Análisis de sentimientos sobre el posconflicto colombiano utilizando herramientas de minería de texto

Torres Samboni, Lina Andrea.
lina.torres@escuelaing.edu.co
Escuela Colombiana de Ingeniería Julio Garavito
Maestría en Gestión de Información

Resumen— En la sociedad actual, el volumen de información generada de forma digital gracias a la Web 2.0 es de gran tamaño, por ello se crea un interés por analizarla y obtener conocimiento nuevo con la ayuda de la minería de texto. Así es como en este proyecto se analizan 27.194 comentarios generados desde Twitter y páginas web en formato RSS para obtener valoraciones sobre el proceso de paz colombiano y lo que será el posconflicto una vez se firme el acuerdo de paz; el proceso de minería de texto se realiza con la ayuda de la herramienta Knime.

Índice de Términos— Generadores de opinión, medios de comunicación, minería de texto, posconflicto, procesamiento del lenguaje natural, proceso de paz, Twitter.

Abstract— In today's society, the volume of information generated digitally thanks to the Web 2.0 is large, so interest is created by analyzing and gain new knowledge with the help of text mining. That's how this project analize 27,194 comments generated from Twitter and web pages in RSS format for getting assessments of the Colombian peace process and what will be the post-conflict once a peace agreement is signed; the text mining process is done with the help of the Knime tool.

Keywords— Opinion makers, social media, text mining, posconflict, natural language processing, peace process, Twitter.

I. INTRODUCCIÓN

Actualmente Colombia está en el conflicto armado y el Gobierno estableció una mesa de conversación con la guerrilla de las FARC (Fuerzas Armadas Revolucionarias de Colombia), donde ya han transcurrido 3 años de diálogos y existen algunos acuerdos que a Diciembre de 2015 no se habían hecho públicos. A partir de la existencia de dichos acuerdos se han generado múltiples interrogantes sobre lo que será el posconflicto y la capacidad que tiene el país de asumirlo, entendiéndolo desde los diferentes frentes: sectores políticos, económicos, sociales, entre otros.

Todos los interrogantes que tiene la población colombiana sobre el posconflicto se reflejan en información subjetiva, es decir que cada persona tiene su propio juicio, valoración e interpretación sobre el tema. Entre esos puntos de vista, la valoración se puede clasificar en sentimientos positivos, negativos o neutros; siendo los sentimientos "la evaluación consciente que los seres humanos hacen de la percepción de su estado corporal durante una respuesta emocional" [5].

Dada la naturaleza y sensibilidad del tema en Colombia, se realizó un análisis de sentimientos a personajes provenientes de diferentes sectores de la vida pública porque día a día manifiestan su opinión a temas de la actualidad y además cada uno tiene un perfil o postura sobre las negociaciones que está haciendo el Gobierno. Los perfiles se examinan en los siguientes sectores de interés: políticos, medios de comunicación, medios de comunicación, generadores de opinión y guerrilleros.

El proyecto se desarrolla con la metodología de minería de opinión o como también es conocida, análisis de sentimientos [2], cuyo objetivo es analizar información no estructurada como el lenguaje de los humanos, para obtener información de la cual no existe un registro escrito. Luego, el objetivo del proyecto es perfilar la opinión positiva o negativa del posconflicto a partir de información no estructurada proveniente de los sectores que se acaban de mencionar, como método de análisis de tendencias en opinión.

II. Minería de datos y minería de texto

La minería de datos consiste en el análisis de grandes volúmenes de información estructurada que está contenida en bases de datos. Su esquema de organización permite una extracción de información más fácil para adquirir conocimiento de los datos originales. Luego, la minería de texto es una aplicación de la minería de datos y consiste en descubrir o hallar, a partir de cantidades de información no estructurada el conocimiento del cual no existe ningún registro escrito.

La información estructurada es aquella que no está contenida en un "almacén" o base de datos, de forma organizada para luego ser encontrada y utilizada fácilmente para distintos propósitos, lo cual dificulta su extracción. Esta información puede estar representada en textos como los mensajes de correo electrónico, presentaciones en power point, documentos en word, mensajes instantáneos (Twitter, Whatsapp), software de colaboración (conferencias de video); o en forma no textual en imágenes de formato JPEG, archivos de audio MP3, correos de voz, etc.

La principal característica de la minería de texto es que trabaja en base al lenguaje natural y éste es el que hablamos los humanos todos los días, es espontáneo, no es artificial y no ha sido programado de ninguna manera. Es así, como en los textos ya descritos se representa el lenguaje natural y la minería de texto se enfoca en el descubrimiento de patrones interesantes o sucesos recurrentes, su objetivo es descubrir tendencias, desviaciones y asociaciones en la gran cantidad de información textual disponible.

III. Minería de opinión (Opinion mining)

El análisis de sentimientos o minería de opinión (opinion mining) se encarga de analizar el procesamiento del lenguaje natural, que puede encontrarse representado en opiniones, sentimientos, puntos de vista, emociones, etc. Una de sus principales características consiste en que el tipo de información analizada además de ser subjetiva, generalmente es representada de forma escrita y sin un estándar o un formato pre establecido (información no estructurada). Dentro de este campo, existen aplicaciones que realizan un análisis más o menos profundo de los contenidos textuales, en función de la tarea o problema que se quiera resolver.

