Discussion paper / Artículo de reflexión / Documento de discussão - Tipo 2

Election analysis in Colombia and Venezuela 2015 through sentiment analysis and Twitter

Sonia Ordoñez Salinas, Ph.D / sordonez@udistrital.edu.co

Juan Manuel Pérez Trujillo / jmperezt@correo.udistrital.edu.co

Romario Albeiro Sánchez Montero / rasanchezm@correo.udistrital.edu.co

Grupo de investigación Gesdatos

Universidad Distrital Francisco José de Caldas, Bogotá-Colombia

ABSTRACT This paper presents an analysis of the accounts of the main candidates in the regional elections on October 25, 2015 in Colombia (Bogotá, Medellin and Cali) and the official hashtags of the two main parties for parliamentary elections on December 6, 2015 in Venezuela (PSUV and MUD) in order to determine the positive or negative trends and compare them with the results of the respective elections. To develop the analysis, we resorted to the technique of sentiment analysis own of data mining and the use of descriptive statistics, concluding that sentiment analysis for estimating trends requires processes to monitor retweets, if you want acceptable results.

KEYWORDS Sentiment analysis; elections; natural language; Twitter; Apicultur; spammer.

Análisis de las elecciones en Colombia y Venezuela 2015 a través de análisis de sentimiento y Twitter

RESUMEN Este artículo presenta un análisis de las cuentas de los principales candidatos de las elecciones regionales del 25 de octubre de 2015 en Colombia (Bogotá, Medellín y Cali) y los hashtag oficiales de los dos partidos pincipales para las elecciones parlamentarias del 6 de diciembre de 2015 en Venezuela (MUD y PSUV), con el fin de determinar las tendencias positivas o negativas y compararlas con los resultados de las respectivas elecciones. Para el desarrollo del análisis se recurrió a la técnica de análisis de sentimiento, propio de la minería de datos, y al uso de estadísticas descriptivas; se concluye que el análisis de sentimiento para la estimación de tendencias requiere de procesos que permitan controlar los retweets, si se quieren resultados aceptables.

PALABRAS CLAVE Análisis de sentimiento; elecciones políticas; lenguaje natural; Twitter; Apicultor.

Análise das eleições na Colômbia e Venezuela 2015 através de análise de sentimento e Twitter

RESUMO Este artigo apresenta uma análise das contas dos principais candidatos às eleições regionais de 25 de Outubro, 2015, em Colômbia (Bogotá, Medellín e Cali) e os hashtag oficiais dos dois principais partidos para as eleições parlamentares de 06 de dezembro de 2015 na Venezuela (MUD e PSUV), a fim de determinar as tendências positivas ou negativas e comparará-las com os resultados das respectivas eleições. Para o desenvolvimento da análise foi utilizada a técnica de análise de sentimento, típica da mineração de dados, bem como estatísticas descritivas; conclui-se que a análise de sentimento para estimar tendências, requer processos que permitam monitorar os retweets, a fim de esperar resultados aceitáveis.

PALAVRAS-CHAVE Análise de sentimento; eleições políticas; linguagem natural; Twitter; Apicultor.

Received / Recepción: October 31, 2016 - Accepted / Aceptación: December 2, 2016 *doi:10.18046/syt.v14i39.2349*

I. Introduction

The traditional way of knowing the opinion of a group of people is through surveys validated by means of sampling and statistical studies, but today, with the use of Internet –and with it the use of large collections of information and own methods of artificial intelligence–, have been possible to generate estimates and patterns with levels very similar or higher than those obtained with surveys. Advertising campaigns have seen an opportunity to conduct their analysis using information provided by social networks and especially Twitter.

Marketing companies and political analysts widely use Twitter content to predict electors trends in presidential campaigns, especially those of young people, as they are the largest users of networks (Pasek, 2006), as an example the analysis of the US elections (Stirland, 2008; Pasek, 2006). Information from social networks related to presidential campaigns is used, not only for political purposes, but also for academic purposes, and has served as a source to analyze different themes, especially those related to the sentiment analysis.

In Colombia, these analyses are performed by the marketing companies, as a request from candidates or groups; in contrast, at an academic level there are few studies that perform this type of analysis. Thus, in order to corroborate how approximate trends are in relation to reality, in this work we used Twitter information from a week before the voting to the mayors of the main cities of Colombia, and the official groups and opponents of Venezuela that participated in the elections of representatives to the National Assembly.

Although the development of this article did not take into account factors such as design, representation of the sample, side effects that may occur given the penetration of the Internet in different socio-economic sectors and in general in the country (in 2014 Colombia had an Internet penetration of 52.6%, compared with the United States, 87.4%; Germany, 86.2%, and the Netherlands, 93.2%), the results, as will be seen below, are very close to reality.

Twitter appears in 2006 in order to allow users to post short messages, up to 140 characters, called tweets and visible to the public. Registered users can read and post tweets, but unregistered users can only read them. According to figures published on the official Twitter page, by the end of 2015, 500 million tweets were published per day, there are 316 million active monthly users and more

I. Introducción

La forma tradicional de conocer la opinión de un grupo de personas es a través de encuestas validadas por medio de muestreo y estudios estadísticos, pero hoy, con la utilización de la Internet –y con ello con el aprovechamiento de grandes colecciones de información y métodos propios de la inteligencia artificial–, se han logrado generar estimaciones y patrones con niveles muy similares o superiores a los logrados con encuestas. Las campañas publicitarias han visto una oportunidad de realizar sus análisis utilizando la información que proveen las redes sociales y en especial Twitter.

Las empresas de mercadeo y los analistas políticos ampliamente usan el contenido Twitter para predecir las tendencias de los electores en las campañas presidenciales, y en especial las de los jóvenes, pues ellos son los mayores usuarios de las redes (Pasek, 2006), como ejemplo los análisis realizados sobre las elecciones estadounidenses (Stirland, 2008; Pasek, 2006). La información proveniente de las redes sociales relacionada con las campañas presidenciales se utiliza, no solo con fines políticos, sino también académicos, y ha servido de fuente para analizar diferentes temáticas, en especial las relacionadas con el análisis de sentimiento.

En Colombia estos análisis son realizados por las empresas de mercadeo, a petición de los candidatos o grupos, en cambio a nivel académico son escasos los trabajos que realizan este tipo de análisis. Es así que, con el fin de corroborar qué tan aproximadas son las tendencias con relación a la realidad, en este trabajo se utilizó la información de Twitter de una semana antes de las votaciones a las alcaldías de las principales ciudades de Colombia, y de los grupos oficialistas y opositores de Venezuela que participaron en las elecciones de representantes a la Asamblea Nacional.

Pese a que para el desarrollo de este artículo no se tuvo en cuenta factores como el diseño, la representación de la muestra, los efectos secundarios que se pueden presentar dada la penetración de Internet en los diferentes sectores socio-económicos y en general en el país (para 2014 Colombia presentó una penetración a Internet de 52.6%, frente a Estados Unidos, 87.4%; Alemania, 86,2%; y los Países Bajos, 93,2%), los resultados, como se verá más adelante, son muy aproximados a la realidad.

Twitter aparece en 2006 con el fin de que los usuarios puedan publicar mensajes cortos, de hasta 140 caracteres, llamados *tweets* y visibles al público. Los usuarios registrados pueden leer y publicar *tweets*, pero los usuarios no registrados sólo pueden leerlos. Según cifras publicadas en la página oficial de Twitter, para finales de 2015, se publicaron 500 millones de *tweets* al día, existen 316 millones de usuarios activos mensuales y se manejan más de 35 idiomas (Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011). Estas cifras y la posibilidad de que los usuarios puedan expresar sentimientos y reacciones relacionados con

una temática o producto en tiempo real han hecho que Twitter se haya convertido en la fuente más efectiva para los investigadores del análisis de sentimiento (Agarwal et al., 2011 ; Thelwall, Buckley, & Paltoglou, 2011; Bifet & Frank, 2010; Pla & Hurtado, 2013). Los datos de Twitter se han utilizado, no solo para el ámbito político, sino para una amplia gama de estudios, entre ellos para análisis de mercado (Porshnev, Redkin, & Shevchenko, 2013; Kiplinger, 2011; Wolfram, 2010), análisis de finanzas (Mao, Counts, & Bollen, 2011) y movimientos sociales (Qureshi, Memon, Wiil, & Karampelas, 2011).

