2nd International Conference on Recent Innovations in Science, Engineering and Management
JNU Convention Center, Jawaharlal Nehru University, New Delhi (ICRISEM-15)
22 November 2015 www.conferenceworld.in ISBN: 978-81-931039-9-9

SENTIMENT ANALYSIS TECHNIQUES: A REVIEW

Shivani Rana

M.Tech Student, Hindu College of Engineering, Sonipat, Haryana, (India)

ABSTRACT

This paper summarizes the study of different supervised and unsupervised learning techniques of sentiment analysis. The growth of social web contributes vast amount of user generated content such as customer reviews, comments and opinions. This user generated content can be about products, people, events, etc. This information is very useful for businesses, governments and individuals. While this content meant to be helpful analyzing this bulk of user generated content is difficult and time consuming. So there is a need to develop an intelligent system which automatically mine such huge content and classify them into positive, negative and neutral category. Sentiment analysis is the automated mining of attitudes, opinions, and emotions from text, speech, and database sources through Natural Language Processing (NLP). The objective of this paper is to discover the concept of Sentiment Analysis in the field of Natural Language Processing, and presents a comparative study of different techniques used in this field.

Keywords: Sentiment Analysis, TF*PDF algorithm, SVM, F-Measure, EFS algorithm.

I. INTRODUCTİON

Sentiment analysis is a type of natural language processing for tracking the mood of the public about a particular product or topic. Sentiment analysis, which is also called opinion mining, involves in building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets. Sentiment analysis can be useful in several ways. For example, in marketing it helps in judging the success of an ad campaign or new product launch, determine which versions of a product or service are popular and even identify which demographics like or dislike particular features.

There are several challenges in Sentiment analysis. The first is an opinion word that is considered to be positive in one situation may be considered negative in another situation. A second challenge is that people don't always express opinions in a same way. Most traditional text processing relies on the fact that small differences between two pieces of text don't change the meaning very much. In Sentiment analysis, however, "the picture was great" is very different from "the picture was not great". People can be contradictory in their statements. Most reviews will have both positive and negative comments, which is somewhat manageable by analyzing sentences one at a time. However, in the more informal medium like twitter or blogs, the more likely people are to combine different opinions in the same sentence which is easy for a human to understand, but more difficult for a computer to parse. Sometimes even other people have difficulty understanding what someone thought based on a short piece of text because it lacks context. For example, "That movie was as good as its last movie" is entirely dependent on what the person expressing the opinion thought of the previous model. The user's hunger is on for and dependence upon online advice and recommendations the data reveals is merely one reason behind the emerge of interest in new systems that deal directly with opinions as a first-class object. Sentiment

analysis concentrates on attitudes, whereas traditional text mining focuses on the analysis of facts. There are few main fields of research predominate in Sentiment analysis: sentiment classification, feature based Sentiment classification and opinion summarization. Sentiment classification deals with classifying entire documents according to the opinions towards certain objects. Feature-based Sentiment classification on the other hand considers the opinions on features of certain objects. Opinion summarization task is different from traditional text summarization because only the features of the product are mined on which the customers have expressed their opinions. Opinion summarization does not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the main points as in the classic text summarization. Languages that have been studied mostly are English and in Chinese. Presently, there are very few researches conducted on sentiment classification for other languages like Arabic, Italian and Thai.

II. TECHNIQUES

2.1 TF*PDF Algorithm

TF*PDF algorithm[1,2,3] is a supervised learning algorithm adapted in the ETTS which is useful in tracking the emerging topic in a particular information area of interest on the Web, by summarizing the change posted on it. TF*PDF algorithm is designed in a way that it would assign heavy term weight to these kind of terms and thus reveal the main topics since the web became widespread, the amount of electronically available information online, especially news archives, has proliferated and threatens to become overwhelming. It can be used in an information system that will extract main topics in a news archive on a weekly basis. By obtaining a weekly report, a user can know what the main news events were in the past week. In general, related research on subject identification is classified into two types.

First one is term weighting method to extract useful terms that is relevant to collected documents and modelled also. Second is TF-IDF mostly used for term weighting in Natural language processing and information extraction process [1].

