User Generated Content: An Analysis of User Behavior by Mining Political Tweets

Rocío Abascal-Mena, Erick López-Ornelas, and J. Sergio Zepeda-Hernández

Universidad Autónoma Metropolitana – Cuajimalpa
Departamento de Tecnologías de la Información
Avenida Constituyentes 1054, Col. Lomas Altas, Del. Miguel Hidalgo
11950, México D.F. México
{mabascal,elopez,jzepeda}@correo.cua.uam.mx

Abstract. With the emergence of smarthphones and social networks, a very large proportion of communication takes place on short texts. This type of communication, often anonymous, has allowed a new public participation in political issues. In particular, electoral phenomena all over the world have been greatly influenced by these networks. In the recent elections in Mexico, Twitter became a virtual place to bring together scientists, artists, politicians, adults, youth and students trying to persuade people about the candidate: Andrés Manuel López Obrador (AMLO). Our research is based on the collection of all tweets sent before, during and after the presidential elections of July 1, 2012 in Mexico containing the hashtag #AMLO. The aim of this study is to analyze the behavior of users on three different times. We apply SentiWordNet 3.0 in order to know how user behavior changes depending of the political situation and whether this is reflected on the tweets.

Keywords: user behavior, sentiment analysis, social web, Twitter, public participation, web 2.0, user generated content.

1 Introduction

The increasing participation in social media and networks has resulted in an explosion called user-generated content. Researchers use this information in order to answer interesting questions that were impossible to solve before as: how does an idea spreads across users? What topics attract more? What is the user behavior face to a certain event? Some studies have established that it exist a certain correlation between the analysis of sentiments and the public opinion polls for job approval ratings [1]. In this way, we find that the analysis of sentiments is fundamental to give some clues about how users can influence or persuade others, by using a kind of humor, in Twitter.

The aim of this article is to analyze, from the hashtag #AMLO or containing among its 140-character the word *"AMLO"* (acronym for *"Andrés Manuel López Obrador"* that was one of the four candidates with longer follow-up by young people

A.A. Ozok and P. Zaphiris (Eds.): OCSC/HCII 2013, LNCS 8029, pp. 3–12, 2013.

in social networks), the interaction of the users during a political event. Our corpus is composed by the tweets coming from one day before the election (June 30, 2012), the day of the election (July 1, 2012), and one day after the election (July 2, 2012). This dataset should, therefore, provide a broad coverage of the discussions between users that were for or against the candidate. In our approach we created a computer program in order to extract, from Twitter, the 140-character tagged with the hashtag #AMLO. Our program extracts all the contributions done in one day and it makes a copy of the information into a text document that we used later in two different computer programs to analyze discursive elements. These elements are full of humor, most of the time, of the political interaction in Twitter. The first one is Tropes Semantic Knowledge (see: http://www.semantic-knowledge.com/tropes.htm) which is used to make a textual analysis. The second one used was SentiWordNet 3.0 (see: http://sentiwordnet.isti.cnr.it/), to analyze the extracted adjectives after making the proper English translation. It presents numerical indices that allow researchers to subjectively assess the opinions, feelings and sensitivities discharges in the tweets that refer to the candidate Andres Manuel Lopez Obrador. Our objective is to gain some insight of the Mexican election in order to make an interpretation of the attitude or mood expressed in the tweets.

In this paper, Section 2 describes some state of the art around the use of analysis of sentiments in political cyber participation. Section 3 explains a tweet and how our corpus was created. Section 4 presents details of the use of SentiWordNet 3.0 in order to analyze sentiments covered in the tweets and we illustrate our experiment results performed to examine user behavior patterns. In Section 5, we conclude this paper with suggestions for further work.

