International Journal of Inventive Engineering and Sciences (IJIES)
ISSN: 2319-9598 (Online), Volume-5 Issue-6, September 2019

An Enhanced Technique for Analyzing Sentiments of Public Reviews - I

Chintan Panjwani, Rashmi Thakur

Abstract: Sentiment analysis is the process of extracting the opinion expressed in a piece of text to determine the writer's attitude towards a topic, product or any service in general and classify it into classes such as positive, negative or neutral. Bag of Words is the traditional approach for text representation in Sentiment Analysis where text is represented as bag of its words. This approach represents the text by breaking the sentence into words disregarding other semantic information. A problem that occurs due to this representation is Polarity Shift problem. To address polarity shift problem a dual sentiment analysis (DSA) system is created. It looks at the reviews from both the sides i.e. positive and negative. The existing work on dual sentiment analysis includes techniques where dual training and dual prediction is performed. The proposed system is to enhance the classification performance of the existing system by applying different classifiers apart from those used in existing system to obtain better results. After classification of reviews into appropriate classes, various graphs are plotted based on different parameters to validate the results and determine the best classifier from the applied classifiers.

Keywords: Bag of Words, Enhanced Dual Sentiment Analysis, Polarity Shift problem, Sentiment Analysis, Support Vector Machine

I. INTRODUCTION

In the past decade due to the rapid growth of internet a lot of business activities are taking place online. A number of business activities from selling of expensive products or services to the sale of grocery everything is now available online. With the increase of online sales, the number of online reviews available for such products and services have also increased. This makes it very necessary to extract information from the huge amount of textual data. Text mining is a technique of extracting meaningful information from a piece of text. It has many applications including spam filtering, creating recommendations, etc. Sentiment analysis is one such application of text mining. This paper focuses on analysis of public opinions from the large amount of reviews. Sentiment analysis can be done mainly using two techniques viz. Lexicon Based approach and Machine leaning approach. Many approaches have been implemented in literature for sentiment analysis and most of them use the Bag of Words approach to represent the text for analysis. This approach does not maintain the grammar, word order, semantic information of the text which leads to a problem called as Polarity Shift Problem. This problem leads to polarity reversion of a piece of text. In this paper a technique called Enhanced Dual Sentiment Analysis is used which includes Dual training and Dual prediction for the text.

The existing Dual Sentiment Analysis(DSA) framework includes the following: 1. Data expansion technique to create reversed reviews, 2. Dual training and Dual prediction, 3. 3-class sentiment classification and 4. Creation of a corpus based pseudo-antonym dictionary [1]. The classifiers used in dual training and dual prediction phase are Naïve Bayes, Support Vector Machines and Logistic Regression. The Enhanced Dual Sentiment Analysis(EDSA) focuses on improving the classification accuracy of the current system by performing objective and subjective analysis and performs sentiment analysis on only subjective reviews. It also applies Maximum Entropy and Voted classifier apart from Naïve Bayes and Support Vector Machines and compares the results with existing system.

Sentiment analysis uses natural language processing to identify subjective information from a piece of text. The reviews obtained from different websites are first segmented into subjective and objective reviews. Sentiment analysis is performed only on subjective sentences as only subjective sentences express some opinion whereas objective sentences are just statements providing facts of certain topic. A supervised sentiment classification approach is used in the existing system where all the reviews considered for sentiment analysis are labelled. Obtaining labelled data for every domain is a very costly and time consuming. A semi supervised classification technique can be used which does not require the entire dataset to be labelled.

The rest of the paper is organized as follows. Section II provides the literature review where various techniques for sentiment analysis are reviewed and surveyed. Section III presents the proposed system which will give details about the enhancements that will be applied to the existing system. Section IV gives the expected outcome from the proposed system. Section V gives some applications and uses of sentiment analysis in real world. Section VI is the conclusion which summarizes the learnings of the entire paper and provides certain applications of the proposed system.

II. RELATED WORK

Several techniques are present in literature for performing sentiment analysis. To address the Polarity Shift problem, a number of techniques have been proposed in literature. Some methods consider the negation whereas some consider other polarity reversal phrases to detect the Polarity Shift in a piece of text. Other thing that has to be considered for sentiment analysis is the feature set used to represent the text.

Revised Manuscript Received on September 20, 2019.
*Correspondence Author

Chintan Panjwani*, M.E Computer Engineering Student, Thakur College of Engineering & Technology, Mumbai, chintan.panjwani@gmail.com

Mrs. Rashmi Thakur, Assistant Professor, Thakur College of Engineering & Technology, Mumbai, thakurrashmik@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Every method used for sentiment analysis considers a number of feature sets for classification of reviews into different classes.

