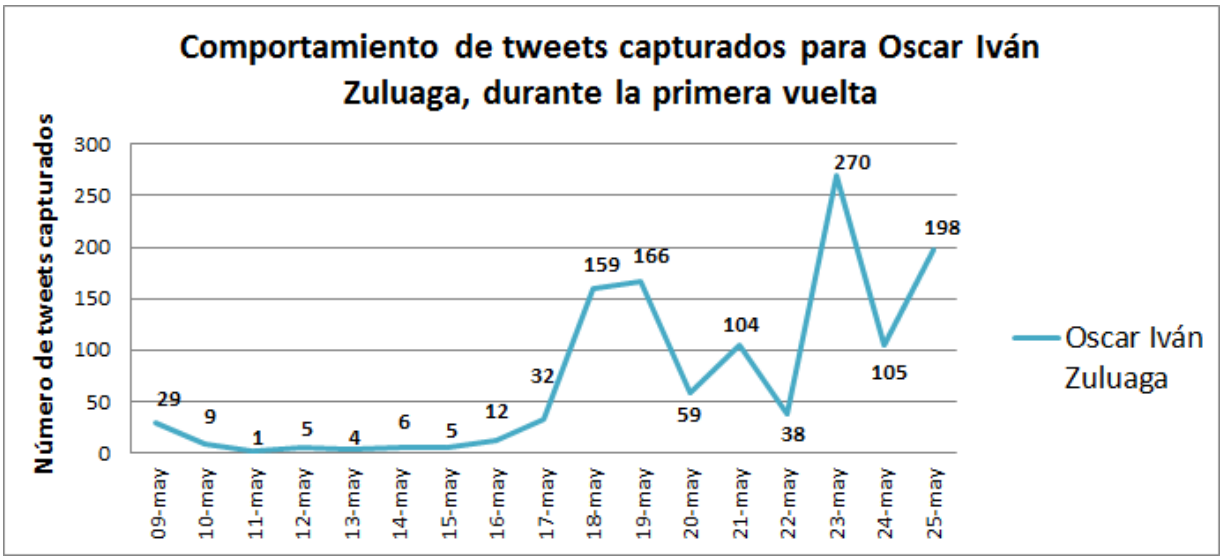
En la tabla 12 se observa cuáles fueron los cinco términos negativos que más se utilizaron para referirse a Juan Manuel Santos durante las elecciones. La palabra "Terroristas" fue la palabra más negativa utilizada, mientras que la palabra que más se repitió fue "FARC". Esto puede explicarse porque el candidato presidente Juan Manuel Santos, como jefe de estado, autorizó el inicio de conversaciones con las FARC, buscando una salida al conflicto armado, mediante negociaciones de paz, razón por la cual se utilizó mucho la palabra FARC para hacer referencia al proceso llevado a cabo en La Habana, Cuba.

Término o palabra utilizada	Puntuación (score)	Frecuencia absoluta
Terroristas	-5	7
Guerrilla	-4	18
FARC	-4	65
Mentiroso	-3	8
Corruptos	-3	4

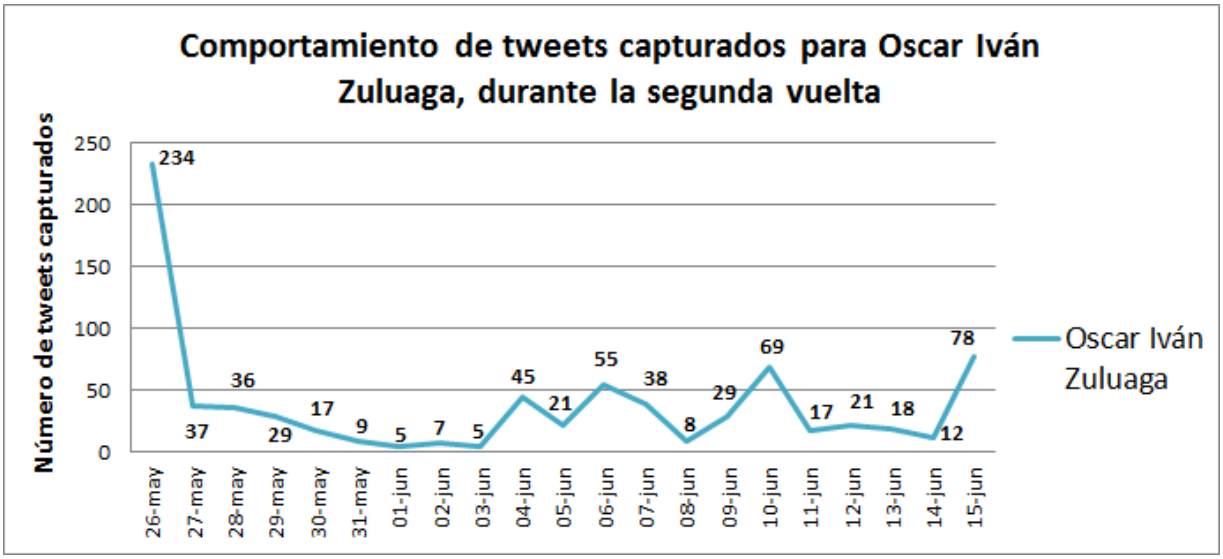
Tabla 12. Términos negativos más utilizados para referirse a Juan Manuel Santos

4.2.5 Resultados de opinion mining para el candidato Oscar Iván Zuluaga

Para comprender de mejor manera los resultados arrojados por la minería de opiniones y el cálculo de la polaridad de los comentarios, obsérvese la figura 17, en la cual se muestra la fluctuación de tweets descargados por fecha para el candidato Oscar Iván Zuluaga:



a) Primera vuelta



b) Segunda vuelta

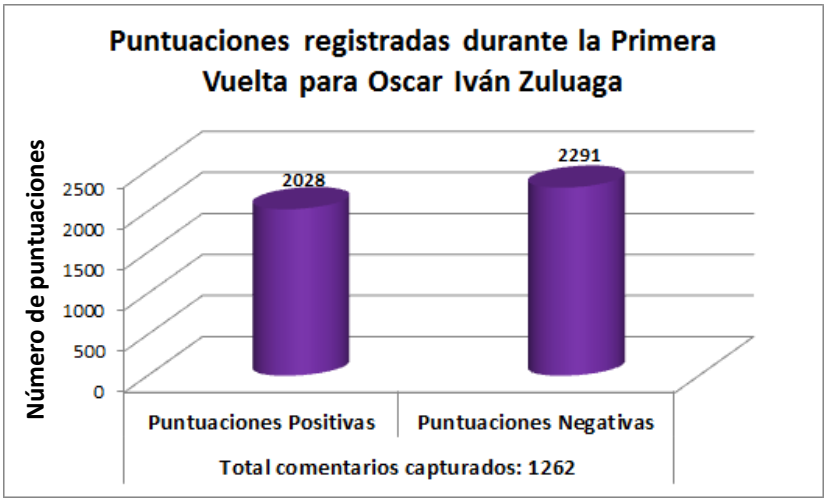
Figura 17. Comportamiento en el tiempo del número de tweets capturados para Oscar Iván Zuluaga, durante primera y segunda vuelta

En la figura 17.a se puede observar que el número de comentarios generados para hacer referencia al candidato Oscar Iván Zuluaga fueron incrementándose

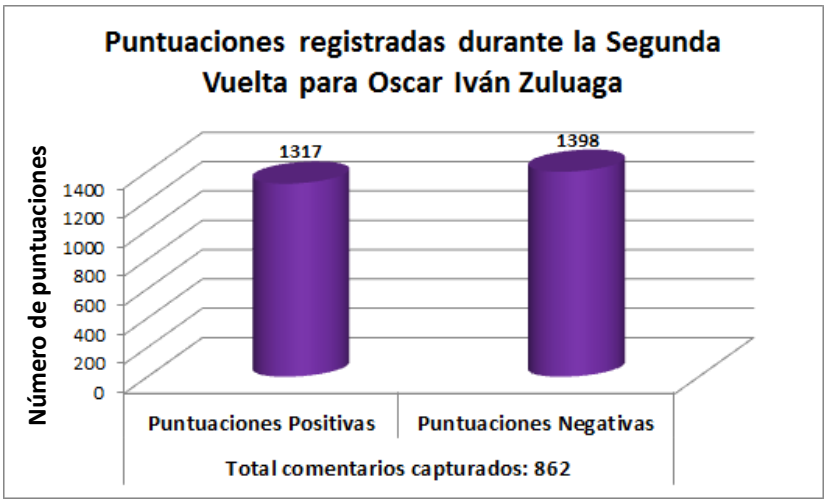
aproximadamente 10 días antes de las votaciones de primera vuelta. Esto indica que varias noticias relacionadas con Oscar Iván Zuluaga y un presunto hacker a sus servicios, contratado para infiltrar el proceso de paz con las FARC, no afectaron significativamente para mal la imagen de este candidato presidencial, sino que por el contrario su imagen se fortaleció durante la primera vuelta.

La figura 17.b indica que Oscar Iván Zuluaga, a pesar de haber obtenido las votaciones más altas en la primera vuelta, fue disminuyendo su aceptación para gran parte de los ciudadanos; la gráfica indica que el número de comentarios fue disminuyendo, y esto se vio reflejado en la pérdida de las elecciones para este candidato en la segunda vuelta.