El análisis de sentimiento trata de identificar la opinión o juicio de valor sobre algún tema o producto, utilizando técnicas propias del procesamiento del lenguaje natural y la lingüística computacional. El análisis textual se puede clasificar según si este se refiere a un hecho o a una opinión (Liu, 2010): "... los hechos son expresiones objetivas sobre las entidades, eventos y sus propiedades, y las opiniones son generalmente subjetivas que describen sentimientos de la gente". Los pioneros en el área tratan dicha temática como Thumbs up or thumbs down … (Qureshi et al., 2011) y Thumbs up … (Pang, Lee, & Vaithyanathan, 2002; Turney, 2002), aprobación o desaprobación. Las opiniones se clasifican en regulares y comparativas, las regulares hacen referencia a una simple opinión, mientras que las comparativas expresen relación de similaridades o diferencias entre dos o mas entidades (Liu & Zhang, 2012).

A través de la minería de opinión se pueden resolver tareas como: reconocimiento y agrupamiento de la entidad; reconocimiento y agrupamiento del aspecto; reconocimiento y agrupamiento de la fuente de opinión; reconocimiento de la opinión; reconocimiento del sentimiento de dicha opinión y tiempo cuando la opinión es emitida. La entidad es una persona, producto, organización u objeto objetivo que es evaluado; el aspecto hace referencia al objeto del cual se opina; y la fuente de opinión es quien emite la opinión. Y el sentimiento que emite la opinión generalmente se clasifica en positiva, negativa o neutra (Liu & Zhang, 2012). Para resolver las tareas del análisis de sentimiento se han utilizado, tanto métodos no supervisados (como en Turney, 2002), supervisados (como en: Pang et al., 2002; Brown, 2012; Agarwal et al., 2011; Jiang, Yu, Zhou, Liu, & Zhao, 2011; Pla & Hurtado, 2013).

II. Revisión de la literatura

Trabajos académicos sobre la opinión positiva o negativa en el ámbito político se han realizados en diferentes países y a partir de diferentes técnicas, para el habla inglesa cabe resaltar los siguientes:

- el aplicado en Alemania (Tumasjan, Sprenger, Sandner, & Welpe, 2010) para la elección federal del Parlamento Nacional, que tuvo lugar el 27 de septiembre de 2009, donde se analizaron 104.003 *tweets* publicados en las semanas previas y además de hacer seguimiento a los candidatos, analizaron el comportamiento de los usuarios deduciendo, por ejemplo, que el 71% de 1.8

than 35 languages are handled (Agarwal, Xie, Vovsha, Rambow, and Passonneau, 2011). These figures and the possibility that users can express feelings and reactions related to a thematic or product in real time have become Twitter the most effective source for the researchers of the sentiment analysis (Agarwal et al., 2011; Thelwall, Buckley, and Paltoglou, 2011; Bifet and Frank, 2010; Pla and Hurtado, 2013). Twitter data has been used, not only for the political field, but for a wide range of studies, including market analysis (Porshnev, Redkin, and Shevchenko, 2013; Kiplinger, 2011; Wolfram, 2010), financial analysis (Mao, Counts, and Bollen, 2011) and social movements (Qureshi, Memon, Wiil, and Karampelas, 2011).

The sentiment analysis tries to identify the opinion or value judgment on some subject or product, using techniques of natural language processing and computational linguistics. Textual analysis can be classified according to whether it refers to a fact or an opinion (Liu, 2010): "... facts are objective expressions about entities, events and their properties, and opinions are generally subjective expressions that describe feelings from the people". The pioneers in the field treat this subject as Thumbs up or thumbs down ... (Qureshi et al., 2011) and Thumbs up ... (Pang, Lee, and Vaithyanathan, 2002; Turney, 2002), approval or disapproval. Opinions are classified as regular and comparative, regular refer to a simple opinion, while comparative express the relation of similarities or differences between two or more entities (Liu and Zhang, 2012).

Through the opinion mining we can solve tasks such as: recognition and grouping of the entity; recognition and grouping of the appearance; recognition and grouping of the source of opinion; recognition of opinion; recognition of the sentiment of said opinion and time when the opinion is issued. The entity is an objective person, product, organization or object that is evaluated; the aspect refers to the object of which an opinion is given; and the source of opinion is the one who issues the opinion. And the sentiment emitted by the opinion is usually classified as positive, negative or neutral (Liu and Zhang, 2012). In order to solve the tasks of the sentiment analysis it have been used both unsupervised (as in Turney, 2002), and supervised (as in: Pang et al., 2002; Brown, 2012; Agarwal et al., 2011; Jiang, Yu, Zhou, Liu, and Zhao, 2011; Pla and Hurtado, 2013) methods.

II. Literature review

Academic work about positive or negative opinion in the political field has been performed in different countries and from different techniques, for English speaking it is worth highlighting the following:

- the one applied in Germany (Tumasjan, Sprenger, Sandner, and Welpe, 2010) for the federal election of the National Parliament, which took place on September 27, 2009, where 104.003 tweets published in the previous weeks were analyzed and, in addition to monitoring the candidates, the behavior of the users was analyzed deducing, for example, that 71% of 1.8 million users had visited Twitter only once and 15% at least three times;
- also in Germany (Tumasjan, Sprenger, Sandner, and Welpe, 2011) 100.000 Twitter messages are used to predict preference for a political group, using sentiment analysis through classification methods;
- for the 2008 US campaign, the system presented in Wang, Can, Kazemzadeh, Bar, and Narayanan (2012), through supervised methods, allowed to perform the sentiment analysis in real-time for nine Republican candidates and Barack Obama, for which was used 17.000 tweet that were previously labeled;
- for the 2011 Singapore presidential election, through surveys and 16.616 tweets that were collected between the nomination period and the campaign period, the percentage of voting per candidate was estimated (Choy, Cheong, Laik, and Shung, 2011);
- for the 2011 Irish general election, Bermingham and Smeaton (2011) explored political sentiment through the mining of tweets;
- Selvan and Moh (2015) designed a framework using the flow of opinions of Twitter in real-time and a dictionary of weighted sentiments;
- Anjaria and Guddeti (2014) present studies for the 2012 presidential elections and the 2013 Karnataka elections, using neural networks and support vector machines [SVM];
- Nguyen, Varghese, and Barker (2013) present a software that analyzes the opinions made on Twitter and take as a case study the actual birth of 2013 in the United Kingdom, the results are presented using different statistical techniques;
- Razzaq, Qamar, and Bilal (2014) analyze the 2013 Pakistan elections using tweets.

millones de usuarios habían visitado Twitter solo una vez y el 15% por lo menos tres veces;

- también en Alemania (Tumasjan, Sprenger, Sandner, & Welpe, 2011) se utilizan 100.000 mensajes de Twitter para predecir la preferencia hacia un grupo político, para lo que utilizaron análisis de sentimiento a través de métodos de clasificación;
- para la campaña del 2008 en Estados unidos, el sistema presentado en Wang, Can, Kazemzadeh, Bar, & Narayanan (2012), a través de métodos supervisados, permitió hacer el análisis de sentimiento en tiempo real hacia nueve candidatos republicanos y Barack Obama, para lo que utilizaron 17.000 *tweet* que previamente fueron etiquetados;
- para la elección presidencial 2011 de Singapur, a través de encuestas y 16.616 *tweets* que se recogieron entre el período de nominación y el período de campaña, se estimó el porcentaje de votación por candidato (Choy, Cheong, Laik, & Shung, 2011);
- para la elección general irlandesa del 2011, Bermingham y Smeaton (2011) exploraron el sentimiento político a través de la minería de los *tweets*;
- Selvan y Moh (2015) diseñaron un *framework* utilizando el flujo de opiniones de Twitter en tiempo real y un diccionario de sentimientos ponderados
- Anjaria y Guddeti (2014) presentan los estudios para las elecciones presidenciales del 2012 y las elecciones de Karnataca del 2013, para lo que utilizan redes neuronales y máquinas de vectores [SMV];
- Nguyen, Varghese, y Barker (2013) presentan un software que analiza las opiniones realizadas en Twitter y toman como caso de estudio el nacimiento real de 2013 en el Reino Unido, los resultados se presentan utilizando varias técnicas estadísticas;
- Razzaq, Qamar, y Bilal (2014) analizan las elecciones de Pakistán de 2013, utilizando *tweets*.

