Characterization of sentiment groups on Twitter

Nistor Grozavu & Nicoleta Rogovschi
LIPN UMR CNRS 7030, Paris 13 University, France
nistor@lipn.univ-paris13.fr
LIPADE, Paris Descartes University, France
nicoleta.rogovschi@parisdescartes.fr

***Abstract* — Opinion Mining is the field of computational study of people's emotional behavior expressed in text. The purpose of this article is to introduce a new framework for characterization of the groups of emotions extracted from tweet data. In contrast to supervised learning, the problem of clustering characterization in the context of opinion mining based on unsupervised learning is challenging, because label information is not available. The proposed framework uses topological unsupervised learning and hierarchical clustering, each cluster being associated to a prototype and a weight vector, reflecting the relevance of the data belonging to each cluster. The proposed framework requires simple computational techniques and is based on the double local weighting self-organizing map (dlw-SOM) model and Hierarchical Clustering.**

The proposed framework has been used on a real dataset issued from the tweets collected during the 2012 French election campaign.

***Index Terms* — emotions mining, clustering, twitter, topological learning, feature weighting**

I. INTRODUCTION

Opinion Mining is a recent research field in science that combines informational retrieval and computational linguistics. This field is an emerging problem in data mining and only some work on this subject can be found in the literature, especially using unsupervised machine learning techniques.

In recent years, there is a growing interest in sharing personal opinions on the Web, such as product reviews, photos, videos, economic analysis, political polls, etc. This information can be found in discussion forums, tweets, social networks, etc. These opinions cannot only help independent users make decisions, but also obtain valuable feedbacks. The opinion mining research field, including sentiment classification, opinion extraction, opinion question answering, and opinion summarization, etc. are receiving growing attention [1; 2; 3; 4; 5].

Opinion retrieval from text data is very different to classical informational retrieval approaches. Typical sources are blogs that generally reflect personal opinions, forums that present group opinions and tweets data where the messages are more shortly represented and the analysis became more difficult. Although web retrieval pays more attention to precision, opinion retrieval attaches extra importance to recall, since further sentiment mining relies heavily on the coverage of the opinion collection [4; 6].

Finally, the greatest challenge for opinion retrieval approaches lies in the difficulty in representing the user's information need and to characterize the opinions group in an automatic way by detecting the relevant features.

Sentiments are central to almost all human activities because they are key influencers of our behaviors. Whenever we need to make a decision, we want to know others opinions. In the real world, businesses and organizations always want to know more about the public opinions about their products and services in order to better organize their offers. The opinions are also important for the individual consumers what want to know the opinions of other users about a product before purchasing it, or about a discussion before to make a conclusion. In a political election, the individuals can be also interested in the others opinions about political candidates before making a voting decision [1; 2 ; 3].

With the explosive growth of social media (e.g., reviews, forum discussions, blogs, micro-blogs, Twitter, comments, and postings in social network sites) on the Web, individuals and organizations are increasingly using the content in these media for decision-making.

In this paper, we focus on sentiments retrieval, whose goal is to find a set of tweets containing not only the similar query keyword(s) but also the relevant emotions and to make an automatically characterization of the opinions' groups (clusters).

One of the challenges in this case is the representation of information needs for effective opinion retrieval.

In recent years, we have witnessed that opinionated postings in social media have helped reshape businesses, and sway public sentiments and emotions, which have profoundly impacted on the social and political systems. Such postings have also mobilized masses for political changes such as those happened in some Arab countries in 2011. It has thus become a necessity to collect and study opinions on the Web [1; 2; 6].

In this work, we are interested in methods, which aim at automatically finding attitudes or opinions about specific targets, in our case the opinions about the candidates in 2012 French elections.

The rest of the paper is organized as follows: the proposed framework for the opinion mining is presented in Section II. We introduce the weighted topological learning in section II.B after the Preprocessing step presented in section II.A. In Sections III, we present the validation of the proposed approach on tweets data sets and finally the paper ends with a conclusion and some future works for the proposed framework.

II. OPINION MINING

Data Clustering is the main task of knowledge discovery in databases [14; 15]. It aims to group a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

The approaches allowing the extraction of emotions from the text can be categorized into two main groups: lexicon-based and classification-based.

The lexicon-based approaches uses a manually or automatically built list of subjective words, such as `good' and `like', and assumes that the presence of these words in a document (tweet) is the evidence of document opinionatedness. A term's opinion score can be used in different ways to assign an opinion score to the whole document [1; 3; 5].

