# ANÁLISIS DE LA OPINIÓN DE PERSONALIDADES INFLUYENTES DE LA POLÍTICA COLOMBIANA A TRAVÉS DE TÉCNICAS DE ANÁLISIS DE SENTIMIENTO

**JESÚS MARIO RODRÍGUEZ VELÁSQUEZ**

**PROGRAMA DE INGENIERIA DE SISTEMAS**
**FACULTAD DE INGENIERIA**


**UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA**
BUCARAMANGA, mayo 31 de 2019

# ANÁLISIS DE LA OPINIÓN DE PERSONALIDADES INFLUYENTES DE LA POLÍTICA COLOMBIANA A TRAVÉS DE TÉCNICAS DE ANÁLISIS DE SENTIMIENTO

**JESÚS MARIO RODRÍGUEZ VELÁSQUEZ**
**Email: jrodriguez862@unab.edu.co**

**Director: MARITZA LILIANA CALDERÓN BENAVIDES DIRECTOR**
**Email: mcalderon@unab.edu.co**

**PROGRAMA DE INGENIERIA DE SISTEMAS**
**FACULTAD DE INGENIERIA**
**UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA**

Bucaramanga, mayo 31 de 2018

**TABLA DE CONTENIDO**

**LISTA DE TABLAS**

## LISTA DE ILUSTRACIONES

## LISTA DE GRÁFICOS

# 1. INTRODUCCIÓN

En el presente documento se presenta trabajo en el que se pretende realizar un Análisis de Sentimiento a las personalidades más influyentes de la política colombiana para determinar el sentimiento con el que se trasmite la información en las redes sociales pues se está generando confusión entre los ciudadanos al momento de la toma de decisiones debido a la desinformación.

## 2. PLANTEAMIENTO DEL PROBLEMA

En los últimos años, con el auge de redes sociales, algunas plataformas como Twiter [1] se han convertido en importantes distribuidores de información debido a la facilidad de creación y difusión de información ya que cada usuario que pertenece una red social puede postear cualquier tipo de información, sea de su autoría o no [2], y tener cierto impacto en la población que recibe dicha información; hay que tener en cuenta que según algunos estudios la mayoría de retweets de un tema en específico se producen en la primera hora, y en un 35% de los casos en los primeros 10 minutos, generando un alto impacto entre los usuarios [3]. El problema surge cuando se desconoce el sentimiento con el que se difunde este tipo de información, pues no siempre se puede asegurar que la información trasmitida es real; además, hay que tener en cuenta que muchos usuarios de las redes sociales, en su afán de obtener una popularidad efímera (retweets, likes, etc.) y así afianzar sus burbujas de autoafirmación [4] buscan trasmitir información que genere polémica y debate para que sea difundida entre los usuarios rápidamente, usuarios que ignoran totalmente el sentimiento con el que está trasmitiendo dicha información, y así estas personas obtienen la popularidad que tanto buscan. Finalmente, como la política genera ansias de poder entre los mismos políticos [5], estos buscan trasmitir todo tipo de información- generalmente en contra de sus "rivales" políticos y a favor de sus "aliados"-, valiéndose de su influencia y sus miles de seguidores en las redes sociales generando, en muchas ocasiones difamación debido a que la información que trasmiten está viciada y no corresponde a la realidad [6], estos miles de seguidores, en su gran mayoría ignoran el sentimiento con el que se conformó dicha información y simplemente se dedican a difundirla, generando polarización entre los ciudadanos debido a que el objetivo de esta información es generar una respuesta emocional en los usuarios que ignoran que la información es verídica o no [7]. Además, el acto de difundir una información falsa es relativamente fácil y rápido, sin embargo, desacreditar esta información falsa es un proceso sumamente lento, tedioso y complicado [4] debido a que las personas no siempre terminan confiando en la desacreditación que se hace, sino que se quedan con la primera información que recibieron [8]. Lo anterior está generando polarización política, conflictos sociales y finalmente confusión de los ciudadanos en la toma de decisiones [7]. En el marco colombiano, según la web www.socialbakers.com, se encuentra que las personalidades de la política colombiana con más número de seguidores son [9]:

> Juan Manuel Santos Calderón (@JuanManSantos), presidente de la República (2010 – 2018) con 5'303.917 seguidores.
>
> Álvaro Uribe Vélez (@AlvaroUribeVel), presidente de la República (2002 – 2010) y senador de la República (2014 – presente) con 4'767.865 seguidores.
> Gustavo Francisco Petro Urrego (@petrogustavo), alcalde Mayor de Bogotá (2012 – 2015), candidato presidencial en las elecciones de 2018 y senador de la República (2018 – presente) con 3'280.436 seguidores.

Antanas Mockus Šivickas (@AntanasMockus), alcalde Mayor de Bogotá (2001 – 2003) y senador de la República (2018 – presente) con 1’877.870 seguidores.

Enrique Peñalosa Londoño (@EnriquePenalosa), alcalde Mayor de Bogotá (2016 – presente) con 1’851.468 seguidores.

Sergio Fajardo Valderrama (@sergio_fajardo), gobernador de Antioquia (2012-2015), alcalde de Medellín (2004-2007) y candidato presidencial en las elecciones de 2018 con 1’368.757 seguidores.

Claudia López Hernández (@ClaudiaLopez), senadora de la República (2014-2018) y candidata vicepresidencial en las elecciones de 2018 con 1’181.839 seguidores.

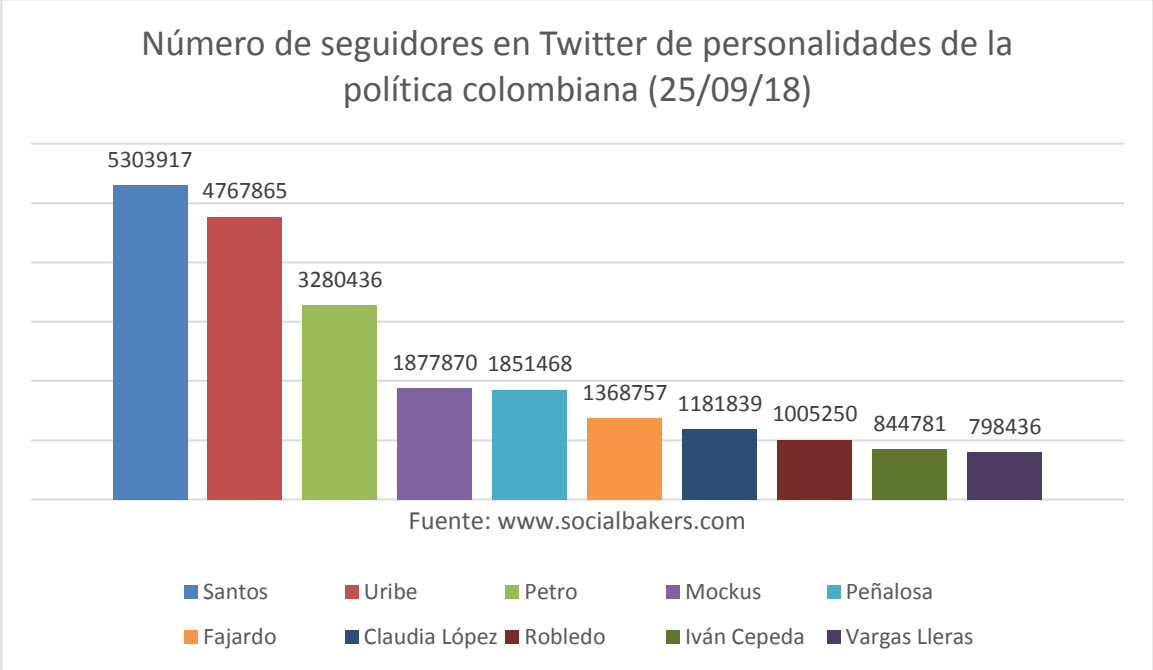
Jorge Enrique Robledo Castillo (@JERobledo), senador de la República (2002-presente) con 1’005.250 seguidores.

Iván Cepeda Castro (@IvanCepedaCast), senador de la República (2014-presente) con 844.781 seguidores.

Germán Vargas Lleras (@German_Vargas), Vicepresidente de la República (2014-2017) y candidato presidencial en las elecciones de 2018 con 798.436 seguidores.

Datos tomados al 25 de septiembre de 2018. Ver gráfico 1.

**Gráfico 1. Número de seguidores en Twitter de personalidades de la política colombiana (25/09/2018).**

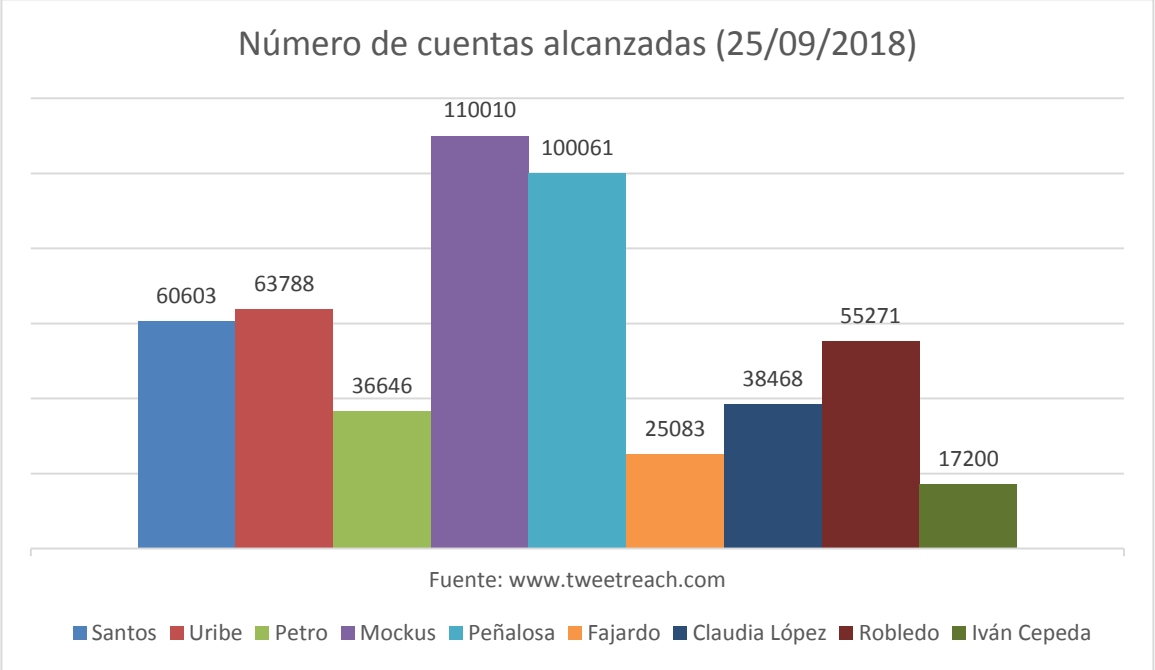



Fuente: www.socialbakers.com [9].

El hecho de ser influyente se refleja con estadísticas tales como el número de cuentas de Twitter alcanzadas por un usuario y el número de impresiones que tiene un usuario, la principal diferencia entre ambos radica en que el alcance es entendido como el número de personas que ven el contenido publicado, es la medida de audiencia efectiva; y la impresión es el número de veces en el que el contenido se muestra en la Timeline de algún usuario, un espectador no tiene que interactuar con una publicación para que esto cuente como una interacción [10]. Para efectos prácticos se entiende que a mayor número de alcance e influencia un usuario será más influyente.

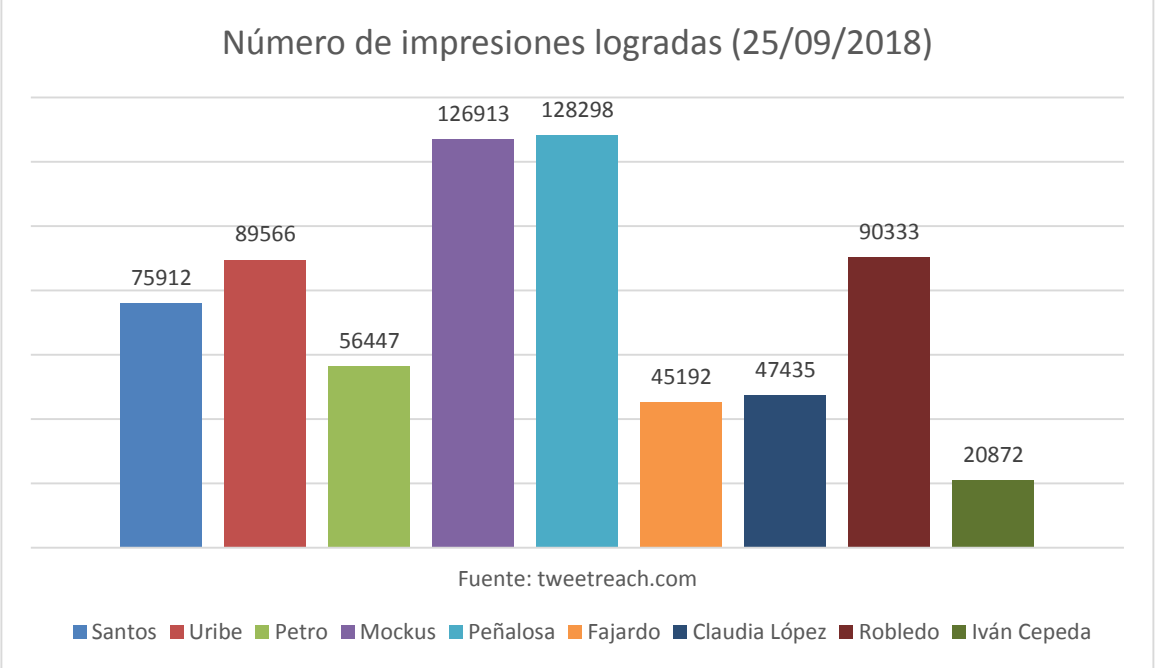
Al realizar la consulta, mediante la herramienta www.tweetreach.com [11], del número de cuentas alcanzadas e impresiones de las personalidades de la política colombiana se encuentra que Juan Manuel Santos Calderón (@JuanManSantos), personalidad política colombiana con más seguidores en Twitter actualmente, ha obtenido con un alcance en el último mes de cerca de 60.673 cuentas y obteniendo aproximadamente 75,912 impresiones en los tweets publicados. A continuación, se muestran dos gráficos con los alcances e impresiones aproximados para las personalidades de la política colombiana con más seguidores en Twitter. Ver gráfico 2 y ver gráfico 3.

**Gráfico 2. Número de cuentas alcanzadas en Twitter de personalidades de la política colombiana (25/09/2018).**




Fuente: www.tweetreach.com [11].

**Gráfico 3. Número de impresiones logradas en Twitter por personalidades de la política colombiana (25/09/2018).**




Fuente: www.tweetreach.com [11].

Por lo anterior se propone que las personalidades a las que se les va a realizar el Análisis de Sentimiento sean las que cuentan con más influencia en Twitter en relación al alcance e impresiones. Se escogen a cinco políticos:

Juan Manuel Santos Calderón (@JuanManSantos).

Álvaro Uribe Vélez (@AlvaroUribeVel).

Gustavo Francisco Petro Urrego (@gustavopetro).