Una de las tareas relacionadas con la minería de opinión es la detección de la polaridad, cuya capacidad consiste en determinar si una opinión es positiva o negativa. Más allá de una polaridad básica, también se puede obtener un valor numérico dentro de un rango determinado, que de una determinada forma trate de obtener una valoración objetiva asociada a determinada opinión. Su principal característica es el uso de diccionarios semánticos y de la estructura sintáctica de las oraciones para clasificar un texto, sin embargo está técnica es muy dependiente de la calidad, el tamaño y dominio de los datos de entrenamiento [3].

La creciente importancia del análisis de sentimientos coincide con el crecimiento de los usuarios de internet y más específicamente de las redes sociales. Los medios sociales aportan cada día un gran conjunto de datos, que por primera vez en la historia humana permiten tener acceso a enormes volúmenes de datos almacenados en un formato digital para un análisis [4]; simultáneamente, se ha incrementado la atención por identificar el contenido emocional.

IV. Antecedentes del análisis de sentimientos

La minería de opinión ha despertado el interés de los investigadores para su análisis y es por ello que actualmente se encuentran varios estudios enfocados en diversas especialidades. A continuación se mencionan algunos de ellos, en los que se destaca el análisis de sentimientos con información que proviene de la Web 2.0 y que se distinguen por hacer un procesamiento del lenguaje natural en español:

"Linguistic Inquiry and Word Count" (LIWC) es una herramienta de software que analiza textos y fue diseñado por James W. Pennebaker, Roger J. Booth, y Martha E. Francis. LIWC es capaz de calcular cómo las personas usan diferentes categorías de palabras a través de una gran diversidad de textos. Ya sea en correo electrónicos, discursos, poemas o la trascripción de cualquier diálogo cotidiano, LIWC permite determinar el grado en que autores/hablantes usan palabras que

connotan emociones positivas o negativas, auto-referencias, palabras extensas o palabras que se refieren a sexo, comer o religión. El programa fue diseñado para analizar simple y rápidamente más de 70 dimensiones del lenguaje a través de cientos de muestras de texto en segundos.", [5].

La investigación "Estudio de las categorías LIWC para el análisis de sentimientos en español" consiste en el análisis de varias dimensiones lingüístico- psicológicas obtenidas desde LIWC para clasificar opiniones en español en cinco categorías. LIWC tiene un diccionario en español con 7.515 palabras y cada palabra se puede clasificar en 72 categorías, además las categorías se clasifican en cuatro dimensiones: procesos lingüísticos estándar, procesos psicológicos, relatividad y asuntos [6]. Este trabajo estuvo a cargo de investigadores españoles quienes definieron un corpus que fue obtenido de opiniones de productos electrónicos tales como dispositivos móviles. El corpus se procesó en LIWC [5] y la evaluación de resultados se hizo mediante clasificadores con los siguientes algoritmos J48, SMO, y BayesNet de WEKA. Estos algoritmos se basan en árboles de decisiones, que consisten en representación de funciones lógicas (if-then).

Por medio de aprendizaje automático se infiere un árbol de decisión a partir de un conjunto de instancias o ejemplos. El algoritmo "J48" de Weka utiliza un método heurístico para inferir el árbol, donde se realiza la selección del atributo en cada nivel del árbol en función de la calidad de la división que produce. Por otro lado, el algoritmo SMO es un clasificador hace uso del algoritmo secuencial de optimización minimal de John Platt para entrenar el clasificador de soporte vectorial.

Finalmente los resultados expresaron que la clasificación de opiniones con dos categorías (positiva, negativa) obtuvo mejores resultados, siendo el clasificador SMO el que tuvo un mejor comportamiento.

Por otro lado, la investigación "Análisis de Sentimientos y Minería de Opiniones: el corpus EmotiBlog Sentiment Analysis and Opinion Mining: The EmotiBlog Corpus" se basó en la carencia de recursos, métodos y herramientas para un análisis efectivo de la información subjetiva. El software utilizado fue EmotiBlog, el cual es una colección de entradas de blogs creado y anotado para detectar expresiones subjetivas en los nuevos géneros textuales nacidos con la Web 2.0. El principal reto fue demostrar que la creación de un corpus y con el uso de la implementación tecnológica EmotiBlog [8], se podían superar esos desafíos. Además, demostrar que con la ayuda de: SentiWordNet y WordNet [7], se aumentaría la cobertura de los resultados sin disminuir la precisión, incluyendo métodos de procesamiento del lenguaje natural (PNL) como stemming o lematización, bolsa de palabras, stop words, etc.

En la medida que se recolectó información sobre estudios previos en minería de opinión, se encontró que en gran parte de los experimentos se genera el corpus por medio de comentarios de Twitter, haciendo un filtro de cada palabra en el corpus de entrenamiento y prueba, dejando únicamente aquellas palabras que cumplan con una y solamente una etiqueta morfológica y el proceso de etiquetado lo hacen usando el etiquetador TreeTagger [9].