The classification-based approaches imply the use of the word occurrence and sometimes-linguistic features and build a classifier based on positive (opinionated) and negative (non-opinionated) documents using Machine Learning techniques.

Nevertheless, most of the early research in this area ignores the problem of retrieving documents that are related to the topic of the user's interest.

For this work, we propose to use the linguistic knowledge and the topological clustering in order to obtain clusters of opinions and to automatically characterize the opinions.

The figure 1 shows the proposed framework.

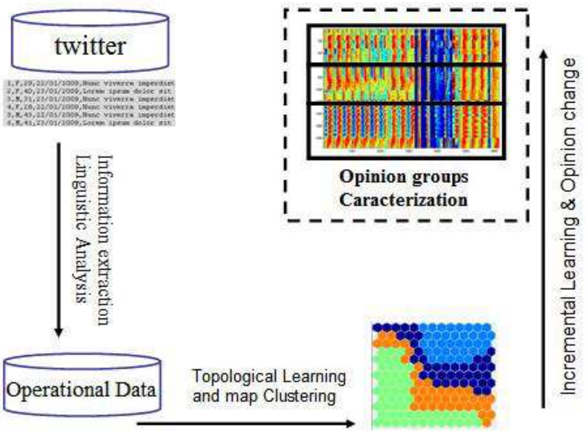

FIGURE1. THE PROPOSED FRAMEWORK

Also, the proposed approach can be used in incremental way as the topological map can be updated using new data (tweets) after the learning process.

In the next sections we describe the three steps used in the proposed framework: preprocessing, topological learning and opinion clustering and characterization. Note, that all these steps are linked and cannot be used separately for this problem.

A. Preprocessing

For the preprocessing step, we, firstly start by annotating the tweets using a morphosyntactic tag that allows to assign to each term of a tweet a part of speech (POS) tag [5, 6]. Then, the principle of Bag of Words is used in order to create a bag of words from the tweets by extracting the words (terms) from each tweet (document).

And, the last part of the preprocessing step is the use of the TF-IDF.

The TF-IDF weight (term frequency-inverse document frequency) is a weight often used in text mining [6]. This weight criterion is a statistical measure used to evaluate the importance of a term from a document in a corpus. The importance increases proportionally to the number of times a term appears in the document but is offset by the frequency of the word in the respective collection.

$$tf_{i,f} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where n_{ij} represents the number of occurrences of the term t_i in the document d_j.

The Inverse Document Frequency (IDF) is a measure that computes the importance of the term t_i in the respective collection (corpus) which is obtained by computing the logarithm of the inverse of the proportion of documents in the corresponding collection. The IDF is defined as follows:

$$IDF_i = \log_2 \frac{|D|}{|d_j : t_i \in d_j|}$$

where |D| is the total number of documents presented in the corpus, and d_j : t_i x d_j represents the documents containing the term t_i.

And, finally, the TF-IDF weight of a term t_i is the product of TF and IDF:

$$TF-IDF_{i,j} = TF_{i,j} \times IDF_i$$

B. Topological clustering

Data mining, or knowledge discovery in databases (KDD), an evolving area in information technology, has received much interest in recent studies. The aim of data mining is to extract knowledge from data.

The data size can be measured in two dimensions, the size of features and the size of observations. Both dimensions can take very high values, which can cause problems during the exploration and analysis of the dataset. Models and tools are therefore required to process data for an improved understanding [14, 15].

Indeed, datasets with a large dimension (size of features) display small differences between the most similar and the least similar data. In such cases it is thus very difficult for a learning algorithm to detect similarity variables that define the clusters.

Topological learning is a recent direction in Machine Learning, which aims to develop methods grounded on statistics to recover the topological invariants from the observed data points. Most of the existed topological learning approaches are based on graph theory or graph-based clustering methods.

The topological learning is one of the most known technique, which allow clustering, and visualization simultaneously. At the end of the topographic learning, the "similar" data will be collect in clusters, which correspond to the sets of similar observations. These clusters can be represented by more concise information than the brutal listing of their patterns, such as their gravity center or different statistical moments. As expected, this information is easier to manipulate than the original data points. The neural networks based techniques are the most adapted to topological learning as these approaches represent already a network (graph).

The models that interest us in this paper are those that could make at the same time the dimensionality reduction and clustering, i.e. using Self-Organizing Maps (SOM) for dimensionality reduction and Hierarchical Clustering to cluster the map [8, 9, 17].