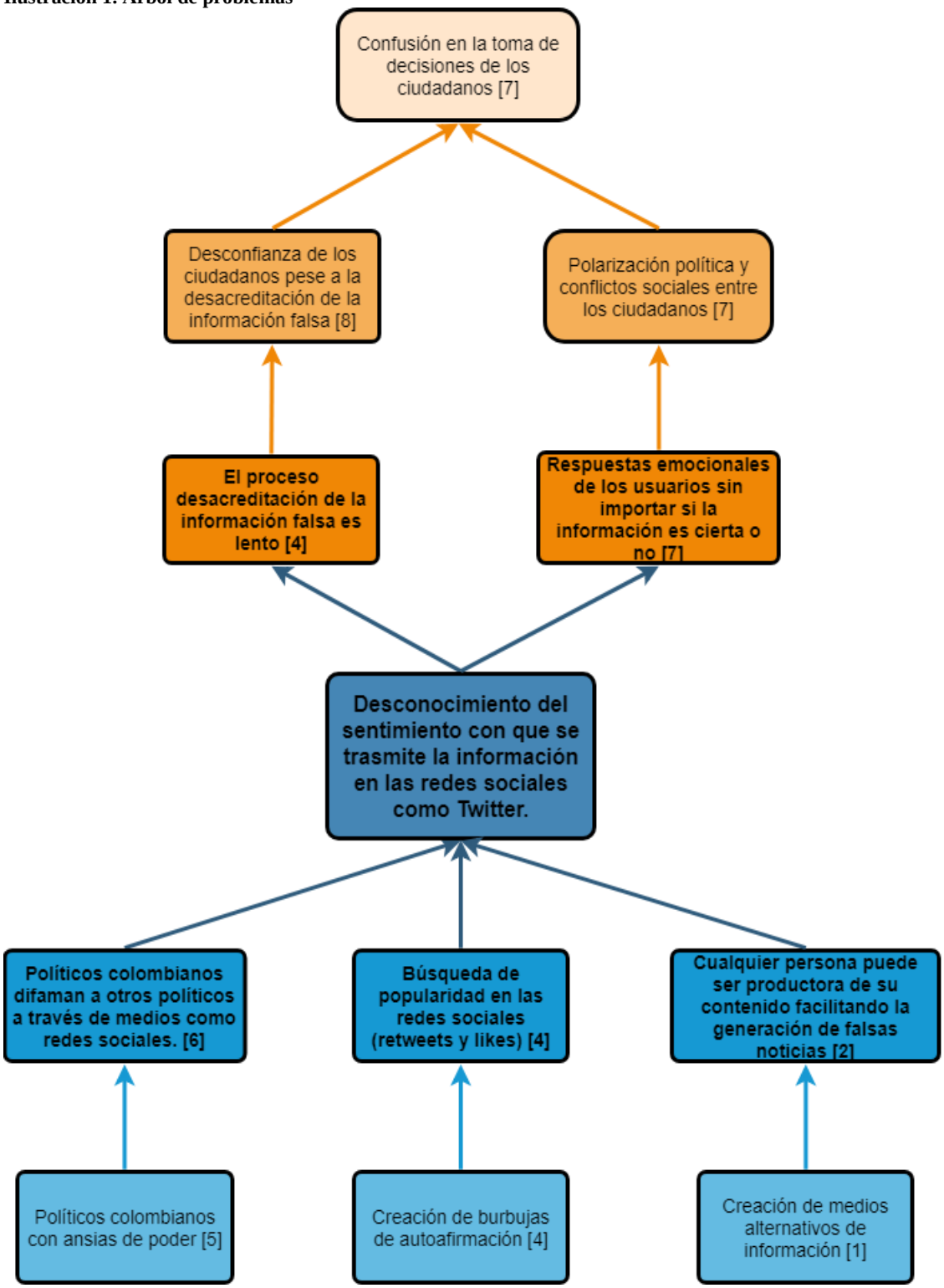
Antanas Mockus Šivickas (@AntanasMockus).

Claudia López Hernández (@ClaudiaLopez).

El descarte de otros políticos, pese a que tienen más seguidores o influencia, se da para la inclusión de Claudia López, porque también se quiere analizar qué tipo de sentimiento suscita una mujer en la política colombiana, un campo que generalmente se ha visto representado en su mayoría por hombres.

## 2.1 ÁRBOL DE PROBLEMAS

**Ilustración 1. Árbol de problemas**



## 3. JUSTIFICACIÓN

El desconocimiento del sentimiento con el que se trasmite la información de carácter político en las redes sociales está generando polarización política, debates innecesarios, conflicto ideológico, fomentado por las respuestas emocionales [7] entre los usuarios de estos medios. Lo anterior dictamina la necesidad de determinar el sentimiento que tienen las opiniones trasmitidas por personalidades influyentes de la política colombiana en redes sociales, en este caso Twitter, esto se pretende lograr usando técnicas de Análisis de Sentimiento. Es importante estudiar este fenómeno y proponer una forma de conocer el sentimiento con el que se trasmite la información en las redes sociales pues se está generando confusión entre los ciudadanos al momento de la toma de decisiones.

## 4. PREGUNTA DE INVESTIGACIÓN

¿Cómo se puede determinar el sentimiento que tienen las opiniones trasmitidas por personalidades influyentes de la política colombiana en redes sociales como Twitter?

## 5. HIPÓTESIS

A través del uso de técnicas de Análisis de Sentimiento es posible determinar el sentimiento que tienen las opiniones trasmitidas por personalidades influyentes en la política colombiana en redes sociales como Twitter.

# 6. OBJETIVOS

## 6.1 OBJETIVO GENERAL

Determinar el sentimiento que tienen las opiniones trasmitidas por personalidades influyentes en la política colombiana en redes sociales como Twitter, a través del uso de técnicas de Análisis de Sentimiento

## 6.2 OBJETIVOS ESPECÍFICOS

- Elaborar un estado del arte de herramientas, métodos y aplicaciones que implementen Análisis de Sentimiento.
- Diseñar una herramienta de software que implemente Análisis de Sentimiento.
- Desarrollar una herramienta de software que implemente Análisis de Sentimiento.
- Realizar el Análisis de Sentimiento de opiniones expresadas por personalidades influyentes de la política colombiana.

## 7. RESULTADOS ESPERADOS

- Documento con el estado del arte de herramientas, métodos y aplicaciones que implementen Análisis de Sentimiento.
- Documento con el diseño de una herramienta de software que implemente Análisis de Sentimiento.
- Una herramienta de software que implemente Análisis de Sentimiento.
- Documento con el Análisis de Sentimiento de personalidades influyentes de la política colombiana.

# 8. MARCO TEÓRICO

A continuación, se presentan algunos términos y conceptos relacionados con el desarrollo del presente trabajo.

## 8.1 MARCO CONCEPTUAL

### 8.1.1 Sentiment Analysis [12].

El Análisis de Sentimiento, también llamado opinión mining, es el campo de estudio que analiza opiniones, sentimientos, evaluaciones, actitudes y emociones de las personas hacia entidades tales como productos, servicios, organizaciones, individuos, problemas, eventos, temas y sus atributos

El término Análisis de Sentimiento apareció por primera vez en Nasukawa y Yi de 2003, y el término opinión mining apareció por primera vez en Dave, Lawrence y Pennock de 2003.

Para determinar el sentimiento de un texto se utilizan diferentes niveles, a continuación, se describen.

#### 8.1.1.1 Niveles de análisis.

##### 8.1.1.1.1 Nivel de documento.

El objetivo en este nivel es clasificar si un documento de opinión expresa un sentimiento positivo o negativo. Por ejemplo, dada una revisión de un producto, el sistema determina si la revisión expresa una opinión positiva o negativa sobre el producto. Este nivel de análisis supone que cada documento expresa opiniones sobre una sola entidad (por ejemplo, un único producto). Por lo tanto, no es aplicable a los documentos que evalúan o comparan varias entidades.

##### 8.1.1.1.2 Nivel de oración.

Determina si cada oración de un texto expresó una opinión positiva, negativa o neutral.

##### 8.1.1.1.3 Nivel de entidad y aspecto.

El nivel de aspecto mira directamente la opinión en sí misma. Se basa en la idea de que una opinión consiste en un sentimiento (positivo o negativo) y un objetivo (de opinión). Por ejemplo:

"*Aunque el servicio no es tan bueno, todavía me encanta este restaurante*"

Tiene un tono positivo, pero no se puede decir que esta frase sea completamente positiva. De hecho, la oración es positiva sobre el restaurante (enfatizado), pero negativa sobre su servicio (no se enfatiza). Por lo tanto, el objetivo de este nivel de análisis es descubrir los sentimientos sobre las entidades y/o sus aspectos. Por ejemplo:

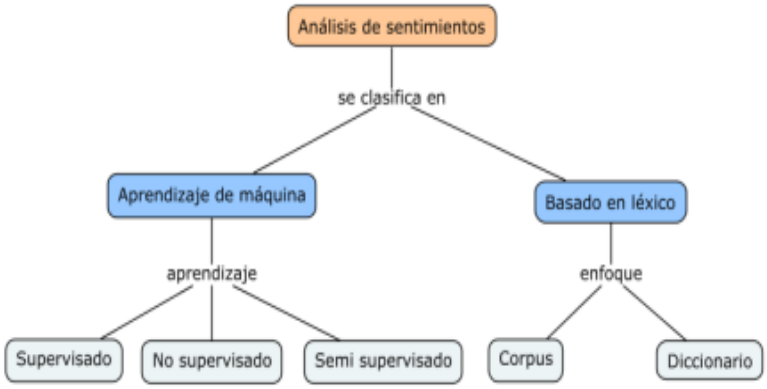
> "*La calidad de la llamada del iPhone es buena, pero su duración de la batería es corta*"

Evalúa dos aspectos, la calidad de la llamada y la duración de la batería, del iPhone (entidad). El sentimiento sobre la calidad de las llamadas de iPhone es positivo, pero el sentimiento sobre la duración de la batería es negativo. La calidad de la llamada y la duración de la batería del iPhone son los objetivos de opinión.

8.1.1.2 Técnicas de Análisis de Sentimiento [13].

Las investigaciones en el campo de Sentiment Analysis emplean diferentes técnicas para la clasificación de sentimientos, ver figura __

**Ilustración 2. Técnicas para la clasificación de sentimientos.**




Fuente: Tomado de [13].

8.1.1.2.1 Aprendizaje de máquina o machine learning [13].

Los métodos de clasificación que utiliza el enfoque ML se pueden dividir en:

> Aprendizaje supervisado [14]: Técnica para predecir o clasificar etiquetas a partir de unos datos de entrenamiento.
>
> Aprendizaje no supervisado [15]: Técnica para determinar cómo se agrupan los datos naturalmente.
>
> Aprendizaje semi-supervisado [16]: Un hibrido entre los dos anteriores.

8.1.1.2.2 Basado en léxico [13].

El enfoque basado en el léxico [17] utiliza el método de diccionario y el método basado en el corpus.

> El método del diccionario, hace uso de un diccionario existente, es decir, de una colección de palabras de opinión que tienen un Score positivo (+) o negativo (-) el cual determina la fuerza del sentimiento.
>
> El método basado en el corpus, se basa en la probabilidad de ocurrencia de palabras de opinión con orientaciones específicas en un texto.

Esta clasificación se puede hacer mediante el uso de métodos estadísticos o semánticos.

8.1.2 Text mining o minería de texto [18].

Se refiere al proceso de extracción de patrones interesantes y no triviales de documentos de texto.

Como la forma más natural de almacenar información es el texto, se cree que la minería de textos tiene un potencial comercial más alto que el de la minería de datos.

Sin embargo, la minería de textos también es una tarea mucho más compleja que la minería de datos, ya que implica tratar con datos de texto que son intrínsecamente desestructurados y difusos.

La minería de textos es un campo multidisciplinario, que incluye recuperación de información, análisis de textos, extracción de información, agrupación, categorización, visualización, tecnología de bases de datos, aprendizaje automático y extracción de datos.

8.1.3 Opinión [12].

Formalmente una opinión consta de dos componentes clave: un objetivo (g) y un sentimiento (s) en el objetivo, es decir:

> *(g, s),*

Donde g puede ser cualquier entidad o aspecto de la entidad sobre la cual se ha expresado una opinión y s es un sentimiento positivo, negativo o neutral, o una calificación numérica que expresa la fuerza o intensidad del sentimiento (por ejemplo, 1 a 5 estrellas). Las orientaciones de sentimiento, opinión o polaridades son clasificadas como positivas, negativas o neutras.
Ejemplificando:

"*(1) Compré una cámara Canon G12 hace seis meses. (2) Simplemente me encanta. (3) La calidad de la imagen es sorprendente. (4) La duración de la batería es larga. (5) Sin embargo, mi esposa piensa que es demasiado pesada para ella*".

Posteado por: John Smith. Fecha: septiembre 10 de 2011.

De la anterior oración se pueden notar algunas cosas importantes:

(1) La revisión tiene una serie de opiniones, tanto positivas como negativas, sobre la cámara Canon G12.
(2) Expresa una opinión positiva sobre la cámara Canon en su conjunto.
(3) Expresa una opinión positiva sobre la calidad de imagen.
(4) Expresa una opinión positiva sobre la duración de la batería.
(5) Expresa una opinión negativa sobre el peso de la cámara.

A partir de estas opiniones, se pueden hacer las siguientes observaciones:

Por ejemplo, el objetivo de opinión en la oración (2) es la cámara Canon G12, y el objetivo de opinión en la oración (3) es la calidad de imagen de la cámara.

Este ejemplo tiene opiniones de dos personas, llamadas fuentes de opinión. El titular de las opiniones en las oraciones (2), (3) y (4) es el autor de la revisión "John Smith", pero para la oración (5), es la esposa del autor.

La fecha de la revisión es el 10 de septiembre de 2011. Esta fecha es importante en la práctica porque a menudo se quiere saber cómo las opiniones cambian con el tiempo y las tendencias de opinión.

8.1.4 Entidad [12].

Formalmente una entidad e es un producto, servicio, tema, problema, persona, organización o evento. Se describe con un par, e: (T, W), donde T es una jerarquía de partes, sub-partes, etc., y W es un conjunto de atributos de e.

"*(1) Compré una cámara Canon G12 hace seis meses. (2) Simplemente me encanta. (3) La calidad de la imagen es sorprendente. (4) La duración de la batería es larga. (5) Sin embargo, mi esposa piensa que es demasiado pesada para ella*".

Ejemplo: Un modelo particular de cámara es una entidad, por ejemplo, Canon G12. Tiene un conjunto de atributos: calidad de imagen, tamaño y peso, y un conjunto de partes: lente, visor y batería. La batería también tiene su propio conjunto de atributos, por ejemplo, duración de la batería y peso de la batería.

En el ejemplo anterior la oración (2) expresa una opinión positiva sobre la entidad de la cámara Canon G12 como un todo. La oración (3) expresa una opinión positiva sobre el atributo de calidad de imagen de la cámara.

### 8.1.5 Natural Languaje Procesing [19].

Es un área de investigación y aplicación que explora cómo se pueden usar las computadoras para comprender y manipular el texto o el habla del lenguaje natural para hacer cosas útiles.

Los investigadores de NLP buscan recopilar conocimiento sobre cómo los seres humanos entienden y usan el lenguaje para que las herramientas y técnicas apropiadas puedan desarrollarse para hacer que los sistemas informáticos entiendan y manipulen los lenguajes naturales para realizar las tareas deseadas.

Las aplicaciones de NLP incluyen varios campos de estudios, como traducción automática, procesamiento y resumen de texto en lenguaje natural, interfaces de usuario, recuperación de información multilingüe e intralingüística (CLIR), reconocimiento de voz, inteligencia artificial y sistemas expertos, y más.

### 8.1.6 Twitter [20].

Twitter es una red social y un servicio de comunicación en tiempo real lanzado en 2006 y utilizado por millones de personas y organizaciones para compartir y descubrir información con rapidez. La palabra Twitter proviene del sonido frecuente de los pájaros, por lo tanto, su logotipo es un pájaro. Los usuarios pueden acceder al sitio a través de la web y dispositivos móviles para intercambiar actualizaciones de información llamadas 'tweets'. Estos mensajes o tweets son públicos por defecto y visibles para aquellos que están siguiendo al usuario. Twitter permite seguir a otros usuarios que sean del interés del usuario para que pueda ver sus tweets en su página de inicio. Los usuarios comparten estos tweets que pueden contener cosas como fotos, videos, citas, enlaces de artículos y más. Cada tweet también puede tener respuestas de otras personas que crean conversaciones en tiempo real sobre diversos temas, noticias de última hora y nuevos contenidos. Twitter ha evolucionado desde algo más que una herramienta de comunicación en tiempo real hasta una de las principales fuentes mundiales de descubrimiento social y eventos de interés periodístico.

### 8.1.7 Tweet [21].

Un tweet es una publicación en Twitter de 280 caracteres o menos. Es probable que se lo denomine "tweet" porque se parece al mismo tipo de chirrido corto y dulce que puede escuchar de un pájaro. Ver figura _.

Hay diferentes tipos de tweet:

Tweet regular: solo texto y poco más.

Tweet con imagen: se pueden cargar hasta cuatro imágenes en un tweet junto con un mensaje. También se pueden etiquetar a otros usuarios de Twitter en las imágenes.

Tweet con video: se puede subir un video y publicarlo.

Tweet con Media-rich link: cuando se incluye un enlace en el tweet.

Tweet con ubicación: al redactar un tweet aparecerá una opción que permite compartir la ubicación geográfica del usuario que puede activar o desactivar según su preferencia.

Tweet con mención: cuando se quiere nombrar a otro usuario en un tweet. Se agrega el signo "@" y seguidamente el nombre del usuario al que se quiere nombrar.

Tweet con encuesta: permite hacer una pregunta y agregar diferentes opciones que los seguidores pueden elegir para responder. Se pueden ver las respuestas en tiempo real.

8.1.8 Retweet [22].

Es una nueva publicación de un Tweet. La función de retweet ayuda al usuario y a otros a compartir rápidamente ese Tweet con todos los seguidores. Se pueden retwittear los tweets propios o los de otro usuario. Ver [C] en Figura __.

8.1.9 Like o me gusta en Twitter [23].

Se usa para mostrar aprecio por un tweet, están representados por un corazón pequeño. Ver [D] en Figura__.

8.1.10 Timeline [24].

Es la página de inicio de Twitter donde se muestra una secuencia de tweets de las cuentas que el usuario ha elegido seguir en Twitter. Es posible que se vea contenido sugerido basado en las características del usuario.

