Review Paper on Sentiment Analysis Using Multi Class Classification to Acronyms

[1]Mr. Ganesh T. Mahalkar, [2]Prof. P.M. Mohite, [3]Prof. P.D. Sathya
[1]M.E Student, Department of Computer Sci. & Engineering, SYCET Aurangabad
[2]Assistant Professor, Department of Computer Sci. & Engineering, SYCET Aurangabad
[3]Associate Professor & HoD, Department of Computer Sci. & Engineering, SYCET Aurangabad
Email – [1] ganeshmahalkar87@gmail.com

Abstract: *The advance technique for sentiment analysis has been fast and direction to explore the opinions or text present on different twits of social media through machine-learning techniques. Sarcasm is a sort of sentiment where public expresses their negativity using positivity within the text. For humans it is very tough to acknowledge and analysis or unfriendliness calculations. To deal with these challenges, the contribution of this paper includes the confirm sentiment analyzer that includes machine learning. This paper also provides a comparison of techniques of sentiment analysis in the analysis when the user uses the shortcut or acronyms like "ILU, 143(happy), KMN (kill me now), MYOB (mind your own business), B3 (blah, blah, blah)"the machine unable to perform the classification, so in Proposed system may be able to machine learning algorithm can classify the acronyms and find the appropriate result to improve the sentiment analysis.*

Keywords: *Twitter; sentiment analyzer; machine learning etc.*

1. INTRODUCTION:

Due to the presence of excessive amount of data available on web, various cooperative started taking interest in this as mining this information can be very valuable to them. This gives birth to an entirely different and expansive field of study known as Sentiment Analysis. Various names have given to this field as opinion mining, opinion extraction etc. However there is slight difference in meaning between these various terms. Before automatic mining of sentiments traditional survey techniques were highly biased as they were taken individually by users thus a need of an automatic system arose that can directly deal with hundreds of thousands of opinions hidden in users' posts in the form of reviews, online journal etc. Various applications of sentiment analysis are as in product reviews, movie reviews, business, politics, recommender system etc. Based on the opinion about a product or about different condition of a product, an organization can make changes accordingly.

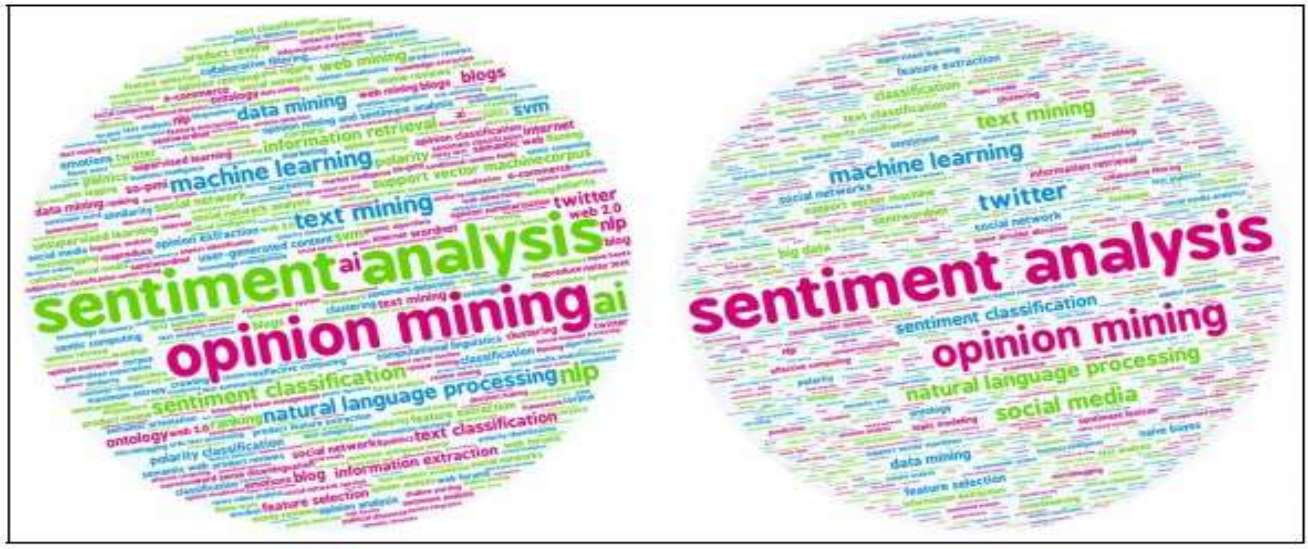

Fig1. Sentiment Analysis Opinion

Similarly based on the opinion about a particular political party, government policies' changes can be made accordingly. Two main techniques used for sentiment analysis are machine learning based and lexicon based. Supervised, Unsupervised and Semi-supervised comes under Machine Learning. Supervised approaches e.g.