Para Oscar Iván Zuluaga, el cálculo de la polaridad arrojó un acumulado de 2028 puntuaciones positivas, frente a 2291 negativas, obteniéndose una polaridad *Negativa* (figura 18.a) en la primera vuelta, determinada por lo expresado en los comentarios. Para la segunda vuelta arrojó un acumulado de 1317 puntuaciones positivas, frente a 1398 negativas, obteniendo una polaridad *Negativa* (figura 18.b).



a) Primera vuelta



b) Segunda vuelta

Figura 18. Comportamiento de puntuaciones registradas para el candidato Oscar Iván Zuluaga, durante la primera y la segunda vuelta

Finalmente, el cálculo de la polaridad para la primera y segunda vuelta, arrojó un acumulado de 3345 puntuaciones positivas, frente a 3689 negativas, obteniéndose de esa manera una polaridad global **Negativa** para el candidato Oscar Iván Zuluaga, teniendo en cuenta un total de **2124** comentarios capturados.

En la tabla 13 pueden verse cuáles fueron los cinco términos positivos que más utilizaron los colombianos para referirse a Oscar Iván Zuluaga durante las elecciones. Las palabras “Paz” e “Inteligente” fueron las más positivas utilizadas, mientras que la palabra “Paz” fue la que más se repitió.

Término o palabra utilizada	Puntuación (score)	Frecuencia absoluta
Paz	+4	64
Inteligente	+4	6
Ganar	+3	31
Esperanza	+3	6
Ganador	+3	34

Tabla 13. Términos positivos más utilizados para referirse a Oscar Iván Zuluaga

La tabla 14 indica cuáles fueron los cinco términos negativos que más se utilizaron para referirse a Oscar Iván Zuluaga durante las elecciones. Las palabras “FARC” y “Paraco” fueron las palabras más negativas utilizadas, pero la palabra que más se repitió fue “Hackers”.

Término o palabra utilizada	Puntuación (score)	Frecuencia absoluta
FARC	-4	10
Paraco	-4	9
Hackers	-3	70
Chuzadas	-3	10
Títere	-3	25

Tabla 14. Términos negativos más utilizados para referirse a Oscar Iván Zuluaga

En la tabla 14 se observan varios términos que fueron utilizados solamente para dirigirse negativamente a este candidato durante las elecciones presidenciales: *hackers, chuzadas, títere*. Esta podría ser una razón de peso para que haya perdido las elecciones: falta de credibilidad y transparencia para con el país.

4.2.6 Resultados de opinion mining obtenidos mediante otros hashtag

Además de los comentarios acerca de los candidatos presidenciales, se capturaron tweets relacionados con las elecciones presidenciales a partir de la segunda vuelta, por medio de los hashtag *#Elecciones2014* y *#SegundaVuelta*, con el objetivo de obtener un punto de vista global acerca de lo que los colombianos pensaban acerca de esta segunda fase de las elecciones presidenciales en el país.

4.2.6.1 Resultados obtenidos para el hashtag *#Elecciones2014*

El cálculo de la polaridad global del fichero en el cual fueron almacenados los tweets capturados mediante el hashtag *#Elecciones2014*, arrojó un acumulado de 22057

puntuaciones positivas, frente a un acumulado de 22112 puntuaciones negativas. En este caso se obtuvo una polaridad global **Negativa** (figura 19), teniendo en cuenta un total de **15918** comentarios capturados.

Figura 19. Cálculo de la polaridad global del fichero contenedor de tweets capturados por medio del hashtag #Elecciones2014

4.2.6.2 Resultados obtenidos para el hashtag *#SegundaVuelta*

Con respecto a los tweets capturados por medio del hashtag *#SegundaVuelta*, la polaridad global obtenida también fue **Negativa** (figura 20), tras obtener un acumulado de 9279 puntuaciones positivas, frente a 9569 negativas, teniendo en cuenta un total de **6253** comentarios capturados.

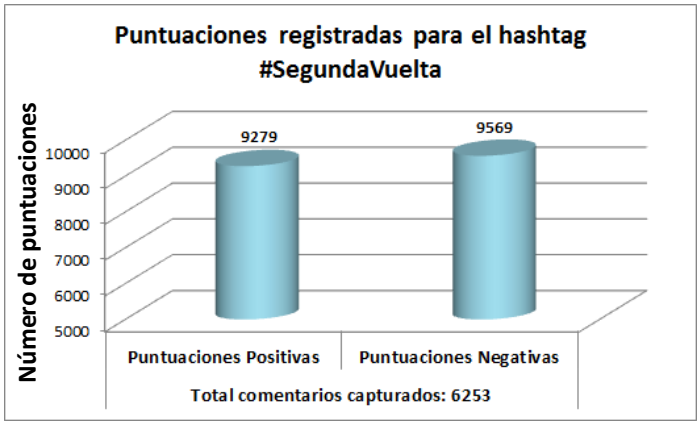


Figura 20. Cálculo de la polaridad global del fichero contenedor de tweets capturados por medio del hashtag #SegundaVuelta

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

Este análisis general, construido a partir de los resultados obtenidos gracias al proceso de minería de opiniones, se divide principalmente en dos secciones relacionadas con las dos fases (o vueltas) que tuvo el proceso de elecciones presidenciales en Colombia en el año 2014, analizando el papel que jugó cada candidato involucrado en cada una de estas dos fases.

El análisis también se apoya en la fluctuación o comportamiento de tweets capturados para cada candidato presidencial, desde el 9 de mayo hasta el 15 de junio de 2014, tal como se muestra en la figura 21.

En dicha figura 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.

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, pues ellos además de pasar a la segunda vuelta, siempre generaron polémicas noticias durante las elecciones presidenciales.

Puede observarse también que la candidata Marta Lucía Ramírez fue quien menos tweets generó a lo largo de las elecciones presidenciales, y tuvo su máxima mención en Twitter el día de las elecciones presidenciales en primea vuelta.

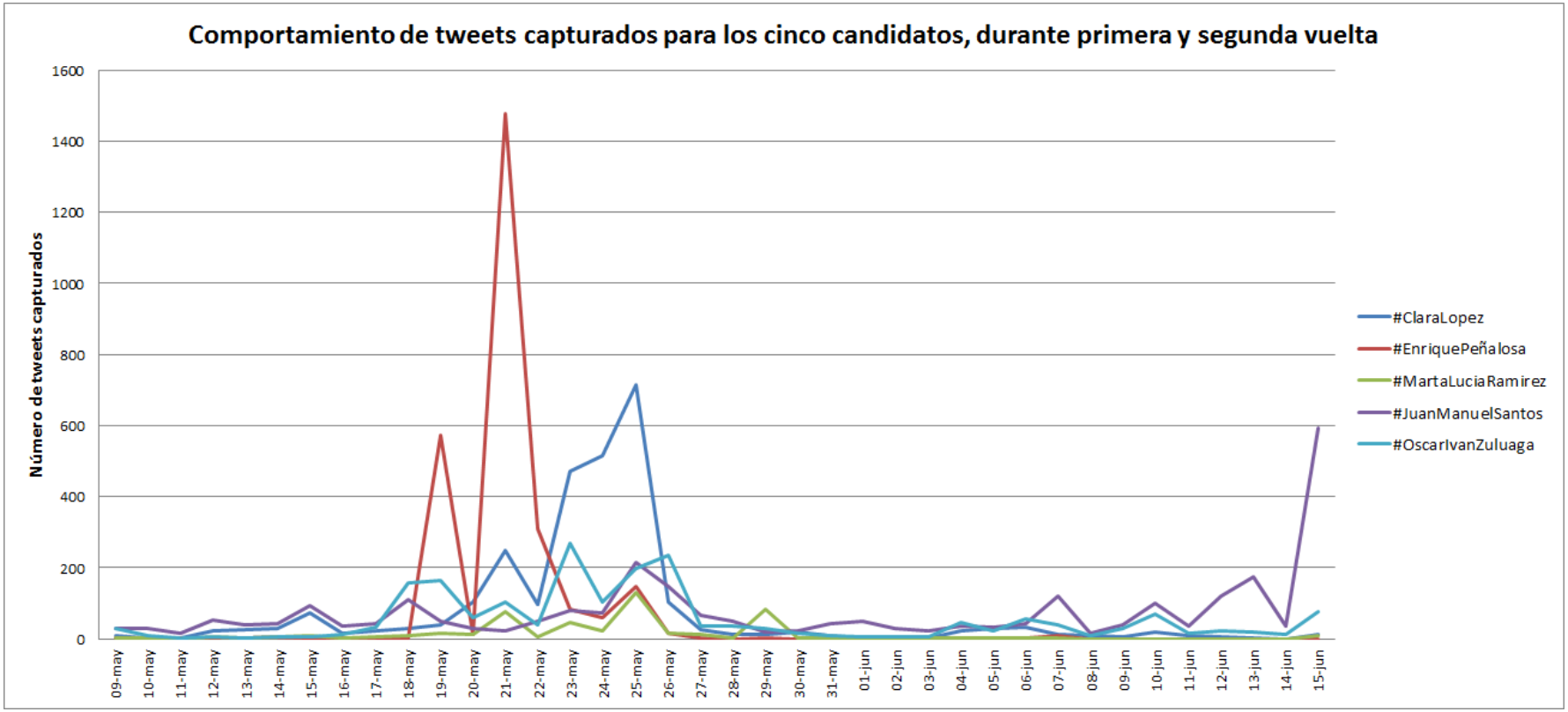


Figura 21. Comportamiento en el tiempo del número de tweets expresados para cada candidato presidencial, durante primera y segunda vuelta

Cabe mencionar que este análisis se construyó con base en los comentarios recolectados y los resultados de la minería de opiniones, y no con base en los resultados reales de las elecciones presidenciales. Los comentarios recolectados conformaron la fuente de información y son la materia prima a partir de la cual se realizaron varios procesos en este trabajo de investigación, entre ellos este análisis general. Por lo tanto, la concordancia o discordancia de los resultados explicados en este análisis en comparación con los resultados reales de las elecciones presidenciales, se analizará en el transcurso de este apartado.