8.1.11 Ejemplo de un tweet.

**Ilustración 3. Ejemplo de un tweet.**


Fuente: Tomada de twitter.com.

[A]: El texto del tweet.
[B]: Usuario que escribe el tweet.
[C]: Cantidad de retweets que tiene el tweet.
[D]: Cantidad de likes o me gustas que tiene el tweet.
[E]: Cantidad de respuestas que tiene el tweet.
[F]: Fecha de publicación del tweet.

## 8.2 MARCO LEGAL

En el marco del desarrollo del proyecto se considera pertinente relacionar la temática del mismo con el Habeas data debido a que se va a realizar tratamiento de información pública.

### 8.2.1 Habeas data [25][26]:

El derecho de habeas data comprende dos derechos principales: el derecho de acceso para solicitar y obtener la información existente sobre la persona (sea física o jurídica) y eliminación o corrección si la información fuera falsa o estuviera desactualizada.

En Colombia el marco regulatorio del Habeas Data se ha venido consolidando a partir de la incorporación de su reconocimiento en la Constitución de 1991, en cuatro fases:

- Primera, la jurisprudencia nacional, como lo es la Corte Constitucional.
- Segunda, con la aparición de normas especiales, la Ley 1266 de 2008 para información financiera, la que surge de la mano de

normativa penal de los delitos informáticos consagrados en la Ley 1273 de 2009.

- Tercera, con el desarrollo de la normativa general de datos personales de personas naturales con la Ley 1581 de 2012.
- Cuarta, con la reglamentación administrativa, la puesta en funcionamiento del sistema de registro de bases de datos personales o registro de ficheros (RNBD) ante la Superintendencia de Industria y Comercio.

## 8.3 ESTADO DEL ARTE

Para el presente trabajo se tomaron las siguientes características al momento de hacer la búsqueda de documentos relacionados.

Keywords: sentiment analysis; text mining; natural languaje procesing.

Se hace una búsqueda usando las keywords mencionadas anteriormente arrojando un total de 675 documentos.

La búsqueda se concentró en documentos que fueron publicados del 2010 en adelante, documentos en español e inglés y se usaron las siguientes bases de datos académicas: ACL, AAAI, SAGE, Semantic Scholar, arXiv.org, DORAS (DCU Online Research Access Service), Dialnet, RUA, academia.edu, scielo, Google Scholar, IEEE, entre otros.

Los criterios de inclusión usados fueron los siguientes:

Artículos que hablan de “text mining” y “sentiment analysis”.
Artículos que hablan de “text mining” y “natural languaje procesing”.
Artículos que hablan de “sentiment analysis” y “natural languaje procesing”.
Artículos que hablan de “text mining”, “sentiment analysis” y “natural languaje procesing”.
Artículos que hayan sido escritos desde el 2010.

Los criterios de exclusión usados fueron los siguientes:

Artículos que hablan de “text mining”, pero no hablan de “sentiment analysis”.
Artículos que hablan de “text mining”, pero no hablan de “natural languaje procesing”.
Artículos que hablan de “sentiment analysis”, pero no hablan de “text mining”.
Artículos que hablan de “sentiment analysis”, pero no hablan de “natural languaje procesing”.
Artículos que hablan de “natural languaje procesing”, pero no hablan de “text mining”.

Artículos que hablan de “natural languaje procesing”, pero no hablan de “sentiment analysis”.
Artículos que hayan sido escritos antes del 2010.

A continuación, se presenta una revisión de los trabajos relacionados a la temática del proyecto en el ámbito internacional, y regional.

### 8.3.1 Ámbito internacional

**Tabla 1. A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. (2012) [27]**

| | | | | |
|---|---|---|---|---|
| **1.** | **Título** | A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. (2012) [27] | | |
| | **Autores** | Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, Shrikanth Narayanan. | | |
| | **Institución** | University of Southern California. | **País** | Estados Unidos. |
| | **Keywords** | **sentiment analysis**; real-time transcription | | |
| | **¿Qué se hizo?** | Una herramienta para el análisis en tiempo real del sentimiento expresado en Twitter, buscando explorar si Twitter proporciona información sobre el despliegue de las campañas e indicaciones de los cambios en la opinión pública.<br><br>Se capacitó al modelo para comprender el lenguaje que usado en Twitter y sarcasmo.<br><br>La infraestructura de procesamiento de datos en tiempo real se hizo en la plataforma InfoSphere Streams de IBM.<br><br>Se diseñó un panel HTML basado en Ajax para mostrar el volumen y el sentimiento por cada candidato, así como las palabras clave y las estadísticas del sistema, se actualiza cada 30 segundos. | | |
| | **¿Datos usados?** | Reunieron más de 36 millones de tweets sobre los candidatos, un cuarto de millón por día en promedio.<br><br>Se recopilaron todos los tweets relevantes en tiempo real de todo el tráfico de Twitter a través de Gnip Power Track, un proveedor comercial de datos de Twitter, usando filtros o reglas para rastrear los tweets. | | |
| | **¿Técnicas usadas?** | Análisis automatizado del sentimiento en tiempo real.<br><br>Usando prácticas NLP se tokenizan los tweets. Se usó el tokenizador básico de Twitter de Christopher Potts. | | |

| | | |
|---|---|---|
| | | Se usó un enfoque crowdsourcing para hacer anotaciones de opinión sobre los tweets mediante Amazon Mechanical Turk (AMT) y cerca de 800 turkers. Las categorías usadas fueron (positivo, negativo, neutral, inseguro, sarcástico o humorístico). De 17,000 tweets (16% positivos, 56% negativos, 18% neutrales, 10% inseguros).<br><br>Se usó un clasificador bayesiano ingenuo.<br><br>Se usó la métrica TF-IDF para identificar efectivamente las palabras más destacadas cuando ocurre un evento. |
| | **Resultados** | Se presenta un sistema para el análisis del sentimiento de Twitter en tiempo real de las elecciones presidenciales estadounidenses de 2012.<br><br>Se usa Twitter "firehose", las reglas y palabras clave seleccionadas por expertos para obtener una imagen completa y precisa del panorama político en línea.<br><br>Cuenta con una infraestructura de procesamiento de datos en tiempo real y modelo de sentimiento estadístico que evalúan los cambios en el sentimiento público en respuesta a eventos políticos emergentes y noticias a medida que se desarrollan.<br><br>La arquitectura y el método son genéricos, y pueden adoptarse fácilmente y extenderse a otros dominios. |

**Tabla 2. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. (2010) [28]**

| | | | | |
|---|---|---|---|---|
| **2.** | **Título** | Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. (2010) [28] | | |
| | **Autores** | Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe | | |
| | **Institución** | Technical University of Munich. | **País** | Alemania. |
| | **Keywords** | twitter; microblogging; politics; elections; **sentiment analysis**; **data mining**. | | |
| | **¿Qué se hizo?** | Examinar si Twitter es un vehículo para la deliberación política al observar cómo la gente lo usa para intercambiar información sobre asuntos políticos. | | |

| | | |
|---|---|---|
| | | Se evalúa si los mensajes de Twitter reflejan el sentimiento político actual fuera de la red.<br><br>Se analiza si la actividad en Twitter se puede usar para predecir la popularidad de los partidos políticos en el mundo real. |
| | **¿Datos usados?** | Se usaron 104,003 tweets publicados en las semanas previas a la elección. |
| | **¿Técnicas usadas?** | Para extraer el sentimiento de forma automática se usó LIWC2007 (software de análisis de texto desarrollado para evaluar emocional, cognitivamente, y los componentes estructurales de muestras de texto utilizando un diccionario interno validado psicométricamente). Este software calcula el grado en que una muestra de texto contiene palabras que pertenecen a categorías psicológicas y estructurales definidas empíricamente. Específicamente, determina la medida a la que ciertas cogniciones y emociones (por ejemplo, orientación futura, emociones positivas o negativas) están presentes en el texto. Para cada dimensión psicológica, el software calcula la frecuencia relativa con la que ocurren las palabras relacionadas con esa dimensión en una muestra de texto dado.<br><br>Se enfocaron en 12 dimensiones para perfilar el sentimiento político: orientación hacia el futuro, orientación pasada, emociones positivas, emociones negativas, tristeza, ansiedad, enojo, tentativa, certeza, trabajo, logros y dinero.<br><br>Se concatenaron todos los tweets publicados en una muestra de texto para ser evaluada por LIWC.<br><br>Los tweets fueron descargados en alemán y traducidos automáticamente al inglés para ser procesados por el diccionario inglés LIWC. |
| | **Resultados** | Se encuentra que Twitter se utiliza como plataforma para la deliberación política. La cantidad de tweets refleja las preferencias de los votantes y se acerca a las encuestas electorales, el sentimiento de los mensajes de Twitter se corresponde con los programas políticos, los perfiles de los candidatos y la evidencia de la cobertura de los medios de la campaña electoral. |

**Tabla 3. Every tweet counts? How sentiment analysis of social networks can improve our knowledge of citizens' policy preferences. An application to Italy and France. (2012) [29].**

| | | | | |
|---|---|---|---|---|
| **3.** | **Título** | Every tweet counts? How sentiment analysis of social networks can improve our knowledge of citizens' policy preferences. An application to Italy and France. (2012) [29]. | | |
| | **Autores** | Andrea Ceron, Luigi Curini, Stefano M. Iacus, Giuseppe Porro | | |
| | **Institución** | Università degli Studi di Milano | **País** | Italia |
| | **Keywords** | analysis of public opinion; electoral campaign; political forecast; **sentiment analysis**; social media; text analytics; **text mining**. | | |
| | **¿Qué se hizo?** | Rastrearon la popularidad en línea de los líderes políticos italianos en 2011 y la intención de voto de los usuarios de internet franceses en la votación presidencial de 2012 y en la elección legislativa. | | |
| | **¿Datos usados?** | Para descargar los datos se usaron dos fuentes:<br>• La plataforma de monitoreo y análisis de redes sociales Crimson Hexagon<br>• El motor de Internet Voices from the Blogs.<br>El análisis de los tweets se ha realizado usando dos fuentes:<br>• La plataforma ForSight proporcionada por Crimson Hexagon<br>• En el lenguaje R | | |
| | **¿Técnicas usadas?** | Se usó análisis de sentimiento supervisado y automatizado. El primer paso involucra codificadores humanos y consiste en codificar una sub-muestra de los documentos descargados de alguna fuente de Internet. La sub-muestra representa un conjunto de entrenamiento que será utilizado por el algoritmo HK para clasificar todos los documentos no leídos, en la segunda etapa. Este método permite superar uno de los principales problemas del SA anticuado basado en diccionarios ontológicos, es decir, el hecho de que tienden a clasificar erróneamente las publicaciones que no adoptan un lenguaje sencillo. Por el contrario, los codificadores humanos pueden ser más precisos durante dicha. Entonces, el análisis estadístico automatizado proporcionado por el algoritmo HK extiende tal precisión a toda la población de publicaciones, lo que permite capturar adecuadamente las opiniones expresadas en la web. | | |
| | **Resultados** | El análisis muestra una correlación consistente entre los resultados de los medios sociales y los que se podrían | | |

| | | |
|---|---|---|
| | | obtener de encuestas más tradicionales, así como una notable habilidad de los medios sociales para pronosticar los resultados electorales.<br><br>El análisis de sentimiento de las redes sociales parece proporcionar predicciones más precisas cuando se enfoca en los líderes más populares o en los principales partidos.<br><br>Cuando un creciente número de ciudadanos expresan en línea su opinión y/o elección de voto, la exactitud del análisis de las redes sociales aumenta, siempre que los usuarios de Internet actúen de manera coherente sobre eso, confirmando su preferencia declarada en redes y emitiendo un voto. |

**Tabla 4. Political Tendency Identification in Twitter using Sentiment Analysis Techniques. 2014. [30].**

| | | | | |
|---|---|---|---|---|
| **4.** | **Título** | Political Tendency Identification in Twitter using Sentiment Analysis Techniques. 2014. [30]. | | |
| | **Autores** | Ferran Pla, Lluís-F. Hurtado | | |
| | **Institución** | Universitat Politecnica de València | **País** | España |
| | **Keywords** | **sentiment analysis**; text corpus; entity | | |
| | **¿Qué se hizo?** | Se presenta un sistema para abordar la tarea de identificación de tendencias políticas de usuarios de Twitter basada en técnicas de SA. | | |
| | **¿Datos usados?** | Se usó un Corpus general (TASS20131) que contiene aproximadamente 68000 mensajes de tweets escritos en español (noviembre de 2011 y marzo de 2012) por 158 conocidas personalidades del mundo de la política, economía, comunicación, medios de comunicación y cultura.<br><br>El corpus está codificado en XML. Cada tweet incluye su ID, la fecha de creación y la ID de usuario. Se etiqueta con su polaridad global usando etiquetas N y N + para polaridad negativa con diferente intensidad, etiquetas P y P + para polaridad positiva con diferente intensidad, y la etiqueta NEU para polaridad neutra. La etiqueta NONE se usó para representar tweets sin ninguna polaridad. | | |
| | **¿Técnicas usadas?** | El sistema desarrollado consta de 4 módulos.<br>• El primer módulo es el módulo de procesamiento previo que realiza el reconocimiento de | | |

| | | |
|---|---|---|
| | | tokenización, lematización y entidades nombradas en el tweet.<br>• El segundo módulo permite obtener la polaridad de las entidades contenidas en el tweet. Si se omite este paso se obtiene la polaridad global del tweet.<br>• El tercer módulo es el módulo Feature Extraction que selecciona las características del tweet pre-procesado (o de los segmentos de tweets) y se obtiene un vector de características.<br>• El cuarto módulo es el módulo Polarity Classifier que usa un clasificador (aprendido de los vectores de características del conjunto de entrenamiento) para asignar una etiqueta de polaridad al tweet.<br><br>Se decide utilizar y adaptar las herramientas disponibles para el etiquetado de tokenización y lematización. Se adapta el paquete Tweetmotif 2 para procesar tweets en español. También se usa Freeling 3 para derivar, reconocimiento de entidad y etiquetado POS.<br><br>Se necesita determinar la polaridad de las entidades políticas que contiene un tweet. Porque la polaridad de cada entidad podría ser diferente de la polaridad global del tweet. Se elegido un enfoque que consiste en definir un conjunto de heurísticas para determinar qué segmento del tweet se refiere a cada una de las entidades presentes en él. Se definen algunas reglas para hacer esta segmentación. Si el tweet contiene solo una entidad, el contexto considerado es todo el tweet.<br><br>Debido a la corta longitud de los tweets, el contexto de una entidad a menudo es tan pequeño que no contiene información suficiente para clasificar correctamente la polaridad de la entidad. En tal caso, la opción que se eligió es establecer un umbral de contexto, y si está por debajo de este umbral, se les asignó la misma polaridad a todas las entidades del tweet. Cuando el número de entidades es mayor que dos en muchos casos, se asigna la misma polaridad a todas las entidades del tweet porque no se tenía suficiente contexto. |

| | | |
|---|---|---|
| | | Se usó WEKA5, que es una herramienta que incluye una colección de algoritmos de aprendizaje automático que se pueden usar para tareas de clasificación.<br><br>Específicamente, se usó un enfoque basado en SVM. En la tarea de Análisis de Sentimientos, SVM (Support Vector Machines) ha demostrado su capacidad para manejar grandes espacios de características y determinar las características relevantes. Se usó el algoritmo NU-SVM de Scho'lkopf (2000) desde una biblioteca externa llamada LibSVM6, que es un software muy eficiente para construir clasificadores SVM. Es fácil integrar este software con WEKA, lo que permite utilizar todas las funciones de WEKA. Se usó el enfoque de bolsa de palabras para representar cada tweet como un vector de características que contiene la frecuencia de las características seleccionadas del conjunto de entrenamiento.<br><br>Para realizar la clasificación de usuarios se asume la siguiente hipótesis:<br>las opiniones positivas sobre un partido político son una orientación política similar a la del usuario que realiza la revisión para este partido; por el contrario, una opinión negativa sobre un partido es una orientación política opuesta a eso mostrado por esta parte.<br><br>De esta forma, para clasificar a los usuarios por su orientación política, primero se identifican las entidades asociadas a los partidos políticos y, en segundo lugar, se analiza la polaridad de estas entidades en los tweets de cada usuario.<br><br>Se detectaron un total de 864 entidades relacionadas con partidos políticos y líderes políticos. Se asignaron los valores para cada polaridad. |
| | **Resultados** | Se ha descrito un enfoque para la identificación de tendencias políticas de los usuarios de Twitter. Se ha definido una métrica, denominada Tendencia política, que tiene en cuenta la polaridad de las entidades relacionadas con los partidos políticos que aparecen en los tweets del usuario. También se presentó el sistema |

| | | |
|---|---|---|
| | | de Análisis de Sentimiento desarrollado para obtener la polaridad de estas entidades. |