2 State of the Art

Social networks have become in recent years, an excellent media that works in parallel to the traditional ones. In many countries, social media like Twitter, blogs and Facebook, are playing important roles in politics. Specially, Twitter prompts the user to express their thoughts and feelings and share them instantly and globally by trying to condense information. This has been very attractive to many users in the way they can read rapidly and have an idea of what is happening by not spending many time. Twitter is increasingly attracting a lot of users all over the world and it has experienced an extraordinary growth over the past years becoming the number one microblogging tool.

We find that Twitter has been, also, used as an important social medium in some critical social actions such as the Mumbai terror and the post-election protests in Iran. In another example, in the absence of democratic elections, an estimated of 70 million bloggers in China have become the voice of the people [2]. Also, a growing number of Pakistanis turned to YouTube, Flickr, Facebook and short message service (SMS) text messages as alternative media during the *"Pakistan Emergency"* from 2007 to 2008. This began after Pakistani President General Pervez Musharraf suspended the

chief justice of the Supreme Court and the government canceled cell phone networks and some news channels were blocked [3]. Similarly, in Arab countries like Iran large-scale protests were coordinated from Twitter. In the same way the Arab countries did the *"Arab Spring"*, Mexico had its *"Mexican Spring"* in 2012 during election campaigns for President of the Republic. In this way, the use of social digital media presents a new opportunity to study how the user interacts with others knowing that the anonymity has a great importance in free expression. It is important to say that, in one year, the social network Twitter, in Mexico, grew from 4.1 million users to 10.7 million, thus Mexico is among the 10 countries *"most twiteros"* such as the U.S., England, Japan and Brazil.[1].

In Mexico, as in many other countries, most politicians and opinion leaders have an Internet presence and make use of social networks, especially Facebook and Twitter. However, this has not led into a closer relationship between government and citizens, a better democracy or reliable information. The marginal social groups, minorities, and civic organizations have been the most benefited from the advantages that digital media offers. The youth organization *"yosoy132"* was born during the elections in Mexico in 2012, using social networks and bringing together youth universities groups from all the country regardless of their social conditions. Digital media helped these groups to have greater presence in public space because traditional media generally didn't pay much attention to them. They have learned to create collaboration networks, and to share information and knowledge.

Bollen presents the first work around the analysis of the use of humor in Twitter data [4]. In this research Bollen uses POMS (Profile of Mood States) to distill, from Twitter messages, 6 time series corresponding to different emotional attributes (for example, tension, depression, anger, vigor, fatigue and confusion). POMS is a psychometric instrument that provides a list of adjectives for which the patient has to indicate the level of approval. Each adjective is related to a state of mind and, therefore, the list can be exploited as the basis for a mood analyzer from textual data. It is important to mention that although the authors argue that the data from Twitter could be used to predict the future of an election campaign, they do not present any predictive method. Although this article used the 2008 presidential campaign and the election of Obama as a stage, we don't find inferred conclusions regarding the predictability of elections. On the other hand, O'Connor employs a subjective lexicon coming from the Opinion Finder in order to determine a positive and negative score from every tweet corresponding to each data set [1]. In this case, the relation between the number of positive and negative tweets about a given topic are used to calculate a confidence score. O'Connor et al., clearly indicates that for the simple manual inspection we can find many examples that have been incorrectly classified according to a feeling. This method is used, by the authors, to measure issues such as consumer

[1] El Economista, March 21, 2012. http://eleconomista.com.mx/tecnociencia/2012/03/21/twitter-alcanza-mexico-107-millones-usuarios

confidence, presidential approval and the 2008 presidential election in the United States. According to O'Connor et al., the consumer confidence and the approval of presidential elections exposed some correlation between the feelings analyzed from the data coming from Twitter. However, they didn't find a correlation between the polls and the sentiments contained in the data from Twitter. In this case, there was no evidence to say that it is possible to make a prediction about the presidential candidates by analyzing the preferences expressed in Twitter.