4.3.1 Análisis basado en la primera vuelta presidencial y sus candidatos

Los cinco candidatos presidenciales con los cuales se inició el proceso de elecciones contaron con el apoyo de equipos de trabajo mediante los cuales se realizó el despliegue de campañas para cada uno de ellos, con el objetivo de dar a conocer sus programas de gobierno.

Con base en los programas y propuestas de cada candidato, los ciudadanos colombianos fueron formándose un perfil personal de cada uno de ellos, e inclinándose por la opción que más les parecía adecuada para gobernar al país. Sin embargo, además de los programas de gobierno, todos los candidatos se vieron involucrados en situaciones que favorecieron o perjudicaron su imagen frente a los electores, quienes asumieron en su gran mayoría una actitud crítica para las elecciones, actitud que se vio reflejada en muchos comentarios expresados a través de sitios web 2.0, entre ellos los recolectados para esta investigación.

Los resultados obtenidos mediante minería de opiniones en este trabajo han sido utilizados para comprender la imagen que, para los ciudadanos de quienes se recolectaron comentarios, reflejaron los cinco candidatos presidenciales y el proceso de elecciones presidenciales en general.

Para iniciar con el análisis de la imagen que reflejaron los candidatos durante la primera vuelta presidencial, obsérvese la siguiente figura:

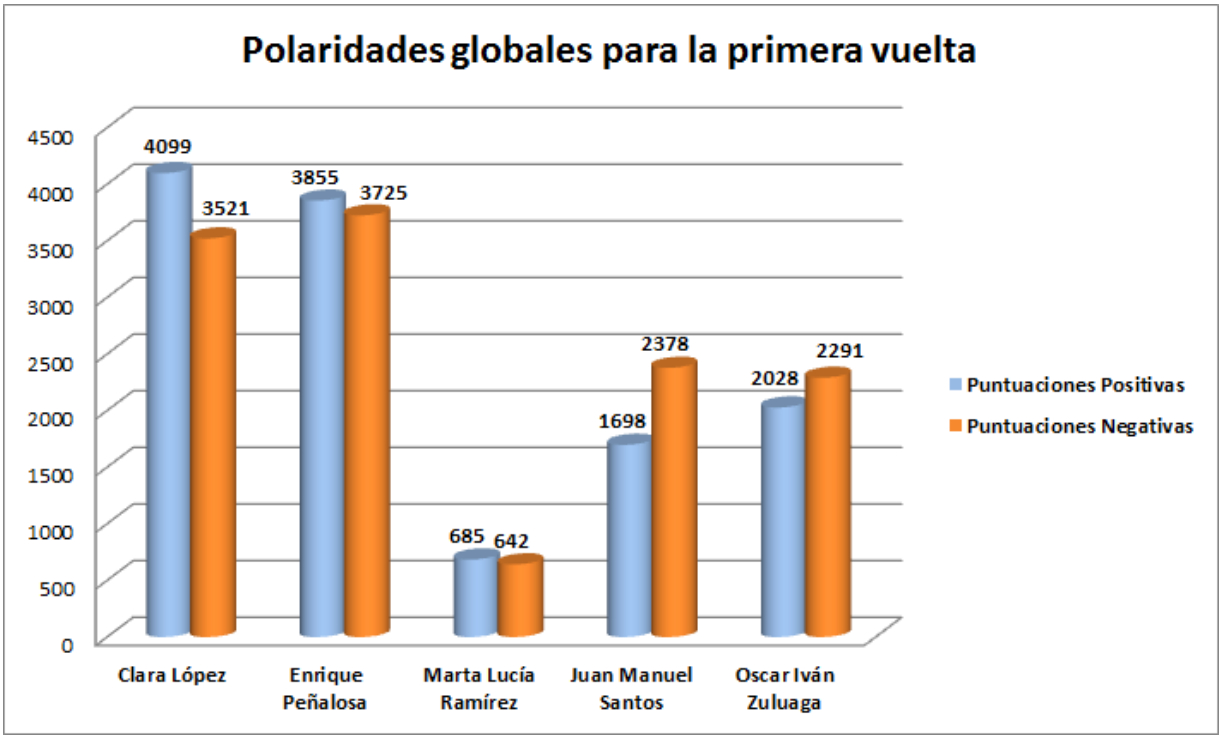


Figura 22. Resumen de polaridades globales calculadas con base en los cinco candidatos para la primera vuelta

En la figura 22 se puede observar que los candidatos: Clara López, Enrique Peñalosa y Marta Lucía Ramírez, transmitieron una imagen positiva durante la primera vuelta. Es decir, según la minería de opiniones, los colombianos mayoritariamente demostraron preferencia por estos tres candidatos.

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 presidencial, por lo cual transmitieron una imagen negativa frente a los demás candidatos.

Para determinar cuál de los cinco candidatos reflejó la mejor imagen durante 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 o imagen positiva (según la figura 22).

La candidata presidencial que obtuvo la mayor diferencia entre las puntuaciones positivas y negativas fue Clara López, con una diferencia de 578, en comparación con las diferencias de 130 y 43 que obtuvieron Enrique Peñalosa y Marta Lucía Ramírez, respectivamente. Esto indica que, según los comentarios recolectados y sometidos a opinion mining, Clara López reflejó la imagen más positiva durante la primera vuelta presidencial.

Aunque según los resultados de la minería de opiniones, la candidata Clara López reflejó la imagen más positiva durante la primera vuelta, en las elecciones del 25 de mayo no logró obtener una votación que le permitiera pasar a la segunda vuelta. De igual manera, los otros dos candidatos que según la minería de opiniones reflejaron una imagen positiva (Enrique Peñalosa y Marta Lucía Ramírez) tampoco lograron pasar a la segunda vuelta.

Esto podría explicarse porque los comentarios recolectados para realizar la minería de opiniones generalmente son expresados por ciudadanos que han adoptado una posición crítica frente a temas relacionados con la política, lo cual los convierte en electores con una idea sólida sobre qué candidato apoyar, en contraposición con una gran parte de la población, especialmente la que carece de acceso constante a diferentes fuentes de información, que puede presentar inseguridad para elegir a sus gobernantes, o peor aún, que se inclina siempre por el(la) candidato(a) que puntee en las encuestas en determinados momentos.

En la figura 16 puede observarse que la candidata Clara López generó la mayor cantidad de tweets relacionados con su candidatura el día de las elecciones en primera vuelta (25 de mayo): 716 tweets, pero al no pasar a la segunda vuelta, su mención en los tweets disminuyó rápidamente.

La figura 21 también muestra una fluctuación de tweets considerable para el candidato Enrique Peñalosa los días 19 de mayo (572 tweets) y 21 de mayo (1478 tweets); sin embargo este candidato fue dejando de generar tweets a partir de la segunda vuelta, a tal punto que en los últimos 8 días antes de las elecciones de celebrarse las votaciones de segunda vuelta, no generó ningún tweet.

Por su parte la candidata Marta Lucía Ramírez, según la figura 21, siempre mostró un bajo número de tweets expresados sobre su candidatura, pues el máximo número de tweets que se capturaron para ella fue de 130 (el 25 de mayo, día de las votaciones en primera vuelta). A partir de entonces la fluctuación de tweets relacionados con ella disminuyó, generando cero tweets los días previos a las votaciones de segunda vuelta, y solamente se capturaron 9 tweets sobre esta candidata el día 15 de junio, día final de las elecciones presidenciales de 2014.

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 (25 de mayo) fueron: Juan Manuel Santos y Oscar Iván Zuluaga, por lo tanto pasaron a la segunda vuelta llevada a cabo el 15 de junio de 2014.