Thus, in order to fulfill the objective to recognize the terms that explain the hot topics, TF*PDF is innovated to count the significance (weights) of the terms. Different from the conventional term weight counting algorithm TF*IDF, in TF*PDF algorithm, the weight of a term from a channel is linearly proportional to the term's within channel frequency, and exponentially proportional to the ratio of document containing the term in the channel. The total weight of a term will be the summation of term's weight from each channel as follows.

$$Wj = \sum_{c=1}^{c=D} |Fjc| \exp(njc/Nc) \quad (1)$$

where, Wj=Weight of term j; Fjc=Frequency of term j in channel c; njc=Number of document in channel c where term j occurs; Nc=Total number of document in channel c; k=Total number of terms in a channel; D=number of channels

There are three major compositions in TF*PDF algorithm. The first composition that contributes to the total weight of a term significantly is the "summation" of the term weight gained from each channel, provided that the term deems to explain the hot topic discussed generally in majority of the channels. In other words, the terms that deem to explain the main topic will be heavily weighted. Also, larger the number of channels, more accurate will

be this algorithm in recognizing the terms that explain the emerging topic. The second and third compositions are combined to give the weight of a term in a channel in many documents compare to the one occurs in just a few containing certain terms of significant weight, the results would be deviated from having terms that explain the hot topics in majority channels. In short, TF*PDF algorithm give significant weights to the terms that explain the common hot topic in majority channels.

2.2 SVM

Support Vector Machine model[4,5] a supervised learning approach usually performs well on various text categorization tasks. Derived from the vector-space model, it is a classical technique to weight each term through applying the *tf idf* formula, in which the component *tf* represents the occurrence frequency within the text. The *idf* (= log (*df*/*n*)) mainly corresponds to the logarithm of the inverse document frequency (denoted *df*), while *n* indicates the total number of texts.

As an alternative, normalize both components such that the only possible values would fall in [0 - 1]. For the *tf* part, we select the augmented *tf* weighting scheme defined as *atf* = 0.5 + 0.5 . (*tf* / *max tf*), where *max tf* corresponds to the maximal occurrence frequency for the underlying text and *nidf* is obtained by simply dividing the *idf* value by log(*n*). Based on this representation we use the freely available SVM *light* model[6,7] which determines the hyperplane that best separates the examples belonging to the two categories. In this case the best hyperplane refers to the one having the

largest separation (or margin) between the two classes (and of course together with a reduction for the number of incorrect classifications). This first version belongs to the linear classifier paradigm and we have also considered nonlinear kernel functions (polynomial, sigmoid). The use of non-linear kernel functions did not improve the quality of the classification, at least in our classification task.

2.3 F-Measure

In unsupervised technique, classification is done by comparing the features of a given text against sentiment lexicons whose sentiment values are determined prior to their use. Sentiment lexicon contains lists of words and expressions used to express people's subjective feelings and opinions. For example, start with positive and negative word lexicons, analyze the document for which sentiment need to find. Then if the document has more positive word lexicons, it is positive, otherwise it is negative. The lexicon based techniques to Sentiment analysis is unsupervised learning because it does not require prior training in order to classify the data.

The F-measure[8,9] is a unsupervised learning technique is not the F-score or F measure used in text classification or information retrieval for measuring the classification or retrieval effectiveness (or accuracy).F-measure explores the notion of implicitness of text and is a unitary measure of text's relative contextuality (implicitness), as opposed to its formality (explicitness). Contextuality and formality

can be captured by certain parts of speech. A lower score of F-measure[10,11] indicates contextuality, marked by greater relative use of pronouns, verbs, adverbs, and interjections; a higher score of F measure indicates formality, represented by greater use of nouns, adjectives, prepositions, and articles. F-measure is defined based on the frequency of the POS(Part of speech) usage in a text (*freq.x* below means the frequency of the part-of-speech *x*): $F = 0.5 * [(freq.noun + freq.adj + freq.prep + freq.art) – (freq.pron + freq.verb + freq.adv + freq.int) + 100]$.

2.4 EFS Algorithm

Few research techniques have indicated that the combination of both the machine learning and the lexicon based approaches improve sentiment classification performance[12,13]. The main advantage of their hybrid approach using a lexicon/learning symbiosis is to attain the best of both worlds-stability as well as readability from a carefully designed lexicon, and the high accuracy from a powerful supervised learning algorithm.EFS[14,15] takes the best of both worlds. It first uses a number of feature selection criteria to rank the features following the filter model. Upon ranking, the algorithm generates some candidate feature subsets which are used to find the final feature set based on classification accuracy using the wrapper model. Since our framework generates much fewer candidate feature subsets than the total number of features, using wrapper model with candidate feature sets is scalable. Also, since the algorithm generates candidate feature sets using multiple criteria and all feature classes jointly, it is able to capture most of those features which are discriminating. The algorithm takes as input, a set of n features $F = \{f1, \ldots, fn\}$, a set of t feature selection criteria $\Theta = \{\theta 1, \ldots, \theta t\}$, a set of t thresholds $T = \{\tau 1, \ldots, \tau t\}$ corresponding to the criteria in Θ, and a window w. τi is the base number of features to be selected for criterion θi. w is used to vary τi (thus the number of features) to be used by the wrapper approach.