Tumasjan et al., presented at 2010 a work that consists of two distinct parts: the first one used LIWC (Linguistic Inquiry and Word Count) to make a superficial analysis of the tweets that are related with the different political parties that were competing for the German Federal election of 2009 [5]. Is in the second part, however, in which the authors claim that the count of tweets that mention one of the parties or any candidate, accurately reflects the election results. On the other hand, they contend that the MAE (Mean Absolute Error) of the *"prediction"* based on Twitter data was very close to the real surveys. The study of Twitter combined with politics is a field of research that is just beginning and that allows not only the envision of situations that were not very frequent before, like the participation of youth in politics, but it also provides a sea of information, from technology, that can be of great interest to the society.

In the next section we are going to explain the constitution of our corpus and the process done to analyze the data.

3 Extraction of Tweets in Order to Constitute the Corpus

The role of social media during the presidential campaign of 2012 in Mexico gained great importance because Twitter became the principal media for the youngest people. Our analysis is based on all the tweets collected before, during and after the 2012 presidential elections in Mexico.

3.1 What Is a Tweet and How It Is Composed?

In our approach we used tweets extracted from Twitter. A tweet is a little message of no more than 140 characters that users creates in order to communicate thoughts, feelings, or even participate in conversations. The tweet allows the communication of texts, videos or pictures by providing a link to it. Some words of the tweet are preceded by the pound sing # (hashtag). By using the hashtag, users can recover, reply (known as retweet) or follow conversations about a certain subject because this hashtag becomes automatically a hyperlink on Twitter. Everyone who clicks on a hashtag has the possibility to view the sear results of all other tweets that contains the same hashtag. In our case, we used the hashtag AMLO, #AMLO, to recover all the conversations, ideas, phrases that were produced during the Mexican elections of 2012.

The tweets that we recovered have different structures. For example:

- A simple phrase like: *"2012 AMLO even though the others"* (in Spanish *"¡AMLO 2012 le peje a quien le peje!"*)
- A phrase containing name(s) of the user(s). For example: *"The document of AMLO seems very real. What's your opinion @XochitlGalvez?"* (In Spanish: *"El documento de AMLO parece muy real. ¿Tú cómo ves? @XochitlGalvez"*)
- A phrase with links, for example: *"They disobey the IFE, the promotional against AMLO is still in the air http://t.co/kEEhosEA."* (In Spanish: *"Desobedecen al IFE, promocional contra AMLO sigue al aire http://t.co/kEEhosEA"*)
- A phrase with retweet RT and the name of the user who originally sent the phrase. For example: *"RT @epigmenioibarra: @ IFEMexico has enabled that the lies of PAN against AMLO, that were reported and verified in the media, are still present on radio and TV. That is outrageous."* (In Spanish: *"RT @epigmenioibarra: Ha permitido @IFEMexico que mentiras del PAN vs AMLO, denunciadas y comprobadas en los medios, salgan al aire en radio y TV. Que indignante"*)
- A phrase with hashtag(s), like #elpejehaceturismoelectoral, #elecciones2012, #NiUnVotoAlPRI."

3.2 Extraction of Tweets and Constitution of the Corpus

Our corpus was composed with tweets recovered before, during and after the elections of 2012. In order to recover the tweets we have done a program using Processing (see: http://processing.org) which is an open source programming language and integrated development environment (IDE) built to teach the fundamentals of computer programming in a visual context. Our program recovers all the tweets containing a #AMLO, and produces a text were the tweets are listed. Part of the pseudocode is listed below:

```
1. Use of Twitter API to communicate with Processing by a
   statement indicating that the search is done by #AMLO and
   that we are going to get a maximum of 100 responses each
   time.
2. Generate the file in which the information is going to be
   saved by declaring a variable of type PrintWriter.
3. The information that is retrieved from Twitter is in XML
   format.
4. Assign a new element XMLElement to explore structures in
   XML and to save tweets in order to recover the xml struc-
   ture.
5. Start a loop to read the file and retrieve the contents
   of all output lines.
```

For this analysis, we used Tropes (see http://www.semantic-knowledge.com/tropes.htm) to extract the main concepts around one of the four candidates for President named Andrés Manuel López Obrador (AMLO). We have also made an English translation of tweets in order to apply a sentiment analysis using SentiWordNet 3.0. [6].