La imagen que para los ciudadanos reflejaron los dos candidatos que pasaron a la segunda vuelta, determinada por los resultados obtenidos mediante minería de opiniones, se puede apreciar en la figura 23:

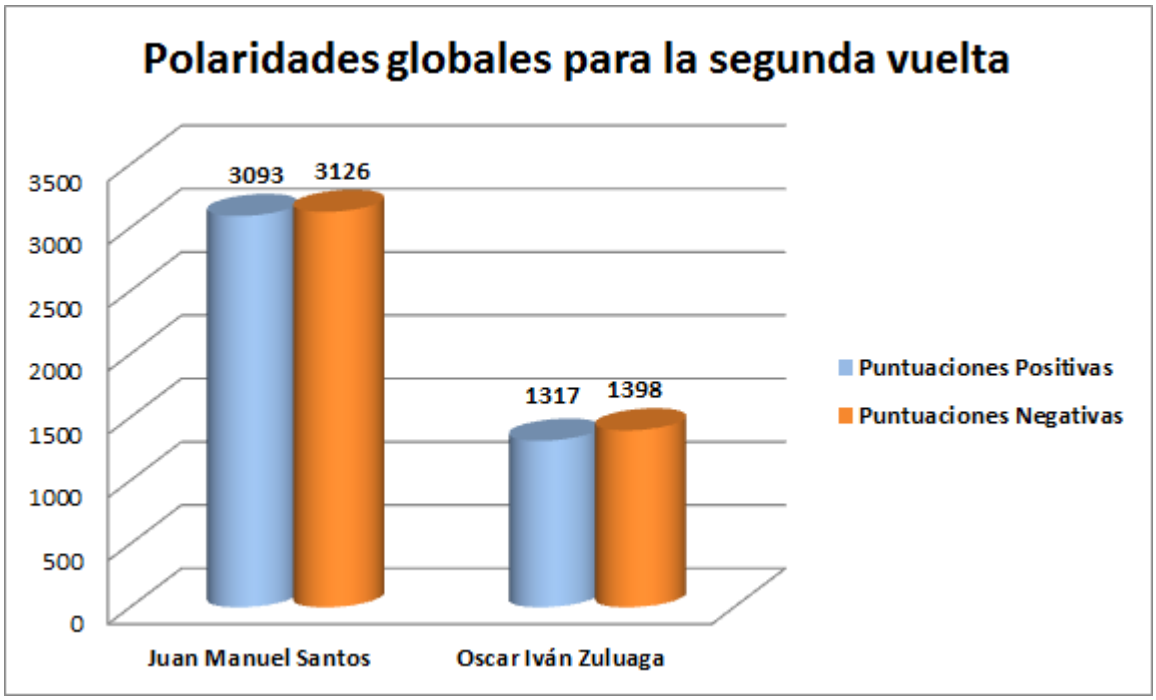


Figura 23. Resumen de polaridades globales calculadas con base en los dos candidatos para la segunda vuelta

Según los resultados arrojados por la minería de opiniones, ambos candidatos reflejaron una imagen negativa durante la segunda vuelta.

Como en este caso ambas imágenes reflejadas fueron negativas, se procede a determinar cuál de los dos candidatos reflejó la imagen menos negativa durante la segunda vuelta, de acuerdo a la minería de opiniones realizada, obteniendo esta vez la menor diferencia (resta) entre las puntuaciones negativas y positivas para cada uno de los candidatos.

El candidato Juan Manuel Santos fue quien obtuvo la menor diferencia entre las puntuaciones negativas y positivas, con una diferencia de 33, en comparación con una diferencia de 81 obtenida por el candidato Oscar Iván Zuluaga. Esto quiere decir que con base en los comentarios recolectados y sometidos a opinion mining, Juan

Manuel Santos reflejó 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, entre los dos candidatos que pasaron a segunda vuelta y los que no. Dichas alianzas quedaron de la siguiente manera:

- *Juan Manuel Santos*: formó alianza con Clara López, candidata a las elecciones para la primera vuelta por la coalición de los partidos polo democrático alternativo (PDA) y unión patriótica (UP).
- *Oscar Iván Zuluaga*: formó alianza con Marta Lucía Ramírez, candidata a las elecciones para la primera vuelta por el partido conservador colombiano (PCC).

Es importante mencionar que 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, los cuales dieron como ganador al candidato Juan Manuel Santos, se deben analizar las polaridades globales que tuvieron todos los candidatos durante la primera y la segunda vuelta. La figura 19 resume esa información:

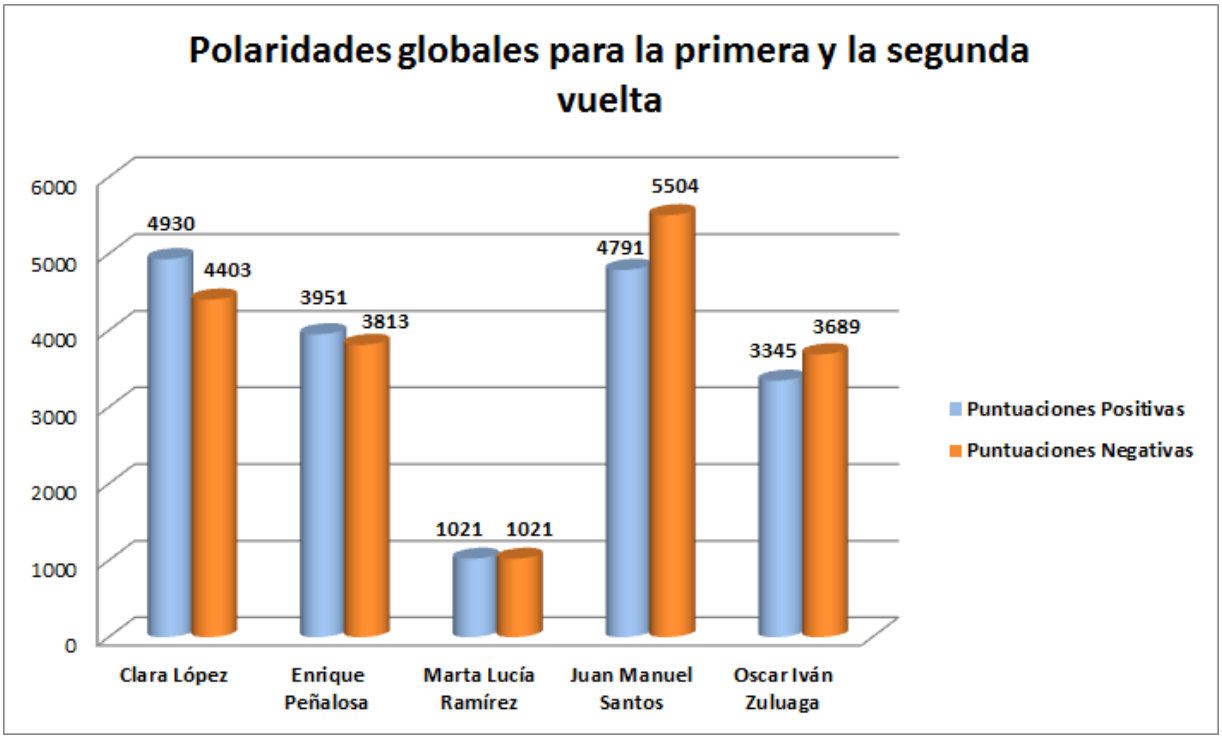


Figura 24. Resumen de polaridades globales, calculadas con base en los cinco candidatos, para la primera vuelta y la segunda vuelta

En la figura 24 se puede observar que la candidata Clara López, quien hizo alianza con Juan Manuel Santos, reflejó una imagen global (primera y segunda vuelta) positiva, lo cual puedo haber favorecido enormemente (entre otros factores) la candidatura de Juan Manuel Santos, quien finalmente ganó las elecciones para la presidencia de Colombia en el año 2014.

La figura 24 también muestra que la candidata Marta Lucía Ramírez, quien para la segunda vuelta se vinculó a la campaña de Oscar Iván Zuluaga, reflejó una imagen global (primera y segunda vuelta) 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[34].

El candidato Enrique Peñalosa, al dejar en libertad a sus seguidores, fue el candidato del que menos se expresaron comentarios en los sitios web 2.0 a partir del 26 de mayo (ver figura 21), de tal manera que para este trabajo sólo se recolectaron 54 comentarios acerca de él durante la segunda vuelta, en comparación con los 2750 comentarios que se recolectaron en la primera vuelta. Esto se puede comprender mejor si se tiene en cuenta que los electores seguían generando comentarios, pero acerca de los candidatos que pasaron a la segunda vuelta y a sus ex contendores con quienes realizaron alianza. Por tal motivo se puede decir que el candidato Enrique Peñalosa prácticamente desapareció para los ciudadanos que expresaron comentarios a partir de la segunda vuelta.

Si también se observa la polaridad global o imagen durante todas las elecciones (primera y segunda vuelta) para cada uno de los candidatos (figura 24), se puede analizar que los dos candidatos que pasaron a la segunda vuelta reflejaron 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, y no continuar con el hasta entonces muy cuestionado proceso de paz llevado a cabo por el presidente candidato Juan Manuel Santos, ni regresar a las políticas del uribismo, encarnadas por el candidato Oscar

[34] http://www.elespectador.com/noticias/politica/acusan-zuluaga-y-uribe-de-chuzar-y-sabotear-el-proceso-articulo-490746

Iván Zuluaga, de quien se dijo en el transcurso de todas las elecciones que era un títere del ex presidente Álvaro Uribe Vélez[35].

La figura 21 indicó una fluctuación de tweets mayor para el candidato Juan Manuel Santos a partir de la segunda vuelta, en comparación con su rival Oscar Iván Zuluaga, a tal punto que durante los 8 días previos a las votaciones de segunda vuelta, Juan Manuel Santos siempre superó en número de tweets expresados a su contendor en las elecciones de segunda vuelta.

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

Los hashtag *#Elecciones2014* y *#SegundaVuelta* marcaron tendencia en Twitter durante la segunda vuelta presidencial, debido a que los ciudadanos los utilizaron mucho para referirse al proceso de elecciones en general, incluyendo los cinco candidatos.

