From Tweet to Graph: Social Network Analysis for Semantic Information Extraction

Rocio Abascal-Mena, Rose Lema, Florence Sèdes

From Tweet to Graph: Social Network Analysis for Semantic Information Extraction

Rocío Abascal-Mena, Rose Lema
Universidad Autónoma Metropolitana – Cuajimalpa
México, D.F. Mexico
mabascal@correo.cua.uam.mx, rose@xanum.uam.mx

Florence Sèdes
Institut de Recherche en Informatique
Université Paul Sabatier
Toulouse, France
sedes@irit.fr

***Abstract*—**This paper represents a study along the cutting edge of the current analysis of online social network in relation with the contents communicated among users. Twitter data is carefully selected around a fixed hash-tag in order to study the specified content in relation with other contents that users bring to connection. A separate network of hash-tags related (in tweets) is constructed for different days; the networks are analyzed within advanced Gephi package, providing several measures ---degree, betweenness centrality, communities, as well as the longest path, by which the evolution of communication around specified concepts is quantified. Our study is absolutely in the current trend of analysis of online social networks that, going beyond mere topology, reveals relevant linguistic and social categories and their dynamics.

Keywords—text mining; Social Web; Social Network Analysis; Theory of Graphs; community detection; Twitter.

I. INTRODUCTION

The communication through the use of Information and Communication Technologies (ICT) leads today to ask the question about the writing and the meaning of the messages. From the academic language to the phonetic language, through complex processes, the linguistic practices are evolving. The emoticons, the "SMS language" or the alternations between different codes are several elements that allow us to open the discussion about the meaning of the relationship between orality and writing (where the oral and the written form the contextualization of the speech). Currently, we find new ways of Computer-Mediated Communication (CMC) such as the social networks that are composed by nodes (persons, groups or organizations linked to others according to their interests and/or points of view in common). Virtual social networks have led to a new way of communication that is different from the oral one, where the restriction of time and space generates new linguistic practices. It is not enough to use only an electronic media, such as the computer, to communicate and to interact in a virtual social network. Therefore, it is necessary to have prior knowledge and language skills in these new media to follow and be part of virtual conversations.

Twitter is a website that is registered under the category of social networking and microblogging. The main feature of Twitter is that messages that are sent have a maximum length of 140 characters, called tweets. These tweets can be stored and categorized into themes by the use of *hashtags* (keywords or abbreviations) that are preceded by the symbol '#' (called pad) and that allow to follow, search and find conversations related to a common theme. We can say that the tag has become essential and fundamental in the understanding of a new form of communication in which the tags evolve over time and mark rhythms and themes (Trending Topics, TT) while they are mass used. The conversations or messages, tweets, take a nonlinear way including tags (#hashtags), usernames (using the symbol '@' before the name of the user) and links to images and websites. This delinearization is an approach to the way in how we think and talk at the same time: going from one idea to another to return to the initial conversation.

Beyond denominations and differences that exist, Twitter has changed the way in which users participate in Internet. The idea of sharing in order to become key actors, producers and generators of information and content, leaving a passive stance of exclusive consumption to become active participants and key elements is a trend that can not be disputed and that looms irreversible. In this context, we find the use of Twitter as a medium for political discussion in local and national contexts (such as the 2010 elections in Australia, 2011 in Mexico and 2012 in the United States, to name a few) and also in protests and activist movements such as those in Egypt, Tunisia and Yemen. Twitter has become a legitimate channel of communication in the political arena as a result of the 2008 presidential campaign conducted in the United States [1].

For researchers from various disciplines, Twitter is a key element to be studied and analyzed in order to understand and learn new ways of communication and language through the use of *techno-discursive elements* that exist in the tweets. These elements such as the shortening of words or the assumption of knowledge on a given topic leads to the encryption and the encoding of the message that in some cases can only be understood by those who are in contact, all the days, with Twitter.

Honeycutt and Herring [2] showed that Twitter is not only used in one direction of communication but generally functions as a media for conversation. In their study, they found that randomly selecting a set of tweets, 31% will contain the '@' symbol and 91% will be sent to a specific user. Cunha *et al.*, 2011 [3] studied the propagation of hashtags, containing the symbol #, within speech communities which are groups of people whose members linguistically influence each other. In this way, they analyzed in which cases the hashtags are adopted and reused in future messages. In their research they found a relationship between the distribution of hashtags within the rankings of frequency and the length of the hashtag in order to be adopted by the users. According to [3], the hashtags that are more used correspond to those containing more characters and that are more explicit or that are already known by the speech community. In 2012, Cunha *et al.* included the gender, in their research, as a social determinant that influences the user in selecting hashtags about a particular topic [4]. Similarly, Romero *et al.* [5] studied how information spreads on Twitter by analyzing the variations of the diffusion characteristics in different themes. To do this, they calculated the probability of adopting a hashtag based on the number of exposures and the speed with which the curve of influence descends.

Chew and Eysenbach [6] found the existence of variability in the use and context of hashtags when they studied, in a period of time, the use of '# H1N1' and '# swine flu' in the tweets generated during 2009.