Apart from feature sets a number of classification algorithms are used in all such different techniques. The following sections review all such techniques for performing sentiment analysis.

A. Techniques to address polarity shift problem

Councill et al. [2] proposes a system for detecting the scope of negation to address the polarity shift problem. Negations may occur in two forms: morphological negations, where word roots are modified with a negating prefix (e.g., “dis-”, “non-”, or “un-”) or suffix (e.g., “less”), and syntactic negation, where clauses are negated using explicitly negating words or other syntactic patterns that imply negative semantics. Only explicit negations are considered while detecting negation scope. Ikeda et al. [3] proposes a machine learning based method that models the polarity-shifters. The model can be trained in two different ways: word-wise and sentence-wise. While the word-wise learning focuses on the prediction of polarity shifts, the sentence-wise learning focuses more on the prediction of sentence polarities. The improvement over other methods is significant when a limited amount of training data is available. New feature sets can be explored to improve the performance of the system. A. Kennedy and D. Inkpen [4] proposed a system where the effect of valence shifter on classifying the reviews is examined. There are three types of valence shifters: negations, intensifiers, and diminishers. Negations are used to reverse the semantic polarity of a particular term, while intensifiers and diminishers are used to increase and decrease, respectively, the degree to which a term is positive or negative. It depends on external dictionary to detect the negations in the text. Two dictionaries used are General Inquirer and Choose the Right Word(CTRW). This increases the time of computation thereby reducing overall performance. S. Li and C. Huang [5] proposed a model where negation and contrast transition were considered as structures that shift the polarity of a text. First, sentences are classified into sentiment reversed and non-reversed parts. Then, represent them as two different bags-of-words. Third, present three general strategies (remove, shift and joint) to do classification with two-bag-of-words modelling. Many other structures that reverse the sentiment polarity can be incorporated to improve the efficiency and performance of the system. S. Li et al. [6] proposed a method used for document level sentiment classification. It first divides the sentences in the document into polarity shifted and polarity unshifted sentences using a polarity shift detector and then polarity classification is done. The polarity shifting training data is created automatically due to which the data contains noise which degrades the overall performance of the system. T. Wilson, J. Wiebe and P. Hoffmann [7] proposed a system which considers contextual polarity for sentiment classification. The contextual polarity of the phrase in which a particular instance of a word appears may be quite different from the word’s prior polarity. Rui Xiaa et al. [8] proposed a three-stage model, namely Polarity Shift Detection, Elimination and Ensemble (PSDEE), to address polarity shift for document-level sentiment classification. First the polarity shift is detected using various techniques, then the negations are eliminated and finally the polarity shifted sentences are separated in different categories. Nilam V. Kolekar [9] proposed a system for performing sentence level and phrase level sentiment analysis. It first segments the compound sentences and then parsing is done to identify negation polarity shifters, if negation is found then the system is trained to remove and modify negations to address the polarity shift problem. It is a dictionary based approach which uses WordNet 2.1 to get the antonyms of words in polarity shifted sentences.

B. Types of feature sets

Basant Agarwala & Namita Mittal [10] proposed a model for sentiment analysis where initially various features are extracted such as unigrams, bi-grams and dependency features from the text. In addition, new bi-tagged features are also extracted that conform to predefined part-of-speech patterns. Furthermore, various composite features are created using these features. Information gain (IG) and minimum redundancy maximum relevancy (mRMR) feature selection methods are used to eliminate the noisy and irrelevant features from the feature vector. Experimental results show that composite features created from prominent features of unigram and bi-tagged features perform better than other features for sentiment classification. mRMR is a better feature selection method as compared with IG for sentiment classification. Rui Xia et al. [11] proposed techniques for feature extraction where two types of feature sets are designed for sentiment classification, namely the part-of-speech based feature sets and the word-relation based feature sets. POS information is supposed to be a significant indicator of sentiment expression. Word relation features such as higher order n-grams and word dependency relations, have been widely employed in text representation. Farhan Hassan Khan et al. [12] proposed a method in which POS such as adjectives, adverbs, nouns and verbs were used as feature sets for feature extraction. Fersini E. et al. [13] proposes the use of various expressive signals to better capture the sentiment orientation of the messages. Several valuable expressive forms explored are (1) adjectives, (2) emoticon, emphatic and onomatopoeic expressions and (3) expressive lengthening. Jalilvand, Abbas, and Naomie Salim [14] proposed a method in which feature unionization was applied on repetitive features. It results in dimension reduction which accounts for decreased computation time and an increased classification accuracy. Using union operator, this approach combines several features to construct a more informative feature. María del Pilar Salas-Zárate, et al. [15] proposed a method in which polarity of the features in each document is calculated by taking into account the words from around the linguistic expression of the feature. These words are obtained by using the ‘N_GRAM After’, ‘N_GRAM Before’, ‘N_GRAM Around’ and ‘All_Phrase’ methods.

