ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 12, December 2017

Social Network Structure – Twitter

[1] Abirami.M, [2] Santhiya.M, [3] Sumithra
[1][2] III Year
[1][2][3] B.Sc Computer Science, Holy Cross Home Science College,Thoothukudi,Tamil Nadu, India.

***Abstract:*-- As user interact with social media spaces, like twitter they form connections that emerge into complex social network structures. This article proposes a conceptual and practical model for the classification of topical Twitter networks based on their network-level structures. These connections are indicators of content sharing, and network structures reflect patterns of information flow. As current literature focuses on the classification of users to key positions, this study utilizes the overall network structures In order to classify Twitter conversation based on their patterns of information flow. Four network level metrics which have been established as indicators of information flow characteristics density, modularity, centralization and the fraction of isolated users-are utilized in a three step classification model. This process led us to suggest six structures of information flow, divided, inified , fragmented ,clustered in and out hub and spoke networks. We demonstrate the value of these network structure by segmenting 60 Twitter topical social media network dataset's into these six distinct patterns of collective connections. We discuss conceptual and practical implications of each structure.**

Key words: - Social Network, Hierarchical Classification.

INTRODUCTION

An increasing number of people is progressively approaching to the social networking sites, which become more and more popular and complex: within their context many and various communities are originated by users with common interests or with similar ways to feel part of the community. The kinds of analysis as well as information that can be extracted from the social networking sites are varied and increasingly appealing both to the world of marketing and to the social or political one. The classical approach to Social Network Analysis allows to study the topology of a network through the connections that develop within it, giving rise to a hierarchy of communities within the main topic. Furthermore, certain types of social networks, like Twitter, allow to track relationships also in those cases in which knowledge is not mutual: simply a node is a follower of another node. The number of followers defines in part the popularity of a node within the network, but it is not able to point out if this popularity is positive or negative.

On the other hand, the explosion of data on the Web has made the research in automatic cataloging of texts increasingly interesting, as well as the extraction of information or meta-information and the Sentiment Analysis of a review, an emotion, a tweet. Moreover, in this area the explosion of microblogging, and the use of a simple "like" or a retweet as a form of acceptance or sounding board for information as well as the dynamism and the speed with which everyone reads and writes content make the analysis of these opinions hard, if you use the methods of text mining, while they introduce, or amplify, new issues and problems for sentiment analysis (such as citations, irony, role of emoticons) that are difficult to deal with regardless the context in which they are written.

This paper presents a combined approach between Social Network and Sentiment Analysis. In particular we have tried to introduce some kind of information about sentiments on the graphs showing the results of the Social Network Analysis (SNA): in this way we hope to highlight other potential correlations among the nodes of net under examination. The idea behind it is that, on the one hand, the network topology and the selected topics of the network can contextualize and then, in part, unmask some incorrect results of the Sentiment Analysis (SA), and, the other hand, the polarity of the feeling on the network can highlight the role of semantic connections, as a possible foundation for the organization and the hierarchy of the communities highlighted by the Social Network Analysis.

In the following, after a brief description of the background, the system architecture will be showed, together with the choices which we made, then some results obtained from the initial evaluation of the system will be discussed.

2. BACKGROUND

SNSs are a collection of web-based services that allow users to build a profile within the system and define a list of other users with whom they have some kind of connection [7]. The architecture of social networking platforms is very differentiated. While the most popular platforms are built as essentially centralized systems, other platforms have a distributed architecture. The decentralized systems try to address some of the risks associated with online social networking, which are often perceived as quite serious by

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 12, December 2017

many users and have already led to serious incidents [6]. SNA has the objective to model social structures with different properties, starting from the mathematical theory of graphs and the use of matrix algebra, and is often augmented though computer-based simulations. SA is a branch of Opinion Mining, that aims to listen and process the data that users post on social media. Generally SA classifies web comments into positive, neutral, and negative categories. To make these systems more intelligent ad flexible, a deeper analysis of affective knowledge could be incorporated. In some case an ontology driven approach is used.

In this research work, we built a system for social network and sentiment analysis, which can operate on Twitter data, one of the most popular social networks. The analysis of large amount of data is an exciting challenge for researchers, but it is also crucial for all those who work at different levels in the current information society: Twitter has been the subject of attention from researchers as early as 2009.

