International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 6, June 2016

Impact Analysis of Features on Classification for Sentiment Analysis of Tweets

[1]Jasleen Kaur, [2]Sukhjit Singh Sehra, [3]SumeetKaurSehra
[1] Research Scholar, [2] [3] Assistant Professor,
[1][2][3]Department of Computer Science and Engineering, Guru Nanak Dev Engineering College, Ludhiana, Ludhiana

Abstract: **Sentiment Analysis is a sub domain in computational linguistics and information retrieval that is concerned with the topic as well as opinion expressed by a document. Prompted by the growth in use of Twitter, Companies are seeking new ways to extract the opinion expressed by people about their products and services. Tweets represent abridged thoughts which pose a challenge for excavation due to the incredible breadth of covered topic. This paper presents an impact analysis of selection of features on the performance of classifiers for opinion mining of tweets. A model is formulated for classification of tweets and comparison of various existing approaches. N grams approach is used for feature extraction and tf-idf is used as weighing criteria. Maximum accuracy is achieved by employing a hybrid method of SVM and Naïve Bayes attaining 79.80% accuracy employing unigrams and 82.60% employing bigrams as features.**

Index terms: **Feature extraction, Machine learning approaches, Opinion mining, Sentiment analysis.**

I. INTRODUCTION

Popularity of Web 2.0 has made Web a more common platform for interactivity[1]. The credibility of user generated data using Web 2.0 and Flash in extracting opinionative contents about products and services has prompted many researchers to work in the field of sentiment analysis. Microblogging sites like Twitter have attracted many users due to their free formatting of messages as well as better interactivity options as compared to traditional blogs. Users have been empowered to comment and discuss about various social issues on social networking platforms by the growth in mobile applications and Internet[2].

Sentiment Analysis or Opinion Mining basically involves classification of the polarity of any document or piece of text. Determination of polarity of any term or document is the process of classifying it as positive or negative in accordance with the opinion it expresses. Hence sentiment analysis can be described as a system which summarizes the sentiment or opinion of any input document after analyzing it. Automatic labeling of product and movie reviews with their polarity is used by intelligent or recommending systems for scrutinizing these reviews to estimate popularity of their products or services. Also it aids consumers to make more informed decisions by giving them insights into the sea of data generated by crowds. The major challenge in classification of these tweets is their short length (140 characters per tweet) owing to which authors use slang words, abbreviations, smileys which ultimately lead to noise thus making the task of classification more arduous.

Features in reference to sentiment analysis include the terms or words which emphasize the positive or negative opinions expressed in the text. These strongly influence the inclination of given text. Feature extraction is the process of extracting the main characteristics present in the text. Term frequency, its presence and its relative position in the given text are most widely used as features [3]. This paper is concerned with the evaluation of different machine learning approaches for automatic analysis of polarity or sentiment analysis. The accuracy of these approaches is evaluated and compared using different n grams as features.

The rest of this paper is organized as follows: Section 2 gives an account of existing approaches in this area. Section 3 describes the Methodology adopted for the current work. Results are presented in Section 4 followed by Conclusion in Section 5.

II. SENTIMENT ANALYSIS APPROACHES

There are many feature extraction techniques that have been utilized in the past by various researchers to extract features from the text. Different features have different impact on the accuracy of machine learning algorithms for classification.

Marking a word of corpus according to its corresponding part of speech, on basis of the context in which it is used, is referred as POST or Part-of-speech tagging. Bermingham and Smeaton[4] found that syntactic patterns using POS and n-gram features together are more reliable than using only unigrams. Support Vector Machine outperformed Naïve Bayes in blogs but mining of microblogs was found to be more accurate using the latter. However, Go et al.[5] did not find POS tags useful and used both bigrams and unigrams for extracting features. Their approach increased the accuracy of Naïve Bayes (from 81.3% to 82.7%) and MaxEnt (from 80.5% to 82.7%) but declined the same for SVM (from 82.2% to 81.6%). Pak and Paroubek[6] employed n grams and POS tags as features. The distribution of POS tags was analyzed to classify the tweets as subjective or objective. They found Naïve Bayes using bigrams to be most effective as bigrams provide coverage as well as capture the expressing patterns.

