Comparison on Three Different Approaches on Sentiment Analysis

Mrs. Madhuri Agrawal Gupta
Assistant Professor, Geethanjali College of Engineering and Technology,Hyd.
Madhuriagrawal2000@gmail.com

Mrs S. Guru Jyothi
Assistant Professr, Sphoorthy Engineering College,Hyd
Fareenajyothi12@gmail.com

Abstract- This paper introduces the comparisons in between the three different approaches of sentiment analysis. There are three main academic streams on conducting the sentiment analysis task: Symbolic Approach, Supervised Learning Approach and Clustering approach. It is obtained that classification approach is efficient and no manual participation is required for solving the sentiment analysis problems.

Index Terms—opinion mining, sentiment analysis, clustering, supervised learning.

I. INTRODUCTION

Nowadays, social network sites contains a very vast amount of opinion expressing contents such as feedback, reviews, critiques, blogs , comments and so on. All content consists of full valuable information and helps the people to make decision. For example, Movie reviews help the viewers to take decision to go theater or not. Product review helps an enterprise to promote their products. Comments can help to clarify the strategy, etc. However, content is very huge and expressed in natural language. It is very difficult to read and analyze all the content by human. It helps to determine the positive or negative sentiment direction of online text contents and developing such task of technique is called opinion mining or sentiment analysis. It comes under a part of text mining and natural language processing. Sentiment Analysis is important to understand the test business KPIs, to improve customer service, to improve any campaign success or product messaging and to generate the leads. In this paper, different approaches are compared with respect to accuracy, effectiveness and human participation.

II. SENTIMENT ANALYSIS

To determine the positive or negative attitude direction of a writer with respect to a topic based on natural language processing is the main aim of sentiment analysis. The positive or negative attitude may be their mind state, emotional communication, evaluation on the basis of behavior or judgment, opinions, feelings, satisfaction ratings, the quality of shares, re-tweets, comments, replies, rating and also the quality of engagement over time. For example - An opinion is an expression that consists of two key components: target and sentiment. A target is one which we call as topic and sentiment is on the target or topic.

Such as -“I love this office”. Here “this office” is the topic and sentiment is expressed by the verb that is “love”, which is positive. There are major three types of sentiment analysis.

a. Manual Processing

Most mature and accurate judge of sentiment is done by human interpretation but that also not 100% accurate.

b. Keyword Processing

It assigns a degree or term of positivity, negativity to an individual word then it gives percentage score to text. For example: excellent, great, like, love can treated as positive words while terrible, dislike, not interested are considered as negative. This is very fast process to calculate, easily predictable, cheaper to implement and run as well. There is a major drawback is to deal with double meaning words means dealing with double negatives or positives.

c. Natural Language Processing

Natural language processing dictates a computer system that process human language in terms of its meaning. NLP understands several words. From them make a phrase, from several phrases make a sentence and from several sentences convey ideas.NLP is for analyzing the language for its meaning. Major drawback with NLP is to finding or detecting exaggerated statements and social media acronyms such as omg, b/w etc.

First, need to identify the attitude of the text means the opinion is positive or negative, even few also be classified as neutral. Second, is the identification of text as subjectivity or objectivity class? An objective sentence presents factual information whereas subjective sentence expresses personal feelings, view or their beliefs. For objective sentences positive, negative or neutral classification is helpful while opinion expressing

words and phrases indicates subjective sentence. By using the above two identification, retrieves a lot of important information from the social networking texts and the opinion expressing texts.

Paper focuses on the opinion classification as positive, negative or neutral and is concentrated mainly for the document analysis. It means it can be researched at word, phrase, sentences and document level.

A. Symbolic Approach

This approach requires pre-conversion of raw document into text vectors. Then constructs a feature- document matrix of m*n. It states a collection of n documents with m unique feature. This matrix was previously used for the automatic indexing. The main criterion is to assign each feature a sentiment score. The score is a measurement of the direction and intensity of the feature on scale of positive or negative. Once score of every feature is provided, the score of full document is calculated by using aggregation functions, usually average or sum. The core step is to provide score feature. There are three major types to score features.

a. Score by human subjects: Simple but not reliable and costly, even for large data it is not suitable. Scores are given on opinion expressing document by human then applied pseudo-expected value to establish scored word bank.

b. Score by Word Net: For English language, Word Net is a lexical database. Here only adjectives are provided as score, then finding the shortest path between two adjectives and it is considered that adjective with a shorter distance to 'good' be more positive and closer to 'bad' are more negative.

c. Score by Web Search: It is considered that the term co-occurred frequently is having same meaning and the distance between two words measured by statistics. It is having low accuracy rate and the method of integrating the score by average or sum is very simple.

B. Supervised Machine Learning Approach

It is a classification approach for extracting objective sentences than assigning scale value. To eliminate the negation words negation processing technique adopted. To perform classification three classifiers are selected Naive Bayes classification, Maximum Entropy Classification and Support Vector Machine. Testing can be done by cross validation. Accuracy level is high as compared to symbolic approach but costly because it needs a large data to pre-define by classes manually. It is a time consuming approach.

C. Clustering Approach

It uses K- mean algorithm for clustering the documents. No need to identify the class of document and to go for training process and free from human participation .Thus it is a time saving approach. Documents are clustered into two clusters positive and negative. To improve accuracy term frequency – inverse document frequency is applied on raw data then voting mechanism used to provide stable cluster. Then symbolic approach score are provided to enhance the result.