El último estudio a considerar es titulado "Herramienta para el Análisis de Sentimiento en el Proceso de Paz Colombiano", que consistió en la construcción de una herramienta para hacer el seguimiento de los comentarios generados en Twitter, con respecto al tema específico del proceso de paz colombiano. Los autores utilizaron las metodologías CRISP-DM y SCRUM, debido a que además de hacer minería de texto también se involucraba construcción de un software. Cabe anotar que SCRUM es una metodología ágil y flexible para gestionar el desarrollo de software, cuyo principal objetivo es maximizar el retorno de la inversión. Se basa en construir primero la funcionalidad de mayor valor para el cliente y en los principios de inspección continua, adaptación, auto-gestión e innovación.

Los autores construyeron un lexicón o un diccionario de términos de forma semiautomática a partir de palabras semillas, estas palabras son aquellas que para el autor están cargadas de significado, en este caso son las palabras definidas por un experto de dominio y, utilizaron calificaciones obtenidas de un diccionario

suministrado por la Sociedad Española para el Procesamiento del Lenguaje Natural. Para medir los resultados utilizaron las métricas de precisión y exhaustividad, que son empleadas en la medida del rendimiento de los sistemas de búsqueda, recuperación de información y reconocimiento de patrones; En este contexto se denomina precisión como a la fracción de instancias recuperadas que son relevantes, mientras recall o exhaustividad es la fracción de instancias relevantes que han sido recuperadas.

De esa forma, el análisis concluyó en un modelo que alcanzó una exhaustividad promedio aproximada de 57% y una precisión promedio de 61%, cuyos valores son aceptables para el modelo hibrido construido [10].

V. Diseño metodológico

El desarrollo del proyecto se hizo utilizando metodologías de minería de datos (CRIPS-DM) [11] y de minería de textos. A continuación se describe detalladamente cada etapa que se llevó a cabo:

A. Comprensión del negocio

En esta etapa inicial se comprende el objetivo del proyecto y se definen las fuentes de información. En esta metodología se seleccionaron personajes que generan interés de acuerdo a las referencias dadas por los siguientes artículos “21 Personajes de la Política Colombiana con más de 50.000 Followers” [12], el artículo de la revista Semana “20 mejores líderes de Colombia” [13] y el artículo de la revista Dinero “Los más influyentes” [14]; sin embargo, se trató de usar un criterio imparcial teniendo la certeza de que este listado se puede complementar en trabajos futuros. De esta forma se agruparon los personajes dentro de 5 categorías que fueron consideradas relevantes de acuerdo al contexto nacional en el que se producen las noticias e información importante del país:

Medios de comunicación en formato RSS: en esta categoría se encuentran los medios que publican noticias diarias en el formato RSS, el cual es un formato de datos que sirve para el envío de contenidos a quienes están registrados en un determinado sitio de Internet. Esta estructura permite que la distribución del contenido se realice sin que sea necesario valerse de un navegador, ya que la acción se lleva a cabo a través de un software creado especialmente para leer esta clase de datos que se conoce como agregador. En ese sentido, se escogieron las noticias generadas por los periódicos El Espectador y El Tiempo, ya que ofrecen este formato y además son diarios muy reconocidos a nivel nacional.

Comentarios en Twitter (Tweets) de guerrilleros colombianos: en la definición de este listado se seleccionaros guerrilleros de las FARC y el Ejército de Liberación Nacional (ELN) que tuvieran una cuenta en Twitter, puesto que ninguna publicación formal ha hecho un artículo o un comunicado donde filtren los personajes de este tipo con gran importancia e influencia en el país.

Comentarios en Twitter de políticos colombianos: para hacer esta selección se consultó un experto en Ciencias Sociales quien dio su concepto sobre los personajes de la vida política que despiertan interés para el análisis sobre el tema de estudio, adicional se utilizó la información del artículo de Caracol radio “Revelan ranking de los 10 políticos más influyentes en Colombia por Twitter” [15].

Comentarios en Twitter de los generadores de opinión: los generadores de opinión son aquellos periodistas o personalidades públicas que son líderes en opinión y están al tanto de las situaciones del país para expresar un comentario. Los personajes fueron seleccionados de acuerdo a la información del artículo “Los 8 periodistas colombianos más influyentes en Twitter” [26] y al artículo “¿Quiénes son los tuiteros más influyentes de Colombia?” de la revista Semana [16].

Comentarios en Twitter de medios de comunicación: en esta categoría se eligieron los principales medios de comunicación colombianos con un cubrimiento de todo el territorio nacional, de acuerdo a lo mencionado por el periódico el Tiempo en su artículo “EL TIEMPO, el perfil más influyente de Twitter en Colombia” [17], la lista de los medios de comunicación a nivel nacional del Centro de Prensa Internacional [18] y el artículo de Kien y Ke “Top 10 de Medios Colombianos en la Web” [19]. Asimismo, es importante aclarar que en esta categoría se les da un tratamiento a los medios de comunicación como institución y no como

persona natural, esto con el propósito de evitar relacionarlos con los periodistas que trabajan ahí.