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).

[35] http://es.wikipedia.org/wiki/%C3%81lvaro_Uribe

Estos resultados demuestran que la mayoría de los ciudadanos que utilizaron Twitter para expresar sentimientos, opiniones y emociones acerca de las elecciones presidenciales de Colombia en 2014, tuvieron una perspectiva negativa acerca del proceso de las elecciones en general, por supuesto incluyendo a los principales actores involucrados: los candidatos presidenciales.

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, mala administración, ideales políticos que según los colombianos no le favorecían al país, entre otros.

4.4 Contraste entre el análisis general y encuestas de Gallup y Datexco

Los resultados de la minería de opiniones obtenidos en este trabajo fueron comparados con varias encuestas realizadas por dos de las más prestigiosas firmas encuestadoras con presencia en Colombia, las cuales son: Gallup Colombia Ltda., y Datexco Company S.A.

4.4.1 Comparación con encuestas realizadas para la primera vuelta

La firma encuestadora Gallup Colombia Ltda. realizó la encuesta llamada “La gran encuesta de los medios #6”, recolectando datos desde el 10 hasta el 13 de mayo de 2014. Esta encuesta, incluyendo su ficha técnica, puede ser consultada accediendo a la siguiente URL: http://media.caracoltv.com/LA-GRAN-ENCUESTA-DE-LOS-MEDIOS-ELECCIONES-2014--6.pdf

Los resultados de dicha encuesta se resumen en la figura 25. Estos resultados indicaron que el candidato Oscar Iván Zuluaga sería el candidato con 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 estos dos serían los candidatos que pasarían a la segunda vuelta. Los resultados de esta encuesta se compaginaron con la realidad, pues en efecto, el candidato Oscar Iván Zuluaga fue quien obtuvo la mayor votación en las elecciones de primera vuelta, seguido por el candidato Juan Manuel Santos, pasando ambos a la segunda vuelta.

En cuanto a la imagen reflejada por los candidatos durante la primera vuelta, obtenida mediante minería de opiniones en este trabajo, los resultados indicaron que la candidata Clara López obtendría la votación más alta en la primera vuelta (figura 26), seguida por el candidato Enrique Peñalosa, por lo tanto ellos dos pasarían a segunda vuelta. Estos resultados no coincidieron con los resultados descritos con base en la encuesta de Gallup para la primera vuelta.

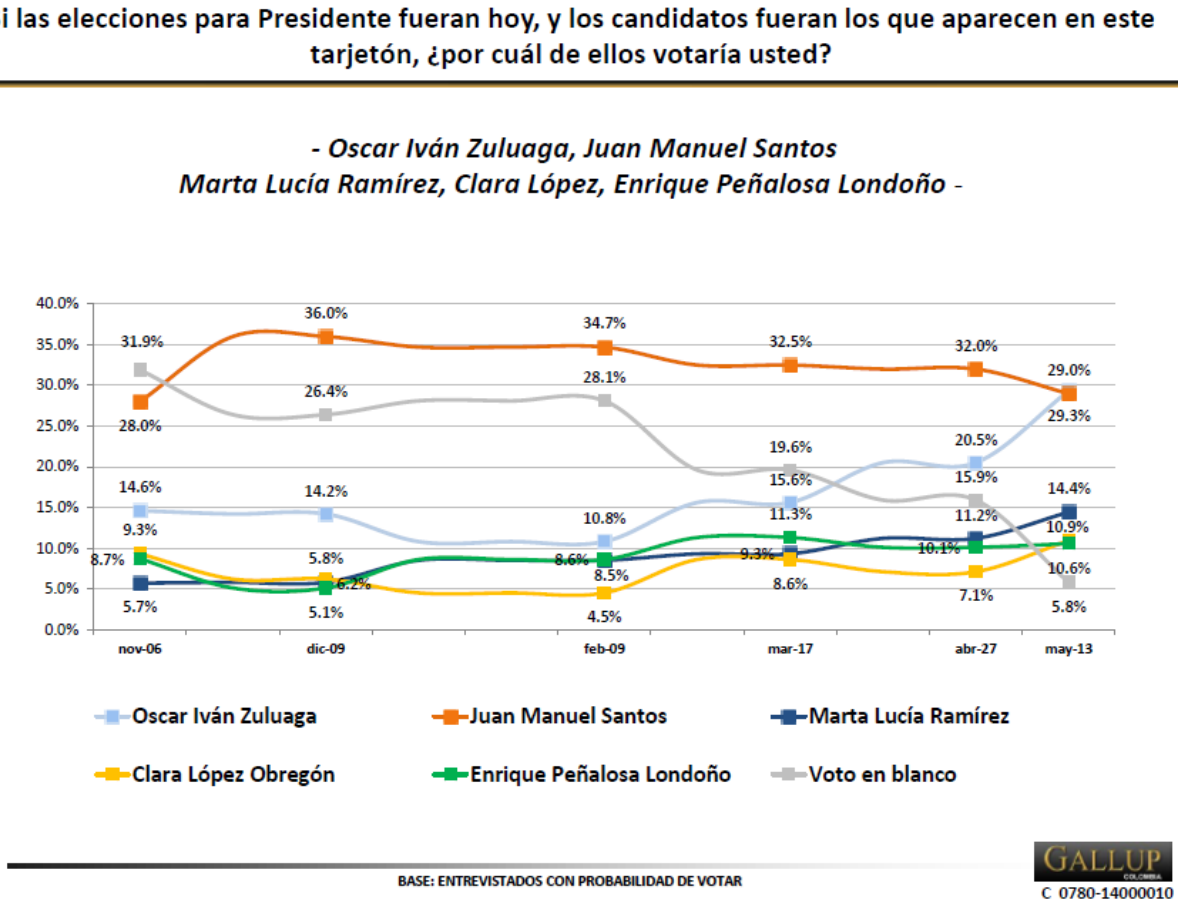


Figura 25. Intención de voto de los colombianos para la primera vuelta, según encuesta realizada por Gallup

Por otra parte, la firma Datexco Company S.A. en su "Tercera gran encuesta para las elecciones presidenciales 2014", recolectó datos desde el mes de enero hasta el mes de marzo de 2014, con el objetivo de conocer la intención de voto de los colombianos en la primera vuelta presidencial. Todos los detalles de esta encuesta pueden ser consultados accediendo a la siguiente URL: http://www.eltiempo.com/Multimedia/infografia/encuestapresidencial/

Los resultados de esa encuesta se muestran de manera resumida en la figura 27, en donde se puede observar que los resultados generales indicaron que el candidato Juan Manuel santos, con una intención de voto del 25,5% sería el candidato con

mayor votación en la primera vuelta, mientras que el candidato Enrique Peñalosa, con una intención de voto del 17.1%, ocuparía el segundo lugar, por lo cual ambos pasarían a la segunda vuelta. Los resultados de esa encuesta realizada por Datexco no se compaginaron con la realidad, pues como ya se mencionó, en las elecciones reales el primer puesto lo ocupó el candidato Oscar Iván Zuluaga, seguido por el candidato Juan Manuel Santos.

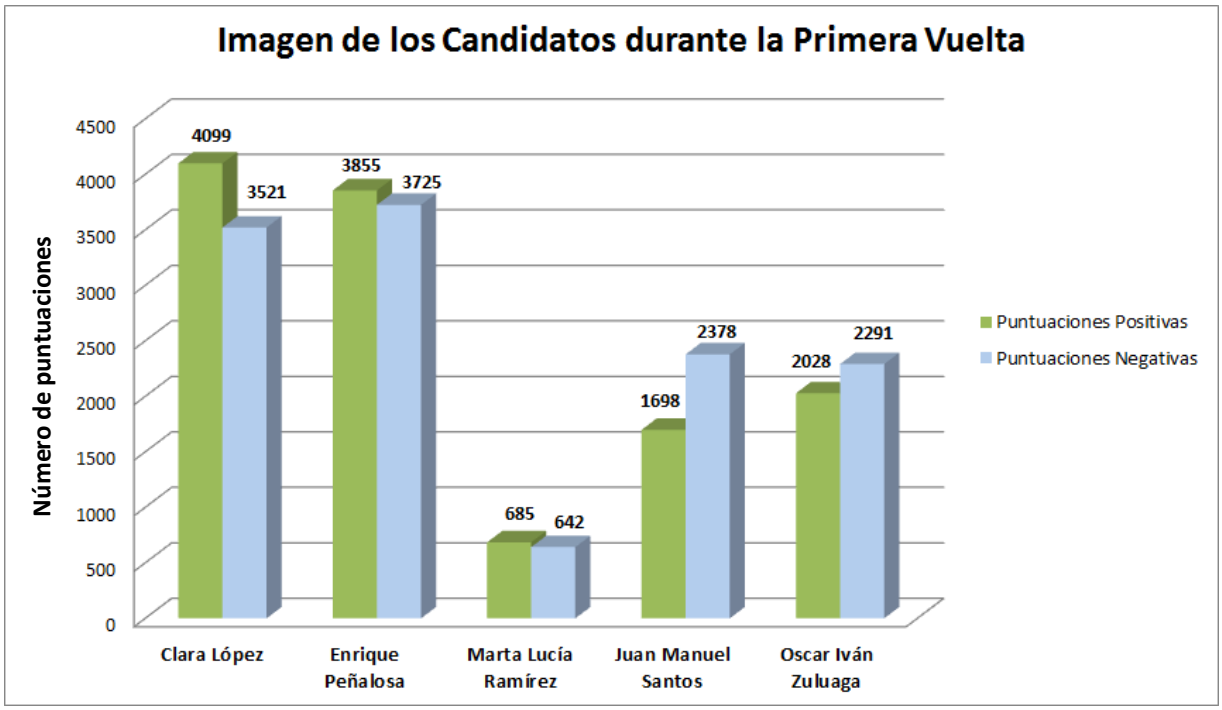


Figura 26. Imagen reflejada por los candidatos durante la primera vuelta, según resultados obtenidos mediante minería de opiniones

Comparando los resultados de la encuesta de Datexco para la primera vuelta (figura 27), con la imagen reflejada por los candidatos durante la primera vuelta, obtenida mediante minería de opiniones en este trabajo (figura 26), 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 en este trabajo estuvieron relacionados con la mencionada encuesta realizada por Datexco.