The hashtags have also been studied in order to determine the feelings issued within tweets. Rodrigues *et al.*, [7] presented a study of a corpus of tweets recovered during the 2010 elections where about ten hashtags were categorized into positive, negative or neutral. In this way, they count the number of tweets that are included to analyze if the hashtags can be used in order to determine feelings issued in a tweet. And, they found that the hashtag has already an intention that reflects a sentiment.

Also, Twitter has been used to create communities through the use of hashtags [8, 9] that are applied in order to help the formation of ad-hoc public around specific issues. The hashtags are used to quickly expand events and emerging issues. Such is the case of, for example, #yosoy132 (Mexican motion generated during the presidential campaign of 2012) or # 5-M (protests in Spain from 2011 to 2013). In this context, we find similarities in the labels used among various movements around the world, as a way to show the same ideology as concrete connections and similarities despite their geographic situation or cultural context.

While the previously presented works give us an idea of the way in which Twitter is analyzed in order to understand the new way of communication, we don't find studies that explore more *semantic* aspects of the use of the hashtags and their relationship with other elements. In this paper we present the results of our research around the use of hashtags as basic elements in the composition of the message trying to understand and interpret the meaning of the relations between concepts and hashtags over time. Our research was carried out from the recovery of tweets tagged with *#noalospluris* in mid-December 2013. This hashtag was used at the end of 2013 in Mexico to express an upset by the existence of multimember in the Chamber of Deputies and the Senate. The purpose of our research is to find visualizations of the information as *networks* (nodes and links between information, in this case between the hashtags) in order to understand the meaning of the use of different elements in the tweet. How hashtags are related within other concepts and words used? Is it possible to visualize the central problematic of the movement #noalospluris by using a network? How clusters and subgroups of information can help to understand the meaning of the movement? Our work aims to help the analysis of political and social movements in Twitter through a quantitative analysis by extracting the frequency of words and hashtags, the application of the theory of graphs to model the tweets like a network and the use of basics of Social Network Analysis (SNA) to extract semantic information. We have carried out, also, a qualitative analysis of the tweets by using the results of our quantitative analysis in order to make an interpretation of our results.

The rest of this article is organized as follows. In the next section we briefly describe the main characteristics of a tweet. In section 3 we present a background in the theory of graphs and Social Network Analysis (SNA). Section 4 presents a background of some works that have for purpose the analysis of tweets to understand social behavior. Section 5 is devoted to the presentation of the methodology used to extract and analyze our corpus. This methodology is based on techniques provided by the SNA field. The application of these techniques and the results are presented in section 6. Finally, we provide insights for future work and our conclusions.

II. Characteristics of a Tweet: the Case of the Hashtag

At the launch of Twitter in 2006, the main objective was that users respond to the question *"What are you doing?"* being a kind of *"life streaming"* in order to make public the day a day of the people in Internet and opening the doors of the individuality and the privacy. However, it is from the daily use and trends of Web 2.0 (where the tagging of information is encourage in order to make more easily to recover information not only from its syntax but also from its meaning) that Chris Messina (technologist located in San Francisco) proposed, in 2007, from his blog and his Twitter space the use of the symbol '#' to improve filtering and contextualization in Twitter. Overall, the proposal was aimed for the creation of interest groups. However, using the pad as a means of labeling can be considered as a *linguistic innovation* that allows the creation of jumps within a sentence and it is a reference, always present, supplementing the information (meta-information) that is being read.

The use of labeling extends the ability of the message making the hashtag an identifier to (a) document the tweet using hypertextualization, (b) categorize the tweet in ad-hoc groups and subject areas, (c) generate metadiscursive comments, (d) create expressive marks, (e) create personal interpretations and (f) identify a thematic as generic.

For the generation of metadiscursive comments we use the definition of metadiscourse provided by Hyland and Tse [10]: *metadiscourse refers to all the strategies used by the writer to organize the text and allow the connection of different ideas*. However, the metadiscourse is a crucial element in the argumentative writing (found on Twitter) that facilitates persuasion in messages [11]. There is no more research on the use of metadiscourse in Twitter except for the work of Russell [12] and Poell [13] that confirm the use of metadiscourse processes to persuade tweeters around raised movements of the Arab Spring. In the case of Egypt, the tweets that were sent intended to build the history instead of being only transmitters of information in which the criticism and the scrutiny were present. Russell [12] argues that the awareness and the metadiscourse that is inspired by these tweets have their precedents: *it is a recurring phenomenon in the history of professional journalism, which originates, particularly in times of political crisis or polarization* [14]. According to Georgakopoulou [15], what stands out in these new forms of digital storytelling is the potential for greater interactivity with the user and the range of forms of involvement of the reader, who may even decisively change the course and interpretation of the story told.

Currently, the use of the generic label (hashtag) is not given at all because we still find that the messages, in Twitter, contain multiple hashtags in order to reach a wider audience. Bruns and Burgess [9] argue that the use of a generic label would categorize tweets within a variety of themes making it much easier the disambiguation of tweets that have nothing in common. Similarly, the generic tag allows an alternative explanation and emphasis without the use of visual elements (the use of bold, italic, etc.) or the use of emoticons.