III. PERFORMANCE ANALYSIS

Directly document is having high dimensional vector space which results in inefficient working. So, reduce the dimension slightly by applying Porter's algorithm. Then extract all adjectives and adverbs from the document. Remove all other words from the document and convert remaining into a vector space in both frequency and presence form. Validate the data by using SVM then obtain the accuracy rate which comes nearly equal to Pang's result. With same data remove the class tags and apply k- mean algorithm with Mat lab toolbox to cluster documents in two groups. Cosine distance method is used to measure distance. As the actual class of document is known confusion matrix is constructed. Positive group if satisfied (a+d) > (b+c). Otherwise, negative group. Accuracy is calculated as (a+d) /no. of documents or (b+c)/ no. of documents.

Table1. Confusion Matrix

	positive	negative
Actual positive	a	b
Actual negative	c	d

For frequency of data accuracy is in between 50-60 percentage and for presence of data is in between 50- 65 percentage means results in low accuracy and very unstable as compared with classification approach. So improve accuracy rate by using TF-IDF Weighting Method. It is used to evaluate importance of term in document. TF-IDF can be calculated as:

$$wi = tfi*\log(D/dfi)$$

D=no. Of documents,

dfi= document frequency,

tfi=term frequency.

To improve stability in clustering approach voting mechanism is used and results in presence of data more competent than frequency of data. Along with that performance is enhanced by importing term scores from Word Net and by simple substitution directive accuracy is obtained. In order to create a weighting Vector based on term score combining the term score with clustering approach. The execution time of clustering approach is proportional to the number of dimensions.

IV. CONCLUSION

Accuracy is highest in supervised learning approach, acceptable in case of clustering approach and low in symbolic approach. Efficiency from the time point of view is very fast in symbolic approach, and for supervised learning approach it is very slow on the training data and fast on the test data while the cluster approach gives fast on the data. Symbolic approach and clustering approach are not required human participation at all but supervised learning approach does. So, overall the performance of cluster approach is most balanced in terms of efficiency, accuracy and human participation. Thus, it is suitable and good for real time applications. There are two major challenges with cluster approach. First, outcomes can be influenced on the size of document set and second, Word Net generate 70% accuracy in generating term score, if we find another better way to obtain term score ,better results can be obtained.

V. REFERENCES

[1] B. Pang, L. Lee and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” in Conference on Empirical Methods in Natural Language, pp. 79–86, 2002.

[2] P. Turney, “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews”, in Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 417–424, 2002.

[3] B. Pang and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts”, in Proceedings of the Association for Computational Linguistics, pp. 271–278, 2004.

[4] H. Yu and V. Hatzivassiloglou, “Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences”, in Conference on Empirical Methods in Natural Language, 2003.

[5] E. Boiy, P. Hens, K. Deschacht, M. F. Moens, “Automatic Sentiment Analysis in On-line Text”, in Proceedings ELPUB2007 Conference on Electronic Publishing – Vienna, Austria, June 2007.

[6] N.O. Andrews and E.A. Fox, "Recent Developments in Document Clustering," Technical Report TR-07-35, Computer Science, Virginia Tech, 2007.

[7] G. Salton, A. Wong and C.S. Yang, “A vector space model for automatic indexing”, Communications of the ACM, vol. 18, iss. 11, pp. 613-620, Nov 1975.

[8] E. Brill, “Some advances in transformation-based part of speech tagging”, in Proceedings of the Twelfth National Conference on Artificial Intelligence Menlo Park, pp. 722-727, 1994.

[9] C. Cesarano et al., “OASYS: An Opinion Analysis System”, in AAAI 2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, pp. 21–26, 2006.

[10] G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross and K. Miller, “WordNet: An online lexical database”, International Journal of Lexicography, vol. 3, iss. 4, pp. 235-244, 1990.

[11] J. Kamps, M. Marx, R. J. Mokken and M. D. Rijke, “Using WordNet to measure semantic orientation of adjectives”, in International Conference on Language Resources and Evaluation, vol. IV, pp. 1115-1118, 2004.

[12] P. Turney, “Mining the Web for synonyms: PMI-IR versus LSA on TOEFL”, in Proceedings of the Twelfth European Conference on Machine Learning in Springer Verlag, Berlin, pp. 491-502, 2001.

[13] B. Pang and L. Lee, “A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts”, in Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL) in Barcelona, Spain, pp.271–278, July21-26 2004.

[14] B. Pang and L. Lee, “Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales”, in Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL) in University of Michigan, USA, pp.115–124, June25–30 2005.

[15] S. Das and M. Chen, “For Amazon: Extracting market sentiment from stock message boards”, in Proceeding of the 8th Asia Pacific Finance Association Annual Conference (APFA), 2001.

[16] V. Hatzivassiloglou and K. R. McKeown, “Predicting the semantic orientation of adjectives” , in Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Madrid, ES, pp. 174–181, 1997.

[17] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval, Information Processing & Management, vol. 24, issue. 5, pp :513–523, 1988.

[18] F. Benamara et al., “Sentiment Analysis: Adverbs and Adjectives Are Better than Adverbs Alone”, in Proceedings of 2007 International Conference Web-logs and Social Media (ICwsm 07), 2007;

[19] M.F. Porter, “An Algorithm for Suffix Stripping”, Program, vol. 14, issue. 3, pp. 130-137, 1980.

[20] K. Toutanova and C. D. Manning, “Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger”, in Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corporations (EMNLP/VLC- 2000), pp. 63-70, 2000.

[21] Gang Li, Fei Liu, “A Clustering based approach on sentiment analysis”,IEEE Transaction,pp: 331-337,2010