IMPORTANCE OF THE NEUTRAL CATEGORY IN FUZZY CLUSTERING OF SENTIMENTS

Lawrence Nderu[1], Nicolas Jouandeau[2]and Herman Akdag[2]

[1]Jomo Kenyatta University of Agriculture and Technology, Kenya

[2]University of Paris 8-LIASD, France,

ABSTRACT

Social media is said to have an impact on the public discourse and communication in the society. It is increasingly being used in the political context. Social networks sites such as Facebook, Twitter and other microblogging services provide an opportunity for public to give opinions about some issues of interest. Twitter is an ideal platform for users to spread not only information in general but also political opinions, whereas Facebook provides the capability for direct dialogs. A lot of studies have shown that a need exists for stakeholders to collect, monitor, analyze, summarize and visualize these social media views. Some authors have tended to categorize these comments as either positive or negative ignoring the neutral category. In this paper, we demonstrate the importance of the neutral category in the clustering of sentiments from the social media. We then demonstrate the use of fuzzy clustering for this kind of task.

KEYWORDS

Sentiments Analysis, Fuzzy Clustering, Social Media, Neutral Category, Unsupervised Classification

1. Introduction

Clustering is a type of unsupervised learning that automatically forms clusters of similar things [1]. During cluster analysis the aim is to put similar things in one cluster and dissimilar things in a different cluster [2]. Several types of clustering have been developed and a lot of literature exists on clustering methods. In fuzzy clustering, data elements can belong to more than one cluster and associated with each of the data points are membership grades which indicate the degree to which a data point belongs to the different clusters [3].

Sentiments found within comments, feedback or critiques provide useful indicators for many different purposes. Sentiment analysis provides policy makers and politicians with an opportunity to estimate the public sentiments with respect to policies, public services or political issues [4]. With the massive growth of the social media a lot of outlets for publishing one’s opinion (s) exist. A number of authors have considered sentiments analysis as a two class question, i.e. positive or negative [5]. We assert the importance of treating neutrals as a fully qualified category.

The rest of the paper is organized as follows: first a discussion of related work on public sentiment analysis or opinion mining is presented followed by a discussion of clustering algorithms. We then present our model in this study and then conclusion and future work.

DOI : 10.5121/ijfls.2014.4201

2. Related Work

2.1 Sentiments Analysis

Sentiment analysis has attracted a lot of interest in the recent past. An increasing number of scientists are tackling the task of automated sentiments detection and classification. This has resulted in a computing domain called effective computing [4]. More and more users are posting about products and services they use, or expressing their political and religious views on issues. Such data can be efficiently used for marketing, making political decisions or social studies.

The computational treatment of opinion, sentiment and subjectivity has recently attracted a great deal of attention [6],[7]. This is in part due to the potential of application areas [8][7]. In sentiment analysis neutrality is handled in various ways, this depends on the technique that is being used [9]. A number of papers tend to ignore the neutral category under the assumption that neutral texts lie near the boundary of the binary classifier. It is also assumed that not much can be learnt from neutral texts comparing to the ones with clear positive or negative sentiment [9],[4]. Sentiment about a subject is the orientation (or polarity of the opinion on the subject that deviates from the neutral state [10]. The reasoning behind having the neutral sentiment as a category on its own is based on the fact that some opinions are facts. Thus, the statement “the song was well done” has a positive polarity. The statement “the song fails to impress” has a negative polarity, but the statement “the song is in the higher note” has a neutral polarity and indeed provides further information about the song. It is therefore not uncommon to have statements that are neutral.

2.2 Clustering

Clustering is a technique used in unsupervised learning. The aim is to cluster similar data points in one cluster and dissimilar points in a different group. Most clustering algorithms do not rely on assumptions common to conventional statistical methods, such as the underlying statistical distribution of data. This means that clustering algorithms are useful in situations where little prior knowledge about the distribution of the data is known [2]. Since clusters can formally be seen as subsets of data set, one possible classification of clustering methods can be according to whether the subsets are fuzzy or hard.