III. Background: Theory of Graphs and Social Network Analysis

The study of social networks is of great interest to the scientific area because the understanding of connections allows, at the same time, to represent and understand various problems of the nature that contain similar patterns to other kinds of problems. However, the study of social networks is not new, the analysis of society based on the concept of network was a significant changing approach on the science in the twentieth century. Before that, the studies were limited to the analysis of the phenomena of social networking and examine only certain parts in detail. Only from the last century the work begins to change the focus to the social phenomenon constituted by the interaction between the parties that compose the entire network.

In 1736, the mathematician Leonhard Euler Paul published an article about the enigma of the bridges of Königsberg [16]. He was the first to use the metaphor of networks to resolve a problem. Köningsberg was a Prussian town in the middle of islands in the middle of a river. In total, the city had seven bridges and for his habitants it had become a distraction to cross the town passing only once every bridge. Euler demystifies the idea of the citizens, demonstrating that to cross those bridges without repeating the path it was impossible and he presents a possible route for the habitants of the region. The mathematical connect four land parts (*nodes*) with seven bridges (*edges*), confirming the absence of a route desired by the residents and the creation of the first theorem of graph theory. This theorem assumes that to enter and leave the city without having to use the same bridge would require to each part to have at least two bridges.

A network can be mathematically represented by a graph consisting of a set of points, nodes or vertices connected by lines connecting pairs of nodes, the edges. When two points are directly connected by an edge they are called *adjacents*. In the case of social networks, the vertices represent the actors, and the edges represent the relationships between the actors involved. Thus, the actors become important when they are connected with other actors. In 1984, Sowa introduced the modeling of a text as a graph, called a *semantic network*, in which the main objective was to model human knowledge using graphs of concepts that are related to others by the existence of conceptual relations [17].

In this article, we present the constitution of graphs from the analysis performed to tweets in order to find how the labels evolve over time and what position they take within the subject. Thus, in this study we apply graph theory to complement the linguistic analysis from the interpretation, over a period of time, of the existing relationships between the concepts used in the tweets. Our analysis responds to the need of studying visual elements that could help us to understand the formation and variation of language, as its meaning, when it is mediated by the use of current technologies.

The application of graphs in textual analysis is not intended that the graph models the whole meaning of the text; rather it focuses on capturing the dependence that exists between the concepts referred to in the text. In this case, the graph is seen as a social network of words where the meeting points are the grammatical contexts in which these words appear. When performing a text analysis using graphs, modeled as a social network, it is possible to use techniques that come from the Social Network Analysis (SNA). In our case we use three techniques that allow us to identify key concepts and grammatical contexts: the *degree*, the *centrality* and the generation of *communities*. The degree is the number of points that is adjacent to a particular point. That is, the number that indicates how many connections (relationships) exist between a node and others nodes.

In 1948, Bavelas introduced the idea of centrality applied to human communication in small groups and hypothesized about the relationship that exist between structural centrality and the influence in group processes [18]. One point is central if it has a high degree, which corresponds to the intuitive idea of centrality in which a point is central if it is well connected to other points of their environment [19]. In our analysis we use the idea of *betweenness* that determines in what amount a point is an intermediary between other points because it is located in the path between them [19]. The centrality index based on *betweenness* is often used, with more success, to extract the words that are relevant in unstructured texts. This index has shown superiority over the use of word frequency to extract those words that are relevant to a set of texts [20].

Communities commonly called subgroups (clusters) are associated with an area of the graph that has a relatively high density. In some cases, these clusters allow to derive a family of related concepts.

IV. Background: The Use of Twitter to Analyze Social Behavior

The rise of social web, particularly Twitter, provides new opportunities to collect real time data in large quantities directly from users. Big data can be analyzed in various ways in order to examine patterns in a wide range of subjects. In this way, tweets can be analyzed in order to track reactions to events. Since Twitter provides the possibility to extract tweets and compose actual corpus there have been a lot of linguistic research applied in tweets. An increasing number of empirical analyses of sentiment and mood are based on textual collections of data generated on Twitter as they used sophisticated algorithms to preprocess, apply grammatical rules and classify them in mood categories. In this way we find, for example, the use of a lexicon based classifier as a dataset that is also classified using SVM and/or Naïve Bayes [21, 22]. A classifier is developed specifically tuned for tweets, using key words, phrases and emoticons to determine the mood of each tweet [23]. Several methods have been already proposed for exploiting tweets in order to detect people's mood changes throughout the day [22, 24].

In general, studies analyzing tweets by combining different sentiment analysis algorithms have been able to give new insights into human behavior as a result [e.g., 25, 26, 27 and 28].

In our previous work [29] we showed some results of our analysis of tweets in order to study the behavior of users before, during and after the presidential elections of 2012 at Mexico by applying semantic techniques.

In this way, several approaches to analyze, extract and automatically detect meaning in tweets have being proposed in the last years leaving evidence of the limitations that already exist in the analysis of messages when applying only semantic approaches.