Figura 27. Intención de voto de los colombianos para la primera vuelta, según encuesta realizada por Datexco (Foto: eltiempo.com)

4.4.2 Comparación con encuestas realizadas para la segunda vuelta

Con relación a la segunda vuelta presidencial, la firma Gallup Colombia Ltda. realizó una nueva gran encuesta de los medios, cuya recolección de datos se efectuó desde el 31 de mayo hasta el 3 de junio de 2014. Todos los detalles de esta encuesta más reciente pueden ser consultados accediendo a la siguiente URL: http://www.elespectador.com/noticias/infografia/santos-y-zuluaga-diferencia-apenas-de-08-articulo-496690

Los resultados de la encuesta indicaron que el candidato Oscar Iván Zuluaga sería el ganador de las elecciones presidenciales de Colombia en 2014, porque según la encuesta revelaba una intención de voto del 48,5%, aventajando al candidato Juan Manuel Santos, quien según la encuesta revelaba una intención de voto del 47,7% (figura 28). Sin embargo los resultados de esta encuesta no se compaginaron con la realidad, debido a que el candidato Juan Manuel Santos fue quien obtuvo la mayor votación en las elecciones de segunda vuelta, lo cual finalmente lo convirtió en el nuevo presidente de Colombia para el período 2014 - 2018.

Con respecto a la imagen reflejada por los candidatos durante la segunda vuelta, obtenida mediante minería de opiniones en este trabajo, los resultados indicaron que el candidato Juan Manuel Santos tuvo la imagen más favorable (figura 29), pues la diferencia entre puntuaciones negativas y positivas obtenidas por este candidato en la minería de opiniones indicó que podría superar en las urnas al candidato Oscar Iván Zuluaga.

Aunque para este caso los resultados anunciados por la nueva gran encuesta de los medios realizada por Gallup para la segunda vuelta no coincidieron con los resultados de la minería de opiniones, se puede observar que dichos resultados de minería de opiniones para la segunda vuelta sí coincidieron con los resultados reales obtenidos en las urnas el 15 de junio de 2014.

Figura 28. Intención de voto de los colombianos para la segunda vuelta, según encuesta realizada por Gallup (Foto: elespectador.com)

La firma Datexco Company S.A. también realizó una nueva encuesta para las elecciones presidenciales 2014 en segunda vuelta, recolectando datos desde el 31 de mayo hasta el 4 de junio, con el objetivo de conocer la intención de voto de los

colombianos para la segunda vuelta presidencial. Todos los detalles de esta encuesta puede ser consultados accediendo a la siguiente URL: http://ow.ly/d/2equ y posteriormente se debe descargar el fichero .xlsx que contiene la encuesta completa.

Los resultados de la nueva encuesta realizada por Datexco se muestran resumidamente en la figura 30. Los resultados generales indicaron que el candidato Juan Manuel santos, con una intención de voto del 41,9% ganaría las elecciones presidenciales de Colombia en 2014, 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.

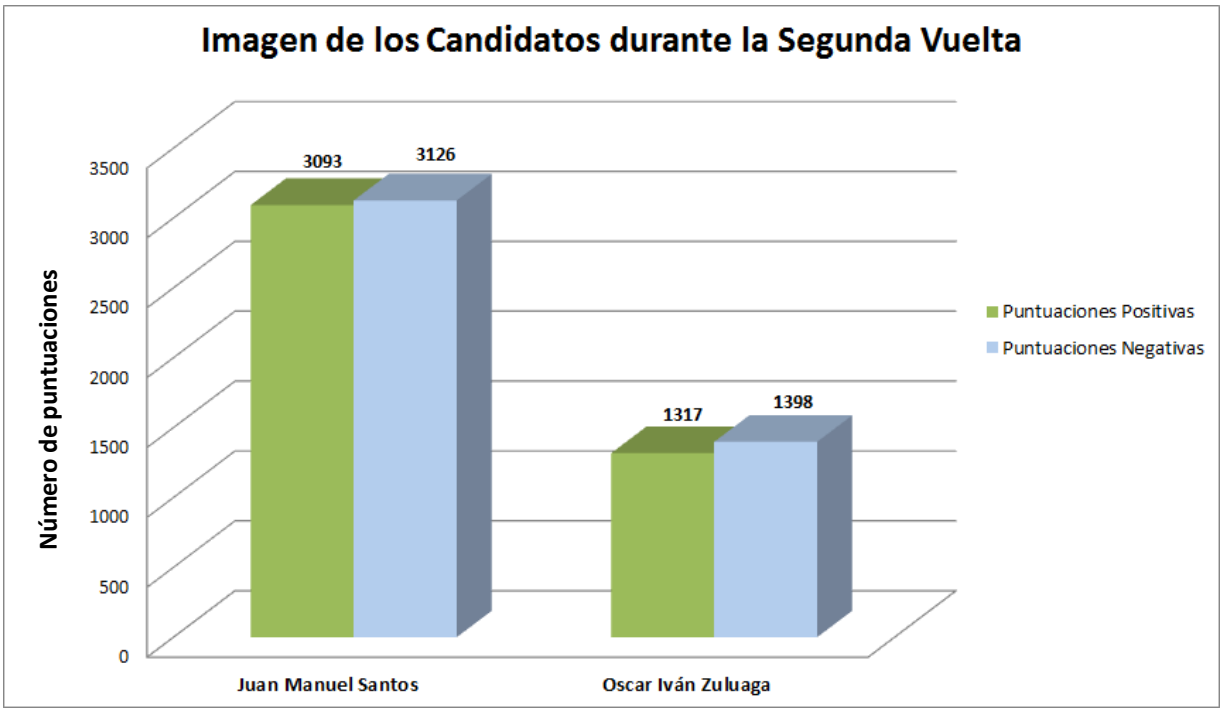
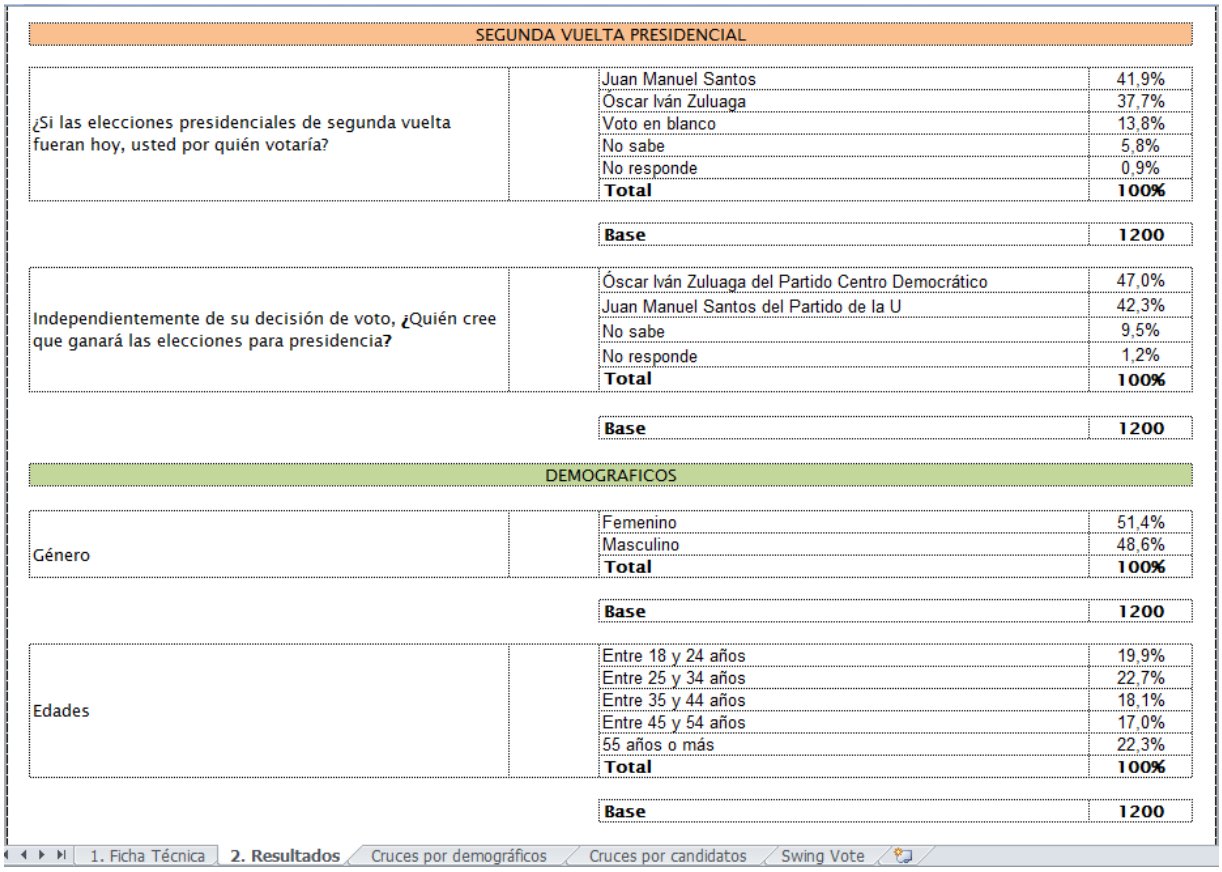