3. SYSTEM ARCHITECTURE

In this paragraph we describe our system for social network and sentiment analysis, which can operate on Twitter data.
Twitter is a platform which may contain opinions, thoughts, facts, references to images and other media and, recently, stream video filmed live and put online by users. So it is more than just a SNCs in which a user displays and increases their social relationships, it is a real communication channel in which a user can choose its topics and its node of reference according to his interests and culture.

A study of the network topology and the number of interconnections of a node are able to highlight the communities in the network and also in part to how the information is propagated, but they are not able to say anything about the degree of agreement and cohesion of members of a community. To solve this task you need to carry out an investigation into the semantic content of the messages.

Compared to the problems of classic data mining, sentiment analysis shows many difficulties in terms of effectiveness. This is mainly due to the subtle distinction that exists between positive and negative sentiment or between neutral and positive one. Let us suppose for example a sentence containing irony or sarcasm, where the interpretation of the meaning is strictly subjective. In this case, two human beings may be in disagreement about the real feeling that it expresses. Furthermore, not always the opinions are expressed through the use of opinion words, in many cases the special language constructs (such as the figures of speech) come into play.
Difficulties also are due to the use of non-formal expressions and slangs that do not belong to the vocabulary of a language. These terms are often used in an intensive way to express a particular opinion or a certain mood.

Additional problems are due to the domain of the subject: in particular we note that the feelings that are expressed by a word are often dependent on the topic. We look at this sentence as an example: “It's quiet!”. It shall render a positive opinion if we are talking of a car engine, but it reveals a disapproval if the matter of discussion is a phone.

As a microblogging service, Twitter is used to publish short messages counting a maximum of 140 characters (tweets). This characteristic if one side it may seem easier because it forces people to take a position, on the other side the few words not allow the user to repeat concepts or emotions: he rather uses slangs shared by the community, emoticons and punctuation.

Besides the ease of retweet increases the difficulty in perceiving what is the real feeling of the user who runs it and the intense use of citations can also distort the sentiment enclosed in the tweet.

However, by combining the information of SA with those of the SNA we can hope to disambiguate some actual cases and the opportunity to know the slang of the channel under examination can improve the efficiency of machine learning algorithms for the SA. Some recent studies about American candidates are important for understanding how public sentiment is shaped and its polarization. In geo-spatial information related to tweets is used for estimating happiness in the Italian cities. Being Twitter a microblogging service, the techniques used generally in SA and Text Classification must be adapted to the famous 140-character tweet and this opens the way for new issues.

Another quite important problem to work on Twitter data is how to automatically collect a corpus for SA and, in general, Opinion Mining purposes: example of how to perform this task is in, for example.

3.1 Social Network Analysis: data selection

As a social networking platform, Twitter is structured as a directed graph, in which each user can choose to follow a number of other users (followees), and can be similarly followed by other users (followers). Thus, the “follow” relationship is asymmetrical, it does not require mandatory

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 12, December 2017

acknowledgement, and it is essentially used to receive all public messages published by any follower user.

3.2 Sentiment Analysis

As a communication medium, tweets have a quite peculiar nature. Some distinguishing features of communication on Twitter are related to technical aspects; those include length of text, tags, urls etc. Other features may be classified as idiomatic use of the medium, and create a sort of Twitter culture.

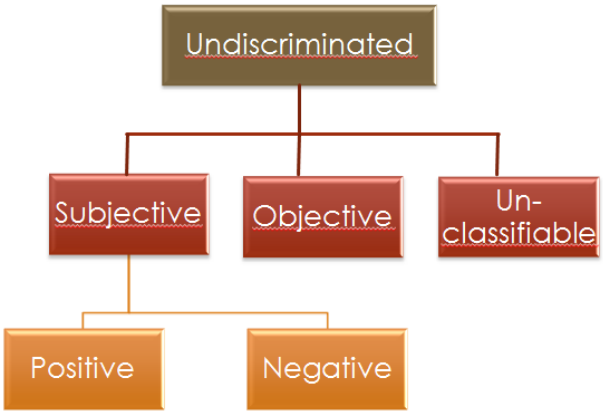
A first filter eliminates useless tokens such as: the "RT" sequence; the @ character and the whole following user name; the # symbol, but not the following topic name, which is kept in the message. The topic name is also removed, though, when it coincides with the name of the channel where tweets are collected from.

A second filter applies the language specific rules. It includes an orthographic correction of the message, which is used to remove unknown words (in the example: "icantbelievit") and other filtering processes for stemming and removal of stopwords.

Finally, another filter separates all punctuation symbols from the text, and organizes them as single-character words. Even if smiles sequences, repeated question and exclamation marks are kept as aggregates because they are important patterns for the classification.