Adverbs and Adjectives impart additional weightage to the expressing strength of succeeding noun or verb. Benamara et al. [7] used both adjectives and adverbs and compared the result with those obtained by using only adjectives. They found that adjectives are better indicators of polarity than adverbs. Hu and Liu [8]used a set of adjective words to identify opinion words and used WordNet to identify the semantic orientation of these. The results for whole document were summarized using individual orientations and thus obtained an average accuracy of 84%. Singh et al. [9] evaluated performance of SentiWordNet using adverbs and adjectives for feature extraction and compared the results with SVM and Naïve Bayes. They found adjectives to be more effective. Also, SVM and NB outperformed SentiWordNet in classification. Eirinaki et al. [10] used High Adjective Count which assignsa score to each noun on basis of adjectives found in that corpus. It provided more reliable results than tf or tf-idf approaches with an accuracy of around 97%.

Tweets are characterized by the extensive use of tags and emoticons which can be used as feature extractors. Kouloumpis et al. [11] used hashtags in Twitter data in construction of training set for sentiment analysis of tweets. They found maximum accuracy using n grams together with lexicon and Microblogging features. Use of POS tags resulted in a drop in accuracy. Amolik et al. [12] compared the performance of SVM and NB for sentiment analysis of movie reviews using Twitter specific data for extraction of features. SVM resulted in better accuracy and recall than NB but the latter was found to have better precision.Neethu and Rajasree [13] used hashtags and emoticons to extract relevant tweets and then constructed the feature vector using unigrams. They found machine learning techniques to be more effective than symbolic techniques.

A continuous occurrence of n terms from any text or document is known as N gram. Akaichi et al. [14] extracted features using unigrams, unigrams with bigrams and unigrams with trigrams as features. They employed two classifiers SVM and Naïve Bayes which obtained their highest accuracy using unigrams with bigrams and least by using unigrams with trigrams. They reported that SVM outperformed NB in classification. Zhai et al. [15] performed an in-depth analysis of all features for exploiting their effectiveness. They found bigrams to be the most effective features with the potential to improve performance of classification. They also found that features at varying length are more effective than those at fixed length. Abbasi et al. [16] utilized the syntactic information from the n grams and proposed a rule-based method of selecting features using relations. Its manner of selecting features was found to be more efficient than existing univariate and multivariate selection methods.

TF-IDF (term frequency–inverse document frequency) reflects the significance of a given word in a corpus and assigns it a weight which in turn helps in accurate classification. Ghag and Shah [17] focused on the average frequency count distribution calculated on the basis of proportional frequency and presence count distribution whereas approaches liketf-idfare based on overall frequency as well as proportional presence count distribution of a term. They achieved an accuracy of 71.3% which showed considerable improvement from 66.5% obtained using tf-idf. Haddi et al. [18] computed three feature matrices based on feature frequency, feature presence and tf-idf. Chi-squared test was then used to filter the extracted features. They reported 92.3% accuracy using tf-idf as feature selection approach. Deng et al. [19] outperformed the traditional approaches using combination of importance of term in document (ITD) and for expressing sentiment (ITS) for analysis.

Machine learning algorithms like SVM, NB, MaxEnt, Decision tree etc. have been extensively used in the field of sentiment analysis. These employ two sets: an already labeled training set which is used in the learning process and a test set which contains text to be classified using the learned model. Mukwazvure and Supreethi [20] proposed a hybrid approach of SVM and kNN for classification of news comments. SVM provided better accuracy than kNN but increasing the value of k increased accuracy of the classifier. Jiang et al. [21] proposed an improved kNN algorithm by using both one pass clustering algorithm and KNN categorization. Their classifier performed better than existing SVM and NB classifiers. Zhu et al. [22] used artificial neural network (ANN) for sentiment analysis and found that adding the prior knowledge to the model increased the accuracy of ANN. Sun et al. [23] conducted a study to compare the effectiveness of different strategies in sentiment analysis of imbalanced text using SVM. Resampling and instance weighting were found to be non-effective and SVM could learn better without using these techniques. Ye et al. [24] compared three machine learning algorithms: SVM, Naïve Bayes, and N-gram model for sentiment analysis of travel blog reviews. They found that SVM and N-gram model outperformed NB when the data set was large.