Hard clustering methods are based on classical set theory and require that an object either does or does not belong to a cluster. Hard clustering involves partitioning the data into specified number of mutually exclusive subsets. Fuzzy clustering methods, however allow the objects to belong to several clusters simultaneously, with different degrees of membership. In many situations, fuzzy clustering is more natural than hard clustering. Objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees 0 and 1 indicating their partial membership [2].

The potential of clustering algorithms to reveal the underlying structure in data can be exploited in a wide variety of applications [11]. Clustering techniques can be applied to data that is quantitative (numerical), qualitative (categorical) or a mixture of both. The data is typically observations of some physical process. Each observation consists of n measured variables, grouped into an n-dimensional column vector Zk = [z1k, z2k, …,znk]T , Zk □ □n . A set of N observation is denoted by Z = {Zk | k=1, 2, ..., N } and is represented by an n X N matrix. Various definitions of a cluster can be formulated, depending on the objective of clustering. The discussion on hard clustering is covered by [12][11], below we present a discussion of fuzzy clustering.

2.2.1 Fuzzy Clustering

The generalization of the hard partition to the fuzzy case follows directly by allowing μik to attain real values in [0, 1]. Conditions for a fuzzy partition matrix are given by[8].

$$\mu ik \in [0, 1], 1 \leq i \leq c, 1 \leq k \leq N$$

$$\sum_{i=1}^{c} \mu ik = 1, 1 \leq k \leq N$$

$$0 < \sum_{k=1}^{N} \mu ik < N, 1 \ \leq i \ \leq c$$

The ith row of the fuzzy partition matrix U contains values of the ith membership function of the fuzzy subset Ai of Z. The total membership of each zk in Z equals one. Considering the range of values from [0, 1]. A number of fuzzy clustering algorithms exist; below we discuss the fuzzy C-means clustering.

2.2.1.1 Fuzzy C-Means Clustering

Fuzzy C-means (also called C methods) is an example of fuzzy clustering which allows one point to belong to one or more clusters [13]. Fuzzy C-means was developed by Dunn and improved by Bezdek and is frequently used in many areas including pattern recognition [14].

2.3 Neutral class

Neutrality in sentiment analysis is handled differently, by different authors [15]. This mostly depends on the technique that is being used. In lexicon- based techniques then eutrality score of the words is taken into account in order to either detect neutral opinions or filter them out and enable algorithms to focus on words with positive and negative sentiment [16][19]. The inclusion of the neutral class in our study is based on the fact that in sentiment analysis we believe that not everything is positive or negative.

2.3.1. From Computing with Words to Computing with Numbers

Computing in its usual sense is centered on manipulation of numbers and symbols. In contrast, computing with words is a methodology in which the objects of computation are words and propositions drawn from a natural language [17]. Computing with words is inspired by the remarkable human capability to perform a wide variety of physical and mental tasks without any measurements and any computations [18]. The basic difference between perception and measurements is that measurements are crisp (hard), whereas perception is fuzzy [12]. Many aspects of different activities in the real world cannot be assessed in a quantitative form, but in a qualitative one [19]. Fuzzy linguistic modelling has been widely used and has provided very good results, it deals with qualitative aspects that are presented in qualitative terms by means of linguistic variables [20].

When sentiments are expressed in general as either positive or negative some important information is lost. Since sentiments are expressed using the natural language, we consider the interpretation of those sentiments using a set of seven terms (more could be used) and their semantics. Fig. 1, demonstrates the computing with words interpretation of numerical values [21]. For sentiments to be beneficial, it is important that we extract a lot of information from them. For example, consider the answer to the question “how are people responding to this ad campaign”. An answer that says negative could lead us into changing the whole ad campaign and yet maybe if we knew it is at a value of 0.67 (as shown in Figure 1), we could just carry out some improvements and the sentiments could change.