SVM[2][11][13][14][19][30][32], KNN[20][21], Naive Bayes[3][4][6][7] etc. requires a good quality training set and thus are highly domain dependent but provide better results if trained properly. Unsupervised approaches e.g. K-Means, Self-Organizing Maps (SOM) etc. do not make use of training set. Semi supervised approaches require partial identify of data and are of two types: a) Transductive Learning b) Inductive Learning Lexicon based approach makes use of dictionary consist of labelled words and with the help of these words, a text is intermediate whether it is subjective or objective[16][23][25]. This approach is further divided into a) Dictionary based which does not take into account the context of word within a text, and b) Corpus based which expands the dictionary with taking associations between different words into account. A complete survey in this field is provided in [9]. This research analyze sentiment of tweets [1][2][3][4][5][8]. As tweets are very unstructured in nature this research converts them into useful information so that better features can be used for machine learning. Hence in this research we provide a good data preprocessing to tweets followed by hybrid classifier. With the help of processed tweets or data features are generated and fed to the two machine learning algorithms KNN and SVM in a hybrid manner. Different feature have tried by authors for improving the results as in [2][3[4][5][18].

TABLE I: Structure of the Dataset Used [34]

Class	Training set	Test set
Happiness	4000	225
Love	2000	219
Sadness	2000	223
Anger	2500	201
Hate	2500	157
Sarcasm	3000	199
Neutral	1000	176
Total	17000	1400

2. LITERATURE SURVEY:

Various researchers have been working on twitter [3][4][5][11][12][16] and from time to time they are publishing their researches . They have used various sentiment analysis techniques for improving the results of classification their work is also helpful in this research as the sentiment analysis techniques they have used, feature selection techniques, different pre-processing steps they have used is taken care of in this research. This research mainly focuses on supervised approach for sentiment analysis task and has surveyed researches both for twitter and non- twitter data and also for both supervised and lexicon based approaches for better clarification and understanding of the topic chosen. Many researches defined multiple faces of sentiment analysis as opinion orientation, feature extraction etc. Machine learning classifiers need various features for learning so different researchers from time to time have selected different features for comparing results. Agarwal et al.[02], Pak and Paroubek [03], Spancer and Uchyigit [04], Koloumpis et al.[05] selected various features as unigrams, bigrams, pos tagging, hash tags, ngrams etc. and found mixed response in classification results. Different features and feature selection methods as semantic features and concepts, information gain, chi-square etc. has been used by Hassan Khan et al.[13], Agarwal et al.[14]. Hassan Khan et al.[13] approach includes rigorous data pre-processing followed by supervised machine learning. They collected labelled datasets of different domains so that machine learning will not be limited to a particular domain. To learn SVM classifier they make use of different training sets each make SVM learn different feature sets -1) Information gain(IG) with feature presence and 2) feature frequency 3) Cosine similarity with feature presence and 4) feature frequency. They found that feature presence is better than feature frequency. Agarwal et al.[14]], found that for better results using machine learning approaches, finding good features is a challenging task. They gave the concept of "Semantic Parser" and treated concepts as features. They used the minimum Redundancy and Maximum Relevance (mRMR) feature selection mechanism. They used different feature sets for their classification task e.g. unigrams, bigrams, bitagged and dependency parse tree along with their proposed scheme so that results can be compared with. Various approaches and classifiers such as lexicon based approach, Naive Bayes(NB), Support Vector Machines(SVM), Maximum Entropy(MaxEnt) etc. have been used time to time with various parameters for evaluating the results as accuracy, precision, recall, f-measure etc. Narr et al[06] concluded 71.5% accuracy with mixed language NB classifier on unigrams. Saif et al.[07] concluded Ankita Gupta et al, International Journal of Computer Science and Mobile Computing, Vol.6 Issue.4, April- 2017, pg. 444-458 © 2017, IJCSMC All Rights Reserved 447 that semantic features used by NB classifier increase f1-measure against unigram by 6.47% and pos+unigram by 4.78%. Asmi [24], Hutto[25], Neviarouskaya[26] proposed rule based approaches for increasing the accuracy. Swati[27], Chikersal[29] and Prabowo and Thelwall[10] proposed hybrid approach consisting of rule based and machine learning classifiers. Hybrid approaches consisting of machine learning classifiers have been underexplored in the literature with very few researches in this approach as in Revathy[28], F.F. da Silva et al.[11]. F.F. da Silva et al.[11] proposed an ensemble based classification in which various classifiers e.g. SVM, Multinomial Naive Bayes, Random Forest, Logistic