**Tabla 5. Can Collective Sentiment Expressed on Twitter Predict Political Elections? (2011) [31]**

| | | | | |
|---|---|---|---|---|
| **5.** | **Título** | Can Collective Sentiment Expressed on Twitter Predict Political Elections? (2011) [31]. | | |
| | **Autores** | Jessica Chung, Eni Mustafaraj | | |
| | **Institución** | Wesley College | **País** | Estados Unidos |
| | **Keywords** | **sentiment analysis**; social media | | |
| | **¿Qué se hizo?** | Se muestra cómo se aplican las herramientas y metodologías presentadas por Tumasjan (2010) y O'Connor (2010) a un conjunto de datos que consiste en los tweets de las elecciones especiales del Senado de EE.UU. en Massachusetts. | | |
| | **¿Datos usados?** | Usando la API de transmisión de Twitter, se recopilaron tweets que contenían uno o ambos nombres de los candidatos a las elecciones del Senado de EE. UU. de 2010 en Massachusetts (MAsen10). Hubo 234.697 tweets de 56.165 usuarios únicos recopilados entre el 13 de enero y el 20 de enero de 2010. | | |
| | **¿Técnicas usadas?** | A los datos recopilados se les eliminó material extraño: Hashtags, nombres de cuenta y enlaces a sitios web. Las contracciones fueron reemplazadas por su forma completa y los emoticones como ":)" fueron reemplazados por etiquetas con nombre, por ejemplo: <feliz><br><br>Se probó el método predictivo utilizado en Tumasjan (2010). Su estudio demostró con éxito que la proporción de tweets para cada candidato en las elecciones correspondía directamente al porcentaje de votos recibidos por cada candidato. Sin embargo, cuando se aplicó este método a los datos del estudio los resultados fueron diferentes. La proporción de tweets para el candidato perdedor en los días previos a las elecciones es mayor que la del ganador. Por lo tanto, depender de la cantidad de menciones que recibe un candidato no es suficiente para predecir el resultado de una elección. El sentimiento de un tweet debe ser examinado, ya que muchos tweets reflejan oposición en lugar de apoyar a un candidato.<br><br>Para determinar el sentimiento de un tweet, se siguió el método descrito en O'Connor (2010), fue usado para | | |

| | | |
|---|---|---|
| | | detectar la correlación entre tweets que contenían la palabra "obama" y encuestas tradicionales que seguían las calificaciones de aprobación del presidente Obama en 2009. Su método de utilizó el léxico de subjetividad de OpinionFinder Wilson, Wiebe y Hoffmann (2005), una lista de palabras que contiene 2.800 palabras que se anotan manualmente como positivas, negativas, neutrales o ambas. Se implementó su algoritmo, que encuentra qué palabras en un tweet tienen una etiqueta en el léxico, y calcula el sentimiento general de un tweet basado en la suma de polaridades de palabras distintas. El algoritmo se aplica luego a un subconjunto de datos de MAsen10, con algunas etiquetas ya definidas.<br><br>La precisión general es de 41.41%, mejor que un clasificador aleatorio, pero no confiable para las predicciones. Esta baja precisión se puede explicar con la baja cobertura de nuestro conjunto de datos del léxico de OpinionFinder. |
| | **Resultados** | Se ha demostrado que los métodos simples actuales para predecir los resultados electorales basados en el análisis del sentimiento del texto de los tweets no son mejores que los clasificadores aleatorios.<br><br>Con el fin de mejorar la precisión del análisis de sentimiento, es necesario ir más allá de los métodos que dependen únicamente de la polaridad de las palabras.<br><br>Las técnicas de pre-procesamiento como el etiquetado POS y la desambiguación del sentido de la palabra pueden ser necesarias, así como las características no léxicas.<br><br>Por último, se necesita una forma de aprender la polaridad de las palabras en el contexto y dominio en el que aparecen. |

**Tabla 6. A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. (2011) [32].**

| | | |
|---|---|---|
| **6.** | **Título** | A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. (2011) [32]. |
| | **Autores** | Murphy Choy, Michelle L.F. Cheong, Ma Nang Laik, Koo Ping Shung |

| | | | | |
|---|---|---|---|---|
| | **Institución** | Singapore Management University | **País** | Singapur |
| | **Keywords** | twitter; **sentiment analysis**; presidential election; Singapore; census. | | |
| | **¿Qué se hizo?** | Se evalúa la cantidad de información con respecto a los eventos electorales en Singapur (país conservador). El objetivo fue desarrollar la metodología para ver qué tan bien se refleja la información en línea con los eventos políticos que ocurren. Por último, se usa la información para predecir al nuevo presidente. | | |
| | **¿Datos usados?** | Se recopilaron 16.616 tweets de Twitter entre el 17 de agosto y el 25 de agosto y entre el período de nominación y de campaña. La información se recopiló del motor de búsqueda de Twitter y la ayuda de la API de Google. Los datos recopilados se basan en el nombre de los candidatos (solo se utiliza el nombre en inglés, esto se debe a las dificultades para evaluar los tweets chinos). Los tweets repetitivos se depuran para garantizar que se pueda usar una colección adecuada y no adulterada de tweets para el análisis. | | |
| | **¿Técnicas usadas?** | Para extraer los sentimientos automáticamente de los datos, se creó y desarrolló un corpus personalizado. Creado a partir de fuentes en línea, diccionario y elecciones anteriores. Este corpus fue desarrollado estrictamente para esa campaña y no se aplica a otras. También se enfocan en el sentimiento de cada tweet y se ignoran algunos de los aspectos más complicados del tweeting. Debido al posible sesgo en los datos, se recopila información adicional en forma de censo, así como una encuesta de los organismos gubernamentales.<br><br>Para estimar los votos, se desarrolló un marco de registro del censo. En este marco, hay varias informaciones claves que se requieren. Al mismo tiempo, hay 2 suposiciones sobre el marco:<br>1. Las personas que votaron en las elecciones generales probablemente voten según las líneas del partido al que votaron.<br>2. El sentimiento en línea es representativo de las personas que expresan sus opiniones.<br><br>Para calcular el porcentaje de votos que cada candidato debería recibir se usaron algunas fórmulas. | | |

| | | |
|---|---|---|
| | | A partir de los resultados se puede ver la diferencia entre el valor predicho y el real. Si bien se calculó con éxito el pequeño margen entre los dos candidatos principales, el modelo no logró predecir el ganador correcto. El primer problema es que el modelo actualmente no puede abordar es el porcentaje de votantes indecisos. Este grupo de votantes con frecuencia cambia su posición. Al mismo tiempo, algunos de los votantes indecisos podrían no haber votado como lo han hecho anteriormente. Esto podría explicar las diferencias en el valor real y el predicho.<br><br>El segundo problema es el problema con el sentimiento falso. Esta situación se puede atribuir a 2 causas principales. La primera fuente de ese sentimiento falso es el astroturf. La otra fuente puede estar relacionada con el escenario en el que los votantes realmente no reflejan sus sentimientos en línea de la elección de su candidato. |
| | **Resultados** | El marco ha sido capaz de predecir los dos mejores contendientes. Si bien el ganador previsto no aparece como el presidente debido al pequeño margen, la estimación del pequeño margen de votos entre los contendientes indicó la capacidad del modelo para modelar de manera realista el escenario político.<br><br>El marco ha sido capaz de convertir la información de Twitter en predicción realista.<br><br>Fue la primera vez que se recopila información de Twitter en un entorno conservador en el que una parte importante de la población no participa políticamente en Twitter.<br><br>El análisis ha demostrado que la información de Twitter puede traducirse en información bastante precisa sobre el panorama político, aunque los usuarios de Twitter no sean tan comunes.<br><br>Se necesitará un trabajo más importante para manejar el tema del astroturf, así como la psicología de los votantes. |

**Tabla 7. Modeling Indian General Elections: Sentiment Analysis of Political Twitter Data. (2015) [33].**

| | | | | |
|---|---|---|---|---|
| **7.** | **Título** | Modeling Indian General Elections: Sentiment Analysis of Political Twitter Data. (2015) [33]. | | |
| | **Autores** | Kartik Singhal, Basant Agarwall, Namita Mittal | | |
| | **Institución** | LNM Institute of Information Technology<br>Malaviya National Institute of Technology | **País** | India |
| | **Keywords** | **sentiment analysis**, social media, political sentiment | | |
| | **¿Qué se hizo?** | Se propone un enfoque para detectar el sentimiento en tweets políticos basados en reglas semánticas propias. El modelo de análisis de sentimiento político propuesto no está supervisado y no requiere ningún conjunto de datos de entrenamiento previo. | | |
| | **¿Datos usados?** | Los tweets se descargaron manualmente entre el 28 de febrero y el 28 de marzo de 2014. | | |
| | **¿Técnicas usadas?** | Se confió en un sistema que fuera capaz de resolver diferentes problemas como sarcasmo, entre otros. Para esto, se propone un enfoque híbrido no supervisado de análisis basado en léxico y basado en reglas que analizará palabras relacionadas con otras palabras, dando así un sentimiento general de la oración. Para el léxico, se usa SentiWordNet que puede dar los puntajes de opinión de una palabra. Un puntaje negativo significa connotación negativa y un puntaje positivo significa connotación positiva de la palabra.<br><br>Se usa SDP (Sentiment community detection) para extraer reglas de los tweets. Se eliminan palabras extrañas que no están relacionadas con el sentimiento general o contribuyen muy poco al sentimiento general. A partir de estas reglas, se extraen los tweets que contienen verbos, adjetivos, adverbios, sustantivos, conjunciones, negaciones y se descarta el resto. Al analizar las oraciones de twitter se descubre que, debido a formaciones gramaticales incorrectas, la eficiencia de SDP disminuye afectando el sistema. Cuando SDP no puede detectar la relación entre dos palabras, usa una regla que muestra una dependencia desconocida entre esas palabras. Para mejorar esto, se usa el etiquetador Ark Twitter POS (ATP) que permite determinar la parte del discurso de las dos palabras, lo que da la dependencia. | | |

| | | |
|---|---|---|
| | | La polaridad de las palabras se determina con SentiWordNet y se procede a calcular los puntajes. |
| | **Resultados** | Entre el total de tweets, 116 son positivos, 92 son negativos y los restantes 51 son tweets objetivos. Se utilizó la precisión como medida de evaluación y se calcula dividiendo los tweets clasificados correctamente con el número total de tweets. El enfoque predijo correctamente 76 tweets positivos y 55 tweets negativos.<br>En el artículo se propusieron varias reglas basadas en la estructura semántica de la oración, los resultados experimentales muestran la efectividad del enfoque propuesto sobre los métodos existentes. |

**Tabla 8. On Using Twitter to Monitor Political Sentiment and Predict Election Results. (2011). [34].**

| | | | | |
|---|---|---|---|---|
| **8.** | **Título** | On Using Twitter to Monitor Political Sentiment and Predict Election Results. (2011). [34]. | | |
| | **Autores** | Adam Bermingham, Alan F. Smeaton | | |
| | **Institución** | Dublin City University | **País** | Irlanda |
| | **Keywords** | monitor political sentiment; predict election result; volume-based measure; political sentiment; underlying content; final election result; political insight; case study; sample size; election campaign; social medium; diverse range; time period; **sentiment analysis**; social analytics; public sentiment; approach combine sentiment analysis; conventional election poll; recent irish general election; voting intention | | |
| | **¿Qué se hizo?** | Se revisa un sistema en vivo desarrollado para la Elección General de Irlanda, 2011. El sistema utilizó una variedad de técnicas para proporcionar una interfaz en tiempo real de Twitter durante las elecciones. Utilizando los datos de volumen y sentimiento de este sistema, se revisaron una cantidad de enfoques de muestreo y métodos para modelar el sentimiento político, replicando el trabajo de otros e introduciendo nuevas medidas. Se evalúa el error con respecto a las encuestas, así como con respecto al resultado de la elección.<br><br>El objetivo del "rastreador de Twitter #GE11" era permitir a los usuarios aprovechar el contenido en Twitter relacionado con las elecciones, a través de una interfaz accesible. | | |

| | | |
|---|---|---|
| | **¿Datos usados?** | Entre el 8 de febrero y el 25 de febrero de 2011 se recopilaron 32.578 tweets relevantes para los cinco principales partidos irlandeses. Se definieron como tweets relevantes, tweets que contuvieran los nombres de candidatos y sus abreviaturas, junto con el hashtag de las elecciones, #ge11. |
| | **¿Técnicas usadas?** | Se usó el MAE (Error Absoluto Medio) para comparar las predicciones basadas en Twitter con las encuestas, así como también con los resultados de las elecciones. Se define la medida basada en el volumen como la porción proporcional de las menciones del partido en un conjunto de tweets para un período de tiempo dado.<br><br>Se usa aprendizaje supervisado. Se usaron nueve anotadores para anotar el sentimiento en tweets. Los tweets se tomaron de diferentes períodos de tiempo con el fin de desarrollar un corpus de entrenamiento lo más diverso posible. Las categorías de anotaciones constaban de tres clases de sentimiento (positivo, negativo, mixto), una clase de no sentimiento (neutral) y otras 3 clases (unánime, no relevante, poco claro). Aproximadamente la mitad de las anotaciones contenían sentimientos de algún tipo.<br><br>El bajo nivel de sentimiento positivo observado es sorprendente, representando solo el 12%. Durante estas elecciones, Irlanda se encontraba en un período de crisis económica y el sentimiento político negativo dominaba los medios de comunicación y el estado de ánimo del público. Para elegir el algoritmo de aprendizaje automático se optimizó F-measure, que equilibra la precisión y el recuerdo en todas las clases. El tokenizador usado en Laboreiro (2010) está optimizado para contenido generado por el usuario, por lo que conserva todas las características sociolingüísticas como los emoticones ((":-)") y la puntuación no convencional ("!"). Estas características se utilizan a menudo para añadir tono al texto y por lo tanto susceptibles de contener información de sentimiento. Se eliminan todos los nombres de usuario y URL para evitar cualquier sesgo hacia estos. Se usó un clasificador Adaboost MNB.<br><br>Se usó la métrica Sentiment TF-IDF. |

| | | |
|---|---|---|
| | **Resultados** | En general, se concluye que Twitter parece mostrar una calidad predictiva que se ve aumentada por la inclusión del análisis de sentimiento.<br><br>Se derivan dos métodos diferentes para monitorear el sentimiento dentro de un partido. Ajustando las características a una regresión se observa que el volumen es la variable predictiva más grande seguida por el sentimiento. Con datos suficientes, el sentimiento parece ser menos valioso como medida predictiva. La especulación es que el éxito relativo del sentimiento hacia los partidos políticos se debe a la naturaleza cerrada del sistema.<br><br>También se observa un cambio dramático en el sentimiento en los dos días previos al día de la votación.<br><br>Quizás haya dos razones por las que el volumen es un indicador más fuerte que el sentimiento. El primero es que el volumen puede simplemente ser un indicador razonable de popularidad en una población de personas, y en este caso, la intención de voto. El otro es que el sentimiento en comparación es reactivo y es difícil discriminar entre el sentimiento que reflejan las preferencias internas de las personas y lo que refleja una respuesta inmediata a una noticia o evento dado. Se ven casos donde el sentimiento es necesario. Por ejemplo, el Partido Verde en esta elección tenía un volumen relativamente alto, pero una mirada más de cerca al contenido revela que esto se debió a que las personas comentaban sobre los bajos niveles de apoyo, un aspecto no captado adecuadamente por nuestro análisis de sentimientos. |

**Tabla 9. Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tweets basada en aprendizaje automático (2017) [35].**

| | | | | |
|---|---|---|---|---|
| **9.** | **Título** | Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tweets basada en aprendizaje automático (2017) [35]. | | |
| | **Autores** | Carlos Arcila Calderón, Félix Ortega Mohedano, Javier Jiménez Amores, Sofía Trullenque | | |
| | **Institución** | Universidad de Salamanca | **País** | España |
| | **Keywords** | análisis supervisado de sentimientos; opinión política; twitter; aprendizaje automático; big data; datos masivos; tweets políticos. | | |

| | | |
|---|---|---|
| | **¿Qué se hizo?** | Se analizó el potencial de estos métodos y técnicas, explicando la evolución y aplicación de un clasificador en tiempo real de opiniones políticas en español con técnicas de aprendizaje automático, implementadas tanto en un ordenador en servicio local como usando computación distribuida comercial de mayor capacidad e inmediatez de análisis.<br><br>Se expone de forma específica cómo se puede implementar y evaluar la técnica de análisis supervisado de sentimientos en el campo de la comunicación política. Esta metodología y técnica representa un instrumento único para el contraste predictivo de los resultados electorales futuros en cualquier país o región, y en particular en el entorno de habla hispana La metodología que se presenta en este artículo ha sido implementada como prototipo bajo el nombre de Autocop.<br><br>El prototipo permite la realización de análisis longitudinales para detectar cambios en los indicadores tendenciales asociados a los partidos políticos y sus candidatos, así como comparar estos cambios con los acontecimientos cotidianos. |
| | **¿Datos usados?** | |
| | **¿Técnicas usadas?** | Análisis supervisado de sentimientos |
| | **Resultados** | |