III. METHODOLOGY

The methodology used by the current work for Sentiment Analysis of Twitter data has been described below:

A. Data Collection

The present work uses Twitter, a prominent micro-blogging site, to analyze opinions or sentiments regarding a specific topic. Twitter provides an Application Programming Interface (API) to programmatically access its tweets by using a term as query. OAuth provides an authorized access to this API. The training and test data sets used in the present work contain only the tweets in English language. The number of tweets downloaded per query has been set to 5000.

B. Preprocessing

Preprocessing is the process of filtering the extracted tweets for removing noise and irrelevant contents which may lead to inaccurate classification. Tweets contain irrelevant content like URLs, slang words, tags which may not contribute in analyzing their orientation but only increase the dimensionality of resulting feature set. Punctuation erasure, Number filter, N chars filter, Case converter, Stemmer and Stopword removal have been incorporated in our project in order to improve accuracy.

C. Feature Creation

As already discussed, features convey the sentiment of the text more robustly and the accuracy of results or classification highly depends on the choice of feature set. Unigrams and Bigrams have been employed in the present work to extract features and tf-idf is used as weighing criteria.

D. Classification

After the creation of feature set, Support Vector Machine (SVM), Naïve Bayes (NB), Decision trees, Neural Networks, and a hybrid method of Support Vector Machine and Naïve Bayes have been employed for classification and their performance is compared.

IV. RESULTS AND DISCUSSION

Accuracy is used as a parameter for comparison of different algorithms using different feature sets. A data set containing 5000 tweets is partitioned for training and testing of these classifiers. The resulting accuracies of these classifiers using unigrams and bigrams for feature extraction are shown in figure 1. Using bigrams in place of unigrams improves the accuracy of SVM (75.90% from 73.40%) and Decision Tree (71.20% from 66.80%) but leads to a decline in case of Naïve Bayes (69.80% from 72.70%) and Neural Network (67.60% from 68.80%). However, the highest accuracy is obtained using a hybrid method of Support Vector Machine and Naïve Bayes (79.80% using unigrams and 82.60% using bigrams).Out of ensemble learning techniques Bagging, Boosting and Random subspace, Bagging was applied and it improved the accuracy of SVM and Decision Tree but decreased the same for NB. When the size of the training set is further increased to 15000 and accuracy of each classifier is compared for unigrams and bigrams, no difference is observed in the accuracies. Same accuracy is obtained on using unigrams and using bigrams.

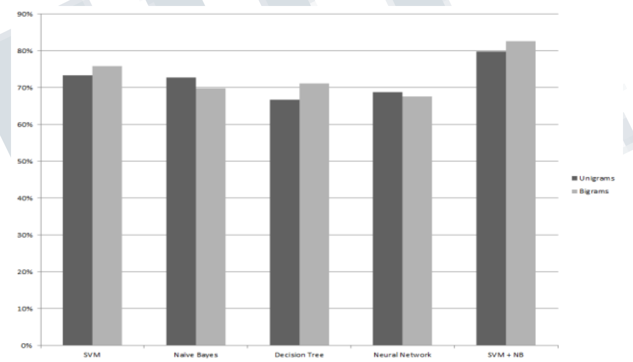

Figure 1 : Accuracy of different Classifiers using Unigrams and Bigrams

V. CONCLUSION AND FUTURE SCOPE

Extraction of user's intentions or sentiments regarding different entities has become worthy accounting to the ever-increasing amount of data by the "sharing-age" of Web 2.0. This paper presents a systematic literature review of various feature extraction and machine learning approaches employed in the field of sentiment analysis techniques. The performance of different classifiers is compared using unigrams and bigrams for feature extraction. Bigrams are proved to be more effective than unigrams. Employing Support Vector Machine as classifier leads to the highest accuracy among the basic classifier. However, the accuracy reaches its highest value of 82.60% when a hybrid approach using SVM and Naïve Bayes using bigrams for feature extraction is employed.