Respecto de estudios en idioma español, específicamente para Colombia, cabe resaltar el presentado por Mamprin (2015) que se realizó para analizar el favoritismo de los candidatos a la alcaldía de Cali de 2015; para dicho análisis utilizaron 40.000 *tweets* que fueron extractados entre el 13 y el 19 de octubre de ese año. En este estudio, con el fin de evitar robots y *spammers* consideraron un solo *tweet* por persona, obteniendo como resultado que el 73.8% de los trinos era positivos; igualmente encontraron el orden en que aparecían los candidatos según el número de trino positivos. Vale la pena además citar el trabajo realizado por Cerón-Guzmán y León (2015) quienes utilizan también los *tweets* de las elecciones presidenciales de 2014, aun cuando su objetivo no es analizar las tendencias de los diferentes candidatos, sino reconocer *spammers*.

III. Metodología

Para el desarrollo de la investigación se utilizaron: las elecciones regionales de Colombia que se llevaron a cabo el 25 de octubre de 2015 para elegir a los gobernadores de

los 32 departamentos y a los diputados de las Asambleas Departamentales, y los alcaldes de sus 1.099 municipios, los concejales municipales y los ediles de las Juntas Administradoras Locales; y las elecciones parlamentarias de Venezuela, celebradas el 6 de diciembre de 2015, para renovar todos los escaños de la Asamblea Nacional. Se definió como alcance incluir los candidatos a las alcaldías de las tres principales ciudades de Colombia –Bogotá, Medellín y Cali– y a los dos principales grupos Venezolanos –MUD y PSUV–.

Se extractaron todos los *tweets* relativos a cada uno de los candidatos o grupos que se sucedieron en la semana previa a las elecciones, para el caso de Colombia se incluyó a los candidatos más representativos, para el caso de Venezuela se utilizaron los *hashtag*: *#VenezuelaQuiereCambio*, para la Mesa de la Unidad Democrática [MUD], y *#LosDeChavezAVotar*, para el oficialismo, el Partido Socialista Unido de Venezuela [PSUV].

Una vez descargados los *tweets*: se filtraron según si su contenido incluía lenguaje natural [LN], descartando los que únicamente incluían un enlace, contenido multimedia o emoticones; se dejó un *tweet* por usuario –con el fin eliminar aquellos generados por robots–; se les aplicó un pre-procesamiento básico; y se sometieron a un proceso de análisis de sentimiento por medio del API Apicultur (APIs Sentiment analysis, 2012). Este API recibe un *tweet* y retorna un valor entre 0 y 5 indicando la intensidad de la opinión y la certeza del resultado, y clasifica el sentimiento del *tweet* en: positivo, negativo, neutro o no procesado.

```
buscar(candidato, cantidad de tweets, fecha limite ):
        buscar tweets candidato "Twespy"
        generar archivo con trinos

analizarSentimiento(archivo con trinos):
        leer archivo de trinos
        por cada trino
                si tiene contenido multimedia
                        ignorar
                sino
                        enviar al analizador de sentimiento "APICULTUR"

separarSentimiento(archivo de análisis):
        leer archivo de análisis
        para cada linea
                verificar si el usuario ya está registrado con esa ponderación
                        registrar análisis de acuerdo a la ponderación

contarRetuit(archivo ponderado):
        leer archivo ponderado
                guardar en un diccionario siendo la clave el texto y la valor las veces que se repite
                guardar diccionario en un archivo

analizarLenguaje(archivo ponderado sin retweets):
        leer archivo ponderado sin retweets
                separar en tokens cada tuit y asignar etiquetas
                guardar en un archivo las etiquetas dadas a cada token
                escribe un archivo con el árbol de cada tweets

ejecutar():
        buscar(candidato, cantidad de tweets, fecha limite )
        analizarSentimiento(archivo con trinos
        separarSentimiento(archivo de análisis)
        contarRetuit(archivo ponderado)
        analizarLenguaje(archivo ponderado sin retweets)
```

Figure1. Pseudocode used for experimentation / Pseudocódigo utilizado para la experimentación

In respect of studies made in Spanish language, specifically for Colombia, it is worth highlighting the one presented by Mamprin (2015) that was performed to analyze the favoritism of the candidates for the mayoralty of Cali 2015; for that analysis it was used 40.000 tweets that were extracted between 13 and 19 of October of that year. In this study, in order to avoid robots and spammers they considered a single tweet per person, obtaining as a result that 73.8% of the trills were positive; they also found the order in which the candidates appeared according to the number of positive trills. It is also worth mentioning the work done by Cerón-Guzmán and León (2015) who also use the tweets of the 2014 presidential elections, even though their objective is not to analyze the trends of the different candidates, but to recognize spammers.

III. Methodology used

To develop the research was take into account: the regional elections of Colombia that were held on October 25, 2015 to elect the governors of the 32 departments and the deputies of the Departmental Assemblies and the mayors of its 1.099 municipalities, the municipal councilors and the councilors of the Local Administrative Assemblies; and the Venezuelan parliamentary elections, held on December 6, 2015 to renew all the seats of the National Assembly. It was defined as the scope to include the candidates for the mayoralties of the three main cities of Colombia –Bogotá, Medellín and Cali– and the two main Venezuelan groups –MUD and PSUV–.

All the tweets related to each one of the candidates or groups that took place in the week prior to the elections were extracted. In the case of Colombia, the most representative candidates were included, in the case of Venezuela the hashtag "*#VenezuelaQuiereCambio*" was used for the Democratic Unity Roundtable [MUD], and *#LosDeChavezAVotar* was used for the ruling party, the United Socialist Party of Venezuela [PSUV].

Once the tweets were downloaded: they were filtered according to whether their content included Natural Language [NL], discarding those that only included a link, multimedia content or emoticons; one tweet was left per user –in order to eliminate those generated by robots–; a basic pre-processing was applied; and were subject to a process of sentiment analysis through API Apicultur (APIs Sentiment analysis, 2012). This API

receives a tweet and returns a value between 0 and 5 indicating the intensity of the opinion and the certainty of the result, and classifies the sentiment of the tweet in positive, negative, neutral or unprocessed.

Both in the pseudocode of **Figure 1** and in the architecture of **Figure 2**, we can observe the software components used: Tweepy, which uses the Twitter API REST and provides real-time reading access over tweets and generates outputs in JSON format; and Huytaca, which is responsible for applying basic pre-processing, interacting with Apicultur, separating tweets by rating, generating statistics and filtering by replica. Finally, comparisons were made with the official results that were produced by the elections for each candidate or group.

IV. Results and analysis

As can be seen in **Table 1**, a total of 130.482 tweets were downloaded for Colombia and 152.010 tweets were downloaded for Venezuela, of which 22% for Colombia and 23% for Venezuela corresponded to NL; the rest contained only multimedia content or emoticons (as anticipated, only those corresponding to NL were used). For the analyses, each group was taken independently, as the search was filtered by the hashtag corresponding to the candidate or group.

Tanto en el pseudocódigo de la **Figura 1**, como en la arquitectura de la **Figura 2**, se pueden observar los componentes de software utilizado: Tweepy, que utiliza el API REST de Twitter y proporciona el acceso de lectura en tiempo real sobre los *tweets* y genera salidas en formato JSON; y Huytaca, que se encarga de aplicar el pre-procesamiento básico, interactuar con el Apicultur, separar los *tweets* por calificación, generar las estadísticas y filtrar por replica. Por último, se realizaron comparaciones con los resultados oficiales que arrojaron las elecciones para cada candidato o grupo.

IV. Resultados y análisis

Como se puede observar en la **Tabla 1**, se descargó un total de 130.482 *tweets* para Colombia y 152.010 *tweets* para Venezuela, de los cuales el 22% para Colombia y el 23% para Venezuela correspondían a LN; el resto contenía únicamente contenido multimedia o emoticones (como se anticipó, se utilizaron solo los correspondientes a LN). Para los análisis se tomó cada grupo de manera independiente, en virtud de que la búsqueda se filtró por el *hashtag* correspondiente al candidato o grupo.