III. COMPARISON AND ANALYSIS

From the study we have done over the above algorithms we found that supervised machine learning techniques have shown relatively better performance than the unsupervised lexicon based methods. However, the unsupervised methods is important too because supervised methods demand large amounts of labelled training data that are very expensive whereas acquisition of unlabelled data is easy. Most domains except movie reviews lack labelled training data in this case unsupervised methods are very useful for developing applications. Most of the researchers reported that Support Vector Machines (SVM) has high accuracy than other algorithms. The main limitation of supervised learning is that it generally requires large expert annotated training corpora to be created from scratch, specifically for the application at hand, and may fail when training data are insufficient. The main advantage of hybrid approach using a lexicon/learning combination is to attain the best of both worlds, high accuracy from a powerful supervised learning algorithm and stability from lexicon based approach.

IV. CONCLUSION

Application of sentiment analysis to mine the huge amount of unstructured data has become an important research problem. Now business organizations and academics are putting forward their efforts to find the best system for sentiment analysis. Although, some of the algorithms have been used in sentiment analysis gives good results, but still no technique can resolve all the challenges. Most of the researchers reported that Support Vector Machines (SVM) has high accuracy than other algorithms, but it also has limitations. More future work is needed on further improving the performance of the sentiment classification. There is a huge need in the industry for such applications because every company wants to know how consumers feel about their products and services and those of their competitors. Different types of techniques should be combined in order to overcome their individual drawbacks and benefit from each other's merits, and enhance the sentiment classification performance.

REFERENCES

[1]. K. Bun and M. Ishizuka."Topic extraction from news archive using TF*PDF algorithm" In Proceedings of Third International Conference on Web Information System Engineering.

[2] K.B. Khoo and M. Ishizuka: "Emerging Topic Tracking System" In: Proc. of Web Intelligent (WI 2001), LNAI 2198 (Springer), pp. 125-130, Maebashi, Japan. 2001 [2] K.B. Khoo and M. Ishizuka: "Information Area Tracking and Changes Summarizing in WWW" In: Proc. of WebNet 2001, International Conf. on WWW and Internet, pp. 680- 685, Orlando, Florida. 2001

[3] S. Dharanipragada, M. Franz, J.S. McCarley, K. Papineni, S. Roukos, T. Ward, W.-J. Zhu: "Statistical Models for Topic Segmentation", In: Proc. of SIGIR '00

[4] F. Sebastiani, "Machine learning in automatic text categorization," ACM Computing Survey, vol. 14(1), 2002, pp. 1-27.

[5] J. Savoy, "Lexical analysis of US political speeches," Journal of Quantitative Linguistics, vol. 17(2), 2010, pp. 123-141.

[6] B. Pang, and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2(1-2), 2008.

[7] Jacques Savoy, Olena Zubaryeva "Classification Based on Specific Vocabulary" published in 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology 978-0-7695-4513-4/11 2011 IEEE

[8] Agrawal, R. and Srikant, R. 1994. *Fast Algorithms for Mining Association Rules*. VLDB. pp. 487-499.

[9] Argamon, S., Koppel, M., J Fine, AR Shimoni. 2003.*Gender, genre, and writing style in formal written texts*. Text-Interdisciplinary Journal, 2003.

[10] Blum, A. and Langley, P. 1997. *Selection of relevant features and examples in machine learning*. Artificial Intelligence, 97(1-2):245-271.

[11] Mukhrjee, A. and B. Liu, 2010. Improving gender classification of weblog authors. Proceedings of Conference on Empirical Methods in Natural Language Processing, (EMNLP' 10), 10RDF Primer. W3C Recommendation. http://www.w3.org/TR/rdf-primer, 2004.

[12] Garganté, R. A., Marchiori, T. E., and Kowalczyk, S.R. W., 2007. *A Genetic Algorithm to Ensemble Feature Selection*. Masters Thesis. Vrije Universiteit, Amsterdam.

[13] Herring, S. C., & Paolillo, J. C. 2006. *Gender and genre variation in weblogs*, Journal of Sociolinguistics,10 (4), 439-459.

[14] Mladenic, D. and Grobelnik, D.1998. *Feature selection for classification based on text hierarchy*. Proceedings of the Workshop on Learning from Text and the Web, 1998

[14] S-W. Lee, "Multilayer Cluster Neural Network for Totally Unconstrained Handwritten Numeral Recognition", *Neural Networks*, Vol. 8, 1995, pp. 783-792.

[15] Ying Chen, Wenping Guo, Xiaoming Zhao, "A semantic Based Information Retrieval Model for Blog "Third International Symposium on Electronic Commerce and Security, 2010, IEEE