Regression are used. They proposed that if we train the different classifiers with different training sets and then by using either average probabilities of different classifiers or maximum voting, we get better results than by using only a single classifier. Moreover they uses two different features for learning the classifiers:- a) Bag of Words(BOW) b) Feature Hashing They used four different datasets for training and testing. They found that Feature Hashing is not better than BOW approach in most of the datasets except one. Our research work mainly focuses on combining the machine learning classifiers and proves that combining gives better results as compared to standalone classifiers. Also this research gives comparative results as against to the feature hashing+lexicon based features used by [11] ,with only a small dataset and few features. Bhadane et al[15], Apple et al[16] proposed combination of sentiment lexicon with machine learning approaches and found increase in accuracy. Muhammad et al.[17] handled word's polarity in terms of local and global context by giving SmartSA system and found that their system is superior to baseline lexicons and systems like SVM, NB etc. with more F1 score. Addlight and Supreethi[20] compared two machine learning methods KNN and SVM and found that SVM outperforms KNN. Saif and He[23] gave the concept of SentiCircles for calculating the context of words. They found that it is necessary for better sentiment classification. Jianqiang et al.[32] discussed the role of rigorous preprocessing in increasing the evaluation measure and gave six different preprocessing methods for the same. Keeping this in mind our approach also uses a good preprocessing to filter the tweets. Khan and Jeong[21] proposed an approach for finding the sentiments about each aspect of a product and this can be a good future work to explore.

3. Conclusion and Scope:

In this paper, a SVM and KNN based hybrid model is presented to update the classification right. The proposed method classified the tweets in positive, negative and neutral sentiments with whereas much of the literature in this field is associated with 2-way classification [10][11]. In existing system the machine is only able to find the 7 type of emotion "Happiness , Love, Sadness, Anger ,Hate, Sarcasm and neutral " but when the user uses the shortcut or acronyms like "ILU,143**,**(happy)**,** KMN(kill me now)**,** MYOB (mind your own business)**,**B3 (blah,blah,blah)"the machine unable to perform the classification, so in Proposed system may be able to machine learning algorithm can classify the acronyms. The comparative observations are taken against the SVM and KNN methods. The comparative results shows that the proposed model has improved the accuracy and f-measure of tweet class prediction. As number of features for learning the classifiers are limited in our approach, we will be using more features and better feature selection methods like Information capture, Chi-Square etc. in our future work. Our comparison with literature [11] shows that increasing our dataset with more tweets and features can also help in increasing reasonable accuracy and f-measure. Other machine learning methods in combined way can also be explored in the future.

REFERENCES:

1. Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1 (2009): 12.
2. A Agarwal, B Xie, I Vovsha, O Rambow. "Sentiment analysis of twitter data." Proceedings of the workshop on languages in social media. Association for Computational Linguistics, 2011.
3. Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREc. Vol. 10. 2010.
4. Spencer, James and Gulden Uchyigit. "Sentimentor: Sentiment analysis of twitter data." Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. 2012.
5. Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. "Twitter sentiment analysis: The good the bad and the omg!." Icwsm 11 (2011): 538-541.
6. Narr, Sascha, Michael Hulfenhaus, and Sahin Albayrak. "Language-independent twitter sentiment analysis." Knowledge Discovery and Machine Learning (KDML), LWA (2012): 12-14.
7. Saif, Hassan, Yulan He, and Harith Alani. "Semantic sentiment analysis of twitter." International Semantic Web Conference. Springer Berlin Heidelberg, 2012.
8. Carpenter, Thomas, and Thomas Way. "Tracking Sentiment Analysis through Twitter." Proceedings of the International Conference on Information and Knowledge Engineering (IKE). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.
9. Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal (2014) 5, 1093–1113. [10] Prabowo, Rudy, and Mike Thelwall. "Sentiment analysis: A combined approach." Journal of Informetrics 3.2 (2009): 143-157.
10. Da Silva, Nadia FF, Eduardo R. Hruschka, and Estevam R. Hruschka. "Tweet sentiment analysis with classifier ensembles." Decision Support Systems 66 (2014): 170-179.
11. Khan, Farhan Hassan, Saba Bashir, and Usman Qamar. "TOM: Twitter opinion mining framework using hybrid classification scheme." Decision Support Systems 57 (2014): 245-257.