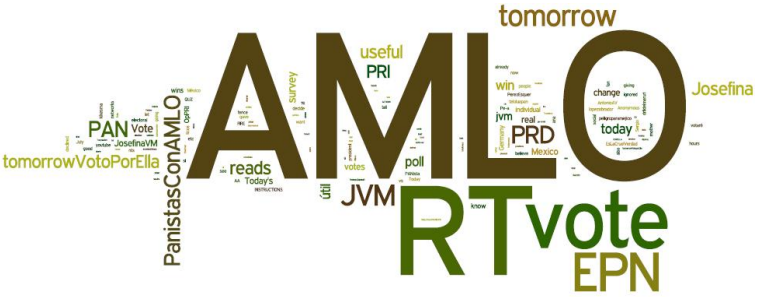

Fig. 1. Cloud of concepts for the corpus BEFORE the elections

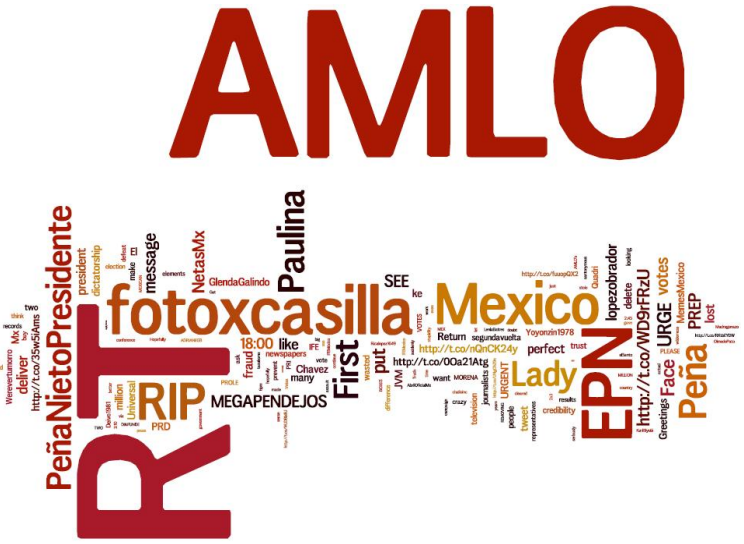

Fig. 2. Cloud of concepts for the corpus AFTER the elections

This analyzer is a lexical resource in which each synset of WordNet [7] is associated with three numerical scores: *O* (how neutral the word is), *P* (how positive the word is), and *N* (how negative the word is). In the next section we are going to present some of the adjectives that were recovered by using Tropes and we are going to explain the process made in the use of SentiWordNet 3.0.

4 Analysis of Tweets by Using SentiWordNet 3.0

To analyze the tweets, our approach was based in the use of adjectives that were used in the tweets. In this case, first of all we extracted the adjectives by using Tropes which is a software that has been useful in the educational field, the discourse analysis, the anthropological and some others sciences. It is used to easily analyze written and oral texts from the semantics, grammar - especially the categories of words - and

mainly the acts of speech. After the extraction made by Tropes we used Wordle (see: http://www.wordle.net) that exhibites by chromatic clouds of words, the most frequent words used in the discourse. They are differentiated by the size of the source. We selected the adjectives that appear in Tropes as well in Wordle. Examples of the clouds of concepts generated by Wordle are shown in Figure 1 (before elections) and Figure 2 (after elections). From the use of Wordle and Tropes we selected some of the adjectives that are more representative to our work. In table 1, we present the adjectives used for our analysis.