The future research in this field should focus on improving the performance of basic classifiers using different optimization techniques like swarm intelligence. Construction of feature set can be more optimal if the effect of presence of negation, irony or sarcasm on the orientation of tweets is further analyzed. Domain dependency is also an issue in the current work and future work should concentrate on development of a model that performs well in all areas irrespective of the domain. The present work involves classification of tweets only in English; multilingual classification field can also be explored.

REFERENCES

[1] H. Lee, P. Ferguson, N. O'Hare, C. Gurrin, and A. F. Smeaton, "Integrating interactivity into visualising sentiment analysis of blogs," in *Proceedings of the first international workshop on Intelligent visual interfaces for text analysis*, 2010, pp. 17–20.

[2] L. Zhang, K. Hua, H. Wang, G. Qian, and L. Zhang, "Sentiment Analysis on Reviews of Mobile Users," *Procedia Comput. Sci.*, vol. 34, pp. 458–465, 2014.

[3] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intell. Syst.*, no. 2, pp. 15–21, 2013.

[4] A. Bermingham and A. F. Smeaton, "Classifying sentiment in microblogs: is brevity an advantage?," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, pp. 1833–1836.

[5] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Proj. Rep. Stanf.*, vol. 1, p. 12, 2009.

[6] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining.," in *LREC*, 2010, vol. 10, pp. 1320–1326.

[7] F. Benamara, C. Cesarano, A. Picariello, D. R. Recupero, and V. S. Subrahmanian, "Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone.," in *ICWSM*, 2007.

[8] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168–177.

[9] V. K. Singh, R. Piryani, A. Uddin, and P. Waila, "Sentiment analysis of Movie reviews and Blog posts," in *Advance Computing Conference (IACC), 2013 IEEE 3rd International*, 2013, pp. 893–898.

[10] M. Eirinaki, S. Pisal, and J. Singh, "Feature-based opinion mining and ranking," *J. Comput. Syst. Sci.*, vol. 78, no. 4, pp. 1175–1184, Jul. 2012.

[11] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!," *Icwsm*, vol. 11, pp. 538–541, 2011.

[12] A. Amolik, N. Jivane, M. Bhandari, and M. Venkatesan, "Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques.," *Int. J. Eng. Technol.*, vol. 7, no. 6, pp. 2038–2044, 2015.

[13] M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," in *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on*, 2013, pp. 1–5.

[14] J. Akaichi, Z. Dhouioui, and M. J. Lopez-Huertas Perez, "Text mining facebook status updates for sentiment classification," in *System Theory, Control and Computing (ICSTCC), 2013 17th International Conference*, 2013, pp. 640–645.

[15] Z. Zhai, H. Xu, B. Kang, and P. Jia, "Exploiting effective features for chinese sentiment classification," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 9139–9146, Aug. 2011.

[16] A. Abbasi, S. France, Z. Zhang, and H. Chen, "Selecting Attributes for Sentiment Classification Using Feature Relation Networks," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 3, pp. 447–462, Mar. 2011.

[17] K. V. Ghag and K. Shah, "ARTFSC–Average Relative Term Frequency Sentiment Classification," *Int. J. Comput. Technol.*, vol. 12, no. 6, pp. 3591–3601, 2014.

[18] E. Haddi, X. Liu, and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013.

[19] Z.-H. Deng, K.-H. Luo, and H.-L. Yu, "A study of supervised term weighting scheme for sentiment analysis," *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3506–3513, Jun. 2014.

[20] A. Mukwazvure and K. P. Supreethi, "A hybrid approach to sentiment analysis of news comments," in *Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), 2015 4th International Conference on*, 2015, pp. 1–6.

[21] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved K-nearest-neighbor algorithm for text categorization," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 1503–1509, Jan. 2012.

[22] J. Zhu, C. Xu, and H. Wang, "Sentiment classification using the theory of ANNs," *J. China Univ. Posts Telecommun.*, vol. 17, pp. 58–62, Jul. 2010.

[23] A. Sun, E.-P. Lim, and Y. Liu, "On strategies for imbalanced text classification using SVM: A comparative study," *Decis. Support Syst.*, vol. 48, no. 1, pp. 191–201, Dec. 2009.

[24] Q. Ye, Z. Zhang, and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 6527–6535, Apr. 2009.