Los contenidos de los *tweets* no se escapan a las particularidades propias del lenguaje utilizado para la interacción a través de redes sociales, correos, juegos, entre otros, como:

- las abreviaciones, como el "+1" que significa apoyo o "me gusta" –y si se le añaden ceros al "+1" enfatiza el agrado–, "tqm" o te quiero mucho, "RT" o *"retweet"*, *"xoxo"* o besos y abrazos y *"xq"* o porque;
- los emoticones;

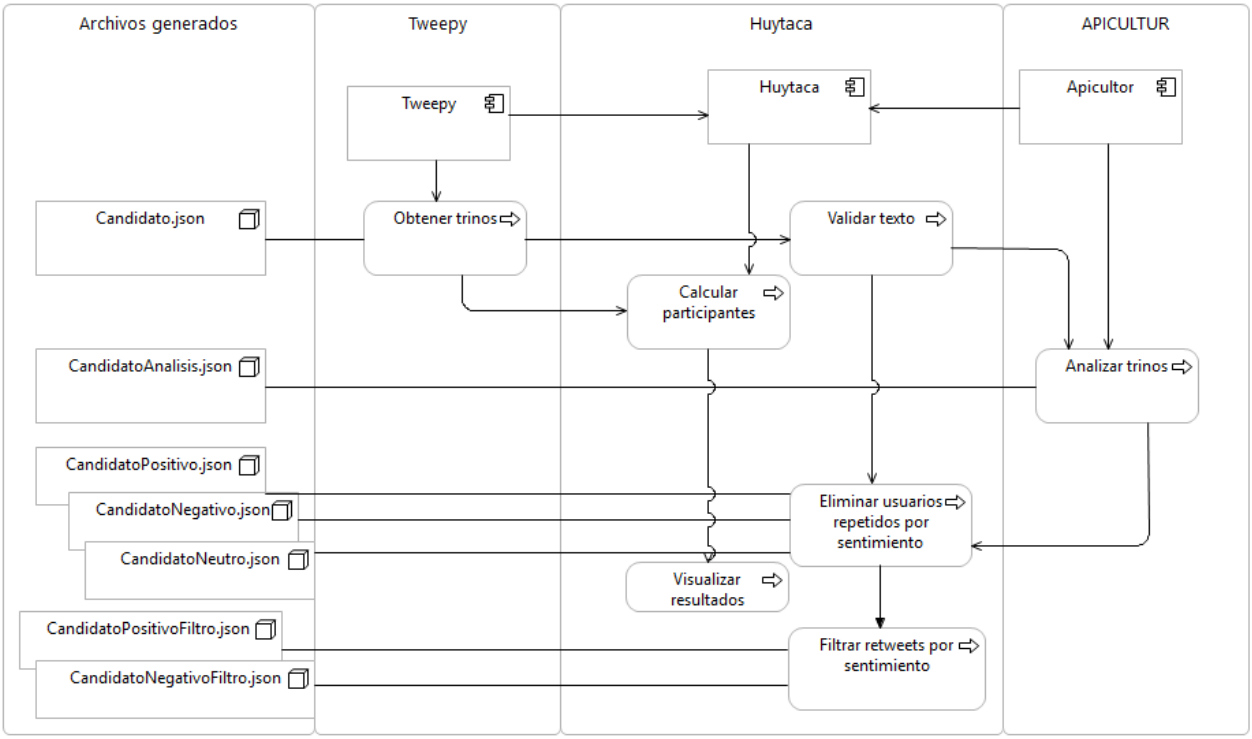

Figure 2. Architecture used for experimentation / Arquitectura utilizada para la experimentación

Table 1. Tweets extracted by candidate or group / Tweets extractados por candidato o grupo

Colombia						
Ciudad	Candidato	Tweets	LN		Enlaces Videos	
		Cantidad	Cantidad	%	Cantidad	%
Bogotá	Enrique Peñalosa	20.000	6.049	30%	13.951	70%
Bogotá	Clara López	15.000	3.317	22%	11.683	78%
Bogotá	Rafael Pardo	23.864	5.426	23%	18.438	77%
Bogotá	Francisco Santos	15.000	4.091	27%	10.909	73%
Medellin	Juan Carlos Velez	20.001	2.403	12%	17.598	88%
Medellin	Gabrile Jaime Rico	4.004	905	23%	3.099	77%
Medellin	Federico Gutierrez	20.001	3.142	16%	16.859	84%
Cali	Maurice Armtage	6.630	1.558	23%	5.072	77%
Cali	Roberto Ortiz	3.456	812	23%	2.644	77%
Cali	Angelino Garzon	2.526	888	35%	1.638	65%
TOTAL		**130.482**	**28.591**	**22%**	**101.891**	**78%**

Venezuela					
Grupo	Tweets	LN		Enlaces Videos	
	Cantidad	Cantidad	%	Cantidad	%
MUD	49.700	10.115	20%	39.585	80%
PSUV	102.310	24.924	24%	77.386	76%
TOTAL	**152.010**	**35.039**	**23%**	**116.971**	**77%**

- la mala escritura, categoría que incluye la mala ortografía, la digrafía digital (invertir u obviar letras) y,
- que el usuario recurra a unir dos o más palabras para que se considere solo una (dado que en la versión analizada de Twitter solo acepta 140 caracteres); y
- los fenómenos propios del lenguaje natural como, el sarcasmo y la influencia del contexto.

De un total de 63.657 *tweets* que se procesaron, el análisis de sentimiento usando la API Apicultur arrojó 24.216 positivos, 18.350 negativos, 14.232 no procesables y 6.859 neutros, para un total de 49.425 tweets procesables (ver **Tabla 2**).

The contents of the tweets do not escape to the particularities of the language used for interaction through social networks, mails, games, among others, such as:

- abbreviations, such as "+1" which means support or "I like" –and if you add zeros to "+1" emphasize the pleasure–, "tqm" or "Te Quiero Mucho" (I Love You a Lot), "RT" or "retweet", "xoxo" or kisses and hugs and "xq" or "porque" (because);
- emoticons;
- bad writing, which includes poor spelling, digital digraph (reverse or skip letters) and,
- the fact that the user could join two or more words to consider only one (since the analyzed version of Twitter only allows 140 characters); and
- natural language phenomena such as sarcasm and context influence.

From a total of 63.657 tweets that were processed, the sentiment analysis using the API Apicultur ranked 24.216 positive, 18.350 negative, 14.232 non-processable and 6.859 neutral, for a total of 49.425 processable tweets (see **Table 2**).

The API used did not escape some of the natural language phenomena, when reviewing the tweets we can easily find examples where they were included in categories that do not correspond, for example:

Table 2. Tweets by sentiment rating / Tweets por calificación del sentimiento

PAIS-CIUDAD	CANDIDATO/GRUPO	POSITIVO		NEGATIVO		NEUTRO		TOTAL
		Cantidad	%	Cantidad	%	Cantidad	%	
Colombia-Bogotá	Enrique Peñalosa	1.815	36%	2.460	49%	788	16%	5.063
Colombia-Bogotá	Clara López	920	38%	1.109	46%	381	16%	2.410
Colombia-Bogotá	Rafael Pardo	1.779	45%	1.537	39%	659	17%	3.975
Colombia-Bogotá	Francisco Santos	1.364	43%	1.444	45%	399	12%	3.207
Colombia-Medellín	Juan Carlos Velez	734	39%	747	39%	422	22%	1.903
Colombia-Medellín	Gabriel Jaime Rico	323	46%	215	31%	166	24%	704
Colombia-Medellín	Federico Gutierrez	1.117	46%	864	35%	467	19%	2.448
Colombia-Cali	Maurice Armitage	465	49%	424	45%	62	7%	951
Colombia-Cali	Roberto Ortiz	222	47%	196	41%	59	12%	477
Colombia-Cali	Angelino Garzon	230	41%	245	44%	81	15%	556
Venezuela	MUD	3.144	40%	3.838	49%	845	11%	7.827
Venezuela	PSUV	12.103	61%	5.271	26%	2.530	13%	19.904

Tweets that were included as "unprocessable" for Clara López candidate for mayoralty of Bogotá:

- "claralopezobre represents the left-wing that has managed the city in recent years with bad results";
- "leszlikalli voting claralopezobre rafaelpardo is voting petro think the vote think in bogotá";
- "claralopezobre aidaavellae gloriacuartas four great women who have recognition";
- "cement education";
- "claralopezobre we filled square last Friday we will fill urns"; and
- "petrogustavo says that poverty is over"

Tweets that were included as "unprocessable" for Enrique Peñalosa candidate for mayoralty of Bogotá:

- "peñalosa does not want a little vip house but asks for votes from social stratum 1 2 3"
- "we can rescue bogota by the hand enrique peñalosa enriquepenalosa"
- "jerobledsorpresa there are 2 big candidates left for mayor of bogotá"
- "at the polls I will support candidate I vote enrique peñalosa enriquepenalosa he will be mayor again"
- "we can rescue bogotá by the hand enrique peñalosa enriquepenalosa"
- "I vote thinking about what the city needs enrique peñalosa enriquepenalosa"

lisis de sentimiento utilizando el API Apicultur clasificó 24.216 positivos, 18.350 negativos, 14.232 no procesables y 6.859 Neutros, para un total de 49.425 *tweets* procesables (ver **Tabla 2**).