**Tabla 10. ElectionMap: una representación geolocalizada de intenciones de voto hacia partidos políticos sobre la base de comentarios de usuarios de Twitter. (2015) [36].**

| | | | | |
|---|---|---|---|---|
| **10.** | **Título** | ElectionMap: una representación geolocalizada de intenciones de voto hacia partidos políticos sobre la base de comentarios de usuarios de Twitter. (2015) [36]. | | |
| | **Autores** | Francisco Agulló, Antonio Guillén, Yoan Gutiérrez, Patricio Martínez-Barco. | | |
| | **Institución** | Universidad de Alicante | **País** | España |
| | **Keywords** | twitter, **sentiment analysis**, map representation | | |
| | **¿Qué se hizo?** | ElectionMap muestra de forma gráfica la opinión de los usuarios de Twitter sobre temas relacionados con la política. | | |
| | **¿Datos usados?** | La aplicación realiza un proceso de recogida y de almacenamiento de datos. | | |

| | | |
|---|---|---|
| | **¿Técnicas usadas?** | Cada vez que un usuario de Twitter escribe uno de los términos que identifican a los partidos a evaluar, se lanza una alerta hacia el sistema notificando el nuevo comentario y se almacena la información relevante al mensaje. Además de la información proporcionada por Twitter para cada uno de los tweets, se realiza un procesamiento del tweet para conocer si el texto del mensaje es positivo, negativo o neutral y se almacena para considerarlo al igual que el resto de la información.<br><br>Para la elaboración del mapa se tuvieron en cuenta las siguientes restricciones.<br>Los votantes solo suman, no restan. Los usuarios pueden expresar tanto su conformidad como su disconformidad con los partidos políticos, ya que cuando un votante acude las urnas vota en positivo, no en negativo, todos los mensajes con valoraciones negativas son ignorados y no quedan representados en el mapa.<br><br>Localización del tweet. Los datos se muestran en un mapa de España, por lo que es necesario saber a qué localidad le corresponde el voto de cada usuario. Para ello en primer lugar se busca en la descripción de usuario, si ´esta tiene una localización válida, se asume que esa es la localización real del usuario. En caso de no obtener ninguna localización en la búsqueda, se comprueba si el usuario tiene activada la geolocalización. En caso de tener algún tweet con localización almacenado, se tiene en cuenta la localización que más veces se ha utilizado, y si no hay ningún tweet geolocalizado, se ignoran los mensajes del usuario ya que no se podrían situar en el mapa.<br><br>Para la representación de los datos se ha utilizado el API de Google Maps al que se han añadido una serie de controles propios.<br><br>Se evalúa si un texto expresa apoyo o rechazo hacia las entidades (partidos políticos) a valorar. El método |

| | | |
|---|---|---|
| | | de evaluación utilizado es el comentado en “The OpAL System at NTCIR 8 MOAT” Balahur (2010).<br><br>El método se basa en un conjunto de lexicones que contienen palabras que expresan opiniones positivas o negativas agrupados según su intensidad. Cuando se evalúa un texto, se buscan las palabras contenidas en los lexicones y se les asigna una valoración base según el lexicón en el que se encuentren. El método además utiliza stemmer para valorar familias de palabras que pudieran estar relacionadas con las distintas palabras de los lexicones. Las palabras que no aparecen en ninguno de los lexicones se consideran neutrales.<br><br>Finalmente se suman las valoraciones de cada una de las palabras y el valor resultante es el que se utiliza para etiquetar el texto como positivo (valoraciones mayores que 0), neutral (valoraciones iguales a 0) o negativo (valoraciones inferiores a 0). |
| | **Resultados** | http://gplsi.dlsi.ua.es/demos/electionmap/ |

### 8.3.2 Ámbito regional

**Tabla 11. Análisis de sentimiento político en twitter para las elecciones de la alcaldía de Bogotá 2.015 (2016) [37].**

| | | | | |
|---|---|---|---|---|
| **13.** | **Título** | Análisis de sentimiento político en twitter para las elecciones de la alcaldía de Bogotá 2.015 (2016) [37]. | | |
| | **Autores** | Jorge Andrés Alvarado Valencia, Angie Carrillo, Julián Forero, Luis Caicedo, Juan Camilo Urueña. | | |
| | **Institución** | Pontificia Universidad Javeriana | **País** | Colombia |
| | **Keywords** | minería de texto; **análisis de sentimiento**. | | |
| | **¿Qué se hizo?** | Realización de un análisis de sentimiento político en el tiempo, basado en Twitter hacia los candidatos a la alcaldía de Bogotá para el año 2015 y una comparación de los resultados obtenidos contra los publicados en los principales sondeos de intención de voto.<br><br>El análisis de sentimiento se dividió en tres etapas fundamentales: detección del sentimiento, recopilación de noticias y sistematización de los datos | | |

| | | |
|---|---|---|
| | **¿Datos usados?** | Se tomaron un total de 870.961 tweets referidos a política mediante una búsqueda de los nombres de los principales candidatos. |
| | **¿Técnicas usadas?** | Para la detección del sentimiento se elaboró un diccionario de sentimiento adaptado al contexto político colombiano. Para el desarrollo del diccionario se tomó una muestra de tweets que hablaran de las elecciones para la alcaldía de Bogotá y se realizó la tokenización y etiquetado de parte de la palabra (POS) de cada tweet mediante la utilización de Treetagger, Schmid (1994). Se tomaron aquellas palabras que fueran adjetivos, sustantivos, verbos y adverbios, obteniendo un total de 9.636 palabras distintas. Posteriormente se halló la frecuencia con la que cada palabra aparecía en los tweets y se creó el diccionario con aquellas palabras que eran más frecuentes en los tweets. Se escogió calificar el 17 por ciento de las palabras base (1.638), correspondientes al 80 por ciento de la frecuencia total de palabras en los tweets.<br><br>Una vez se tuvo el diccionario se procedió a la calificación de las palabras. Cuatro personas distintas calificaron las palabras entre -2 y +2, siendo -2 muy negativo y +2 muy positivo. La calificación se hizo teniendo en cuenta el contexto político colombiano por lo que el diccionario puede ser usado únicamente para el contexto político local.<br><br>Posterior a eso se desarrolló el modelado de los datos mediante un Bag of Words. Se tomaron un total de 870.961 tweets referidos a política mediante una búsqueda de los nombres de los principales candidatos. A estos tweets se les aplicó la herramienta Treetagger para tokenización y POS, se calificó cada palabra de acuerdo al diccionario y se sumó la calificación de cada palabra del tweet para darle el sentimiento correspondiente. Si la suma era menor a cero, el sentimiento era negativo; si era igual a cero, el sentimiento era neutro y si era mayor a cero, el sentimiento era positivo. Por último, se detectó hacia qué candidato iba dirigido cada tweet, mediante un algoritmo de búsqueda dentro de una lista que contenía los nombres y apodos de cada candidato. |

| | | |
|---|---|---|
| | | Se recolectaron noticias de los tres sitios web de noticias más visitados en Colombia tomando un máximo de seis noticias por semana. Posteriormente se escogieron las noticias que más fueron nombradas o compartidas por los usuarios de Twitter. Se dividieron los datos obtenidos en cuanto al sentimiento de los usuarios de manera semanal para la creación de una línea de tendencia en la cual se observa el seguimiento semanal del sentimiento hacia cada candidato y los hitos de los eventos noticiosos.<br><br>La búsqueda de una relación entre los sondeos de intención de voto y la información producida se tomando los sondeos de intención de voto realizados por las cinco firmas más grandes encuestadoras. Se separaron los datos por semana, obteniendo resultados para ocho semanas ya que las publicaciones de los sondeos de intención de voto no se hacen semanalmente. Posteriormente, se realizó una correlación entre los resultados de los sondeos de intención de voto con los datos obtenidos de los sentimientos positivos, sentimientos negativos, menciones de cada candidato y SIM Score para determinar si alguno de estos indicadores tenía relación alguna con los sondeos de intención de voto. |
| | **Resultados** | Se realizó una línea del tiempo en donde se muestra el comportamiento de los usuarios de las redes sociales con respecto a sus sentimientos hacia los candidatos. Se obtuvo un total de 49.040 tweets negativos, 464.853 tweets neutros y 357.068 tweets positivos.<br><br>Para cada candidato se hicieron las correlaciones con los sondeos de intención de voto.<br><br>La utilización de un modelo simple como BOW con un diccionario especializado acerca del contexto del estudio puede llegar a tener igual o mayor efectividad que modelos más complejos utilizados por un software comercial.<br><br>Los eventos noticiosos se ven descriptivamente asociados con el sentimiento de los usuarios hacia los candidatos, sobre todo si estos eventos son expuestos al público mediante algún contenido audiovisual |

| | | |
|---|---|---|
| | | El sentimiento presente en las redes sociales, las menciones y el SIM Score no guardaron relación con los resultados publicados en los sondeos de intención de voto, por lo cual un estudio de este tipo parece ser más útil para la gestión de campañas políticas que para la medición de la intención de voto<br><br>Debido a la gran diferencia en el tiempo de procesamiento de este estudio y el software comercial, es una buena opción aplicar el modelo propuesto con un diccionario especializado para obtener resultados en tiempo real. |

# 9. PLAN DE ACTIVIDADES

**Tabla 12. Plan de actividades.**

| | | | |
|---|---|---|---|
| **Objetivo específico 1** | Elaborar un estado del arte de herramientas, métodos y aplicaciones que implementen análisis de sentimiento. | 1 | Búsqueda de información sobre técnicas de análisis de sentimientos en bases de datos académicas. |
| | | 2 | Análisis de la información para extraer datos relevantes sobre métodos y técnicas de análisis de sentimientos. |
| | | 3 | Elaboración de un documento con el estado del arte. |
| **Objetivo específico 2** | Diseñar una herramienta de software que implemente análisis de sentimiento. | 4 | Establecimiento de la metodología de desarrollo de acuerdo con la arquitectura que soporta el desarrollo de una herramienta que implemente análisis de sentimientos. |
| | | 5 | Definición de las funcionalidades de la herramienta de software. |
| | | 6 | Elaboración de un documento con el diseño de la herramienta. |
| **Objetivo específico 3** | Desarrollar una herramienta de software que implemente análisis de sentimiento. | 7 | Desarrollo de la herramienta siguiendo los pasos de la metodología de desarrollo. |
| | | 8 | Ejecución de pruebas a la herramienta desarrollada. |
| | | 9 | Ajuste de errores y mejoras en la herramienta para análisis de sentimientos. |
| **Objetivo específico 4** | Realizar el análisis de sentimiento de personalidades influyentes de la política colombiana. | 10 | Recolección de datos (tweets). |
| | | 11 | Procesamiento de datos |
| | | 12 | Análisis de los resultados. |
| | | 13 | Elaboración de un documento con el análisis sentimiento de personalidades influyentes de la política colombiana. |

# 10. CRONOGRAMA

**Tabla 13. Cronograma**

| N° | Semanas | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
| 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 7 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 8 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 9 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 10 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 11 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 12 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 13 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 14 | Documentación del trabajo realizado | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

# 11. PRESUPUESTO

## 11.1 PRESUPUESTO GLOBAL

**Tabla 14. Presupuesto global.**

| Rubros | Fuentes | | Total |
|---|---|---|---|
| | Unab | Otra | |
| **Personal** | $ 5’102.000 | $ 4’000.000 | $ 9’102.000 |
| **Equipos y software** | $ 500.000 | $ 2’000.000 | $ 2’500.000 |

| | | | |
|---|---|---|---|
| **Materiales, suministros y bibliografía** | $ 1’000.000 | $ 150.000 | $ 1’150.000 |
| **Total** | $ 6’602.000 | $ 6’150.000 | $ 12’752.000 |

11.1.1 Descripción de los gastos de personal.

**Tabla 15. Descripción de los gastos de personal.**

| **Investigador** | **Formación académica** | **Función** | **Dedicación (semanal)** | **Duración (en semanas)** | **Recursos** | | **Total** |
|---|---|---|---|---|---|---|---|
| | | | | | **Unab** | **Otra** | |
| Jesús Rodríguez | Pregrado | | 8 horas | 32 | $ 0 | $ 4’000.000 | $ 4’000.000 |
| Liliana Calderón | Posgrado | | 2 horas | 32 | $ 3’054.000 | $ 0 | $ 3’054.000 |
| Carolina Suárez | Posgrado | | 2 horas | 32 | $ 2’048.000 | $ 0 | $ 2’048.000 |
| **Total** | | | | | $ 5’102.000 | $ 4’000.000 | $ 9’102.000 |

11.1.2 Descripción y cuantificación de los equipos de equipos y software de uso.

**Tabla 16. Descripción y cuantificación de los equipos de equipos y software de uso.**

| **Equipos - software** | **Cantidad** | **Valor** | | **Total** |
|---|---|---|---|---|
| | | **Unab** | **Otra** | |
| Computador Lenovo | 1 | $ 0 | $ 2’000.000 | $ 2’000.000 |
| Equipos institucionales | 1 | $ 500.000 | $ 0 | $ 500.000 |
| **Total** | | $ 500.000 | $ 2’000.000 | $ 2’500.000 |

11.1.3 Descripción de materiales, suministros y bibliografía.

**Tabla 17. Descripción de materiales, suministros y bibliografía.**

| Materiales | Justificación | Cantidad | Valor | | Total |
|---|---|---|---|---|---|
| | | | Unab | Otras | |
| Bases de datos académicas | Búsqueda de material, trabajos, proyectos relacionados | 4 | $ 1’000.000 | $ 0 | $ 1’000.000 |
| Materiales de oficina | Resmas, lapiceros, demás utensilios | 1 | $ 0 | $ 50.000 | $ 50.000 |
| Transporte | - | 1 | $ 0 | $ 100.000 | $ 100.000 |
| **Total** | | | $ 1’000.000 | $ 150.000 | $ 1’150.000 |

# 12. ANÁLISIS DE SENTIMIENTO EN TWITTER [38]

A continuación, se presentará el paso a paso para la clasificación de textos por su sentimiento en Twitter. El método elegido es usando aprendizaje supervisado. Para lograr esto es necesario contar con un corpus de entrenamiento cuyos ejemplos, en este caso tweets, deben ser etiquetados previamente de forma manual con el sentimiento al que pertenecen. De este corpus ya etiquetado se entrenan varios algoritmos para crear los modelos. Entre más datos se puedan tener en el corpus el entrenamiento del algoritmo será mejor.

Los modelos, scripts, etc.; serán escritos en lenguaje R, y se harán uso de diferentes librerías como Shiny, tm, twitteR, entre otras.

## 12.1. Corpus de entrenamiento

Para comenzar a realizar un análisis de sentimiento usando técnicas de aprendizaje supervisado se tiene que tener un corpus, es decir, un conjunto de datos, tweets, clasificados o etiquetados con su respectivo sentimiento para poder entrenar los algoritmos. Para tener un corpus hay dos métodos, el primero es buscar en la web diferentes corpus ya hechos, surge un problema, la gran mayoría de corpus que hay están para idioma inglés, en español no hay muchos recursos, para escoger como los hay en inglés; el segundo método, consiste en crear nuestro propio corpus, pero es una labor complicada debido a que requiere mucho tiempo y esfuerzo.

Para crear un corpus se debe establecer un criterio para la obtención de datos, si en este trabajo se hubiera construido un corpus desde cero uno de los criterios sugeridos serían tweets que estuvieran relacionados a asuntos políticos de Colombia, dichos tweets podrían ser filtrados si usan algún hashtag o tendencia o mencionan a algún político, o una fecha en especial, la captura de los tweets se haría usando un script en R mediante la API de Twitter, la API de Twitter cuenta con dos formas de obtener los tweets, primero, usando la API Rest, se pueden descargar 3200 tweets aproximadamente que ya se hayan emitido en una cuenta de usuario o con un hashtag en particular, desde el momento que se ejecuta la búsqueda en el script hasta recolectar esa cantidad y usando la API Streaming también se pueden obtener los tweets filtrados por una cuenta de usuario o hashtag, pero en este caso se obtienen los tweets conforme se vayan publicando. En ambos casos se pueden guardar esos tweets en un archivo de valores separados por comas (CSV, por sus siglas en inglés) o en un archivo de texto simple (txt). Cuando ya se tienen los tweets descargados se procede a hacer la clasificación manual de tweets por un grupo de clasificadores humanos.

Actualmente en la Universidad Autónoma de Bucaramanga están trabajando en conjunto el área de Comunicación Social e Ingeniería de Sistemas en la construcción de un corpus para unas primeras exploraciones de Análisis de Sentimiento y otras técnicas de Minería de Texto, pero como se relató líneas atrás

la construcción de un corpus es un proceso muy laborioso y para el momento del desarrollo del presente proyecto aún no se termina de construir, por lo tanto para desarrollar este proyecto se usa un corpus construido en la Universidad de Salamanca [35], se elige este corpus principalmente porque es difícil encontrar un corpus completo en español y, como se dice antes, la construcción de un corpus desde cero es un proceso complejo.