An emergent social network, like Twitter, is a good source of big data to study social interactions as human social behavior. Thereby, an analysis of tweets should not only be based on the linguistics aspects but should also apply SNA to differentiate it from traditional social scientific studies, which assume that it is the attributes of individual actors that matter. SNA produces an alternate view, where the attributes of individuals are less important than their relationships and ties with other actors within the network. This approach has turned out to be useful for explaining many real-world phenomena as we attempt to prove that it could be used to contextualize and to reveal relevant linguistic and social categories and their dynamics.

V. Methodology: Extraction and Composition of the Corpus

In the proposed method, we analyze tweets using three steps. In the first step, we extract tweets from Twitter containing a certain hashtag. In the second step, we parse the extracted tweets in order to clean it and to create two different tables: one with all the labels of the important words contained in the corpus and another containing the relationships between each couple of words. These two steps are automatically performed by using R which is an interpreted computer language designed for statistical data analysis. The third step concerns the creation of the network by using Gephi that is an open source software for graph and network analysis [30].

We have selected R for our approach since it contains some packages (collection of functions) like TwitteR (to connect to Twitter and retrieve tweets) and tm (for text mining) that are useful for data import, corpus handing, preprocessing, data management and for the creation of term-document matrices.

In this way, by using R we have developed a computer program that is able to:

1) Extract tweets containing a certain tag, in our case we extracted tweets containing #noalospluris. It is noteworthy to say that the API of Twitter that is being used only allows the recovery of tweets that have between 6 and 9 days of creation. It's not possible to retrieve tweets older than 10 days, so in some cases where the participation is very high the extraction process should be done daily.
2) Eliminate stopwords and punctuation.
3) Eliminate website addresses (anything that starts with "http").
4) Eliminate actors (all that contains the symbol '@' as the first character).
5) Delete numbers and special characters (except the symbol '#' and letting the numbers and special characters in tags).
6) Change the corpus to lowercase.
7) Eliminate accents.

8) Get the frequency for each word.
9) Generate the nodes for words that are above the third quintile (Q3).
10) Determinate the edges for each node.

To compose our corpus we have selected a movement in Mexico that has a couple of years and that has been promoted through a website (http://revoluciondelintelecto.com/pluri.php) where people enter their details to join the cause: remove from the Chamber of Senators and Representatives all the multi-members who have not gotten their status by the legitimate vote of the people. Thus, to gain a wider audience the movement use the tag #noalospluris in Twitter in order to get the attention of politics and to show them the repudiation and discomfort of people about the existence of multi-members. In late 2013, this movement had its peak in Twitter, the tweets showed not only the labeling of #noalospluris but the main problems afflicting the people. In our corpus we try to show how the movement of #noalospluris is becoming a forum that provides a platform for the expression of the problem of the existence of deputies and senators multi-member but also to show various problems that exist in Mexico. Since January 2014 the movement has replaced the tag #noalospluris for #intelecto (meaning that the only way to gain against political imposition is the use of intellectual and constructive ideas). As confirmed by Heverin and Zack [31] and Bastos *et al*. [32] labels evolve over time, especially in the context of economic crisis and political uncertainty.

Our corpus containing the label #noalospluris is composed by 2,481 tweets sent between the 6 and 16 December 2013. In our analysis we don't include the weekend of 14 and 15 December that have a radical decrease in the number of tweets this because the radio program that promotes the use of the hashtag #noalospluris has its broadcast from Monday to Friday.

In recent years, we have seen the relationship that has Twitter with other traditional media as we find topics on Twitter that occupy the main headlines in the news [33]. This is an essential feature of Twitter: cover and discuss important events in society. The tweet is valid for a couple of days, what is important is what you read here and now. Yesterday is no longer news. The central mechanism used to highlight a tweet is the use of hashtags, as we mentioned before, that place a tweet in a specific subject making it immediately accessible to millions of users. The hashtags, because of brevity of Twitter, are keywords that show the current reality. Before making an automatic analysis of the recollected tweets we were interested in analyzing the meaning of hashtags cited in the corpus in order to be able to understand visualizations of the information.

In this way, the use of hashtags responds to the need to present and discuss, quickly, important events and happenings. Thus, we find in the corpus analyzed some hashtags that disappear within the days as some others that appear only for a couple of days but are specific to an event that happened in those days. Such is the case of #diputadoencueratriz (naked deputy) and #posmeencuero (so I'm going to undress me) that appeared, on 12 and 13 December 2014, to discuss and criticize a Mexican congressman that in a full podium has naked himself in order to protest against the energy reform[1].

Despite that #noalospluris is used for a specific cause we find the inclusion of other hashtags which are pronounced against government actions. However, #noalospluris is a forum that is widely read and the inclusion of hashtags that are out of the context has an especial objective: reach a larger audience, a strategy to increase the visibility. Thus, other labels that make their appearance between #noalospluris are #noaumentocamionjalisco and #noalcamionazo. These tags were used, since December 13, to report an increasing dissatisfaction with the rate of public transport in the State of Jalisco, Mexico. The introduction of these tags is done by the use of phrases that have been said by Mexican political celebrities. The phrases are used with a sense of ridicule to the politicians who have not done anything about the country.