El API utilizado no se escapó de algunos de los fenómenos propios del lenguaje natural, al revisar los *tweets*, fácilmente se pueden encontrar ejemplos donde se incluyeron en categorías que no corresponden, por ejemplo:

Tweets que se incluyeron como "no procesable" para Clara López candidata a la alcaldía de Bogotá

- "claralopezobre representa la izquierda que ha manejado ciudad los últimos años con resultados malos";
- "leszlikalli votar claralopezobre rafaelpardo es votar petro piense el voto piense en bogotá";
- "claralopezobre aidaavellae gloriacuartas cuatro grandes mujeres que tienen reconocimiento";
- "educación cemento";
- "claralopezobre llenamos plaza viernes pasado llenaremos urnas"; y
- "petrogustavo dice q acabo pobreza"

Tweets que se incluyeron como "no procesable" para Enrique Peñalosa candidato a la alcaldía de Bogotá:

- "peñalosa no quiere vivienda vip chica pero pide votos estratos 1 2 3"
- "podremos rescatar bogota de la mano enrique peñalosa enriquepenalosa"
- "jerobledsorpresa quedan 2 grandes candidaturas para alcalde de bogotá"
- "yo en las urnas apoyare candidato voto enrique peñalosa enriquepenalosa será alcalde otra ves"
- "podremos rescatar bogotá de la mano enrique peñalosa enriquepenalosa"

Table 3. Official results vs. positive and negative tweets / Resultados oficiales vs. Tweets positivos y negativos

PAIS-CIUDAD	CANDIDATO/GRUPO	Resultados Oficiales		Tweets Positivos		Tweets Negativos	
		Cantidad	%	Cantidad	%	Cantidad	%
Colombia-Bogotá	Enrique Peñalosa	903.764	36%	1.815	31%	2460	38%
Colombia-Bogotá	Clara López	498.718	20%	920	16%	1109	17%
Colombia-Bogotá	Rafael Pardo	778.050	31%	1.779	30%	1537	23%
Colombia-Bogotá	Francisco Santos	327.852	13%	1.364	23%	1444	22%
	TOTAL	2.508.384		5.878		6.550	
Colombia-Medellín	Federico Gutierrez	244.636	41%	1.117	51%	864	47%
Colombia-Medellín	Juan Carlos Velez	235.633	40%	734	34%	747	41%
Colombia-Medellín	Gabriel Jaime Rico	111.777	19%	323	15%	215	12%
	TOTAL	592.046		2.174		1.826	
Colombia-Cali	Maurice Armitage	264.118	45%	465	51%	424	49%
Colombia-Cali	Roberto Ortiz	175.394	30%	222	24%	196	23%
Colombia-Cali	Angelino Garzon	149.889	25%	230	25%	245	28%
	TOTAL	589.401		917		865	
Venezuela	MUD	7.726.066	58%	3.144	21%	3838	42%
Venezuela	PSUV	5.622.844	42%	12.103	79%	5271	80%
	TOTAL	13.348.910		15.247		9.109	

- "yo voto pensando en lo que necesita la ciudad enrique peñalosa enriquepenalosa"

Tweets que se incluyeron como "positivos" para Enrique Peñalosa candidato a la alcaldía de Bogotá:

- "giordanobrunofi peñalosa es buen gerente dio 95% d utilidad privados distrito 5porciento"
- "que peñalosa regale balones a niños lo hace semejante al candidato a personero del colegio que regala dulces y promete piscina"
- "manoslimpiascsorpresa peñalosa habla experto metro pero no construyó pudo reemplazo hizo transmilenio en la caracas"
- "enriquepenalosa quiere es vender distrito peñalosa quería vender etb $3500 millones 1998 no dejen engañar"
- "lapardita hace años sabemos peñalosa no quiere responder ni entrevistas ni tampoco derechos sociales"

Tweets que se incluyeron como "positivos" para Angelino Garzón candidato a la alcaldía de Cali:

- "creo que roberto está más arriba pero bueno por lo menos reconocen 1er lugar"
- "mano de vejetes vayan a criar nietos y dejen gente joven trabajar"

Con el fin de estimar qué tan acertadas son las tendencias a través de twitter, se incluyeron los resultados oficiales, tanto para Colombia ("2015 Elecciones Regionales," 2015), como para Venezuela (ver **Tabla 3**). Cabe mencionar que los resultados para Venezuela fueron tomados de Wikipedia ("Elecciones parlamentarias...", 2015), dado que el Consejo Nacional Electoral publicó únicamente los porcentajes por puestos en la Asamblea y no por votantes.

Como se pueden observar, las tendencias de favorabilidad (entendida la favorabilidad como el apoyo a un can-

Tweets that were included as "positive" for Enrique Peñalosa candidate for mayoralty of Bogotá:

- "giordanobrunofi peñalosa is a good manager he gave 95% of gross profit private district 5percent"
- "the fact that peñalosa gives balls to children makes him similar to the candidate to representative of high school that gives candy and promises pool"
- "manoslimpiascsorpresa peñalosa speaks expert meter but did not build could replacement made transmilenio in the caracas"
- "enriquepenalosa wants to sell district peñalosa wanted to sell etb $3500 million 1998 do not let deceive"
- "lapardita years ago we know peñalosa does not want to answer neither interviews nor social rights"

Tweets that were included as "positive" for Angelino Garzón candidate for mayoralty of Cali:

- "I think roberto is higher but at least they recognize 1st place"
- "old men you better go raise your grandchildren and let young people work"

In order to estimate how accurate trends are on twitter, official results were included for both Colombia ("2015 Regional Elections", 2015) and Venezuela (see **Table 3**). It should be mentioned that the results for Venezuela were taken from Wikipedia ("Parliamentary Elections ...", 2015), since the National Electoral Council published only the percentages for seats in the Assembly and not per voters.

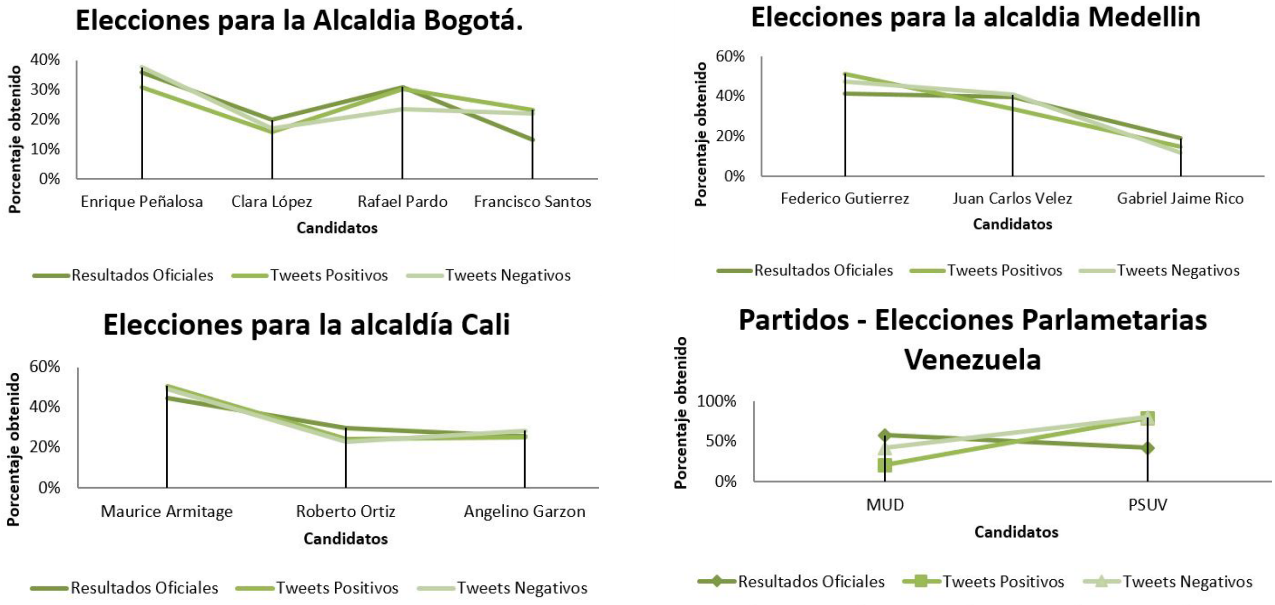

Figure 3. Official results vs. positive and negative tweets / Resultados oficiales vs. tweets positivos y negativos

As can be observed, trends of favorability (understood favorability as supporting a candidate or group, either with the actual vote or with a comment), for the first three cases, i.e. for the election of mayors of the three main cities of Colombia, the results are very similar (actual results vs. positive tweets), while for the Venezuelan parliamentary elections the results are completely reversed, results that may be affected by political propaganda and the use of spammer. Compared to negative tweets it is striking that they exhibit the same behavior as the positive tweets.