## 12.2. Algoritmos de clasificación

Actualmente contamos con diferentes algoritmos de clasificación que se pueden usar en análisis de sentimiento: Naives Bayes, Modelos Lineales, etc. Para este trabajo se tuvieron en cuenta los siguientes algoritmos:

### 12.2.1. Original Naives Bayes [39]

Basado en el Teorema de Bayes. Para estos algoritmos los sucesos excluyentes y exhaustivos son las diferentes clases que se pueden asignar a un mensaje, de forma que no se puede asignar más de una clase simultáneamente, son excluyentes, y esas clases son todos los tipos que existes, exhaustivos. Estos algoritmos suelen ser llamados <<ingenuos>> porque las características en sus cálculos para representar a los ejemplos de entrenamiento son estadísticamente independientes y contribuyen por igual en el proceso de clasificación. En otras palabras, las palabras de un mensaje igual no mantienen ningún tipo de relación entre sí.

### 12.2.2. Bernoulli Naives Bayes [39]

En este modelo las características son booleanos independientes (variables binarias) que describen entradas. Este modelo es popular para clasificar textos cortos porque puede modelar explícitamente la ausencia de términos.

### 12.2.3. Linear Support-Vector Machine [40]

Son modelos que construyen un hiperplanos en un espacio dimensional. Se construyen dos clases con datos de muestra previamente etiquetadas, y estas clases se sitúan en diferentes extremos del hiperplano unidas por un vector soporte y cuando se van incluyendo los nuevos datos pueden ser clasificados a una o a otra clase.

### 12.2.4. Logistic Regression [41]

La regresión logística predice la probabilidad de un resultado que solo puede tener dos valores (es decir, una dicotomía). La predicción se basa en el uso de uno o varios predictores (numéricos y categóricos). Una regresión lineal

no es apropiada para predecir el valor de una variable binaria por dos razones:

Una regresión lineal predecirá valores fuera del rango aceptable (por ejemplo, predicción de probabilidades fuera del rango entre 0 a 1).

Dado que los experimentos dicotómicos solo pueden tener uno de dos valores posibles para cada experimento, los residuos no se distribuirán normalmente alrededor de la línea predicha.

Por otro lado, una regresión logística produce una curva logística, que se limita a valores entre 0 y 1. La regresión logística es similar a una regresión lineal, pero la curva se construye utilizando el logaritmo natural de las "probabilidades" de la variable objetivo, en lugar de la probabilidad. Además, los predictores no tienen que estar distribuidos normalmente o tienen una varianza igual en cada grupo.

12.2.5. Multinomial Naive Bayes [39]

Multinomial Naive Bayes son una versión especializada de Naive Bayes que está diseñada más para documentos de texto. Mientras que el modelo Naive Bayes simple modelaría un documento como la presencia y ausencia de palabras particulares, el MNB modela explícitamente los conteos de palabras y ajusta los cálculos subyacentes.

12.2.6. SGDC Classifier [42]

SGDClassifier es un clasificador lineal (SVM, regresión logística, a.o.) con entrenamiento SGD.

Este estimador implementa modelos lineales regularizados con aprendizaje de descenso de gradiente estocástico (SGD): el gradiente de la pérdida se estima en cada muestra a la vez y el modelo se actualiza en el camino con un programa de fuerza decreciente (también conocido como tasa de aprendizaje). Para obtener los mejores resultados utilizando el programa de tasas de aprendizaje predeterminado, los datos deben tener una media de cero y una variación de unidades.

Esta implementación funciona con datos representados como matrices densas o dispersas de valores de punto flotante para las características.

El rendimiento de los algoritmos es obtenido por medio de métricas que evalúan la eficacia de los anteriores algoritmos al clasificar nuevos datos. Algunas métricas conocidas son la de Exactitud, [] que representa la razón entre las predicciones correctas sobre el total de predicciones realizadas, la de Precisión, [] que es la razón

entre el número de documentos clasificados correctamente como pertenecientes a una clase y el número total de documentos que han sido clasificados por el modelo o algoritmo como de esa misma clase.

## 12.3 Proceso de entrenamiento de los algoritmos

El primer paso para realizar el entrenamiento en un sistema de aprendizaje automático es preparar los datos del corpus. Para esto se deben limpiar los datos y normalizar la información con el fin de reducir el sesgo de la información que pueda influir en el resultado final. Acto seguido, los textos se tokenizan, proceso en el cual divide los textos en unidades más pequeñas, usualmente palabras. Finalmente, estas características son usadas para entrenar los algoritmos:

**Ilustración 4. Fases del entrenamiento de algoritmos**


### 12.3.1 Pre-procesamiento de los datos

Este paso consiste en limpiar y normalizar la información para evitar demasiado sesgo en la información que pueda influir de manera negativa en el resultado final.

Este paso es fundamental pues al estarse trabajando con mensajes publicados en redes sociales pueden estar sujetos a mensajes que contengan faltas de ortografía, repeticiones de caracteres, mezclas entre mayúsculas y minúsculas, uso de jerga y abreviaturas, entre otros.

Las reglas más comunes para normalizar mensajes son las siguientes:

Normalización de mayúsculas y minúsculas:

Pese a que para una persona las palabras: <<casa>> y <<CASA>> sean iguales y signifiquen lo mismo, para un algoritmo no. Un algoritmo reconoce estas palabras como distintas. Para evitar posibles errores en la clasificación todas las letras mayúsculas se reemplazan a su equivalente en minúscula, quedando todos los mensajes en letras minúsculas.

Duplicidad de caracteres:

Al escribir en redes sociales es común repetir las mismas letras en las palabras para dar intensidad a lo que se intenta trasmitir. No es lo mismo escribir: <<Hola>> o <<Hooooola>>. Aunque ambas palabras signifiquen lo mismo en la segunda se da más énfasis al saludo, por lo tanto, es recomendado tratar esa frase eliminando las vocales que se repiten más de dos veces, quedando el ejemplo anterior así: <<Hoola>>.

Eliminación de tildes:

En las redes sociales el uso de las tildes no es generalizado por lo tanto para evitar errores se recomienda eliminar las tildes pues para los algoritmos una palabra con tilde y la misma, pero sin tilde significan lo mismo.

Eliminación de números:

Los números normalmente no tienen información que ayude al proceso de clasificación por lo que se recomienda removerlos.

Normalización de risas:

La risa en redes sociales se puede caracterizar de diferentes formas, emojis, o algunas palabras como <<jajaja>>, <<jejeje>>, <<jijiji>>, etc. La risa puede ser un factor importante al momento de declarar la polaridad de un sentimiento por lo tanto es básico representar la risa en un mismo set de caracteres.

Normalización de jerga:

Comúnmente se suele usar jerga o abreviaciones en los mensajes trasmitidos en redes sociales. No es extraño encontrarse con <<q>> en lugar de <<que>> o <<tb>> en lugar de <<también>>. Un buen proceso de normalización debe tener en cuenta un set de reglas para sustituir esta jerga por sus palabras completas.

### 12.3.2 Tokenización

Al finalizar la normalización se procede a hacer la tokenización. Los textos se dividen en tokens o unidades más pequeñas, normalmente las palabras que conforman una frase serían los tokens.

### 12.3.3 Extracción de las características

Con los tokens extraídos en el paso anterior usando bolsas de palabras donde se representa cada mensaje con tokens.

### 12.3.4 Reducción de las características

Se eliminan stopwords. Una stopword es una palabra que, si bien es necesaria para construir una oración, no cuenta con información para ayudar a clasificar los textos en los que se encuentran. Ejemplo: <<El>>, <<la>>, <<a>>, etc.

Al realizar los pasos anteriores ya se pueden entrenar los algoritmos y luego realizar la clasificación con el resto de datos.

# 13. DISEÑO Y DESARROLLO DE LA HERRAMIENTA

El software desarrollado es una aplicación web que implementa Análisis de Sentimiento para determinar el sentimiento que tienen las opiniones trasmitidas por personalidades influyentes en la política colombiana en redes sociales, en este caso Twitter, la aplicación se encarga de recolectar los datos, en este caso tweets, procesarlos y analizarlos para cumplir con el Análisis de Sentimiento.

## 13.1. Herramientas para hacer Análisis de Sentimiento.

Actualmente existen varias herramientas que sirven para hacer Análisis de Sentimiento [43], conocer estas herramientas puede servir como punto de base para el desarrollo de la herramienta del proyecto:

**Tabla 18. Herramientas de Análisis de Sentimiento**

| Herramienta | Descripción | Precio |
|---|---|---|
| BRAND24 | Brand24 es una herramienta de escucha social y análisis de sentimientos.<br><br>Permite “escuchar” en la web sobre lo que se está diciendo de una marca en específico. Permite revisar desde comentarios en redes sociales hasta críticas profesionales. | $ 49 USD |
| CLARABRIDGE | Desarrollado con inteligencia artificial y aprendizaje automático.<br><br>Clarabridge utiliza análisis de texto para proporcionar un medio altamente efectivo para considerar lo que dicen los clientes de una empresa.<br><br>Recopila llamadas, correos electrónicos, chats, encuestas, interacciones sociales, calificaciones y comentarios, o foros, Clarabridge se conecta a cientos de fuentes y recopila todos los comentarios de sus clientes en un solo lugar. | $ 35 USD |

| | | |
|---|---|---|
|  | Repustate proporciona análisis de texto para empresas en 17 idiomas diferentes. Antes de que tenga lugar el análisis real, la herramienta desencadena un mecanismo conocido como etiquetado de parte del discurso, que se basa en la descomposición de bloques de texto en partes gramaticales: | $ 99 USD |
|  | OpenText Sentiment Analysis Tool es parte de una solución más amplia de OpenText Content Analytics. Es una clasificación especializada que puede usar para identificar y evaluar patrones subjetivos y expresiones de sentimiento dentro del contenido textual. | $ 50 USD |
|  | ParrallelDots es un grupo de investigación de IA aplicada.<br><br>Sirve para codificar las respuestas, clasificar los datos y extraer información de texto y videos utilizando una interfaz altamente optimizada. Se proporcionan soluciones de análisis de contenido poderosas y flexibles impulsadas por la IA que llevan el poder de la NLP y la Visión por Computador a las masas. | $ 79 USD |

13.2. Diseño y desarrollo de la herramienta.

La herramienta debió ser desarrollada siguiendo los siguientes parámetros:

El sistema debe gestionar la recolección de datos (tweets) mediante el uso de la API de Twitter introduciendo un criterio de búsqueda sobre el cual se quiere realizar el Análisis de Sentimiento. Para los criterios de búsqueda se tienen en cuenta dos en especial, la búsqueda por usuario, ejemplo: @usuario; y la búsqueda por tendencia, ejemplo: #hashtag.

El sistema debe encargarse del pre-procesamiento de datos, normalizando los datos, tokenizándolos y haciendo más procesos incluidos en el pre-procesamiento.

El sistema debe aplicar modelos de procesamiento de Natural Languaje Procesing para realizar el Análisis de Sentimiento clasificando los tweets por sentimiento.

El sistema con las clasificaciones de sentimientos deberá graficar los sentimientos obtenidos en tres columnas: Positiva, Negativa y Neutral.

La herramienta se desarrolló en lenguaje R, R es un lenguaje de programación y un entorno de software libre para computación estadística y gráficos [44]. Cuenta con un único script que se encarga de la recolección, procesamiento y visualización de resultados, mediante gráficas, los datos, tweets, se descargan directamente desde la API de Twitter, el texto se procesa usando la librería tm[1] y con los datos procesados se grafican usando tres columnas, una para tweets positivos, otra para tweets negativos y una finalmente para tweets neutrales. Este script se puede ejecutar en el servidor local del computador.

Visualización del script desarrollado:

[1] Paquete de minería de texto de R: https://cran.r-project.org/web/packages/tm/tm.pdf

**Ilustración 5. Script desarrollado**

```
library(shiny)
library(tm)
library(wordcloud)
library(twitteR)
shinyServer(function(input, output, session) {
    setup_twitter_oauth(consumer_key = "hWI7IzxIQibeVEzU9dHMErbpT",
                        consumer_secret = "uNJdWaxMpVSH76RpbCQwPOXZUFSWkANWUj3kuQvftDwEVZm0uJ",
                        access_token =  "238339004-8SeWy66JSTE9wUR0pbugdRRpjPyt8zAkr6sraJrl",
                        access_secret= "SCfU3vRFNXGth8YGXz1L8vzZtxUGbBhYjH1Szfh0aN4cW")
    output$currentTime <- renderText({invalidateLater(100000, session)
    paste("TiempoReal:",Sys.time())})
    observe({
        nvalidateLater(6000000,session)
        count_positive = 0
        count_negative = 0
        count_neutral = 0
        positive_text <- vector()
        positive_split<-list()
        negative_text <- vector()
        negative_split <- list()
        neutral_text <- vector()
        neutral_split <- list()
        vector_users <- vector()
        vector_sentiments <- vector()
        tweets_result = ""
        tweets_result = searchTwitter("@alvarouribevel", n=100)
        for (tweet in tweets_result){
            print(paste(tweet$screenName, ":", tweet$text))
            vector_users <- c(vector_users, as.character(tweet$screenName));
            if (grepl("dÃ-as", tweet$text, ignore.case = FALSE) |
                grepl("buen", tweet$text, ignore.case = FALSE) |
                grepl("abrazo", tweet$text, ignore.case = FALSE)){
                    count_positive = count_positive + 1
                    vector_sentiments <- c(vector_sentiments, "Positivo")
                    positive_text <- c(positive_text, as.character(tweet$text))
                    positive_split<-strsplit(positive_text, split=" ")
                    positive_columnas <- data.frame(unlist(positive_split))
                    tbl_positive <- table(positive_columnas)
            } else if (grepl("aburrido", tweet$text, ignore.case = FALSE) |
                grepl("serio", tweet$text, ignore.case = FALSE) |
                grepl("no", tweet$text, ignore.case = FALSE)) {
                    count_negative = count_negative + 1
                    vector_sentiments <- c(vector_sentiments, "Negativo")
                    negative_text <- c(negative_text, as.character(tweet$text))
                    negative_split=strsplit(negative_text, split=" ")
                    negative_columnas = data.frame(unlist(negative_split))
                    tbl_negative <- table(negative_columnas)
            } else {
                count_neutral = count_neutral + 1
                print("neutral")
                vector_sentiments <- c(vector_sentiments, "Neutral")
                neutral_text <- c(neutral_text, as.character(tweet$text))
                neutral_split<-strsplit(neutral_text, split=" ")
                neutral_columnas<-data.frame(unlist(neutral_split))
                tbl_neutral <- table(neutral_columnas)
            }
        }
        df_users_sentiment <- data.frame(vector_users, vector_sentiments)
        output$tweets_table = renderDataTable({ df_users_sentiment})

        output$distPlot <- renderPlot({
            results = data.frame(tweets = c("Positivo", "Negativo", "Neutral"), numbers =
            c(count_positive,count_negative,count_neutral))
            barplot(results$numbers, names = results$tweets, xlab = "Sentimiento", ylab =
            "Cantidad", col = c("Orange","Gray","Blue"))
            if (length(positive_text) > 0){
```

```
                output$positive_wordcloud <- renderPlot({ wordcloud(names(tbl_positive),
                tbl_positive, min.freq = 1, scale=c(1,.5),random.color=TRUE)  })
            }
            if (length(negative_text) > 0) {
                output$negative_wordcloud <- renderPlot({ wordcloud(names(tbl_negative),
                tbl_negative, min.freq = 1, scale=c(1,.5),random.color=TRUE)  })
            }
            if (length(neutral_text) > 0){
                output$neutral_wordcloud <- renderPlot({ wordcloud(names(tbl_neutral),
                tbl_neutral, min.freq = 1, scale=c(1,.5),random.color=TRUE)  })
            }
        })
    })
})
```

# 14. RESULTADOS

Como se determinó en el planteamiento del problema, después de una investigación sustentada en el número de seguidores e influencia en Twitter se establecieron a los siguientes políticos para realizar el estudio:

Juan Manuel Santos Calderón (@JuanManSantos).

Álvaro Uribe Vélez (@AlvaroUribeVel).

Gustavo Francisco Petro Urrego (@gustavopetro).

Antanas Mockus Šivickas (@AntanasMockus).

Claudia López Hernández (@ClaudiaLopez).

Estos políticos se han caracterizado en los últimos años por ser muy influyentes, polémicos y con gran capacidad para movilizar masas.