Figura 29. Imagen reflejada por los candidatos durante la segunda vuelta, según resultados obtenidos mediante minería de opiniones

Los resultados de esta última encuesta realizada por Datexco Company S.A. (figura 30), comparados con la imagen reflejada por los candidatos durante la segunda vuelta, obtenida mediante minería de opiniones en este trabajo (figura 29), tuvieron una relación muy estrecha, debido a que según la minería de opiniones efectuada el candidato Juan Manuel Santos tuvo la imagen más favorable durante la segunda vuelta presidencial.



SEGUNDA VUELTA PRESIDENCIAL			
¿Si las elecciones presidenciales de segunda vuelta fueran hoy, usted por quién votaría?		Juan Manuel Santos	41,9%
		Óscar Iván Zuluaga	37,7%
		Voto en blanco	13,8%
		No sabe	5,8%
		No responde	0,9%
		Total	**100%**
		Base	**1200**
Independientemente de su decisión de voto, ¿Quién cree que ganará las elecciones para presidencia?		Óscar Iván Zuluaga del Partido Centro Democrático	47,0%
		Juan Manuel Santos del Partido de la U	42,3%
		No sabe	9,5%
		No responde	1,2%
		Total	**100%**
		Base	**1200**
DEMOGRAFICOS			
Género		Femenino	51,4%
		Masculino	48,6%
		Total	**100%**
		Base	**1200**
Edades		Entre 18 y 24 años	19,9%
		Entre 25 y 34 años	22,7%
		Entre 35 y 44 años	18,1%
		Entre 45 y 54 años	17,0%
		55 años o más	22,3%
		Total	**100%**
		Base	**1200**

1. Ficha Técnica | **2. Resultados** | Cruces por demográficos | Cruces por candidatos | Swing Vote

Figura 30. Intención de voto de los colombianos para la segunda vuelta, según encuesta realizada por Datexco

Capítulo 5

Conclusiones

En este capítulo se establecen las conclusiones en relación con los objetivos y los resultados obtenidos durante el desarrollo de este trabajo, con base en la minería de opiniones realizada a partir de información recolectada desde la Web.

5.1 Discusión

Los resultados obtenidos a partir de los instrumentos de recolección de tweets (aplicación web), y del proceso de web crawling son los siguientes:

- Se recolectaron 10781 tweets, directamente relacionados con los cinco candidatos a la Presidencia de Colombia para el período 2014 – 2018.
- De igual manera se recolectaron 22171 tweets, relacionados con el proceso de elecciones presidenciales en general, a través del uso de los hashtag *#Elecciones2014* y *#SegundaVuelta*.
- Fueron recolectados 270 comentarios desde el sitio web *www.elespectador.com*, específicamente comentarios relacionados con cada uno de los cinco candidatos presidenciales.
- Se recolectaron 348 comentarios desde el sitio web *www.semana.com*, también relacionados específicamente con cada uno de los cinco candidatos a la Presidencia de Colombia en 2014.
- En total se recolectaron 33570 comentarios desde *Twitter, elespectador.com* y *semana.com*, gracias a la aplicación de los instrumentos de recolección de datos.

Estos datos comprueban que es posible recuperar información desde la Web, especialmente desde sitios web 2.0, para su posterior análisis.

Mediante el proceso de Opinion Mining, a través del uso del software *SentiStrength*, con base en los comentarios recolectados expresados por los colombianos en los sitios web 2.0 que se han mencionado, se obtuvieron los siguientes resultados:

- La candidata Clara López obtuvo una polaridad global *Positiva (+)* durante todo el proceso electoral, determinada por los comentarios que expresaron los ciudadanos colombianos.
- El candidato Enrique Peñalosa obtuvo una polaridad global *Positiva (+)* durante todo el proceso electoral, también determinada por los comentarios recolectados.
- La candidata Marta Lucía Ramírez obtuvo una polaridad global *Neutral* durante todo el proceso electoral.
- El candidato Juan Manuel Santos obtuvo una polaridad global *Negativa (-)* durante todo el proceso electoral.
- El candidato Oscar Iván Zuluaga también obtuvo una polaridad global *Negativa (-)* durante todo el proceso electoral.
- El análisis de comentarios mediante el hashtag *#Elecciones2014* arrojó una polaridad global *Negativa (-)* durante la segunda vuelta presidencial.
- De igual manera, el análisis de comentarios mediante el hashtag *#SegundaVuelta* arrojó una polaridad global *Negativa (-)* durante la segunda vuelta presidencial.

Estos datos parecen comprobar que la veracidad de los resultados obtenidos mediante minería de opiniones, está fuertemente ligada a la revisión y alimentación de la base de conocimientos sobre la cual las diferentes herramientas de software pueden efectuar opinion mining, y a la preparación del entorno mediante la adecuada disposición de las fuentes (textos) a procesar.

El análisis general realizado con base en la interpretación de los resultados obtenidos mediante opinion mining, permitió obtener los siguientes resultados:

- Para el proceso de elecciones global (primera y segunda vuelta), la candidata Clara López obtuvo durante todo el proceso de análisis una polaridad global *Positiva (+),* determinada por la imagen favorable que reflejó para la mayoría de los ciudadanos que expresaron los comentarios recolectados.
- Para las elecciones en segunda vuelta, el candidato Juan Manuel Santos obtuvo la imagen más favorable, pero la polaridad global de los comentarios relacionados con él, fue *Negativa (-)*.
- El candidato Oscar Iván Zuluaga obtuvo durante la primera y la segunda vuelta una polaridad global de comentarios *Negativa (-)*, al igual que el candidato Juan Manuel Santos, sin embargo este último lo superó por cuanto obtuvo una imagen más favorable, según la minería de opiniones.
- La polaridad global *Negativa (-)* arrojada por el análisis de comentarios mediante los hashtag *#Elecciones2014* y *#SegundaVuelta*, durante la segunda vuelta presidencial, se deben a la predominación de comentarios negativos que fueron expresados por los colombianos, de acuerdo a su perspectiva mayoritariamente negativa con base en el proceso de elecciones presidenciales y sus actores involucrados.

Estos resultados demuestran que la minería de opiniones es actualmente una herramienta clave para la predicción de situaciones o fenómenos, por ejemplo en la política, y que puede ayudar mucho en la toma de decisiones, con base en diferentes opiniones.

El contraste entre el análisis general realizado, y la información pública disponible sobre las elecciones, tales como las encuestas realizadas por Gallup Colombia Ltda., y Datexco Company S.A., arrojó los siguientes resultados:

- Los resultados obtenidos mediante opinion mining en este trabajo, con respecto a la imagen de los candidatos durante la primera vuelta, sugirieron que el candidato Enrique Peñalosa obtendría la segunda mayor votación; estos resultados se relacionaron muy de cerca con los resultados publicados por la "Tercera gran encuesta para las elecciones presidenciales 2014", realizada por Datexco Company S.A. durante la primera vuelta.
- Los resultados obtenidos mediante opinion mining en este trabajo, para la segunda vuelta, estuvieron muy relacionados con los resultados de una nueva encuesta realizada por Datexco Company S.A., en donde se indicó que el candidato Juan Manuel Santos ganaría las elecciones en segunda vuelta, tras derrotar a su contendor Oscar Iván Zuluaga.

Estos últimos resultados parecen demostrar que la minería de opiniones puede ser utilizada en conjunto con las tradicionales encuestas realizadas por reconocidas firmas, para predecir resultados más cercanos a la realidad en determinadas situaciones, y así facilitar la toma de decisiones.

Todos los resultados descritos sugieren la confirmación de la hipótesis de investigación: las opiniones expresadas por los usuarios de servicios de microblogging y otros sitios web 2.0, contienen información necesaria y suficiente para realizar un análisis de los sentimientos de los colombianos.

5.2 Conclusiones

En general los resultados hallados en la presente investigación demuestran que es posible utilizar la minería de opiniones para conocer tendencias de un grupo poblacional, con base en el análisis de textos recuperados desde diferentes medios digitales de expresión.

Por otra parte, los resultados de un proceso de opinion mining siempre dependerán de importantes factores como la calidad del texto analizado, la adaptación (ajustes) de las herramientas de software de acuerdo al contexto en donde se lleve a cabo la investigación, y de la seriedad e imparcialidad del investigador al momento de recolectar las opiniones (textos).

Otras conclusiones obtenidas a través de este trabajo investigativo 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 porque permite a los usuarios comunes generar contenidos susceptibles de ser analizados.
- 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.