Table 1. Adjectives used to analyze each of the stages of the election process: before, during and after the elections

Before the elections	During the elections	After the elections
better	arrogant	absurd
crazy	awesome	perfect
honest	coherent	remarkable
intolerant	damn	sad
temperamental	disrespectful	true
tourist	gay	urgent
true	genius	visceral
worthy	liar	
	lost	
	militant	
	secret	
	serious	
	stupid	

Each of the adjectives presented above was analyzed by using SentiWordNet 3.0. This resource adds three numerical grades for each word: neutrality (how neutral is the word), positivity (how positive is the word) and negativity (how negative is the word). Each score is between 0 and 1 where the sum of the three should give 1. However, it is not as obvious to use SentiWordNet because the user has to carry out a disambiguation in order to know which of the interpretations or meanings of the words, shown in SentiWordNet, correspond with the sense that the user wants to give to the word found in the corpus. For example, if we search the adjective sad[2], in SentiWordNet, it has three senses. Between each of the senses there is a triangle showing a position with respect to the three scores. Below the triangle are the letters P (positive), O (neutral), N (negative) with their respective scores. For the sad example we have selected: *"sad#1 experiencing or showing sorrow or unhappiness; "feeling sad because his dog had died"; "Better by far that you should forget and smile / Than that you should remember and be sad"- Christina Rossetti."* The most important thing in our analysis is the numerical grades given by SentiWordNet 3.0 which we have used to compare the three corpus. In this case, for sad we have selected: P: 0.125 O: 0.125 N: 0.75.

[2] http://sentiwordnet.isti.cnr.it/search.php?q=sad

Table 2. Adjectives and numerical grade for the corpus *before* the elections

Before the elections	**Numerical grade extracted from SentiWordNet 3.0**
better	P: 0.875 O:0.125 N: 0
crazy	P: 0 O: 0.5 N: 0.5
honest	P: 0.375 O: 0.625 N: 0
intolerant	P: 0.125 O: 0.125 N: 0.75
temperamental	P: 0 O: 0.75 N: 0.25
tourist	P: 0 O: 0.75 N: 0.25
true	P: 0.5 O:0.5 N:0
worthy	P: 0.875 O: 0.125 N: 0
AVERAGE	Positivity (P): 0.344 Neutrality (O):0.437 Negativity (N):0.219

Table 3. Adjectives and numerical grade for the corpus *during* the elections

During the elections	**Numerical grade extracted from SentiWordNet 3.0**
arrogant	P: 0.5 O: 0.125 N: 0.375
awesome	P: 0.875 O: 0 N: 0.125
coherent	P: 0.75 O: 0.25 N: 0
damn	P: 0.375 O: 0.25 N: 0.375
disrespectful	P: 0.5 O: 0.125 N: 0.375
gay	P: 0.375 O: 0.5 N: 0.125
genius	P: 0.75 O: 0.25 N: 0
liar	P: 0 O: 0.375 N: 0.625
lost	P: 0.125 O: 0.25 N: 0.625
militant	P: 0.5 O: 0.375 N: 0.125
secret	P: 0.25 O: 0.625 N: 0.125
serious	P: 0.25 O: 0.375 N: 0.375
stupid	P: 0 O: 0.25 N: 0.75
AVERAGE	Positivity (P): 0.404 Neutrality (O):0.288 Negativity (N):0.308

Table 4. Adjectives and numerical grade for the corpus *after* the elections

After the elections	Numerical grade extracted from SentiWordNet 3.0
absurd	P: 0.625 O: 0.375 N: 0
perfect	P: 0.625 O: 0.375 N: 0
remarkable	P: 0.375 O: 0.625 N: 0
sad	P: 0.125 O: 0.125 N: 0.75
true	P: 0.5 O: 0.125 N: 0.375
urgent	P: 0 O: 1 N: 0
visceral	P: 0.25 O: 0.75 N: 0
AVERAGE	Positivity (P): 0.357 Neutrality (O):0.482 Negativity (N):0.161