12. Khan, Farhan Hassan, Usman Qamar, and Saba Bashir. "A semi-supervised approach to sentiment analysis using revised sentiment strength based on Senti Word Net." Knowledge and Information Systems (2016): 1-22.
13. Agarwal, Basant,Soujanya Poria,Namita Mittal,Alexander Gelbukhand Amir Hussain. "Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach." Cognitive Computation 7.4 (2015): 487-499. Bhadane, Chetashri, Hardi Dalal, and Heenal Doshi. "Sentiment analysis: Measuring opinions." Procedia Computer Science 45 (2015): 808-814.
14. Appel, F Chiclana, J Carter, H Fujita. "A hybrid approach to the sentiment analysis problem at the sentence level." Knowledge-Based Systems 108 (2016): 110-124.
15. Muhammad, Aminu, Nirmalie Wiratunga, and Robert Lothian. "Contextual sentiment analysis for social media genres." Knowledge-Based Systems 108 (2016): 92-101.
16. Zhu, Dengya, and Jitian Xiao. "R-tfidf, a Variety of tf-idf Term Weighting Strategy in Document Categorization." Semantics Knowledge and Grid (SKG), 2011 Seventh International Conference on. IEEE, 2011.
17. Anjaria, Malhar, and Ram Mohana Reddy Guddeti. "A novel sentiment analysis of social networks using supervised learning." Social Network Analysis and Mining 4.1 (2014): 1-15.
18. Mukwazvure, Addlight, and K. P. Supreethi. "A hybrid approach to sentiment analysis of news comments." Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), 2015 4th International Conference on. IEEE, 2015.
19. Khan, Jawad, and Byeong Soo Jeong. "Summarizing customer review based on product feature and opinion." Machine Learning and Cybernetics (ICMLC), 2016 International Conference on. IEEE, 2016.
20. Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." Journal of Big Data 2.1 (2015): 5.
21. H Saif, Y He, M Fernandez, H Alani. "Contextual semantics for sentiment analysis of Twitter." Information Processing & Management 52.1 (2016): 5-19.
22. Asmi, Amna, and Tanko Ishaya. "Negation identification and calculation in sentiment analysis." The Second International Conference on Advances in Information Mining and Management. 2012.
23. Hutto, Clayton J., and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." Eighth International AAAI Conference on Weblogs and Social Media. 2014.
24. Neviarouskaya, Alena, Helmut Prendinger, and Mitsuru Ishizuka. "Semantically distinct verb classes involved in sentiment analysis." IADIS AC (1). 2009.
25. Kawathekar, Swati A., and Manali M. Kshirsagar. "Sentiments analysis using Hybrid Approach involving Rule-Based & Support Vector Machines methods." IOSRJEN 2.1 (2012): 55-58.
26. Revathy, K., and B. Sathiyabhama. "A hybrid approach for supervised twitter sentiment classification." International Journal of Computer Science and Business Informatics 7 (2013).
27. Chikersal, Prerna, Soujanya Poria, and Erik Cambria. "SeNTU: sentiment analysis of tweets by combining a rule-based classifier with supervised learning." SemEval-2015 (2015): 647.
28. 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.
29. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.
30. Z. Jianqiang and G. Xiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis," in IEEE Access, vol. 5, no. , pp. 2870-2879, 2017.
31. Ankita Gupta1 , Jyotika Pruthi2 , Neha Sahu3 Sentiment Analysis of Tweets using Machine Learning Approach IJCSMC, Vol. 6, Issue. 4, April 2017, pg.444 – 458
32. "A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter", Graduate School of Science and Technology, Mondher Bouazizi Keio University Yokohama, Japan, Email: bouazizi@ohtsuki.ics.keio.jp
33. Pooja Deshmukh,Sarika Salunke "Sarcasm Detection and User Behaviour Analysis" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 6 Issue: 6.