B. Extracción de información

El flujo del modelo se inició con la extracción de información de cada una de las categorías definidas en la compresión del problema. Se utilizó una conexión en Twitter para poder descargar todos los registros y para ello se creó una cuenta que "siguiera (follow)" a las cuentas oficiales de todos los personajes que fueron seleccionados. Así, se configuró para que dicha cuenta siguiera a:

- 26 cuentas oficiales de guerrilleros.
- 40 cuentas oficiales de políticos.
- 34 cuentas oficiales de medios de comunicación.
- 13 cuentas oficiales de generadores de opinión.

Cada consulta tiene la siguiente estructura: "proceso de paz" + "Cuenta oficial de Twitter", de esta forma el texto: proceso de paz ClaraLopezObre, quiere decir que se va a hacer una consulta sobre la frase "proceso de paz" y que además coincida con la cuenta oficial de Clara López.

Se capturó información desde Junio hasta Octubre de 2015 para un total de 27,194 registros teniendo en cuenta que Twitter establece un límite de 180 consultas por cada 15 minutos y al excederlo se genera un error HTTP 429 "Too many request" y automáticamente bloquea las consultas.

C. Etiquetar el sentimiento y transformación

En esta etapa se etiquetan los sentimientos, es decir que se obtienen las valoraciones positivas o negativas de palabras que provienen de un diccionario de lexicón de emociones en español.

Inicialmente, en archivos de Excel se obtuvieron 844 palabras positivas y 1.194 palabras negativas de un diccionario [6], sin embargo una investigación de estas características es más enriquecedora cuando se obtiene un diccionario semántico más completo. Los siguientes son ejemplos de algunas palabras:

TABLA 1
DICCIONARIO DE PALABRAS POSITIVAS

Palabra	Categoría
abundancia	Alegría
acabalar	Alegría
acallar	Alegría
acatar	Alegría
acción	Alegría
aceptable	Alegría
aceptación	Alegría
acicate	Alegría
aclamación	Alegría
aclamar	Alegría

TABLA 2
DICCIONARIO DE PALABRAS NEGATIVAS

Palabra	Categoría
abominable	Enojo
abominación	Enojo
abominar	Enojo
aborrecer	Enojo
aborrecible	Enojo
aborreciblemente	Enojo
aborrecimiento	Enojo
abusar	Enojo
acometedor	Enojo
acometer	Enojo

Es así es como se reconocen las palabras dadas por el diccionario según la categoría a la que pertenezcan, se identifica que esas palabras estén contenidas en un Tweet y se les asigna la etiqueta "POSITIVE" o "NEGATIVE".

En esta parte del flujo ocurre la transformación y es donde se crea una bolsa de palabras "Bag of words". Todas las palabras de un documento se tratan como términos índices para ese documento o que es lo mismo como un conjunto de palabras claves. Además se asigna un peso a cada término en función de su importancia, determinada normalmente por su frecuencia de aparición en el documento. De este modo, no se toma en consideración el orden, la estructura, el significado, etc. de las palabras [20].

Este concepto consiste en la creación de un vector donde se muestra la frecuencia de palabras de un

texto y este modelo se implementa para obtener predicciones más precisas de opiniones, se representa de la siguiente forma:

D. Preprocesamiento de datos

El propósito del preprocesamiento de datos es principalmente corregir las inconsistencias de los datos que serán la base de análisis en procesos de minería. Con el preprocesamiento de datos se pretende que los datos que van a ser utilizados en tareas de análisis o descubrimiento de conocimiento conserven su coherencia, debido a que generalmente los datos vienen con ruido por diferentes razones, entre las cuales se encuentran [10]:

- Datos incompletos: valores faltantes para algunos atributos o sólo se tienen los datos agregados y no se cuenta con el detalle.
- Ruido: errores en los datos. Por ejemplo, manejar valores negativos para un atributo que maneja fechas.
- Inconsistencias: contiene discrepancias en los datos. Por ejemplo, edad de un empleado = 30 y fecha de nacimiento = 03/07/1998.

En muchas ocasiones, el origen de los problemas de los datos depende de la intervención humana, ya que en su momento pudieron cometer errores en la alimentación de las fuentes originales de los datos, en este caso, errores en la escritura de los Tweets.

Por otra parte, aunque la metodología CRISP-DM en su fase de preparación de los datos se ocupa de la transformación y limpieza de los datos, no desciende hasta el nivel de recomendar técnicas específicas dependiendo de la naturaleza de los datos [21].

Sin embargo en esta etapa se realizaron operaciones o transformaciones sobre un conjunto de documentos objetos de estudio, algunos autores la llaman Text Refining. Este paso es muy importante ya que, dependiendo del tipo de método usado en el preprocesamiento, así mismo es el contenido de los textos que darán origen a los patrones que se descubran. En esta etapa los documentos en el flujo toman el nombre de "Forma Intermedia".