Capítulo 6

Recomendaciones y Trabajo Futuro

Mediante este capítulo se dan a conocer las contribuciones de este trabajo de investigación y la manera cómo éste puede servir de ayuda a otros investigadores, por medio de varias recomendaciones generales.

También se sugiere cómo se podría continuar con el desarrollo de esta investigación y de sus resultados, en el caso de que algunos investigadores deseen realizar algún tipo de extensión.

6.1 Contribuciones

A continuación se presentan las principales contribuciones realizadas mediante el desarrollo de este trabajo de investigación:

- Como aporte al estado del arte, es importante mencionar que este es el primer trabajo que se realiza en Colombia utilizando opinion mining con el objetivo de analizar los sentimientos de los colombianos frente a un proceso electoral.
- En esta investigación se generó una nueva base de conocimientos traducida al idioma español, con el objetivo de poder efectuar opinion mining en el contexto colombiano, mediante el software SentiStrength. Esta base de conocimientos contiene 2878 términos conformados por palabras, combinaciones de palabras, expresiones interrogativas, expresiones de negación y emoticonos, y plantea los cimientos para la creación de bases de conocimientos en castellano con el objetivo de realizar opinion mining frente a

procesos electorales, debido a que a la fecha no existía ninguna base de conocimientos disponible en español.

- Fue desarrollada una aplicación web que permite la captura de tweets relacionados con una persona o tema específico, utilizando hashtag o almohadillas (e.g., *#Elecciones2014*).
- Se diseñó y se alimentó una base de datos que contiene los tweets recolectados durante las elecciones presidenciales de Colombia en el año 2014, disponible para quienes deseen repetir esta investigación. Esta base de datos está conectada a la aplicación web mencionada.
- Se desarrolló un algoritmo que permite calcular la polaridad global de un fichero plano generado por SentiStrength, lo cual permite saber si la mayoría de los comentarios contenidos por dicho fichero son positivos, neutrales o negativos.
- Fue construido un análisis general, con base en la interpretación de los resultados obtenidos mediante opinion mining, que permitió comprender mejor los sentimientos, opiniones, emociones y tendencias de los colombianos frente a las elecciones presidenciales.
- Se realizó un contraste entre el análisis general realizado, e información pública disponible en encuestas realizadas por Gallup Colombia Ltda., y Datexco Company S.A., el cual intenta demostrar que la minería de opiniones puede ser utilizada para predecir resultados muy cercanos a la realidad en determinadas situaciones, facilitando de esa manera la toma de decisiones.

6.2 Recomendaciones

Para repetir esta investigación, o adelantar nuevos trabajos similares basados en sentiment analysis, se presentan a continuación varias recomendaciones:

- Es muy importante contextualizar la investigación y por ende saber escoger las fuentes desde las cuales se realizará extracción de textos para la posterior minería de opiniones.
- Para la captura o recolección de comentarios desde Twitter, es muy aconsejable indagar cuáles son las tendencias (#) que se están manejando en ese momento sobre el tema del que se desea realizar opinion mining.
- Es altamente recomendable almacenar en una base de datos los comentarios recolectados, con el objetivo de exportarlos posteriormente a ficheros con formato plano (e.g., .txt o .csv) para su procesamiento con un software para realizar opinion mining, como es el caso de SentiStrength.
- Se recomienda realizar algún tipo de separación por fechas en cuanto a los comentarios recolectados desde la Web, con el objetivo de analizar posteriormente el comportamiento del tema central de la investigación en determinados períodos de tiempo.
- En cuanto a los términos que conforman la base de conocimientos, es de vital importancia para los resultados de la investigación que sean asignadas con mucho cuidado las puntuaciones (scores) a cada palabra o término conformado por varias. También es importante agregar nuevas palabras a la base de conocimientos, sobre todo aquellas que son parte de un contexto cultural específico, como es el caso de las diferentes regiones de Colombia.

6.3 Trabajo Futuro

Como trabajo futuro se propone revisar y complementar la base de conocimientos, traducida del inglés al español y conformada por 2878 términos, que se entrega junto con este documento, 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 este maravilloso campo del sentiment analysis o minería de opiniones.

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] ARAÚJO, M., GONÇALVES, P., & BENEVENUTO, F. (2013). Measuring Sentiments in Online Social Networks. En *WebMedia 2013, actas del 19° simposio brasileño en Multimedia y la Web*, pág. 97-104.
http://dl.acm.org/citation.cfm?id=2526196

[3] BAEZA-YATES, R. & RIBEIRO-NETO, B. (2011). *Modern Information Retrieval: The Concepts and Technology behind Search (2a edición).* USA: Addison-Wesley.

[4] BALAHUR, A. (2011). *Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types*. Tesis de Doctorado, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Alicante, España.
http://rua.ua.es/dspace/bitstream/10045/19437/1/tesis_alexandrabalahur.pdf

[5] BALAHUR, A. & MONTOYO, A. (2008). Determining the Semantic Orientation of Opinions on Products - a Comparative Analysis. *Procesamiento del Lenguaje Natural,* vol. 41, pp. 201-208.
http://www.sepln.org/revistaSEPLN/revista/41/sec8-art1.pdf

[6] BANFIELD, A. (1982). *Unspeakable Sentences: Narration and Representation in the Language of Fiction*. Routledge & Kegan Paul.

[7] BERRY, M. & KOGAN, J. (2010). *Text Mining: Applications and Theory.* Chichester, Reino Unido: John Wiley & Sons Ltd.
http://gendocs.ru/docs/3/2304/conv_1/file1.pdf

[8] 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

[9] 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

[10] Ekman, P. (1992). An Argument for Basic Emotions. *Cognition & Emotion*, pág. 169-200.
http://www.paulekman.com/wp-content/uploads/2013/07/An-Argument-For-Basic-Emotions.pdf

[11] Ekman, P. (1999). Basic Emotions. En T. Dalgleish and M. Power (Eds.). Handbook of Cognition and Emotion. Sussex, U.K.: John Wiley & Sons, Ltd.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.123.1143&rep=rep1&type=pdf

[12] JOACHIMS, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. En *ECML 1998, actas de la 10ª conferencia europea en Machine Learning,* pág. 137-142.
http://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf

[13] KIM, S., & HOVY, E. (2005). Automatic Detection of Opinion Bearing Words and Sentences. En *IJCNLP, actas de la Conferencia Conjunta Internacional sobre Procesamiento del Lenguaje Natural*, Isla de Jeju, Corea del Sur.
http://acl.ldc.upenn.edu/I/I05/I05-2011.pdf?origin=publication_detail

[14] LIU, B. (2012). *Sentiment Analysis and Opinion Mining.* California, EE. UU.: Morgan & Claypool Publishers.
http://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

[15] 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

[16] NASUKAWA, T., & YI, J. (2003). Sentiment analysis: Capturing Favorability Using Natural Language Processing. En *K-CAP 2003, actas de la 2ª conferencia internacional sobre la captura del conocimiento*, pág. 70-77.
http://tredocs.com/tw_files2/urls_41/40/d-39217/7z-docs/7.pdf

[17] O'CONNOR, B., BALASUBRAMANYAN, R., ROUTLEDGE, B. & SMITH, N. (2010). From tweets to polls: Linking text sentiment to public opinion time series. En *ICWSM '10, actas de la conferencia internacional AAAI en Weblogs y Social Media*, pág. 122-129.
http://www.cs.cmu.edu/~nasmith/papers/oconnor+balasubramanyan+routledge+smith.icwsm10.pdf

[18] PANG, B. (2006). *Automatic Analysis of Document Sentiment*. Tesis de Doctorado, Escuela de graduados, Universidad Cornell, Ithaca, Nueva York.
http://www.cs.cornell.edu/home/llee/extra/pang-thesis.pdf

[19] PANG, B. & LEE, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval,* vol. 2(1-2): pp. 1-135.
http://www.cse.iitb.ac.in/~pb/cs626-449-2009/prev-years-other-things-nlp/sentiment-analysis-opinion-mining-pang-lee-omsa-published.pdf

[20] 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

[21] SARMENTO, L., CARVALHO, P., SILVA, M. & DE OLIVEIRA, E. (2009). Automatic creation of a reference corpus for political opinion mining in user-generated content. En *CIKM '09, actas del primer Workshop internacional sobre Temas de Sentiment Analysis para Opinión Masiva*, pág. 29-36.
http://xldb.lasige.di.fc.ul.pt/xldb/publications/sarmento09_document.pdf

[22] SCHERER, K. (2005). What are emotions? And how can they be measured? *Social Science Information*, vol. 3(44): pp. 695-729.
http://www.affective-sciences.org/system/files/biblio/2005_Scherer_SSI.pdf

[23] VAN DEN BOS, G. (2006). *APA Dictionary of Psychology*. Washington, DC, EE. UU.: American Psychological Association.

[24] WIEBE, J. (1994). Tracking Point of View in Narrative. *Computational Linguistics*, vol. 20(2): pp. 233-287.
http://acl.ldc.upenn.edu/J/J94/J94-2004.pdf?origin=publication_detail

[25] WIEBE, J., WILSON, T., & CARDIE, C. (2005). Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, vol. 39(2-3): pp. 119-122.
http://www.cs.cornell.edu/home/cardie/papers/lre05withappendix.pdf