C. Types of classifiers

Table I shows the various types of classifiers used in for sentiment analysis in literature.

III. PROPOSED SYSTEM

Web technology has grown to a great extent which led to an enormous increase in data present on the internet. A lot of data is generated daily by different users. Several reviews are available expressing the user's experience of the services provided by many websites. Extracting useful information from such enormous amount of data is very important. To automatically extract information and opinion from the reviews, sentiment analysis is performed. Several techniques are used for sentiment analysis as discussed in previous section. This paper adopts the technique proposed in [1] with modifications to obtain better classification accuracy. The phases in the proposed system are explained as follows:

1.Data collection and pre-processing: This phase includes dataset selection to implement the method in training and testing phases. After dataset is selected, the data will be cleaned to remove noise such as stop words which are not necessary for sentiment analysis. Every sentence of a review is checked for subjectivity and only the sentences having subjectivity greater than 0.5 are considered for sentiment analysis. This process eliminates all the objective sentences from reviews thus reducing the length of reviews or completely eliminating some reviews having only objective sentences. This will increase the efficiency of the resulting system.

2.Reversed data creation: In this phase the reversed reviews will be created for the dual training and dual prediction phase. After the reversed data creation, the dataset will be separated into original reviews and reversed review set.

3.Polarity calculation: Polarity calculation will be performed on the original and reversed reviews to obtain positive and negative polarity reviews. Feature extraction (extracting words in the review) will be performed on the reviews obtained which will serve as the training data. 3/4th data will be considered for training and 1/4th for testing purpose in the next phase.

4.Enhanced dual training and dual prediction: In this phase the training and testing data will be used to perform dual training and dual prediction. These techniques perform classification on the original as well as reversed reviews to address the polarity shift problem. Various classifiers such as Naïve Bayes, Support Vector Machines, Maximum Entropy and Voted classifier will be applied for the classification of reviews and calculating the accuracy of the classification along with other parameters for measuring the efficiency of the proposed system.

5.Data visualization: This phase will graphically represent the measuring parameters used for different classifiers in the dual training and prediction phase. This will help in visualization of the results obtained from the proposed system and to measure its efficiency. Fig. 1 shows the various phases of the proposed system.

IV. EXPECTED OUTCOME

This paper focuses on sentiment analysis of public reviews and data visualization to assist businesses to take business decisions more effectively and make improvements to their products and services. In the sentiment analysis task, the expected outcomes are the polarity class of reviews in the dataset used for training and testing. The outcomes of the proposed system are as follows:

1. A set of reversed reviews obtained using the subjective reviews based on the subjectivity test. This will reduce the number of reviews considered for sentiment analysis and it will remove the objective reviews which are not necessary for sentiment analysis.

2. A set of polarity classes of the reviews classified using different enhanced classifiers in the dual training stage. For enhancing the classifier performance, different classification algorithms will be implemented such as Naïve Bayes, Support Vector Machine (SVM), Maximum Entropy and Voted classifier in this stage.

3. A set of polarity classes of the reviews in the testing dataset will be obtained in the dual prediction stage. The testing dataset will contain previously unseen reviews to test the efficiency of the classifier trained in the training stage.

4. A graphical representation of the results obtained. The parameters considered for measuring the efficiency of the proposed system are accuracy, precision, recall and f-measure. The dataset will be separated into balanced and unbalanced reviews where balanced dataset will contain same number of positive and negative reviews and unbalanced dataset will contain unequal number of positive and negative reviews. The resultant parameters will be graphically represented to visually compare efficiency of the classifiers. This graphical representation will help the decision makers in a business to take effective and efficient business decisions. It will help the businesses to understand the most efficient classifier which can be used for sentiment analysis.

Retrieval Number D0926055419/2019©BEIESP
DOI: 10.35940/ijies.D0926.095619
Journal Website: www.ijies.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication
© Copyright: All rights reserved.