Algunas de las técnicas utilizadas para la transformación de documentos en una forma intermedia pueden ser: análisis de texto, categorización, técnicas de procesamiento de lenguaje natural (etiquetado de parte del discurso, tokenización, lematización), técnicas de extracción de información (categorización, adquisición de patrones léxico sintáctico, extracción automática de términos, localización de trozos específicos de texto), técnicas de recuperación de información (indexación) [22].

Según lo anterior, se usaron técnicas de procesamiento del lenguaje natural y de extracción de información, algunas de ellas se mencionan a continuación:

Palabras que contienen números: si un texto tiene un valor numérico se filtra automáticamente.

Signos de puntuación: en los textos que aparezcan signos de puntuación, éstos se filtran.

Palabras con mínima cantidad de caracteres: si una palabra se compone por 3 o menos caracteres se debe eliminar.

Mayúsculas/minúsculas: si los textos son muy similares pero con letras mayúsculas o minúsculas indistintamente ('CASA ' vs' casa'), se convierte todo el texto a una sola forma.

Stop words: Si el texto contiene palabras dentro de la lista de "stop words", éstas se filtran.

Stemmer: hace uso de algoritmos de lematización, los cuales consisten en extraer los sufijos y prefijos comunes de palabras literalmente diferentes pero con una raíz común que pueden ser consideradas como un sólo término.

E. Frecuencias

Se extraen las palabras clave relevantes de los documentos, utilizando el enfoque basado en el grafo que se describe en el KeyGraph de Yukio Ohsawa. El algoritmo de KeyGraph se basa en la segmentación de un grafo, representando la co-ocurrencia entre los términos de un documento, en grupos (clusters). Cada clúster corresponde a los mejores términos calificados por una estadística, basada en cada relación de los términos seleccionados como palabras clave [23].

Luego se calcula la frecuencia de término relativo (tf) de cada término en función de cada documento y se agrega una columna que contiene el valor del

tf. El valor se calcula dividiendo la frecuencia absoluta de un término de acuerdo con un documento por el número de todos los términos de ese documento.

VI. Resultados y conclusiones

El análisis predictivo que se realizó de los perfiles de cada una de las categorías, alcanzó a recolectar un total de 27.194 registros de información entre los que se hizo una calificación de sentimientos positivos y negativos. Para obtener dichos resultados el método que se utilizó fue el aprendizaje no supervisado y fue construido a partir del etiquetado de los sentimientos con las palabras encontradas en el diccionario semántico. Haber realizado el análisis con ese modelo de supervisión hace que no haya una forma de validar con las métricas de exhaustividad y precisión los resultados, ni compararlos con otro corpus que haya sido verificado por juicio de expertos.

Por otro lado, este análisis tiene mayor valor que un pronóstico ingenuo donde se calcula la diferencia aritmética entre palabras positivas y negativas de un Tweet, porque al hacer uso del diccionario semántico se trabaja con la experiencia de los investigadores que han hecho un procesamiento del lenguaje natural serio en construirlo y en el análisis no se da lugar a la subjetividad de quien lo realiza.

En la medida que se quiera realizar la predicción con un método de aprendizaje supervisado, debe dividirse el corpus completo entre un corpus de entrenamiento y de prueba con respecto a un valor de porcentaje definido. El valor de porcentaje se utiliza para determinar el tamaño del corpus de prueba, por lo general, se toma del 70 al 80 por ciento para usarlos como corpus de entrenamiento y los registros sobrantes se toman como el corpus de prueba; de esa manera se minimizan los errores cometidos permitiendo que exista una forma de validar los resultados y obtener métricas que definan la veracidad de los resultados.

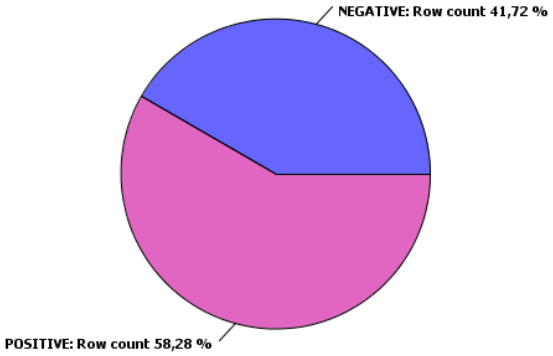
Otra forma de aplicar el aprendizaje supervisado es crear el corpus de entrenamiento a partir del juicio de expertos. El juicio de expertos es un método de validación útil para verificar la fiabilidad de una investigación que se define como “una opinión informada de personas con trayectoria en el tema, que son reconocidas por otros como expertos cualificados en éste, y que pueden dar información, evidencia, juicios y valoraciones” [24]. Tras someter el diccionario semántico a la consulta y al juicio de expertos éste debe reunir dos criterios de calidad: validez y fiabilidad. La validez de contenido se establece con frecuencia a partir de dos situaciones, una que concierne al diseño de una prueba y, la otra, a la validación de un instrumento sometido a procedimientos de traducción y estandarización para adaptarlo a significados culturales diferentes. Es aquí donde la tarea del experto se convierte en una labor fundamental para eliminar aspectos irrelevantes, incorporar los que son imprescindibles y/o modificar aquellos que lo requieran [25].