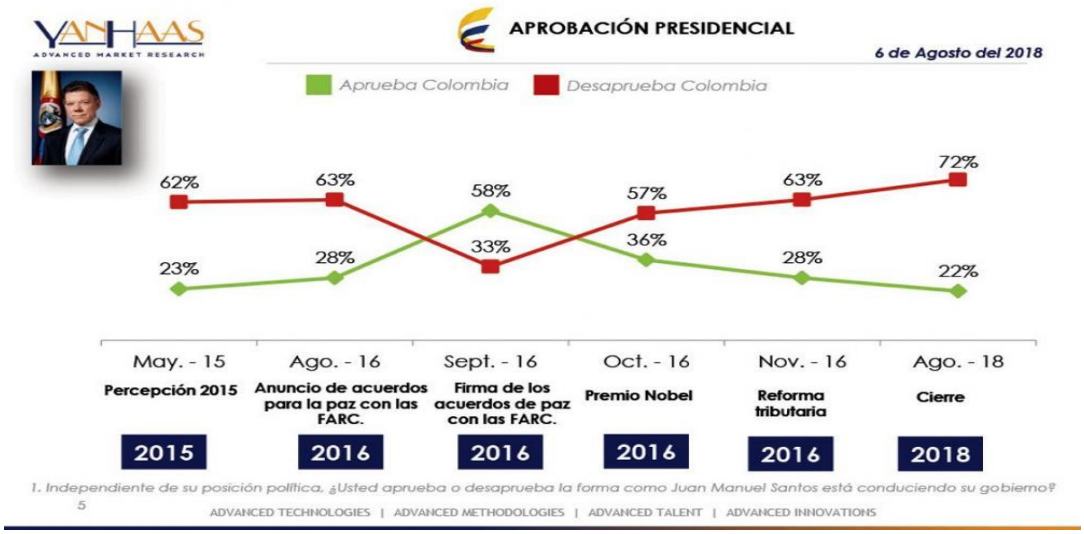
Estos cinco políticos representan corrientes e ideologías distintas, tanto que entre ellos se han presentado roces, discusiones más o menos acaloradas, entre otros.

A continuación, se presentará lo encontrado al realizar el estudio.

## 14.1. Juan Manuel Santos

El 31 de mayo de 2019, al momento de realizar el estudio, Juan Manuel Santos llevaba casi diez meses sin ser Presidente del país, sin embargo, en sus dos administraciones se evidenció el claro descontento de la ciudadanía con su gestión:

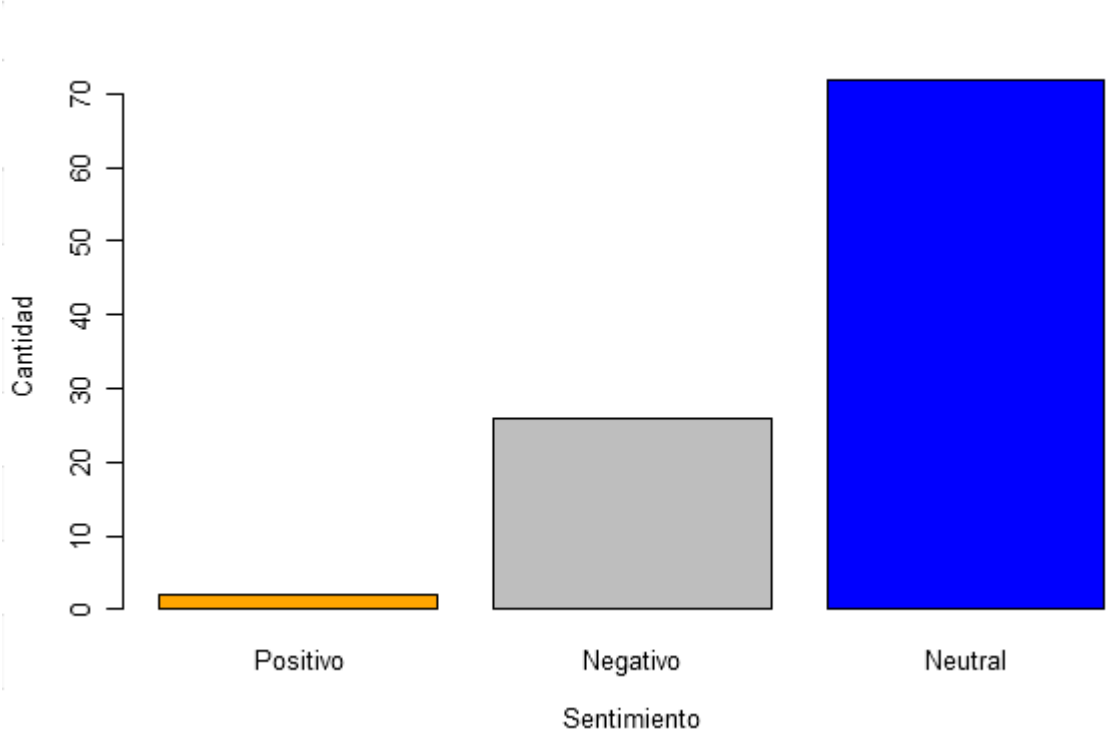
**Ilustración 6. Aprobación del mandato de Juan Manuel Santos. Tomado de Yanhaas.**



La gráfica muestra tendencia a la desaprobación de la gestión de Santos, es de inferir que esta tendencia se mantenga al momento de realizar el análisis, debido a que no lleva mucho tiempo siendo ex presidente y su gran legado de sus ocho años mandato fueron los Acuerdos de Paz con las FARC que surgieron de cuatro años de negociaciones en la Habana entre delegados de su gobierno y de las FARC. Estos acuerdos contaron con una férrea oposición del bloque liderado por el expresidente Uribe, quien desde el primer momento declaró su inconformidad con esos diálogos [45], y esto le pudo haber valido a Santos su gran impopularidad entre los colombianos [46].

Al realizar el Análisis de Sentimiento se realiza la siguiente gráfica:

**Ilustración 7. Análisis de Sentimiento para Juan Manuel Santos**

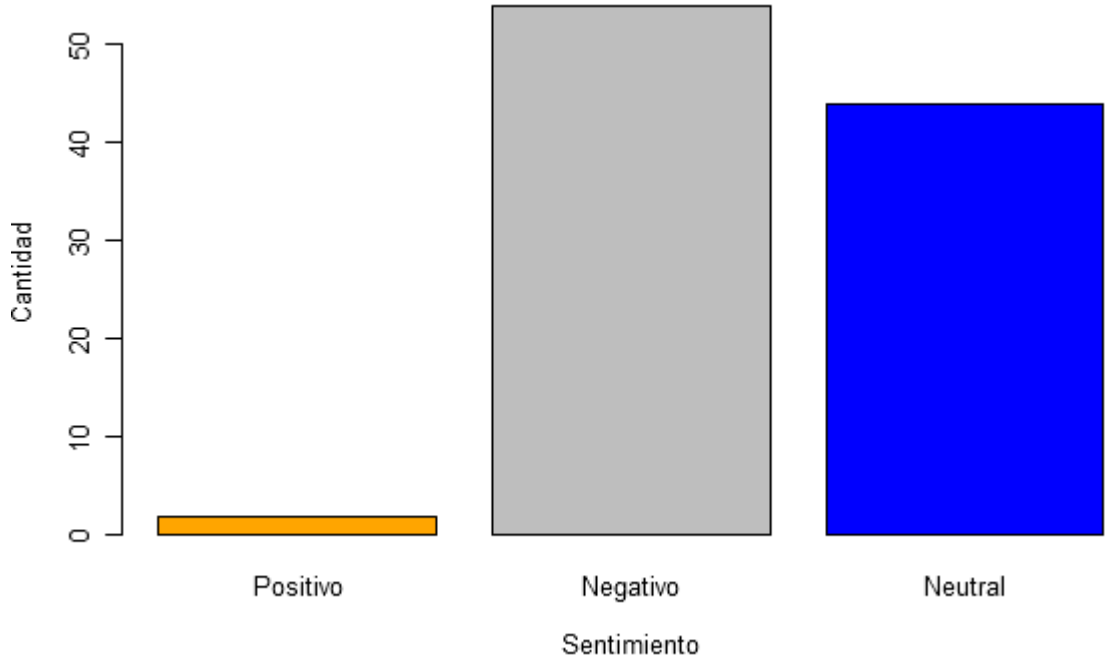



Se evidencia que en los tweets analizados hay un claro sentimiento neutral seguido de un sentimiento negativo, y un poco de sentimiento positivo. Con respecto a lo del sentimiento negativo podemos ver una concordancia con la valoración de su presidencia y respecto a los tweets neutrales; Santos al terminar su mandato anunció su retiro de la política [47] y se ha dedicado a dar conferencias en diferentes universidades del mundo sin referirse a la situación política actual, por lo que se entiende que estos tweets neutrales sean de agencias o medios de información que trasmitan las últimas actividades de Santos.

## 14.2. Álvaro Uribe Vélez

Álvaro Uribe Vélez es quizá el político que más genera polarización en el país. Esto se debe a su rechazo al proceso de paz, sus constantes discusiones y críticas a otros políticos [48]. Esto hace pensar que en su mayoría podría suscitar opiniones negativas entre los usuarios de Twitter.

**Ilustración 8. Análisis de Sentimiento para Álvaro Uribe**




En las gráficas arrojadas al realizar el Análisis muestran que claramente Álvaro Uribe suscita un sentimiento negativo en Twitter, seguido de un sentimiento neutral y por último un muy poco sentimiento positivo. Se ve que las posiciones políticas del senador Uribe influyen en su valoración altamente negativa en Twitter. También tiene un alto componente de sentimiento neutral, es porque al ser un personaje muy mediático muchos medios de comunicación informan día a día de las actividades y declaraciones que él realiza.

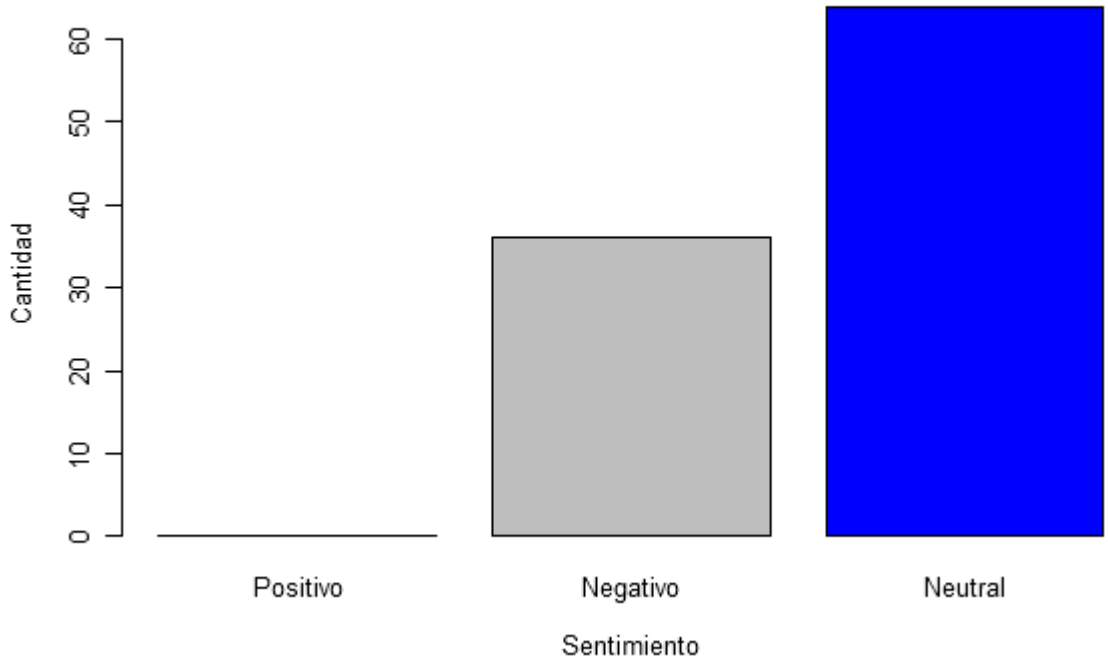
## 14.3. Gustavo Petro

Gustavo Petro es un político de ideología izquierda que siempre se ha caracterizado por sus posturas contrarias a las del senador Álvaro Uribe por lo que en más de una ocasión se ha visto envuelto en discusiones de mayor o menor calibre con el mencionado senador y con otros senadores de derecha.

Petro no es el político más apoyado por personas que tienen ideología derechista por lo que es de esperar que en las redes sociales gran parte de los comentarios de personas con afinidad a la derecha no sean tan positivos para él.

Otro punto que hay que tomar es que para Petro las redes sociales son un bastión para trasmitir sus ideas a su gran número de seguidores, según el informe anual de 2018 presentado por Twitter, Petro fue el personaje político con más relevancia en esta red social [49].

**Ilustración 9. Análisis de Sentimiento para Gustavo Petro**



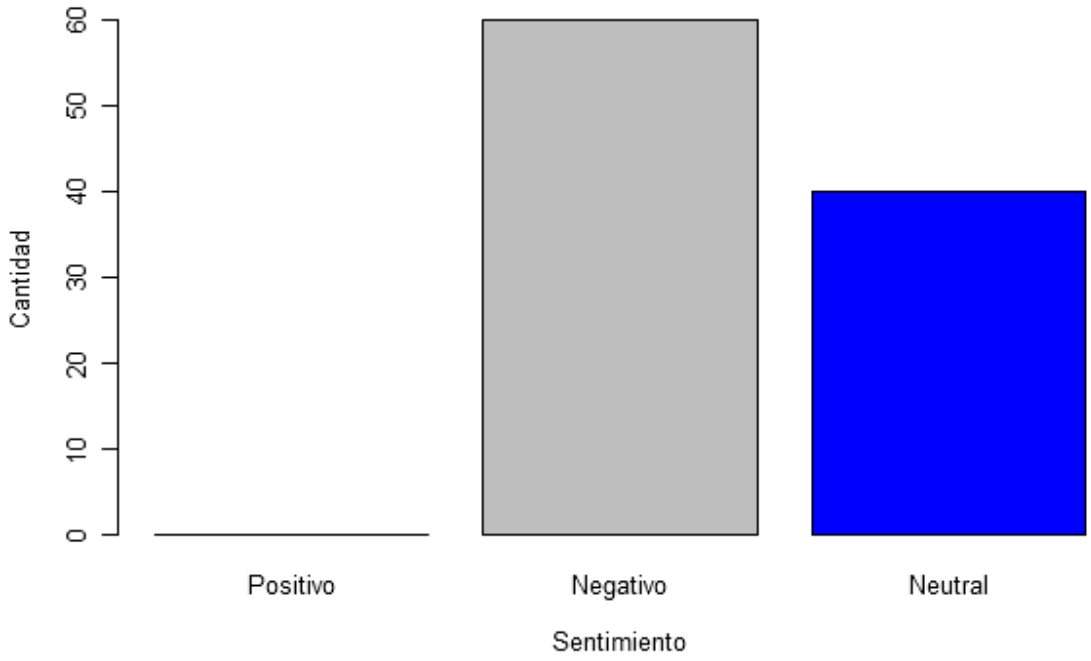

Al revisar la gráfica resultante del análisis de sentimiento se encuentra que el componente neutral es predominante en la valoración de Petro, seguido de un componente negativo y un componente positivo mínimo. Se encuentra una característica especial ya que en este personaje predomina el sentimiento neutral, situación que hasta el momento no se había encontrado en Santos y Uribe, personalidades previamente estudiadas. También es de notar que prácticamente carece de sentimiento positivo en la muestra de tweets tomada.

### 14.4. Antanas Mockus

Antanas Mockus es un político de centro derecha, un personaje que se ha caracterizado por evitar que sus declaraciones o hechos aumenten o generen polarización. Fue candidato a las elecciones presidenciales de 2010 donde cayó en segunda vuelta frente a Juan Manuel Santos, pero esta campaña de ya hace nueve años sirvió como punto de inflexión en la forma de cómo desde las redes sociales se puede abordar la política o unas elecciones, ya desde esa época se hablaba que el éxito de la campaña política de Mockus por Internet había sido impulsado por los mismos usuarios. Es decir, que no obedecía a una estrategia de sus asesores de campaña, sino que la divulgación y difusión del mensaje político es posible gracias

a la participación voluntaria de su electorado, seguidores virtuales, en su mayoría jóvenes [50].

**Ilustración 10. Análisis de Sentimiento para Antanas Mockus**




Al revisar su valoración llama la atención que tiene una valoración supremamente negativa, quizá esta valoración se debe a que hace pocas semanas el Consejo de Estado anuló su elección como senador en las elecciones del 11 de marzo del año pasado[2] y esto pudo generar enfrentamiento entre sus seguidores y contradictores en redes sociales, lo cual explicaría su valoración altamente negativa.

14.5. Claudia López [51]:

En Colombia, según la Registraduría Nacional, hay 36,024,467 de votantes de los cuales 18,605,726 son mujeres. O sea que el 52 % del potencial de votantes son mujeres. A pesar de esta mayoría, arenas políticas como el Congreso cuentan con solo el 18 % de mujeres electas. Esta es la cifra más alta de la historia con 52 mujeres electas de los 268 miembros. Para ONU Mujeres, el problema es estructural: las mujeres siguen estando rezagadas de la mayoría de espacios políticos y económicos por leyes discriminatorias, bajos niveles educativos, estereotipos de género y la pobreza (que afecta a las mujeres de manera desproporcionada).