TABLE I. Types of classifiers

Paper	Classification algorithms	Future Scope
Ensemble of feature sets and classification algorithms for sentiment classification [11]	NB, SVM, MaxEnt Ensemble of classification algorithms and Feature sets individually and in combination using fixed rules, meta classifier, weighted combination.	Hybrid generative/discriminative model for sentiment classification. Hybrid NB/MaxEnt model for text classification.
A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet [12]	Support Vector Machine: A combination of lexicon based approach along with machine learning is employed. It revises the sentiment scores in SentiWordNet using Information Gain and Cosine Similarity	More mathematical models, such as jaccard similarity, correlation coefficient, may be evaluated for association strength computation. We also intend to investigate the sentiment analysis performance via polarity shifting and transfer learning approaches based on the proposed sentiment lexicons
A Multiobjective Weighted Voting Ensemble Classifier Based on Differential Evolution Algorithm for Text Sentiment Classification [16]	Bayesian logistic regression, Naïve Bayes, linear discriminant analysis, logistic regression, and support vector machines. An efficient ensemble classification scheme is used which is based on an optimization technique using a multi objective differential evolution algorithm to efficiently combine the classification algorithms	The performance of metaheuristic methods and evaluation metrics in dynamic classifier selection should be taken into consideration, namely in conjunction with the multi objective differential evolution-based weighted voting scheme. Second, identifying an efficient representation scheme for text documents should also be a priority in additional research.
Expressive signals in social media languages to improve polarity detection [13]	Multinomial Naïve Bayes, Decision Tree, Support Vector Machines, Bayesian Networks and Ensemble methods such as Majority Voting & Bayesian Model Averaging	Considering additional expressive signals such as repeated exclamation marks. Decoding of emojis and symbols which are platform dependent such as those specific to Android or iPhone. Detection of sentiments from hashtags. Other work may include detection of irony and sarcasm in the text.
Feature unionization: A novel approach for dimension reduction [14]	Support Vector Machines, Naïve Bayes, K-Nearest neighbours(KNN) For feature selection Information gain(IG) and Chi-square(CHI2) are used	Other feature selection and classification algorithms can be investigated.(No specific algorithm mentioned)
Recognizing emotions in text using ensemble of classifiers [17]	Naïve Bayes and Maximum Entropy. It also uses a knowledge based tool. Majority voting is used as an ensemble technique.	An extension of system's knowledge base could be made by adding more lexical resources, such as General Inquirer and SentiWordNet resources. Another extension may concern the associations of classifier weights in the voting approach, which could represent to some degree their classification confidence and the strength of the emotion specified.
Sentiment Analysis of Twitter Data: A Survey of Techniques [18]	Naïve Bayes, Maximum Entropy and Support Vector Machine. Lexicon based method is also evaluated for sentiment analysis.	Machine leaning and lexicon based methods can be combined to get better results.
Short text opinion detection using ensemble of classifiers and semantic indexing [19]	Bernoulli Naive Bayes, Multinomial Naive Bayes (NB-M), Gaussian Naive Bayes (NB-G), Linear Support Vector Machines (SVM-L), Radial Support Vector Machines (SVM-R), Polynomial Support Vector Machines (SVM-P), Decision trees (C4.5), K-nearest neighbours (k-NN), Boosted C4.5, Logistic regression. Weighted voting ensemble technique is applied on the above classifiers.	(1) parallelize the most expensive computing process to speed up the model selection and training stages, and (2) merge the proposed method with the one which combines co-training applied to a multi-classifier system.
Text mining for sentiment analysis of Twitter data [20]	Decision tree J48 and Naïve Bayes. Weka tool is used for implementation of decision tree.	Application of the proposed approach to tweet analysis based on different data mining tools.
Using unsupervised information to improve semi-supervised tweet sentiment classification [21]	A version of the Consensus between Classification and Clustering Ensembles(C3E) algorithm is integrated into a Semi-Supervised learning (SSL) framework to perform tweet sentiment classification. The classification power of Support Vector Machines (SVMs), constructed from labelled data is combined with the information provided by the pair-wise similarities between unlabelled data points. The proposed framework is based on an iterative Self-training approach guided by the predictions made by the framework of C3E.	(1) Selection of features should be done on-the-fly as an intrinsic part of SSL. (2) Sarcasm and Irony must be considered while classification. (3) Exploration of feature selection, dimensionality reduction, structure based features and language resources must be done.
Web Service SWePT: A Hybrid Opinion Mining Approach [22]	It combines the Sequential Minimal Optimization (SMO) machine learning algorithm with the use of features obtained by an affective lexicon in Mexican Spanish and a corpus.	1.Lexicons based on a specific domain need be created with the affective word of specific domains. 2.The linguistic rules must be improved for better functioning of the feature extraction module and the verb tense should be taken into account. 3.Discourse analysis should be considered. 4.Some technique to identify irony in the text should be incorporated.