El juicio de expertos consistiría en etiquetar los sentimientos a partir de opiniones que puedan brindar profesionales expertos en la disciplina relacionada al tema que se desarrolla, en este caso se aplicaría un juicio de expertos del lenguaje natural y de ciencias políticas. Una vez se tenga calificado el diccionario se aplicaría el procedimiento explicado anteriormente para realizar un análisis con un método de aprendizaje supervisado. Ahora, en esta investigación no se aplicó el juicio de expertos por la ausencia de los mismos entre los recursos interesados en el desarrollo del proyecto.

Se identificó que la mayoría de registros son originados por los medios de comunicación, sin importar que en el inicio de la extracción de información hubiera mayor cantidad de registros configurados de políticos en Twitter o de guerrilleros que de los medios de comunicación; correspondiendo estos últimos a un 86,36% de la información (23.484 registros). Según lo anterior, se puede afirmar que son la principal fuente de información y debido al volumen que generan tienen una gran presencia en el pueblo colombiano.

En la figura 1 se muestra que la mayoría de los registros producto del análisis fueron identificados como positivos con un 58.28% pero que no se alejan mucho del porcentaje restante negativo.

FIGURA 1
PORCENTAJE DE OPINIONES POSITIVAS Y NEGATIVAS



Y para obtener los perfiles de cada categoría se utilizó "Tag cloud". Un tag cloud ("nube de etiquetas" en español) es una forma de representar la información. En este caso representa las raíces (stemming) de las palabras que el análisis de sentimientos identificó como positivas y negativas de acuerdo a la importancia (score) con que fueron calificadas al calcular la frecuencia. Las palabras se muestran mediante la manipulación de las propiedades visuales, se utiliza la distribución de tamaño de fuente, la transparencia y la audacia de la fuente.

Se obtuvo la gráfica a continuación:

FIGURA 3
TAG CLOUD: TODAS LAS CATEGORÍAS

La raíz que predomina es *final*, sin embargo en general se muestra un equilibrio entre los sentimientos y positivos.

Para cada categoría, la palabra *final* tiene mayor importancia y por el color se puede concluir que es una opinión que se ubica entre los políticos y los guerrilleros, es decir, como un consenso de que para ambos lo más importante es el final del conflicto. Por otro lado, se observa que los generadores de opinión son totalmente escépticos con la situación que se va presentar en el posconflicto, puesto que todas las palabras que los representan son negativas aunque no tienen mayor influencia sobre la generalidad de las palabras porque son muy pocas. Finalmente, quienes tienen mayores opiniones son los medios de comunicación aunque la baja tonalidad del color verde generado demuestra que no tienen mayor importancia, además tienen opiniones mixtas de las que no se logra diferenciar alguna tendencia hacia lo positivo o negativo.

Para el análisis del perfil de los medios de comunicación, el tag cloud muestra una mayoría de palabras negativas concentradas alrededor de la palabra final, lo que indica que esta categoría está orientada hacia una expectativa de que de termine la guerra y la violencia, sin olvidar la reparación de las víctimas.

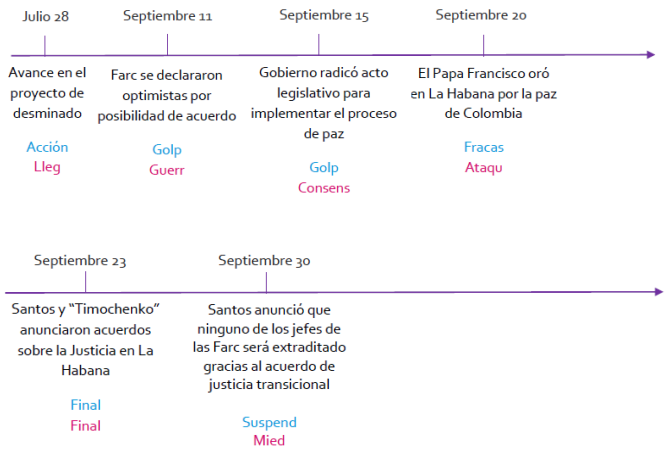
Por último, el análisis de sentimientos de los guerrilleros obtenido de otro tag cloud y de igual forma que la categoría de los generadores de opinión, la poca cantidad de informacion representada no permite extraer un análisis más profundo, pero da unas bases de lo que en principio son las opiniones de esta categoría. En esta nube de etiquetas sólo hay una palabra positiva y las palabras restantes son negativa, las cuales van encaminadas a que los guerrilleros tienen especial cuidado al futuro relacionado con quiénes deben estar en la cárcel y la re inserción a la sociedad de su ejército, teniendo en cuenta el rechazo y odio que pueden recibir de la sociedad civil.