The final result of the filtering process is a word vector, which is then submitted to a set of classifiers.

We use a set of classifier to identifying the following classes of messages: undiscriminated, objective, subjective, positive, negative. Moreover, there is a class in which the system put all the tweets that are too short to be classified. The system is organized as a simple hierarchy of agents, mimicking the hierarchy of sentiment classes. In fact, since objective messages have no polarity by definition, the classifier for positive and negative sentiments is only applied to subjective messages (see Fig. 3). One advantage of this framework for classifiers is the ease with which you can add classifiers trained to identify other emotions. In fact, hierarchical classification has been applied successfully in a number of studies, for information retrieval [23]. It has been proven effective especially in the case of classification over hierarchical taxonomies. Also in the case of sentiment analysis, a hierarchy of classes can be defined [12][5]. Accordingly, hierarchical classification has already been applied to sentiment analysis, too [23].



. Each classifier is based on Multinomial Naive Bayes algorithm, that one of the most popular methods used in SA. We have selected it because it seems to be the most suitable to generate and process large sets of features. In fact, instead of generating a training set by hand, we aimed at realizing an automated (or at least semiautomated) process for obtaining good training sets. In our methodology, the training sets are obtained through the automatic elaboration of some particular streams of tweets and comments, obtained directly from Twitter, without any manual classification. Thus, each training set may contain an important number of wrong data. Nevertheless, we show that they can be used to obtain useful results.

About the objectivity/subjectivity classifier, we adopted a similar strategy to [20]. In fact, to obtain objective content, we gathered messages generated from popular news agencies. In our tests, we used the following list: @ABC, @BBCNews, @BBCSport, @business, @BW, @cnnbrk, @CNNMoney, @fox32news, @latimes, @nytimes, @TIME. To obtain subjective content, instead, we gathered comments directed to the same list of users.

About the polarity classifier, we used different sources, thus generating training sets which do not overlap with those about objectivity/subjectivity. In fact, we used sources of mostly positive or negative messages, respectively. On the one hand, those sources should fit the particular setting of Twitter (short messages, idiomatic expressions, smiles, etc.). On the other hand, they should not be specific to a particular topic or context (sport, music, etc.). Thus, we dropped the idea of collecting messages about particular events, mostly generating either positive or negative sentiments. Instead, we collected messages, using generic yet polar terms as queried

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 12, December 2017

hashtags. In particular, we used the following channels to gather positive content: #adorable, #awesome, #beautiful, #beauty, #cool, #excellent, #great. We used the following channels to gather negative content: #angry, #awful, #bad, #corrupt, #pathetic, #sadness, #shame. Actually, such terms have been chosen quite empirically, taking into account the quality of training sets they generated. But they could be selected from WordNet-Affect [24], SentiWordNet [3], and other affective lexicons, in a more systematic way.

4. EXPERIMENTAL RESULTS

In this section, we will report the results of the classifiers and the analysis carried out on a couple of case studies. Using the methodology and the software which we described in Section 3, it is possible to obtain some generic training sets for the classifiers. This phase was carried out before selecting the final case studies. In our settings, they consist of:
• 86000 instances (polarity);
• 32000 instances (subjectivity).

These instances have been obtained by exploring more than 60 channels on the social network. In the generated models, the selected features are consistent with our expectations: the typical expressions of a certain feeling (such as smileys, or some words that express appreciation or disgust) show a higher probability of belonging to the class of that feeling, rather than to the class of the opposite sentiment.

The results obtained by the classifiers using cross-validation (with folds = 10) on the training sets showed an accuracy of:
• 77,45% (polarity classifier)
• 79,50% (subjectivity classifier)

These results show that the model of the classifiers contains effective features for the recognition of the sentiment of a message.

The case study which was considered in this work is the social network of the #SamSmith channel (the singer who won four awards at the Grammy Awards 2015). The choice of this channel is justified by the strong similarities found between the type of the published tweets and the instances used for training the classifiers. All data were downloaded between 2015-02-02 and 2015-02-10. The awarding of the Grammy took place on 2015-02-08. The social network (shown in Fig. 5) consists of a total of 5570 nodes (users) and 6886 arcs ("follows" relationships). Nodes are deployed according to the ForceAtlas2 algorithm [15], which turns structural proximities into visual proximities, thus highlighting communities. Looking at the figure, it is possible to notice that the network topology is consistent with the nature of the considered case. In fact, most of the channel consists of independent users (or small groups of users) that express their opinion about the artist; however, in the central part of the network there are some major communities. As shown in Fig. 5, the prevailing sentiment detected from the classifier is the negative one. Performing an analysis on a sample of tweets in the network, we noticed that many sentences are actually quotes of songs. These messages contain melancholic and sad phrases, and are therefore classified as negative. Considering that a quote is generally an appreciation for the artist, most users classified as negative are actually positive users. This is a typical example of a classic problem of misunderstanding of the SA: the system, while classifying correctly the tweet, misses the assessment of the feeling because it can not evaluate the tweet together with its context.