The introduction of new hashtags in a certain theme, even if there are not appropriated, is used to arrive to a larger public but also they correspond to new problems that need to widespread. The analysis of this hashtags will let us gain a better comprehension of the network since we know, now, that all the information contained has a certain importance.

Once we got the tweets we make a pre-processing, manipulation, cleaning and formatting in order to compose the corpus. The main structure for managing documents, by using tm package in R, is the *Corpus* that represents a collection of text documents (in our case a collection of tweets). With tm package R, we are able to create a Term-Document Matrix from our corpus. The matrix is exported as a .CSV file in order to be used in Gephi.

In the next section, we present the analysis carried out to our corpus by applying techniques of SNA.

VI. Applying Social Network Analysis

The tables that were generated by R (table of nodes and edges) are used in Gephi in order to:

- Generate a network of the corpus.
- Calculate the degree for each node.
- Calculate the centrality for each node.
- Detect the main communities.

A. Generation of the Network of the Corpus

Some elements, such as the use of the hashtag, can be seen with a naked eye by reading a part or the whole corpus. However, the use of techniques from the SNA will give us

[1]http://www.dailymail.co.uk/news/article-2522868/Mexican-congressman-takes-clothes-angry-protest-historic-energy-privatization-scuffles-break-doors-barricaded.html

greater certainty about the use and positioning that take some words and hashtags over the time.

In the Fig. 1 to 6 we show the network of concepts generated for the each day of our analyzed corpus. As we explained before, a network is composed by nodes and edges (also called links or connections). In our case nodes are the concepts (main words) within the network, and edges are the relationships between the concepts.

Visual representation of the network is important to understand the evolution of concepts used over the time as they tend to concentrate in communities that share the same thematic.

Fig. 1. Network of concepts generated for December 6, 2013 containing the hashtag: #noalospluris

Fig. 2. Network of concepts generated for December 10, 2013 containing the hashtag: #noalospluris.

Fig. 3. Network of concepts generated for December 11, 2013 containing the hashtag: #noalospluris.

Fig. 4. Network of concepts generated for December 12, 2013 containing the hashtag: #noalospluris.

Fig. 5. Network of concepts generated for December 13, 2013 containing the hashtag: #noalospluris.

Fig. 6. Network of concepts generated for December 16, 2013 containing the hashtag: #noalospluris.

B. Degree and Centrality

Using the degree, number that indicates how many nodes are connected to a respective node, for each hashtag we tried to identify the importance of these hashtags for a period of time. Does this degree is indicative of the permanence of the hashtag all along the entire movement?

The centrality based on the intermediation (*betweenness*) measures the frequency with which a node appears on the shortest path between the nodes of the network. Analyzing the results, it is interesting to show the eccentricity of the nodes: the longest path of a node to reach any other in the network. The degree and the eccentricity are shown for the hashtags that appear in at least two days of our corpus even if they are not consecutive days (see Table 1 and 2).

TABLE 1. DEGREE FOR THE PRINCIPALS HASHTAGS CONTAINED IN THE CORPUS OF #NOALOSPLURIS.

Hashtags	**Degree**					
	Dec 6	***Dec 10***	***Dec 11***	***Dec 12***	***Dec 13***	***Dec 16***
#intelecto	7	8	16	8	18	14
#jalisco	-	-	-	-	20	4
#mexico	20	6	62	21	45	27
#noalospluris	165	51	153	97	161	103
#pemex	-	-	-	13	-	10
#piedrasnegras	-	5	5	-	-	-
#posmeduermo	-	6	7	-	-	-
#posmeencuero	-	-	-	5	2	-
#reformaenergetica	-	-	9	14	-	-
#reformapolitica	12	-	-	-	5	-
#revoluciondelintelecto	7	-	8	-	-	-
#saltillo	-	7	9	-	-	7
#teregalo	3	-	4	-	-	-
#torreon	-	9	7	-	-	-
#yoquieroun2014sin	3	-	8	-	7	-

TABLE 2. EXCENTRICITY FOR THE PRINCIPALS HASHTAGS CONTAINED IN THE CORPUS OF #NOALOSPLURIS.

Hashtags	**Excentricity**					
	Dec 6	***Dec 10***	***Dec 11***	***Dec 12***	***Dec 13***	***Dec 16***
#intelecto	8	5	6	4	7	4
#jalisco	-	-	-	-	5	3
#mexico	5	4	4	4	5	3
#noalospluris	4	3	4	3	4	2
#pemex	-	-	-	4	-	4
#piedrasnegras	-	6	5	-	-	-
#posmeduermo	-	4	6	-	-	-
#posmeencuero	-	-	-	7	1	-
#reformaenergetica	-	-	6	6	-	-
#reformapolitica	3	-	-	-	4	-
#revoluciondelintelecto	5	-	5	-	-	-
#saltillo	-	4	6	-	-	1
#teregalo	6	-	1	-	-	-
#torreon	-	4	6	-	-	-
#yoquieroun2014sin	6	-	7	-	5	-

In the Table 3 we present the hashtags that have a centrality greater than zero (number indicating how many nodes must go through to travel a given network node) centrality and that are used repeatedly in more than one day.