Luego de analizar los perfiles de cada categoría, se seleccionaron hitos que se han presentado en el proceso de paz desde Julio hasta Octubre de 2015 según la infografía del periódico El Tiempo "Línea del tiempo de los diálogos de paz", para identificar cuáles sentimientos se producían entre las categorías y observar las diferencias según un acontecimiento. De esta forma se obtuvo lo siguiente:

Para hallar dichos sentimientos sólo se analizaron las fechas específicas según los hitos. Las categorías de las que se extrajo información de esas fechas se resaltan con los siguientes colores:

Medios de comunicación
Políticos

FIGURA 4
HITOS DEL PROCESO DE PAZ



En cada hito se observa que los sentimientos de los medios de comunicación tienden a expresar que cada acontecimiento es una decaída del proceso de paz sin importar que los hitos puedan significar acciones positivas por parte de la Guerrilla o del Gobierno. Además, dan un aire de poca esperanza en el éxito del proceso porque pareciera que cualquier hecho que ocurra va a ser algo negativo o de mucha incertidumbre.

Por otro lado, los políticos se muestran en intervalos de sentimientos negativos, muy negativos y sólo uno positivo. Existe indecisión e inseguridad con respecto al proceso de paz pero según la línea de tiempo, el sentimiento de menor incertidumbre es el consenso sobre el acto legislativo para implementar el proceso de paz, luego, el resto de los sentimientos se interpretan como si estuvieran a la defensiva de cualquier cosa que ocurre y quisieran proteger sus intereses.

Finalmente, es importante mencionar que el propósito de este trabajo es generar una herramienta que a través del tratamiento de información y haciendo uso del análisis de sentimientos de un aporte para profesiones como periodismo, comunicación social y ciencias políticas, de tal manera que la labor de generar una valoración sobre un tema específico tenga mayor soporte investigativo y tengan un mayor acercamiento al problema que están tratando.

VII. TRABAJOS FUTUROS

A partir del conocimiento generado en este trabajo, se puede profundizar en mejorar la extracción masiva de informacion no estructurada desde redes sociales diferentes de Twitter, páginas web dinámicas, links, imágenes, audios y otras fuentes que no fueron contempladas. De igual forma llama la atención incluir en el procesamiento del lenguaje natural el uso de emoticones en los textos, que pueden ser manejados como “Stop words” o que ayuden a alimentar el diccionario semántico asignándoles sentimientos positivos o negativos.

También existe la posibilidad de crear la programación de la extracción de información de tal manera que el proceso sea diario, semanal o mensual de forma automática, teniendo en cuenta los límites de consultas de registros, lapsos de tiempo y permisos que establecen los buscadores.

En cuanto a la próxima selección de un diccionario semántico de sentimientos se puede complementar tomando como referencia un diccionario especializado en inglés que tenga especificado un rango de valores para cada palabra y traducirlo al español, así se construiría un diccionario enriquecido complementando el que existe ahora.

Y por último, es interesante incluir el manejo del sarcasmo en el procesamiento del lenguaje natural porque aunque en este trabajo se hizo una aproximación sobre este aspecto fue muy básico y no se profundizo, así se obtendría un aprendizaje supervisado más acertado.

RECONOCIMIENTO

Este proyecto ha sido fruto del aporte de conocimiento y propuestas de desarrollo por parte del Director de Proyecto.

REFERENCIAS

[1] Española, R. A. (2015). Significado de subjetivo. Recuperado el 17 de Septiembre de 2015, de http://www.significados.com/subjetivo/

[2] Cubero, J. C., & Berzal, F. (s.f.). Sistemas Inteligentes de Gestión - Guía de prácticas de la minería de datos. Granada, España. Recuperado el 13 de Septiembre de 2015, de http://elvex.ugr.es/decsai/intelligent/workbook/D1%20KNIME.pdf

[3] Rodríguez Rojas, O. (s.f.). Metodología para el desarrollo de proyectos en minería de datos CRISP-DM. Recuperado el 04 de Octubre de 2015, de Oldemar Rodríguez, Consultor en minería de datos: http://oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf

[4] Rangel, I. D., Sidorov, G., & Suárez-Guerra, S. (2014). Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. Recuperado el 15 de Agosto de 2015, de http://www.cic.ipn.mx/~sidorov/sel_onomazein_2014_29.pdf

[5] Associates, L. E. (2013). Linguistic Inquiry and Word Count. Obtenido de http://liwc.net/liwcespanol/index.php

[6] Sidorov, G. (07 de Diciembre de 2012). Resource available: Spanish Emotion Lexicon. Recuperado el 16 de Agosto de 2015, de Resource available: Spanish Emotion Lexicon: http://permalink.gmane.org/gmane.science.linguistics.corpora/16865

[7] Baccianella, S., Esuli, A., & Sebastiani, F. (s.f.). SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Istituto di Scienza e Tecnologie dell'Informazione, Pisa.