A phenomenon that usually occurs in social networks is the replicas of the messages from other users or the publication of the content from another person. In a sense, it could be said that if I share a message is because I agree with it or because I will make a comment referring to it. For the Twitter platform, the replica messages are called *retweets* and, in this case, the Twitter platform stores among other data, the user who made the replica and the date. From this phenomenon we can observe (**Table 4**) the number of tweets that are repeated from one to ten times.

According to the results of **Table 4**, for Colombia the percentage of tweets that do not repeat is on average close to 62%, and the percentage of tweets that are repeated more than ten times is on average 14%. On the contrary, Venezuela has a completely inverse situation, the percentage of tweets that do not repeat represents only 25%, while the percentage of tweets that repeat more than ten times

didato o grupo, sea con el voto real o con un comentario), para los tres primeros casos, es decir para la elección de alcaldes las tres ciudades principales de Colombia, los resultados arrojados son muy similares (resultados reales vs *tweets* positivos), mientras que para las elecciones parlamentarias de Venezuela los resultados son completamente inversos, resultados que pueden estar afectados por la propaganda política y el uso de *spammer*. Con relación a los *tweets* negativos llama la atención que presentan el mismo comportamiento que con los *tweets* positivos.

Un fenómeno muy dado en las redes sociales son las réplicas de los mensajes de otros usuarios o la publicación del contenido de otra persona. En cierto sentido, se podría afirmar que si yo comparto un mensaje es porqué estoy de acuerdo con el o porqué haré un comentario refiriéndome a él. Para la plataforma de Twitter, lo mensajes de réplica se llaman *retweets* y, en este caso, la plataforma de Twitter almacena entre otros datos, el usuario que realizó la réplica y la fecha. A partir de este fenómeno se puede observar (**Tabla 4**) el número de *tweets* que se repiten de una a diez veces.

De acuerdo con los resultados de la **Tabla 4**, para Colombia el porcentaje de *tweets* que no se repiten es en promedio cercano al 62%, y el porcentaje de *tweets* que se repiten más de diez veces es en promedio del 14%. Por el contrario, en Venezuela se presenta una situación completamente inversa, el porcentaje de *tweets* que no se repiten representa solamente el 25%, mientras que el porcentaje de *tweets* que se repiten más de diez veces representa el 61%, es decir más de la mitad de la muestra. Para los casos de los *tweets* que se repiten más de dos a diez veces, sus porcentajes no tienen una proporción tan relevante, sin embargo

Table 4. Proportion of repeated tweets per candidate or party / Proporción de tweets repetidos por candidato o partido

PAIS-CIUDAD	CANDIDATO/ GRUPO	Sentimiento	Veces																						Total
			1	%	2	%	3	%	4	%	5	%	6	%	7	%	8	%	9	%	10	%	Más de 10	%	
Colombia-Bogotá	Enrique Peñalosa	Positivo	918	50,6%	68	7,5%	15	2,5%	6	1,3%	6	1,7%	7	2,3%	5	1,9%	3	1,3%	1	0,5%	1	0,6%	20	29,9%	1.815
		Negativo	1071	43,5%	126	10,2%	38	4,6%	13	2,1%	11	2,2%	11	2,7%	5	1,4%	3	1,0%	1	0,4%	2	0,8%	27	31,0%	2.460
Colombia-Bogotá	Clara López	Positivo	569	61,8%	43	9,3%	16	5,2%	4	1,7%	6	3,3%	2	1,3%	3	2,3%	1	0,9%	0	0,0%	1	1,1%	7	13,0%	920
		Negativo	730	65,8%	31	5,6%	18	4,9%	10	3,6%	4	1,8%	0	0,0%	1	0,6%	2	1,4%	0	0,0%	3	2,7%	6	13,5%	1.109
Colombia-Bogotá	Rafael Pardo	Positivo	863	48,5%	37	4,2%	12	2,0%	9	2,0%	4	1,1%	6	2,0%	5	2,0%	7	3,1%	2	1,0%	3	1,7%	20	32,3%	1.779
		Negativo	978	63,6%	30	3,9%	18	3,5%	8	2,1%	6	2,0%	1	0,4%	3	1,4%	1	0,5%	2	1,2%	2	1,3%	11	20,2%	1.537
Colombia-Bogotá	Francisco Santos	Positivo	699	51,2%	42	6,2%	21	4,6%	8	2,3%	8	2,9%	2	0,9%	3	1,5%	5	2,9%	1	0,7%	0	0,0%	9	26,7%	1.364
		Negativo	890	61,6%	34	4,7%	20	4,2%	12	3,3%	5	1,7%	2	0,8%	4	1,9%	2	1,1%	2	1,2%	1	0,7%	9	18,6%	1.444
Colombia-Medellín	Juan Carlos Vélez	Positivo	409	55,7%	29	7,9%	9	3,7%	11	6,0%	6	4,1%	2	1,6%	3	2,9%	3	3,3%	1	1,2%	2	2,7%	3	10,9%	734
		Negativo	331	44,3%	22	5,9%	5	2,0%	2	1,1%	5	3,3%	2	1,6%	2	1,9%	2	2,1%	0	0,0%	2	2,7%	7	35,1%	747
Colombia-Medellín	Gabriel Jaime Rico	Positivo	228	70,6%	20	12,4%	5	4,6%	4	5,0%	1	1,5%	1	1,9%	0	0,0%	0	0,0%	0	0,0%	0	0,0%	1	4,0%	323
		Negativo	131	60,9%	4	3,7%	4	5,6%	2	3,7%	2	4,7%	1	2,8%	0	0,0%	1	3,7%	0	0,0%	0	0,0%	2	14,9%	215
Colombia-Medellín	Federico Gutiérrez	Positivo	753	67,4%	58	10,4%	24	6,4%	10	3,6%	3	1,3%	1	0,5%	0	0,0%	1	0,7%	0	0,0%	1	0,9%	6	8,7%	1.117
		Negativo	514	59,5%	31	7,2%	21	7,3%	6	2,8%	7	4,1%	2	1,4%	2	1,6%	1	0,9%	0	0,0%	0	0,0%	5	15,3%	864
Colombia-Cali	Maurice Armitage	Positivo	329	70,8%	23	9,9%	4	2,6%	7	6,0%	3	3,2%	0	0,0%	1	1,5%	0	0,0%	0	0,0%	0	0,0%	2	6,0%	465
		Negativo	280	66,0%	14	6,6%	9	6,4%	5	4,7%	2	2,4%	2	2,8%	0	0,0%	0	0,0%	2	4,2%	0	0,0%	1	6,8%	424
Colombia-Cali	Roberto Ortiz	Positivo	173	77,9%	13	11,7%	4	5,4%	3	5,4%	0	0,0%	0	0,0%	0	0,0%	0	0,0%	0	0,0%	0	0,0%	0	0,0%	222
		Negativo	160	81,6%	7	7,1%	3	4,6%	0	0,0%	0	0,0%	1	3,1%	0	0,0%	1	4,1%	0	0,0%	0	0,0%	0	0,0%	196
Colombia-Cali	Angelino Garzón	Positivo	157	68,3%	13	11,3%	5	6,5%	4	7,0%	0	0,0%	0	0,0%	0	0,0%	1	3,5%	1	3,9%	0	0,0%	0	0,0%	230
		Negativo	180	73,5%	11	9,0%	8	9,8%	1	1,6%	0	0,0%	1	2,4%	0	0,0%	0	0,0%	0	0,0%	1	4,1%	0	0,0%	245
Venezuela	MUD	Positivo	1291	41,1%	60	3,8%	19	1,8%	24	3,1%	15	2,4%	3	0,6%	7	1,6%	8	2,0%	5	1,4%	2	0,6%	34	41,6%	3.144
		Negativo	1582	41,2%	89	4,6%	38	3,0%	21	2,2%	16	2,1%	13	2,0%	9	1,6%	3	0,6%	3	0,7%	8	2,1%	46	39,8%	3.838
Venezuela	PSUV	Positivo	879	7,3%	107	1,8%	48	1,2%	28	0,9%	25	1,0%	7	0,3%	11	0,6%	3	0,2%	6	0,4%	4	0,3%	59	85,9%	12.103
		Negativo	574	10,9%	49	1,9%	33	1,9%	24	1,8%	12	1,1%	7	0,8%	7	0,9%	6	0,9%	3	0,5%	8	1,5%	36	77,7%	5.271