TABLE 3. HASHTAGS THAT HAVE A CENTRALITY GREATER THAN ZERO.

December 6, 2013		**December 10, 2013**		**December 11, 2013**	
#noalospluris	462	#noalospluris	66	#intelecto	27
#teregalo	49	#saltillo	1	#mexico	48
		#torreon	27	#noalospluris	306
				#revoluciondel-intelecto	262
				#saltillo	26
December 12, 2013		**December 13, 2013**		**December 16, 2013**	
#mexico	10	#jalisco	9	#mexico	24
#noalospluris	603	#mexico	95	#noalospluris	366
#reformaenergetica	90	#noalospluris	1724		

From these results, we found an important relationship between the degree and the centrality: *the centrality is greater as it is higher the degree and there is big chance of having a centrality (bigger than zero) if the hashtag is used repeatedly over time*. If we apply this theory to concepts that are not used as hashtags but have a significant centrality and they are used in at least two different days we can find concepts that appear as important nodes and that are part of clusters (subgroups) for a precise thematic. In the Fig. 7 we show these concepts as we are going to present the clusters in the next section. Some of these concepts, are also used as hashtags as in the case of #mexico.

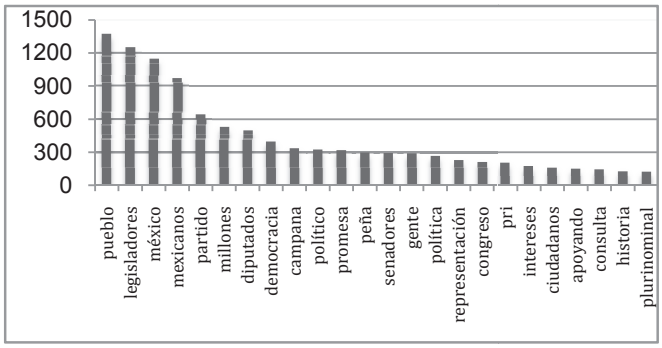

Fig. 7. Average of centralities obtained for each concept.

C. Subgroups

When the tweets of a particular topic are analyzed they might appear like if they are disorganized without a specific center where it prevails only the spontaneity and the anarchy. But, while the tweets are set in our mind like a swarm of birds or insects attacking irrationally, as the meaning of Twitter, we find that the interior of the network is organized, rational and creative, a "*swarm intelligence*". In this logic, according to Arquilla and Ronfeldt [34], the swarm reaches its maximum effectiveness and its greater attack power when the network members aren't together to combat their forces, but they concentrate on the dispersion, sharing important information so it can reach his final destination.

In order to analyze the composition of our corpus we have decided to apply, in Gephi, the modularity algorithm that allows the visualization of subgroups. In this way we have:

- 18 communities for December 6 and December 13,
- 15 communities for December 10,
- 20 communities for December 11 and December 12,
- 12 communities for December 16.

TABLE 4. PERCENTAGE MINIMUM AND MAXIMUM OF WORDS OBTAINED FOR THE GROUP OF COMMUNITIES ANALYZED FOR EACH DAY AS THE PERCENTAGE OBTAINED FOR THE COMMUNITY CONTAINING #NOALOSPLURIS

	Dec 6	**Dec 10**	**Dec 11**	**Dec 12**	**Dec 13**	**Dec 16**
Minimum percentage	0.22	0.6	0.27	0.45	0.34	1.34
Percentage for community containing #noalospluris	**15.5**	**11.98**	**15.85**	**11.71**	**17.12**	**24.83**
Maximum percentage	**15.5**	14.37	**15.85**	13.06	**17.12**	**24.83**

The number of communities in each corpus is not very interesting without reviewing the percentage of words that compose the communities of each day. So, we find that the biggest agglomeration of words in a community is about 24,83% found on December 16, the last day. The minimum percentage is found on the first day. Our results show that at the last day there is a crucial diminution of communities (Table 4).

By analyzing the communities that have a greater concentration of words we find that the hashtag #noalospluris is in this communities. In the Table 4 we show the minimum and maximum percentage of words obtained for the communities analyzed per day. Also, we show in bold the percentage obtained for the community containing #noalospluris which in some days (December 6, 11, 13 and 16) corresponds to the community with maximum percentage of words. See Appendix for a more detailed presentation of the percentage of words obtained by each community per day. The communities containing #noalospluris are in bold in the Table 4 in order to show that *it's not necessary to analyze the entire corpus but only the communities with high percentage in order to be able to know more about the corpus*. In this way, we have found the subgroups (communities) that provide concepts or words that help to disambiguate (with semantic information) the entire corpus. Also, the community with the tag #noalospluris has almost all the concepts shown in Table 4. In this way, we find that concepts with biggest centralities are concentrated in the same community.

Our results show that it is not necessary to read the entire tweets to know about the theme because in some ways we can have tweets that are included with the main hashtag without having something in common. Instead, it is important to concentrate our attention in biggest centralities in order to retrieve the main concepts of the theme. However, communities with a percentage above 10% concentrate important concepts as they provide more semantic information about the tweets.