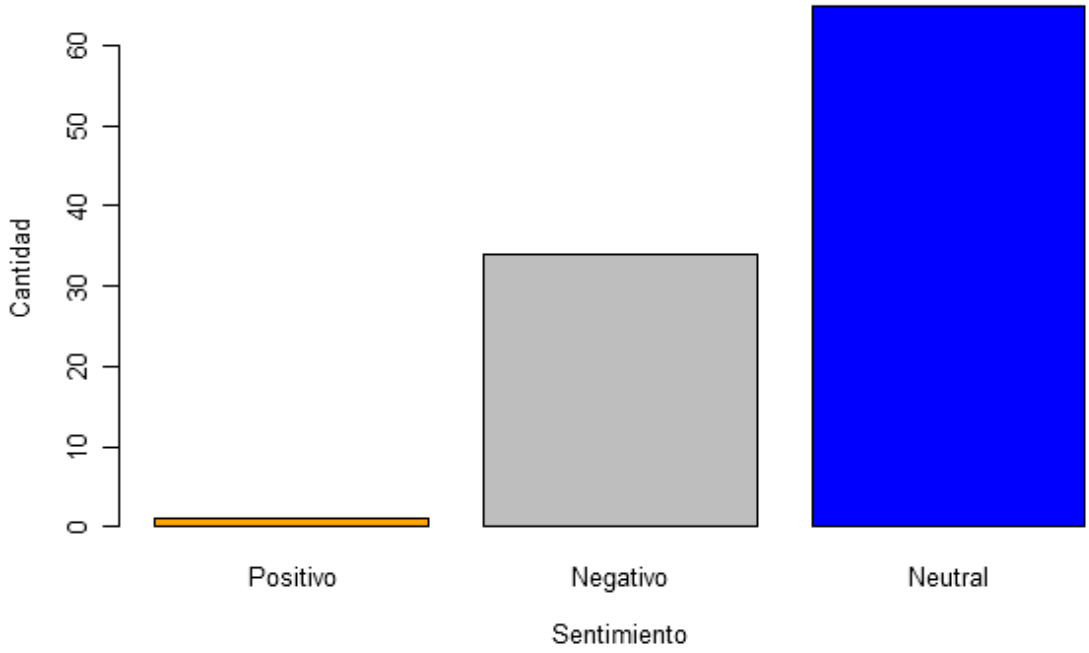
[2] Antanas Mockus pierde su curul en el Congreso: su elección fue anulada: https://www.elespectador.com/noticias/judicial/eleccion-de-antanas-mockus-como-senador-quedo-anulada-articulo-849928

En Colombia una de las mujeres más importantes en la política es Claudia López, senadora de la Alianza Verde. En 2014, López obtuvo la segunda mayor votación, en las mujeres candidatas, con 81.045 sufragios.

Claudia López es una política con gran trayectoria e influencia en el país. Claudia López tiene fama de ser una mujer con temperamento fuerte, impulsiva, eso ha sido bueno para su vida pública pero también le ha creado varios conflictos [52].

Conflictos que pueden verse reflejados en su valoración de sentimientos.

**Ilustración 11. Análisis de Sentimiento para Claudia López**




Al observar los resultados de Claudia López encontramos una situación similar a la de Gustavo Petro, predomina el sentimiento neutral, seguido del negativo y un poco de sentimiento positivo. La razón del sentimiento negativo puede ser los continuos roces que ella ha tenido con personajes como Álvaro Uribe y como se decía anteriormente su fuerte temperamento y su característica al ser muy impulsiva.

## 14.6. Conclusiones de los resultados

Se encuentra al revisar y analizar las gráficas que ninguno de los políticos analizados genera grandes sentimientos positivos, esto puede deberse a la

polarización que enfrenta el país entre ideologías de izquierda y de derecha, ideologías que son representadas entre los diferentes personajes analizados así:

**Tabla 19. Ideologías políticas de las personalidades analizadas.**

| Santos | Uribe | Petro | Mockus | C. López |
|---|---|---|---|---|
| Centro derecha | Derecha | Izquierda | Centro izquierda | Centro izquierda |

Álvaro Uribe y Antanas Mockus son los políticos con peor valoración, ambos pertenecientes a corrientes políticas distintas.

Gustavo Petro es el político que tiene el mayor porcentaje neutral.

No es posible encontrar un político que tenga una alta valoración positiva, para encontrar los políticos basta con ver la columna de sentimiento negativo y se concluye que Juan Manuel Santos y Claudia López lo son ya que tienen los porcentajes más bajos de sentimiento negativo.

# 15. CONCLUSIONES

Se encuentra que al determinar el sentimiento que tienen las opiniones trasmitidas por personalidades influyentes en la política colombiana en redes sociales como Twitter son de carácter negativo para todas las personalidades analizadas, teniendo en cuenta que en todos los análisis el porcentaje de sentimiento negativo supera claramente al negativo, lo que varía es la proporción de sentimiento neutral en el que para algunos políticos es mayor, pero para otros no; lo anterior se ha determinado a través del uso de técnicas de Análisis de Sentimiento.

El análisis concuerda con la radiografía que se tiene de Colombia en el que se encuentra que es un país con altos niveles de polarización que tiene a la población fragmentada entre los que apoyan a ideologías de izquierda vs los que apoyan a ideologías de derecha e incluyendo a las personas que no se alinean con las dos ideologías anteriores, sino que optan por la del centro [53], [54], esto genera choque de opiniones al momento de tomar una decisión como la de emitir un voto por algún candidato en unas elecciones.

El Análisis de Sentimiento clasificando textos en español es un área que todavía no está muy bien explorada pues la mayoría de trabajos, investigaciones, algoritmos, modelos, etc.; están desarrollados para inglés, suponiendo un reto hacerlo en español.

Sumado a lo anterior, el Análisis de Sentimiento en el área de política está muy poco explorado en Colombia, la búsqueda de literatura arrojó que solo se había desarrollado un trabajo de investigación en esta área de investigación[3]. Lo cual constituye un reto a futuro seguir trabajando en esta área debido a que como se mostró las redes sociales constituyen un termómetro para conocer la opinión que tiene parte de la población sobre algún tópico en específico, en este caso la política, más concretamente, actores políticos.

Finalmente, conocer el sentimiento que emite un político en redes sociales puede ser fundamental porque podría servir para replantear su estrategia o discurso político en vista de los buenos o malos resultados encontrados.

[3] J. A. A. Valencia, A. Carrillo, J. Forero, L. Caicedo, and J. C. Urueña, “Análisis de sentimiento político en twitter para las elecciones de la alcaldía de Bogotá 2015,” 2016.

## 16. REFERENCIAS BIBLIOGRÁFICAS


[1] C. Grosvenor, “Social Networking Boom,” 2017. [Online]. Available: https://socialnetworking.lovetoknow.com/Social_Networking_Boom.

[2] H. Wiesenhofer, M. Ebner, and I. Kamrat, “Is Twitter an Individual Mass Communication Medium?,” *C. Crawford al. (Eds.), Proc. Soc. Inf. Technol. Teach. Educ. Int. Conf. 2010*, pp. 1712–1717, 2010.

[3] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a Social Network or a News Media?,” *Proceedings of the 19th international conference on World wide web*, 2010. [Online]. Available: http://an.kaist.ac.kr/~haewoon/papers/2010-www-twitter.pdf.

[4] G. Greenwald, “How Twitter Degrades Discourse and Encourages Distortions: Illustrated by Ex-Pentagon Official Ryan Goodman,” *The Intercept*, 2018. [Online]. Available: https://theintercept.com/2018/07/15/how-twitter-degrades-discourse-and-encourages-distortions-illustrated-by-ex-pentagon-official-and-nyu-law-professor-ryan-goodman/.

[5] B. Caplan, “Political Power-Lust Thrives in a Democracy,” *Foundation for Economic Education*, 2017. [Online]. Available: https://fee.org/articles/political-power-lust-thrives-in-a-democracy/.

[6] C. Weller, “Here’s why Trump can legally get away with saying things that aren’t true,” *Business insiders*, 2016. [Online]. Available: https://www.businessinsider.com/why-politicians-dont-get-sued-over-ads-2016-11.

[7] U. Ecker, “Where does misinformation come from, and what does it do?,” *The Conversation*, 2012. [Online]. Available: https://theconversation.com/where-does-misinformation-come-from-and-what-does-it-do-9885.

[8] G. Jacob and S. Bronda, “How social media can distort and misinform when communicating science,” *The Conversation*, 2016. [Online]. Available: http://theconversation.com/how-social-media-can-distort-and-misinform-when-communicating-science-59044.

[9] S. Bakers, “Social Bakers.” [Online]. Available: www.socialbakers.com.

[10] A. York, “Alcance versus impresiones: entendiendo los términos de engagement,” *Sproutsocial*, 2018. [Online]. Available: https://sproutsocial.com/insights/alcance-versus-impresiones/.

[11] TweetReach, “TweetReach.” [Online]. Available: https://tweetreach.com/.

[12] B. Liu, “Sentiment analysis and opinion mining,” *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.

[13] J. A. Manrique Hernández, “Identificación de la intención de consultas del usuario en búsquedas Web,” 2016.

[14] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, “Supervised machine learning: A review of classification techniques,” *Emerg. Artif. Intell. Appl. Comput. Eng.*, vol. 160, pp. 3–24, 2007.

[15] C. Quan and F. Ren, “Unsupervised product feature extraction for feature-oriented opinion determination,” *Inf. Sci. (Ny).*, vol. 272, pp. 16–28, 2014.

[16] X. Zhu, “Semi-supervised learning literature survey,” *Comput. Sci. Univ. Wisconsin-Madison*, vol. 2, no. 3, p. 4, 2006.

[17] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.

[18] A.-H. Tan, “Text Mining: The state of the art and the challenges,” *Proc. PAKDD 1999 Work. Knowl. Disocovery from Adv. Databases*, 1999.

[19] G. G. Chowdhury, “Natural language processing,” *Annu. Rev. Inf. Sci. Technol.*, vol. 37, no. 1, pp. 51–89, 2003.

[20] D. Zeevi, “Twitter 101: What is Twitter Really About?,” *Social Media Today*, 2013. [Online]. Available: https://www.socialmediatoday.com/content/twitter-101-what-twitter-really-about.

[21] D. Nations, “What Is a Tweet on Twitter?,” *Lifewire*, 2017. [Online]. Available: https://www.lifewire.com/what-is-a-tweet-3486211.

[22] Twitter, “Retweet FAQs,” *Twitter*. [Online]. Available: https://help.twitter.com/en/using-twitter/retweet-faqs.

[23] Twitter, “How to like a Tweet or Moment,” *Twitter*. [Online]. Available: https://help.twitter.com/en/using-twitter/liking-tweets-and-moments.

[24] Twitter, “About your Twitter timeline,” *Twitter*. [Online]. Available: https://help.twitter.com/en/using-twitter/twitter-timeline.

[25] L. F. Cote Peña, “Hábeas data en Colombia, un trasplante normativo para la protección de la dignidad y su correlación con la NTC/ISO/IEC 27001:2013,” *XIX Edición del Premio Protección Datos Pers. Investig. la Agencia Española Protección Datos*, 2015.

[26] P. Palazzi, “What is habeas data?,” *Habeasdata.org*, 2011. [Online]. Available:

http://www.habeasdata.org/wp/2011/06/11/what-is-habeas-data/.

[27] Z. Zhai, B. Liu, J. Wang, H. Xu, and P. Jia, “Product feature grouping for opinion mining,” *IEEE Intell. Syst.*, vol. 27, no. 4, pp. 37–44, 2012.

[28] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, “Predicting elections with twitter: What 140 characters reveal about political sentiment.,” *Icwsm*, vol. 10, no. 1, pp. 178–185, 2010.

[29] A. Ceron, L. Curini, S. M. Iacus, and G. Porro, “Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens political preferences with an application to Italy and France,” *New Media Soc.*, vol. 16, no. 2, pp. 340–358, 2014.

[30] F. Pla and L.-F. Hurtado, “Political Tendency Identification in Twitter using Sentiment Analysis Techniques,” in *Proceedings of the 25th International Conference on Computational Linguistics, COLING*, 2014, pp. 183–192.

[31] J. E. Chung and E. Mustafaraj, “Can collective sentiment expressed on twitter predict political elections?,” in *AAAI*, 2011, vol. 11, pp. 1770–1771.

[32] M. Choy, M. L. F. Cheong, M. N. Laik, and K. P. Shung, “A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction,” *arXiv Prepr. arXiv1108.5520*, 2011.

[33] K. Singhal, B. Agrawal, and N. Mittal, “Modeling Indian general elections: sentiment analysis of political Twitter data,” in *Information Systems Design and Intelligent Applications*, Springer, 2015, pp. 469–477.

[34] A. Bermingham and A. Smeaton, “On using Twitter to monitor political sentiment and predict election results,” in *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, 2011, pp. 2–10.

[35] C. Arcila-Calderón, F. Ortega-Mohedano, J. Jiménez-Amores, and S. Trullenque, “Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tweets basada en aprendizaje automático,” *El Prof. la Inf.*, vol. 26, no. 5, pp. 973–982, 2017.

[36] F. Agulló, A. Guillén, Y. Gutiérrez, and P. Martínez-Barco, “ElectionMap: una representación geolocalizada de intenciones de voto hacia partidos políticos sobre la base de comentarios de usuarios de Twitter,” *Proces. del Leng. Nat.*, no. 55, 2015.

[37] J. A. A. Valencia, A. Carrillo, J. Forero, L. Caicedo, and J. C. Urueña, “Análisis de sentimiento político en twitter para las elecciones de la alcaldía de Bogotá 2015,” 2016.

[38] J. C. Sobrino Sande, “Análisis de sentimiento en Twitter,” 2018.

[39] A. McCallum, “A Comparison of Event Models for Naive Bayes Text Classification,” 1998.

[40] “Support Vector Machines.” [Online]. Available: https://scikit-learn.org/stable/modules/svm.html.

[41] “Logistic Regression.” [Online]. Available: https://www.saedsayad.com/logistic_regression.htm.

[42] “SGD Classifier.” [Online]. Available: https://stackoverflow.com/questions/45455209/is-stochastic-gradient-descent-a-classifier-or-an-optimizer.

[43] K. Rogalski, “The 17 Best Sentiment Analysis Tools,” 2019. [Online]. Available: https://brand24.com/blog/best-sentiment-analysis-tools/.

[44] K. Hornik, “R FAQ,” 2018. [Online]. Available: https://cran.r-project.org/doc/FAQ/R-FAQ.html#What-is-R_003f.

[45] VanguardiaLiberal, “‘Nosotros no elegimos a Santos para que consintiera a las Farc’: Uribe,” 2013. [Online]. Available: https://www.vanguardia.com/colombia/nosotros-no-elegimos-a-santos-para-que-consintiera-a-las-farc-PYVL206889.

[46] ElTiempo, “Los hechos que marcaron los ocho años de Santos en la Casa de Nariño,” 2018. [Online]. Available: https://www.eltiempo.com/politica/gobierno/los-hechos-que-marcaron-los-ocho-anos-de-santos-en-la-casa-de-narino-252072.

[47] ElTiempo, “Juan Manuel Santos anunció que se retira de la política,” 2018. [Online]. Available: https://www.eltiempo.com/politica/gobierno/juan-manuel-santos-dice-que-se-retira-de-la-politica-248220.

[48] C. Rodríguez, M. Romero, L. R. Navarro, A. Saenz, L. Hernández, and O. Páez, “Medios de Redes Sociales y Polarización en Colombia. Uso de Twitter durante el Proceso de Paz Colombiano.,” 2015.

[49] RadioCaracol, “Petro es el personaje del año en redes sociales,” 2018. [Online]. Available: https://caracol.com.co/radio/2018/12/08/tendencias/1544226064_863472.html.

[50] DW, “Mockus: el fenómeno electoral de redes sociales en Colombia,” 2010. [Online]. Available: https://www.dw.com/es/mockus-el-fenómeno-electoral-de-redes-sociales-en-colombia/a-5622105.

[51] J. Ramírez, “Mujeres que transforman la política,” 2018. [Online]. Available: https://cerosetenta.uniandes.edu.co/mujeres-que-transforman-y-politica/.

[52] H. Marta, “Mujer y política: Claudia López va por la Alcaldía de Bogotá,” 2019. [Online]. Available: http://www.mensaje360.com/0/nota/index.vnc?id=2823.

[53] A. Gañán Moreno, “Colombia, un país polarizado,” 2018. [Online]. Available: https://www.elcolombiano.com/opinion/columnistas/colombia-un-pais-polarizado-YB8838313.

[54] ElColombiano, “¿Por qué los colombianos no superamos tanta polarización?,” 2019. [Online]. Available: https://www.elcolombiano.com/colombia/por-que-los-colombianos-no-superamos-tanta-polarizacion-JA10458067.