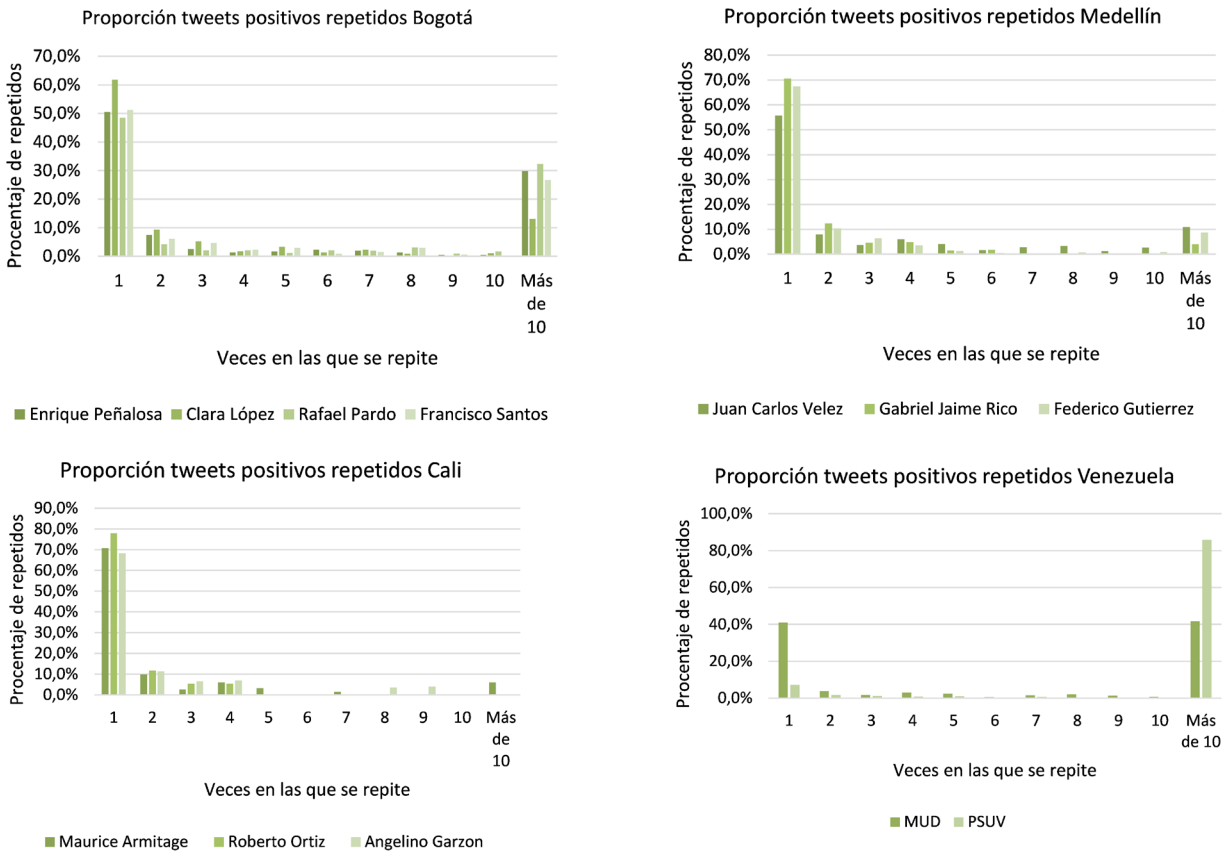

Figure 4. Results of positive repeated tweets per candidate (Colombia) and hashtag (Venezuela) / Resultados de tweets positivos repetidos por candidato (Colombia) y hashtag (Venezuela)

también pueden ser tomados como casos de estudio (este comportamiento se puede visualizar en la **Figura 4**).

A partir de los *tweets* calificados como positivos o negativos y de los *retweets*, se observa que el porcentaje de *retweet* para tres candidatos de la ciudad de Bogotá: Enrique Peñalosa, 49%; Rafael Pardo, 51%; y Francisco Santos, represents 61%, i.e. more than half of the sample. For cases of tweets that repeat more than two to ten times, their percentages do not have such a relevant proportion, however they can also be taken as case studies (this behavior can be visualized in **Figure 4**).

Table 5. Repeated tweets by sentiment / Tweets repetidos por sentimiento

PAIS-CIUDAD	CANDIDATO/GRUPO	POSITIVO			NEGATIVO		
		Cantidad	Retweet	%	Cantidad	Retweet	%
Colombia-Bogotá	Enrique Peñalosa	1.815	897	49%	2.460	1.389	56%
Colombia-Bogotá	Clara López	920	351	38%	1.109	379	34%
Colombia-Bogotá	Rafael Pardo	1.779	916	51%	1.537	559	36%
Colombia-Bogotá	Francisco Santos	1.364	665	49%	1.444	554	38%
Colombia-Medellín	Juan Carlos Velez	734	325	44%	747	416	56%
Colombia-Medellín	Gabriel Jaime Rico	323	95	29%	215	84	39%
Colombia-Medellín	Federico Gutierrez	1.117	364	33%	864	350	41%
Colombia-Cali	Maurice Armitage	465	136	29%	424	144	34%
Colombia-Cali	Roberto Ortiz	222	49	22%	196	36	18%
Colombia-Cali	Angelino Garzon	230	73	32%	245	65	27%
Venezuela	MUD	3.144	1.853	59%	3.838	2.256	59%
Venezuela	PSUV	12.103	11.224	93%	5.271	4.697	89%

Table 5. Repeated tweets by sentiment / Tweets repetidos por sentimiento

PAIS-CIUDAD	CANDIDATO/GRUPO	Positivo		Negativo	
		Cantidad	%	Cantidad	%
Colombia-Bogotá	Enrique Peñalosa	918	30%	1.071	29%
Colombia-Bogotá	Clara López	569	19%	730	20%
Colombia-Bogotá	Rafael Pardo	863	28%	978	27%
Colombia-Bogotá	Francisco Santos	699	23%	890	24%
Colombia-Medellín	Juan Carlos Velez	409	29%	331	34%
Colombia-Medellín	Gabriel Jaime Rico	228	16%	131	13%
Colombia-Medellín	Federico Gutierrez	753	54%	514	53%
Colombia-Cali	Maurice Armitage	329	50%	280	45%
Colombia-Cali	Roberto Ortiz	173	26%	160	26%
Colombia-Cali	Angelino Garzon	157	24%	180	29%
Venezuela	MUD	1.291	59%	1.582	73%
Venezuela	PSUV	879	41%	574	27%

From the positive or negative tweets and *retweets*, we can observe that the percentage of retweet for three candidates from the city of Bogotá: Enrique Peñalosa, 49%; Rafael Pardo, 51%; and Francisco Santos, 49%; represent approximately 50%. For other Colombian candidates, this value is less than 44%. For the PSUV group, *retweets* are equal to 93% and for MUD 59%.

Based on the results of non-repeated tweets, **Figure 5** shows the trends vs. the official results, obtaining a co-

49%; representan aproximadamente el 50%. Para los demás candidatos colombianos, este valor es inferior al 44%. Para el grupo PSUV, los *retweets* equivalen al 93% y para MUD el 59%.

Con base en los resultados de los *tweets* sin repetir, en la **Figura 5** se incluyen las tendencias vs. Los resultados oficiales, obteniendo ya una aproximación corregida y más acorde con la realidad, que muestra que, al descontar los *tweets* repetidos, la tendencia por los mensajes enviados a la red es un muy buen estimador de las preferencias políticas de las personas.