[8] Gálvez-Pérez, J. R., Gómez-Torrero, B., Ramírez-Chávez, R. I., Sánchez-Sandoval, K. M., Castellanos-Cerda, V., García-Madrid, R., . . . Villatoro-Tello, E. (2015). Sistema automático para la clasificación de la opinión pública generada en Twitter. Recuperado el 15 de Agosto de 2015, de Research in Computing Science 95: http://www.micai.org/rcs/2015_95/Sistema%20automatico%20para%20la%20clasificacion%20de%20la%20opinion%20publica%20generada%20en%20Twitter.pdf

[9] Minería de textos. Sistemas avanzados de recuperación de la información. (2010). Recuperado el 2015 de Septiembre de 17, de http://mineriadetextos.tripod.com/

[10] Hernández G., C. L., & Rodríguez R., J. E. (20 de Abril de 2008). Preprocesamiento de datos estructurados. Structured Data Preprocessing. Revista vínculos, 27-48. Recuperado el 18 de Octubre de 2015, de http://revistavinculos.udistrital.edu.co/files/2012/12/Preprocesamiento.pdf

[11] Chapman, P., Clinton, J., & Kerber, R. (2000). CRISP-DM 1.0 Step by step data mining guide. Recuperado el 22 de Septiembre de 2015, de https://the-modeling-agency.com/crisp-dm.pdf

[12] Suarez, W. (14 de Marzo de 2013). Prospectador. Vigía estratégica de Redes Sociales. Recuperado el 30 de Agosto de 2015, de Prospectador. Vigía estratégica de Redes Sociales: http://prospectador.co/21-personajes-de-la-politica-colombiana-con-mas-de-50-000-followers/

[13] Semana. (31 de Agosto de 2013). Semana. Recuperado el 30 de Agosto de 2015, de Semana: http://www.semana.com/nacion/articulo/20-mejores-lideres-colombia/355903-3

[14] Dinero. (04 de Abril de 2014). Dinero. Obtenido de Dinero: http://www.dinero.com/edicion-impresa/pais/articulo/los-personajes-mas-influyentes-colombia/194261

[15] Radio, C. (08 de Julio de 2012). Caracol Radio. Obtenido de Caracol Radio: http://caracol.com.co/radio/2012/08/07/nacional/1344321600_738497.html

[16] Semana. (20 de Julio de 2015). Semana. Obtenido de Semana: http://www.semana.com/tecnologia/articulo/quienes-son-los-tuiteros-mas-influyentes-de-colombia/435580-3

[17] Tecnósfera, R. (21 de Julio de 2015). El Tiempo. Recuperado el 30 de Agosto de 2015, de El Tiempo: http://www.eltiempo.com/tecnosfera/novedades-tecnologia/tuiteros-mas-influyentes-de-colombia/16123675

[18] Internacional, C. d. (18 de Febrero de 2013). Centro de Prensa Internacional. Recuperado el 30 de Agosto de 2015, de Centro de Prensa Internacional: http://wsp.presidencia.gov.co/cepri/medios-colombia/Paginas/default.aspx

[19] KienYKe. (15 de febrero de 2012). Kien y Ke. Recuperado el 30 de Agosto de 2015, de Kien y Ke: www.kienyke.com

[20] Vallez, M., & Pedraza-Jimenez, R. (2007). Anuario académico sobre documentación digital y comunicación interactiva. Recuperado el 12 de Octubre de 2015, de http://www.upf.edu/hipertextnet/numero-5/pln.html

[21] Uribe, I. A., & Jiménez Ramírez, C. (09 de Junio de 2009). Hacia una metodología para la selección de técnicas de depuración de datos. (U. N. Colombia, Ed.) Avance en sistemas e informática, Vol.6 ¬ No.1, 185-190. Obtenido de www.redalyc.org

[22] Iribarra, F. (Junio de 2013). Minería de Datos. Universidad Tecnológica Metropolitana. Recuperado el 18 de Octubre de 2015, de http://mineriadatos1.blogspot.com.co/2013/06/etapas-de-la-mineria-de-texto.html

[23] Ohsawa, Y., Benson, N. E., & Yachida, M. (s.f.). KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor. Tesis, Dept. Systems and Human Science, Graduate School of Engineering Science Osaka University,, Toyonaka. Recuperado el Octubre de 2015, de 06

[24] Escobar-Pérez, J., & Angela, C.-M. (15 de Febrero de 2016). Validez de contenido y juicio de expertos: una aproximación a su utilización. Obtenido de Validez de contenido y juicio de expertos: una aproximación a su utilización: http://www.humanas.unal.edu.co/psicometria/files/7113/8574/5708/Articulo3_Juicio_de_expertos_27-36.pdf

[25] Garrote, P. R., & Rojas, M. d. (06 de Marzo de 2016). Revista Nebrija. Obtenido de Revista Nebrija: http://www.nebrija.com/revista-linguistica/la-validacion-por-juicio-de-expertos-dos-investigaciones-cualitativas-en-linguistica-aplicada

[26] Barrero, J. S. (31 de Mayo de 2015). *Carreta Digital*. Obtenido de Carreta Digital: http://carretadigital.com/index.php/destacado/item/292-los-8-periodistas-colombianos-mas-influyentes-en-twitter

Lina Andrea Torres Samboni
Ingeniero de Sistemas de la Escuela Colombiana de Ingeniería Julio Garavito, con intereses en las áreas de Gestión de información, Gestión del conocimiento, Arquitectura empresarial, entre otras.