Classifying sentiments as either positive or negative, is very general and some information could be lost, adding the category neutral improves, but still does not capture all the information in the sentiments. Below we demonstrate fuzzy clustering of sentiments.

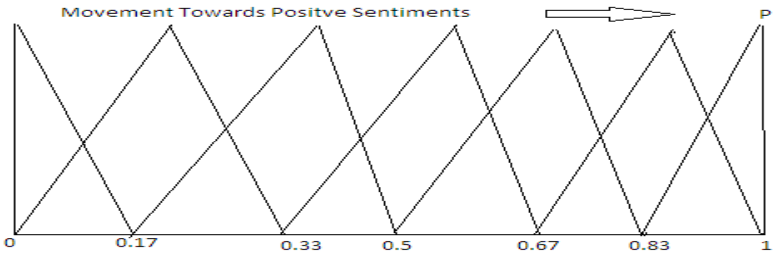

Fig.1. How positive is a sentiment

2.4 The Proposed Model

A total of 140 negative tweets on the International Criminal Court (ICC), in Kenya were collected. The aim is to propose a model that provides more information than just telling us that the sentiments were negative. The choice of the tweets was done by a human being. The Cohen's Kappa which is used to check for inter-rater reliability [22] was used. Generally, a Kappa value > 0.70 was considered to be a satisfactory agreement. Table 1, shows the results of the rating by 2 experts. A value of 0 shows a sentiment that could be considered neutral, while 1 is a sentiment that is highly negative. The two rates agreed on 87 tweets, and this are the ones that were tabulated.

Table 1.The results of the Analyzed Tweets

Numerical values	0	0.17	0.33	0.5	0.67	0.83	1
Count	0	5	3	8	18	24	29

A sentiment could have some parts that are positive, negative or even neutral. Our aim in this mode of sentiment analysis is to gain all the necessary information from the sentiment. In a 3 dimension plot this could be as shown in figure 2, where horizontal axis= positivity of the sentiment, width = negativity of the sentiment and vertical shows the neutrality or the sentiment.

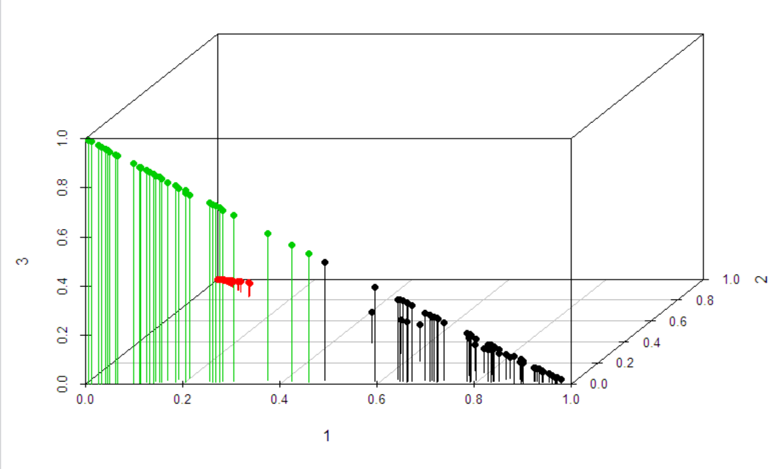

Fig.2. A 3-d plot of the sentiments values

3. Conclusion

Information found in the text data is either: Objective information (facts) or Subjective information (opinions). In this paper we have considered the problem as a three-class classification problem (positive, neutral or negative), with varying levels of positivity, neutrality or negativity rather than a two-class problem (positive or negative). The fuzzy clustering which consistsof optimal grouping of given data points together but in such a way that each data point can be shared by one or more cluster by belonging partially to another has been used. The idea that our model is pursuing here is that in some cases the sentiment holder could be persuaded to change his/her view if for example it is not “a very strong negative” sentiment.

In our future work, we want to implement a sentiment analysis system that is maintained automatically with the exact value of how negative, positive or neutral a sentiment is. Such a system could help opinion leaders for example to know the kind of individual who could be persuaded to change their sentiments over an issue.