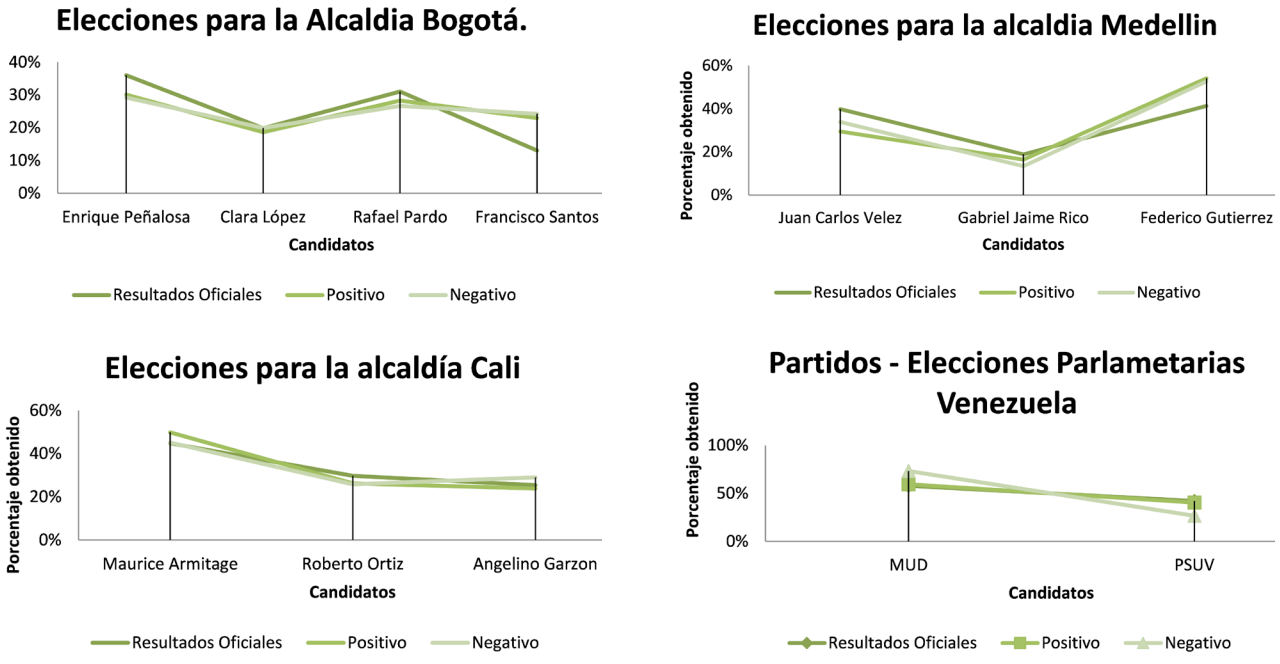

Figure 5. Official results vs. positive and negative non-repeated tweets / Resultados oficiales vs. tweets sin repetir positivos y negativos

V. Conclusiones

Al descontar los *tweets* repetidos, la tendencia por los mensajes positivos enviados a la red es un muy buen estimador de las preferencias políticas de las personas.

Se podría deducir que el efecto de las campañas publicitarias o de los robots se evidencia en los tweets que aparecen más de diez veces.

Para el análisis de sentimiento positivo no es relevante que la muestra sea diseñada, que contenga el mismo de número de observaciones para cada elemento o que sea en el mismo espacio temporal o espacial, pues cada candidato contó con diferentes tamaños y para el caso de Colombia y Venezuela el corte se hizo en diferente época y para diferente país. S&T

rrected approximation and more in line with reality, which shows that by deducting repeated tweets, the trend for messages sent to the network is a very good estimator of the political preferences of people.

V. Conclusions

By deducting repeated tweets, the trend for positive messages sent to the network is a very good estimator of the political preferences of people.

It could be deduced that the effect of advertising campaigns or robots is evident in tweets that appear more than ten times.

For the analysis of positive sentiment it is not relevant that the sample is designed, that contains the same number of observations for each element or to be in the same temporal or spatial space, since each candidate had different sizes and for the case of Colombia and Venezuela the cut was made in different times and for different country. S&T

References / Referencias

2015 elecciones regionales [Colombia.com]. (2015). Retrieved from: http://www.colombia.com/elecciones/2015/regionales/resultados/

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. *In Proceedings of the Workshop on Languages in Social Media* (pp. 30-38). Stroudsburg PA: ACL.

Anjaria, M., & Guddeti, R. M. R. (2014). Influence factor based opinion mining of Twitter data using supervised learning. In Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on (pp. 1–8). http://doi.org/10.1109/COMSNETS.2014.6734907

APIs Sentiment analysis (2012). Retrieved May 20, 2012, from https://store.apicultur.com/

Bermingham, A., & Smeaton, A. F. (2011). On using Twitter to monitor political sentiment and predict election results, In: Sentiment Analysis where AI meets Psychology (SAAIP) Workshop at the International Joint Conference for Natural Language Processing (IJCNLP), 13th November 2011, Chiang Mai, Thailand.

Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in Twitter streaming data. *Lecture Notes in Computer Science, 6332 LNAI, 1–15.* doi :10.1007/978-3-642-16184-1_1

Brown, E. (2012). Will twitter make you a better investor? A look at sentiment, user reputations and their effect on the stock market. *In Proceedings of Southern Association for Information Systems (SAIS)* (pp. 36.42).

Cerón-Guzmán, J. A., & León, E. (2015). *Detecting social spammers in Colombia 2014 presidential election: Lecture Notes in Computer Science, 9414 - Advances in artificial intelligence and its applications,* (pp. 121-141). Switzerland: Springer.

Choy, M., Cheong, M. L. F., Laik, M. N., & Shung, K. P. (2011). *A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction.* Retrieved from: https://arxiv.org/abs/1108.5520

Elecciones parlamentarias de Venezuela de 2015. Retrieved from: https://es.wikipedia.org/wiki/Elecciones_parlamentarias_de_Venezuela_de_2015

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011, June). Target-dependent twitter sentiment classification. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 151-160). Stroudsburg PA:ACL.

Kiplinger. (2011). Market's mood in a tweet. *Kiplingers Personal Finance, (4),* 14.

Liu, B. & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. *In Mining Text Data* (pp. 415-463). US: Springer.

Liu, B. (2010). Sentiment analysis and subjectivity. *In Handbook of natural language processing* (2a ed.), (pp. 627-666). Boca Raton, FL: Chapman and Hall/CRC.

Mamprin, A. (2015, Oct. 21). ¿Cómo les va a los candidatos a la Alcaldía de Cali en Twitter? *El País.* Cali. Retrieved from: http://www.elpais.com.co/elpais/elecciones-2015/noticias/como-les-va-candidatos-alcaldia-cali-twitter

Mao, H., Counts, S., & Bollen, J. (2011). *Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data* [arXiv Preprint, 10]. Retrieved from: https://arxiv.org/abs/1112.1051

Nguyen, V. D., Varghese, B., & Barker, A. (2013). The royal birth of 2013: Analyzing and visualizing public sentiment in the UK using Twitter. *In Big Data, 2013 IEEE International Conference on* (pp. 46-54). doi:10.1109/BigData.2013.6691669

CURRICULUM VITAE

Sonia Ordoñez Salinas, Ph.D. Graduated in Statistics (Universidad Nacional de Colombia, Bogotá) and Systems Engineering (Universidad Distrital Francisco José de Caldas (Bogotá), with studies of specialization, master and doctorate from Universidad Nacional de Colombia. She researches about: natural language processing, data mining, statistics and databases. Currently she is Director of the GESDATOS research group and professor at the Universidad Distrital / Estadística de la Universidad Nacional de Colombia (Bogotá) e Ingeniera de Sistemas de la Universidad Distrital Francisco José de Caldas (Bogotá), con estudios de especialización, maestría y doctorado de la Universidad Nacional, con amplia experiencia profesional y en investigación, en particular en: procesamiento de lenguaje natural, minería de datos, estadística, bases de datos y afines. Es Directora del Grupo de Investigación GESDATOS y docente de la Universidad Distrital.

Juan Manuel Pérez Trujillo Last semester student of the Program of Systems Engineering at the Universidad Distrital Francisco José de Caldas (Bogotá, Colombia), with emphasis in databases and qualitative cybernetics, and member of GESDATOS research group / Estudiante de último semestre del Programa de Ingeniería de Sistemas de la Universidad Distrital Francisco José de Caldas, con énfasis en bases de datos y cibernética cualitativa. Pertenece al grupo de investigación GESDATOS de la misma universidad, desde 2015.

Romario Albeiro Sánchez Montero Last semester student of the Program of Systems Engineering at the Universidad Distrital Francisco José de Caldas (Bogotá, Colombia), with knowledge in databases, artificial intelligence, qualitative cybernetics and software development, and member of GESDATOS research group / Estudiante de decimo semestre del Programa de Ingeniería de Sistemas de la Universidad Distrital Francisco José de Caldas, con conocimientos en bases de datos, inteligencia artificial, cibernética cualitativa y desarrollo de software. Pertenece al grupo de investigación GESDATOS, de la misma universidad, desde el segundo semestre de 2015.