As we can see in the above tables, the average for each of the corpus reflects the tendency of the users in a political process. In this case, *during* the elections we have the adjectives that are more negative. The average of negativity *during* the election is 0.307 against 0.219 for *before* and 0.161 for *after*. Some of the adjectives that characterized the corpus *during* are: *arrogant, damn, disrespectful, liar* and *stupid*. All of these adjectives are supposed characteristics of the candidates. In this article we don't show examples of the tweets but they are, almost all of them, telling a story. In many cases, they tell the story of the user and his opinion. We can notice, also, that users are likely to attack and use strong words when the election is taking place in order to persuade electors.

After the elections, neutrality is the strongest attitude toward the result of the election and even if we find users that are not satisfied by the result they don't participate very much after they know the result of the election. Users who are satisfied are the ones that are making tweets in order to show their feeling. As we can notice one of the adjectives presented in the corpus *after* is: *sad*. Is in this last corpus that we find an adjective for a feeling. In the others two corpus, *before* and *during*, we don't find demonstrations about what the users are feeling and even more we find adjectives to emphasize characteristics of the candidates.

In all the tweets extracted we find the humor in order to persuade others users. When the users want to offend they do it by using the criticism showing cartoony defects. The user behavior is very particular in each of the stages of the election.

5 Conclusions and Further Work

Our paper presents a first interdisciplinary study which analyzes the political participation of young Mexicans during elections for President of the Republic through the tweets sent before, during and after the election of July 1, 2012. In this case, we visualize Mexican youth as social individuals who need others to make decisions, based on the presence of tweets, and instrument that empowers the users and affects the collective.

In our corpus we are able to find sequences than convey, in a remarkable way, mood traits like ethnographic, semantic and psychological elements that are constant throughout the corpus. Ingenuity and inventiveness emerge with particular communicative liberty. The users want to caricature and make ridiculous all the candidates. The criticism is open to everyone, and they do it. They write confidently, cheerfully most of the times, without fear of reprimand.

As in other countries, in Mexico the use of Twitter to create phrases and messages of 140 characters has encouraged the participation of young people creating movements like *"#yosoy132"* or *"#movimientoMx"*. These movements are a manifestation of young people belonging to groups with common interests. In this way, the increasing use of social networking has allowed to dilute geographical boundaries making closer the citizen to a greater democratic participation. Our interpretation through the application of Tropes, and SentiWordNet 3.0 confirms *"moments"* during the election period in which the representation and visualization of tweets reveals a reality of our Mexico: humor present in the political arena.

Further work, should include a comparison between political subjects and events that are produced everyday. Are the users writing tweets to persuade others? Can we compare the same situation (before, during and after elections) in other countries? One of our future works is also dedicated to the incorporation of geographic situation in order to know how this is reflected in the behavior of the users.

References

1. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: linking text sentiment to public opinion time series. In: Proc. of 4th ICWSM, pp. 122–129. AAAI Press (2010)
2. Friedman, T.: Power to the (Blogging) People. New York Times (September 14, 2010), http://www.nytimes.com/2010/09/15/opinion/15friedman.html
3. Yusuf, H.: Old and New Media: Converging During the Pakistan Emergency (March 2007-February 2008). Massachusetts Institute of Technology, Center for Future Civic Media, Cambridge, Mass (2009), http://civic.mit.edu/blog/humayusuf/old-and-new-media-converging-during-the-pakistan-emergency-march-2007-february-2008
4. Bollen, J., Pepe, A., Mao, H.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, pp. 450–453 (2011)
5. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp. 178–185 (2010)
6. Esuli, A., Sebastiani, F.: SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation, LREC 2006, Genova, Italy, pp. 417–422 (2006)
7. Fellbaum, C.: WordNet: An Electronical Lexical Database. The MIT Press, Cambridge (1998)