References

[1] A. Jain, M. Murty, and P. Flynn, “Data clustering: a review,” ACM Comput. Surv., 1999.

[2] W. Wang and Y. Zhang, “On fuzzy cluster validity indices,” Fuzzy Sets Syst., vol. 158, no. 19, pp. 2095–2117, Oct. 2007.

[3] S. Krinidis and V. Chatzis, “A robust fuzzy local information C-Means clustering algorithm.,” IEEE Trans. Image Process., vol. 19, no. 5, pp. 1328–37, May 2010.

[4] F. Dzogang and M. Lesot, “Expressions of graduality for sentiments analysis—A survey,” Fuzzy Syst. (FUZZ …, 2010.

[5] R. Prabowo and M. Thelwall, “Sentiment analysis: A combined approach,” J. Informetr., vol. 3, no. 2, pp. 143–157, Apr. 2009.

[6] X. Ding, S. M. Street, B. Liu, and P. S. Yu, “A Holistic Lexicon-Based Approach to Opinion Mining,” 2008.

[7] F. Breu, S. Guggenbichler, and J. Wollmann, “No Title,” Vasa, 2008.

[8] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, “Syntactic clustering of the Web,” Comput. Networks ISDN Syst., vol. 29, no. 8–13, pp. 1157–1166, Sep. 1997.

[9] V. Vryniotis, “The importance of Neutral Class in Sentiment Analysis | DatumBox,” 2013. [Online]. Available: http://blog.datumbox.com/the-importance-of-neutral-class-in-sentiment-analysis/. [Accessed: 19-Feb-2014].

[10] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, “Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques,” Third IEEE Int. Conf. Data Min., pp. 427–434, 2003.

[11] Y. Sharon, J. Wright, and Y. Ma, “Minimum sum of distances estimator: robustness and stability,” Am. Control Conf. 2009. …, pp. 524–530, 2009.

[12] R. Zass, “A Unifying Approach to Hard and Probabilistic Clustering 2 . Clustering and Complete Positivity,” no. 2, 2007.

[13] S. Paper, “A FUZZY HIERARCHICAL CLUSTERING METHOD FOR CLUSTERING DOCUMENTS BASED ON DYNAMIC CLUSTER CENTERS,” vol. 30, no. 1, pp. 169–172, 2007.

[14] R. L. Cannon, J. V Dave, and J. C. Bezdek, “Efficient Implementation of the Fuzzy c-Means Clustering Algorithms.,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 8, no. 2, pp. 248–55, Feb. 1986.

[15] M. Koppel and J. Schler, “The Importance of Neutral Examples for Learning Sentiment,” 2006.

[16] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, “OpinionFinder : A system for subjectivity analysis,” no. October, pp. 34–35, 2005.

[17] L. a. Zadeh, “The concept of a linguistic variable and its application to approximate reasoning—I,” Inf. Sci. (Ny)., vol. 8, no. 3, pp. 199–249, Jan. 1975.

[18] G. Bordogna and G. Pasi, “A Fuzzy Linguistic Approach Generalizing Boolean Information Retrieval : A Model and Its Evaluation,” JASIS, vol. 44, no. 2, pp. 70–82, 1993.

[19] M. Decision-making, F. Herrera, and L. Martínez, “A Model Based on Linguistic 2-Tuples for Dealing with Multigranular Hierarchical Linguistic Contexts,” vol. 31, no. 2, pp. 227–234, 2001.

[20] L. Nderu, N. Jouandeau, and H. Akdag, “Towards Universal Rating of Online Multimedia Content,” 2013.

[21] F. Herrera, “A 2-tuple fuzzy linguistic representation model for computing with words - Fuzzy Systems, IEEE Transactions on,” vol. 8, no. 6, pp. 746–752, 2000.

[22] I. R. Application, I. Kappa, F. Yellow-bellied, F. Red-bellied, and R. Cooters, “Cohen ’ s Kappa,” pp. 1–3.