The case of Ukraine has been discussed quite largely in traditional media, too, for the supposed role of "trolls" operating on new media to influence the public opinion [25]. In fact, this may represent, as a modern reproposition, the quite classical case of opposing propaganda campaigns, this time carried on through social media. Also for this reason, we analyzed the social communities participating in the channel. We focused on the most active users, who contributed with at least 6 tweets during the whole week we considered (mid July 2014). In fact, among those it is more probable to find candidate opinion makers. The analyzed subnetwork represents around a tenth of the original network, and precisely consists of:
• 3261 nodes
• 84307 edges
We used the community detection algorithm provided with Gephi, at various resolution levels [18]. Quite interestingly, we were able to identify quite clearly two major communities. Additionally, some much smaller communities were found.

5. CONCLUSIONS

This study reports the initial results we obtained from the synthesis of Social Network Analysis and Sentiment Analysis. We experimented our approach on a couple of Twitter channels, as case studies. In particular, we considered the #SamSmith channel during the Grammy Awards in 2015, and the #Ukraine channel during the crisis of 2014. Apart from the particular results, a methodology and some guidelines for the automatic classification of Twitter content have been discussed.

The implemented software allows: (i) to get a training set for the classifiers that deal with Sentiment Analysis, and (ii) to

make a thorough study of the network topology. The study of the global sentiment within the network has highlighted the typical problems of Sentiment Analysis (irony, sarcasm, lack of information, etc.). Additionally, some peculiar problems of the considered channel were also detected (such as the quotes of songs). Also, the analysis of biased channels, may pose additional difficulties.

The performances obtained by the classifiers during tests conducted on the training set and the analysis of the case studies have shown good and promising results.

6. REFERENCES

1.A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. “Sentiment analysis of Twitter data”, Procs of the Workshop on Languages in Social Media (LSM '11), Association for Computational Linguistics, USA, pp. 30-38, 2011.

2.L. Allisio, V. Mussa, C. Bosco, V. Patti, and G. Ruffo, “Felicittà: Visualizing and Estimating Happiness in Italian Cities from Geotagged Tweets,” Proc. of the 1st Int. Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013), Turin, Italy, 2013.

3.A. E. S. Baccianella and F. Sebastiani, “Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining,” Proc. of the 7th Conf. on Inter. Language Resources and Evaluation (LREC’10), ELRA, 2010.

4.F. Bergenti, A. Poggi, and M. Tomaiuolo, “An Actor Based Software Framework for Scalable Applications,” Lecture Notes in Computer Science, 8729, pp. 26-35. 7th International Conference on Internet and Distributed Computing Systems (IDCS), 2014.

5.M. Baldoni, C. Baroglio, V. Patti, and P. Rena, “From tags to emotions: Ontology-driven sentiment analysis in the social semantic web,” Intelligenza Artificiale, vol. 6(1), pp. 41-54, 2012.

6.E. Franchi, A. Poggi, and M. Tomaiuolo, “Information and Password Attacks on Social Networks: An Argument for Cryptography,” Journal of Information Technology Research (JITR), 8(1), 25-42, 2015. doi:10.4018/JITR.2015010103

7.D. Boyd, N. Ellison, “Social Network Sites: Definition, History and Scholarship,” Journal of Computed-Mediated Communication, vol. 13 (1), pp. 210-230, 2008.

8.K. Ca, S. Spangler. Y. Chen; L. Zhang, "Leveraging Sentiment Analysis for Topic Detection," Web Intelligence and Intelligent Agent Technology (WI-IAT '08), IEEE/WIC/ACM Int. Conf. on, vol.1, pp. 265-271, 2008.

9.E. Cambria, B. Schuller, Y. Xia, C. Havasi , "New Avenues in Opinion Mining and Sentiment Analysis", IEEE Intelligent Systems, vol.28, no. 2, 2013.