VII. CONCLUSIONS

Twitter has become a tool with a great social impact transforming the culture and creating a new medium for unfiltered and decentralized participation. McLuhan suggests that the message is the medium and to this extent there is a factor of cultural transformation that each medium causes when it gets implemented. Such is the case of Twitter, with 231.7 million active users around the world, is a medium that develops its own language as it transforms the culture [35]. In this way, we find an interest in applying linguistic analysis on Twitter in order to find grammatical structures and study the use of the hashtag as a means of prediction, grouping and understanding of the "*meaning*" of a certain movement.

This article has registered a methodology, based on the techniques of Social Network Analysis (SNA) to characterize the form in which evolves a topic on Twitter. The methodology allows the identification of groups in which the interaction between the words are particularly intense, allowing to categorize subgroups or communities in certain subjects and to

extract semantic from the network as we study the conformation of clusters.

Our study shows the importance of analyzing subgroups that have more than 10% of words, over a period of time, in order to extract semantic information about the tweets as it provides a better disambiguation of the corpus. In this way, future work should analyze the selection of tweets from a corpus based on centrality and subgroups in order to eliminate tweets that don't have any relation with the movement studied. The composition of corpus with big data should provide mechanisms to enhance the probability of obtaining relevant information.

Future work should consider the role of users (from individual to group behavior) that is relevant for social dynamics. In this way, the study of user groups it's crucial to analyze tweets according to social interactions.

Despite the big quantity of tweets that are written every day they are rarely self-content. So, it is important to find a way to contextualize tweets by providing automatically information about the tweet. This requires the use of crowdsourcing to allow users to classify tweets but also a combination of multiple types of processing from information retrieval to multi-document summarization including entity linking.

REFERENCES

[1] Tumasjan, A., Sprenger, T., et al. (2010). "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," *International AAAI Conference on Weblogs and Social Media*, Washington DC, George Washington University, pp. 178-185.

[2] Honeycutt, C., and Herring, S. C. (2009). "Beyond microblogging: Conversation and collaboration via Twitter," *42nd Hawaii International Conference on System Sciences*, 1-10, Hawaii.

[3] Cunha, E., Magno, G., Comarela, G., Almeida, V., Gonçalves, M. A., & Benevenuto, F. (2011). "Analyzing the dynamic evolution of hashtags on twitter: a language-based approach," *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. pp. 58-65.

[4] Cunha, E., Magno, G., Almeida, V., Gonçalves, M. A., & Benevenuto, F. (2012). "A gender based study of tagging behavior in twitter," *Proceedings of the 23rd ACM conference on Hypertext and social media*. pp. 323-324. ACM.

[5] Romero, D., Meeder, B., and Kleinberg, J. (2011). "Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter," *International World Wide Web Conference (WWW 2011)*. Hyderabad, India.

[6] Chew C., and Eysenbach G. (2010). "Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak," *PLoS ONE* 5(11): e14118. doi: 10.1371/journal.pone.*0014118*

[7] Rodrigues B., G. A., Silva, I. S., Zaki, M., Meira Jr, W., Prates, R. O., & Veloso, A. (2012). "Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment," *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts*. pp. 2621-2626.

[8] Zappavigna, M. (2011). "Ambient affiliation: A linguistic perspective on Twitter," *New media & society*, 13(5), 788-806.

[9] Bruns A. and Burgess, J. E. (2011). "The use of Twitter hashtags in the formation of ad hoc publics," *6th European Consortium for Political Research General Conference*, University of Iceland, Reykjavik.

[10] Hyland, K, & Tse, P (2004). "Metadiscourse in academic writing," *Applied Linguistics*, 25(2), 156–177.

[11] Crismore, A, Markannen, R, Steffensen, M (1993). "Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students," *Writ Commun*, 10(1), pp. 39–71.

[12] Russell, A. (2011). "The Arab Spring Extra-National Information Flows, Social Media and the 2011 Egyptian Uprising," *International Journal of Communication*, 5, 10.

[13] Poell, T., de Kloet, J., & Zeng, G. (2013). "Will the real Weibo please stand up? Chinese online contention and actor-network theory," *Chinese Journal of Communication*, pp. 1-18.

[14] Gitlin, T. (1980). "*The whole world is watching: mass media in the making & unmaking of the new left."* University of California Press.

[15] Georgakopoulou, A. (2013). "Narrative analysis and computer-mediated communication," *Pragmatics of Computer-mediated Communication*. Eds. S. Herring, D. Stein y T. Virtanen. Berlín: Mouton, pp. 695-715.

[16] Euler L. (1741). "Solutio Problematis ad Geometriam Situs Pertinentis," *Commentarii Academiae Scientiarum Imperialis Petropololitanae*, pp. 128-140.

[17] Sowa J. (1984). "*Conceptual structure*." Addison-Wesley Pub., Reading, MA.

[18] Freeman, L.C. (1979). "Centrality in social networks: conceptual clarifications," *Social Network*, 1(3), pp. 215-239.

[19] Herrero, R. (2000). "La terminología del análisis de redes: problemas de definición y de traducción," *Política y sociedad*, (33), pp. 199-206.