SOM models are often used for visualization and unsupervised topological clustering. Its allow projection in small spaces that are generally two dimensional. We find several important research topics in cluster analysis and variable weighting in [11, 13].

In [10], the authors propose a probabilistic formalism for variable selection in unsupervised learning using Expectation-Maximization (EM).

Grozavu et al. proposed two local weighting unsupervised clustering algorithms (lwo-SOM and lwd-SOM) to categorize the unlabelled data and determine the best feature weights within each cluster [7, 13].

Similar techniques, based on k-means and weighting have been developed by other researchers [7, 9, 11, 12, 13].

C. Hierarchical Clustering

Clustering algorithms are generally classified as partition clustering and hierarchical clustering, based on the properties of the generated clusters [17].

Partition clustering divides data samples into a single partition, whereas a hierarchical clustering algorithm groups data with a sequence of nested partitions.

There are two types of the hierarchical clustering methods: agglomerative approach and divide approach. Divide hierarchical clustering method starts from a cluster, which contains all the data, and divide this cluster until obtaining the desired clusters. Contrarily, agglomerative hierarchical clustering method starts from *n* clusters (*n* data) and will merge these clusters until obtaining a cluster containing the whole data.

For this work we used the Hierarchical Clustering algorithm with Wards criterion to avoid merging empty cells. This procedure will allow us to avoid clustering ``cleaning" by eliminating the cells/clusters, which have no captured samples.

Agglomerative clustering starts with *n* clusters, each of which includes exactly one data point. A series of merge operations is then followed that eventually forces all objects into the same group. We can’t apply the HCA on the initial matrix because of the high computational time of this method.

III. Experimental results

The work presented in this paper were tested on a tweets dataset which were obtained as a part of the PoloP Project5 (Political Opinion Mining) which aims to cope with the analysis of the evolution of French political communities over Twitter during 2012 both in terms of relevant terms, opinions, behaviors. 2012 is particularly important for French political communities dues the two main elections:
Presidential and Legislative. The 6th of May was the final Presidential election where F. Hollande has been elected and the legislative elections were finished one month after [16].

The algorithm dlw-SOM allows us to obtain on the one hand, a two-dimensional projection data and on the other hand, a weighting of variables specific to each region of the map. [17] have proposed to segment a topological map by combining the k-means algorithm and Davies-Bouldin index which allows to automatically determine the size of the partition after segmentation. Indeed to use the k-means to cluster the map, we applied the Hierarchical Clustering introduced in section 2, which allows us to obtain stable results compared to k-means. We have applied this approach on referents and on the weights.

We obtained a topological map containing 169 cells, and applying the Hierarchical clustering on the map, we obtained 3 clusters. Note that initially we clustered the map from 2 to 10 clusters and we computed the Davies-Bouldin index [18] for each one in order to choose the best clustering result. The experiences show that the best clustering results is obtaining using 3 clusters (DB index = 0.41).

The DB index [18] is an internal index between two clusters and it's computing as follows: A similarity measure R_{ij} between the clusters C_i and C_j is defined based on a measure of dispersion of a cluster C_i, let s_i, and a dissimilarity measure between two clusters d_{ij}. The R_{ij} index is defined to satisfy the following conditions:

- $R_{ij} \geq 0$
- $R_{ij} = R_{ji}$
- if $s_i = 0$ and $s_j = 0$ then $R_{ij} = 0$
- if $s_j \geq s_k$ and $d_{ij} = d_{ik}$ then $R_{ij} \geq R_{ik}$
- if $s_j = s_k$ and $d_{ij} < d_{ik}$ then $R_{ij} \geq R_{ik}$

So, these conditions impose to R_{ij} to be a non-negative and symmetric. To satisfy the above-mentioned conditions, we have:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

Then, the DB index is defined as:

$$DB_{nc} = \frac{1}{n} \sum_{i=1}^{n_c} R_i$$

The DB_{nc} is the average similarity between each cluster c_i , i=1,...,n_c and its most similar one. So, we seek clustering results that minimize the DB, and thus have minimum possible similarity with the clusters. Some variants of this index were proposed in literature, which are based on Minimum Spanning Tree (MST), Relative Neighborhood Graph (RNG) and the Gabriel Graph (GC) concepts.

In the table 2 we show an example of the pertinent terms from the tweets of the both opinion clusters that characterize them. Note that the cluster situated in the middle of the map (the yellow cluster) contains similar opinions from other two clusters due to the neighborhood of the map, and it seems that tweets belonging to this cluster contains a neutral opinion.