[20] Hotho, A., Nürnberger, A., and Paaß, G. (2005). "A brief survey of text mining," *Ldv Forum*. Vol. 20, No. 1, pp. 19-62.

[21] Wijaya, V., Erwin, A., Galinium, M., & Muliady, W. (2013). "Automatic mood classification of Indonesian tweets using linguistic approach," *2013 International Conference on Information Technology and Electrical Engineering (ICITEE)*. pp. 41-46. IEEE.

[22] Martínez, V., & González, V. M. (2013). "Sentiment characterization of an urban environment via Twitter," *Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction.* pp. 394-397. Springer International Publishing.

[23] Bertrand, K. Z., Bialik, M., Virdee, K., Gros, A., & Bar-Yam, Y. (2013). "Sentiment in New York City: a high resolution spatial and temporal view*," arXiv:1308.5010* (August 20, 2013).

[24] Lampos, V., Lansdall-Welfare, T., Araya, R., & Cristianini, N. (2013). "Analysing mood patterns in the United Kingdom through Twitter content," *arXiv preprint arXiv:1304.5507*.

[25] Chung, J. E., & Mustafaraj, E. (2011). "Can collective sentiment expressed on Twitter predict political elections?," *W. Burgard & D. Roth (Eds.), Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2011).* pp. 1768-1769. Menlo Park, CA: AAAI Press.

[26] Dodds, P. S., & Danforth, C. M. (2010). "Measuring the happiness of large-scale written expression: Songs, blogs, and presidents," *Journal of Happiness Studies*, 11(4), pp. 441-456.

[27] Gruzd, A., Doiron, S., & Mai, P. (2011). "Is happiness contagious online? A case of Twitter and the 2010 Winter Olympics," *Proceedings of the 44th Hawaii International Conference on System Sciences.* Washington, DC: IEEE Computer Society.

[28] Kramer, A. D. I. (2010). "An unobtrusive behavioral model of "Gross National Happiness"," *Proceedings of CHI 2010*. pp. 287-290. New York: ACM Press.

[29] Abascal-Mena, R., López-Ornelas, E., & Zepeda-Hernández, J. S. (2013). "User generated content: an analysis of user behavior by mining political tweets," *Online Communities and Social Computing*. pp. 3-12. Springer Berlin Heidelberg.

[30] Bastian M., Heymann S., Jacomy M. (2009). "Gephi: an open source software for exploring and manipulating networks*," International AAAI Conference on Weblogs and Social Media.*

[31] Heverin, T., & Zach, L. (2012). "Use of microblogging for collective sense making during violent crises: a study of three campus shootings," *Journal of the American Society for Information Science and Technology*, 63(1), 34-47.

[32] Bastos M. T., Puschmann C., Travitzki R. (2012). "Tweeting political dissent: retweets as pamphlets in #FreeIran, #FreeVenezuela, #Jan25, #SpanishRevolution and #OccupyWallSt," *Internet, Politics, Policy 2012: Big Data, Big Challenges? Oxford. Panel 6A: The Arab Spring and Political Dissent*. Oxford: Oxford Internet Institute, 2012. v. 2. pp. 1-20.

[33] Kwak, H., Lee, C., Park, H., & Moon, S. (2010). "What is Twitter, a social network or a news media?," *Proceedings of the 19th international conference on World Wide Web*. pp. 591-600. ACM.

[34] Arquilla J., Ronfeldt D. (2003) "*Networks and netwars: the future of terror, crime, and militancy,"* Santa Monica, Calif.: RAND, 2003, p. 380.

[35] Orihuela J. L. (2006). "La revolución de los blogs. Cuando las bitácoras se convirtieron en el medio de comunicación de la gente," *La Esfera de los Libros*, Madrid, 20fbb06, pp. 283. ISBN: 84-9734-498-7.

APPENDIX

Detailed presentation of the percentage of words obtained by each community per day. For example, for December 11 we have 20 communities where from the total number of concepts extracted for this day we got that the smallest community has 0.27% of concepts while the biggest has 15.85%.

Dec 6	**Dec 10**	**Dec 11**	**Dec 12**	**Dec 13**	**Dec16**
0.22	0.6	0.27	0.45	0.34	1.34
0.44	0.6	0.82	0.45	0.68	3.36
0.66	2.99	0.82	0.45	1.71	4.03
1.31	2.99	1.09	0.45	2.05	4.7
2.62	2.99	1.09	0.9	2.4	4.7
4.15	3.59	1.64	0.9	3.08	6.71
4.15	4.19	2.19	1.8	4.79	8.05
4.8	4.79	2.46	2.25	4.79	8.05
5.02	7.78	3.28	2.25	5.48	10.07
5.02	9.58	3.83	3.15	5.48	11.41
5.24	9.58	4.1	5.41	5.48	12.75
5.46	10.18	4.64	6.76	6.51	**24.83**
6.11	**11.98**	6.28	6.76	6.51	
8.52	13.77	6.83	7.21	7.19	
9.39	14.37	7.38	7.66	7.19	
9.61		8.2	7.66	8.9	
11.79		8.74	9.46	10.27	
15.5		8.74	11.26	**17.12**	
		11.75	**11.71**		
		15.85	13.06		