TABLE I. PERTINENT TERMS FOR OPINION CLUSTERS

cluster	**1**	**2**
terms	lost the Triple A Holland will love Europe growing device change is now!	strong France Europe that defends Europe changing crisis

These results (the relevant terms for each opinion cluster translated from French) are relevant with the real opinion of peoples about this election campaign.

IV. Conclusion

In this study we proposed a framework for opinion mining based on topological unsupervised learning and hierarchical clustering.

The algorithm described in this paper provides topological clustering of the emotions issued from the tweets, each cluster being associated to a prototype and a weight vector, reflecting the relevance of the data belonging to each cluster.

The proposed framework has been used on a real dataset issued from the tweets collected during the 2012 French election campaign and the experimental results have shown promising performance.

Several perspectives can be considered for this work as: to propose an incremental approach in order to analyze the opinion behavior, to validate the framework on different datasets and to compare the method with the existed methods for the opinion mining.

AKNOWLEDGEMENT

This work was supported in part by the AIDEM PEPS HuMains 2013 CNRS Project, France

REFERENCES

[1] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Found. Trends Inf. Retr., vol. 2, no. 1-2, pp. 1–135, Jan. 2008.

[2] X. Huang and W. B. Croft, “A unified relevance model for opinion retrieval,” in Proceedings of the 18th ACM Conference on Information and Knowledge Management, ser. CIKM ’09. New York, NY, USA: ACM, 2009, pp. 947–956.

[3] B. Liu, Sentiment Analysis and Opinion Mining, ser. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[4] B. Liu, M. Hu, and J. Cheng, “Opinion observer: Analyzing and comparing opinions on the web,” in Proceedings of the 14th WWW ’05. New York, NY, USA: ACM, 2005, pp. 342–351.

[5] S. Gerani, M. J. Carman, and F. Crestani, “Proximity-based opinion retrieval,” in SIGIR, 2010, pp. 403–410.

[6] G. Salton, A. Wong, and C. S. Yang, “A vector space model for automatic indexing,” Commun. ACM, vol. 18, no. 11, pp. 613–620, Nov. 1975. [Online]. Available: http://doi.acm.org/10.1145/361219.361220

[7] N. Grozavu, Y. Bennani, and M. Lebbah, “From variable weighting to cluster characterization in topographic unsupervised learning,” in Proc. of IJCNN09, Inter. Joint Conf. on Neural Network, 2009.

[8] T. Kohonen, Self-organizing Maps. Springer, Berlin, 2001.

[9] M. Lebbah, N. Rogovschi, and Y. Bennani, “BeSOM : Bernoulli on Self Organizing Map,” in IJCNN ‘07, Orlando, Florida, 2007.

[10] J. Verbeek, N. Vlassis, and B. Krose, “Self-organizing mixture models,” Neurocomputing, vol. 63, pp. 99–123, 2005.

[11] H. Frigui and O. Nasraoui, “Unsupervised learning of prototypes and attribute weights,” Pattern Recognition 37(3), pp. 567–581, 2004.

[12] J. G. Dy and C. E. Brodley, “Feature Selection for Unsupervised Learning,” JMLR, vol. 5, pp. 845–889, 2004.

[13] N. Grozavu and Y. Bennani, “Simultaneous pattern and variable weighting during topological clustering,” in Proceedings of the 18th ICONIP’11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 570–579.

[14] P. Hansen and B. Jaumard, “Cluster analysis and mathematical programming,” Math. Program., vol. 79, pp. 191–215, 1997.

[15] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” ACM Computing Surveys, vol. 31, no. 3, pp. 264–323, 1999.

[16] F.Bouillot, P.Poncelet, M.Roche, D.Ienco, E.Bigdeli, and S.Matwin, “French presidential elections: What are the most efficient measures for tweets?” in Proceedings of the First Edition Workshop on Politics, Elections and Data, ser. PLEAD ’12. New York, NY, USA: ACM, 2012, pp. 23–30.

[17] J.Vesanto and E.Alhoniemi, “Clustering of the Self-Organizing Map,” Neural Networks, IEEE Transactions on, vol. 11, no. 3, pp. 586–600, 2000.

[18] D. Davies and D. Bouldin, “A cluster separation measure,” IEEE Trans. Pattern Anal. Machine Intell., vol. 1 (4), pp. 224–227, 1974.