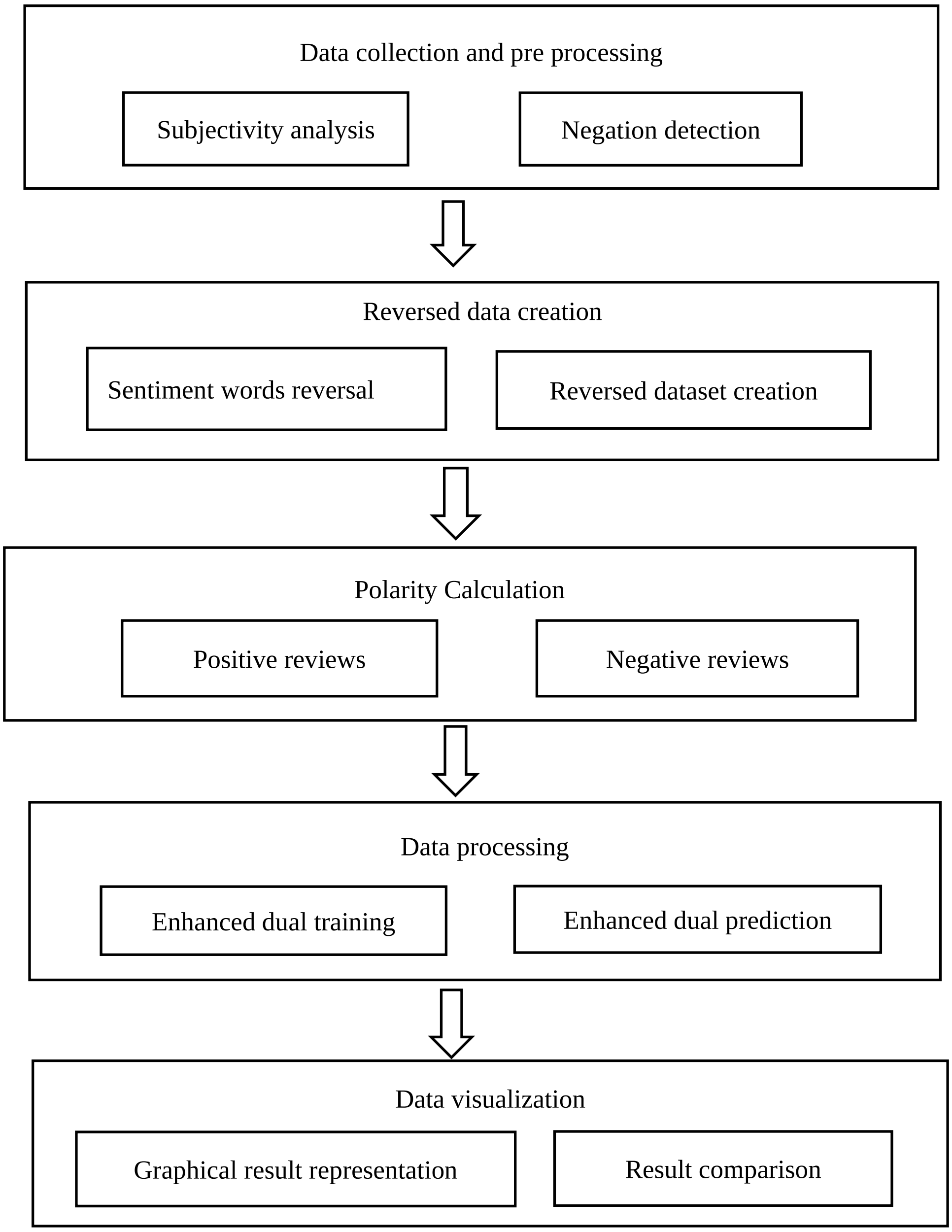

Figure 1. Proposed system

V. APPLICATIONS OF BIG DATA IN SENTIMENT ANALYISIS

Sentiment analysis can be applied on big data as huge amount of data is generated every second on the internet which includes the posts on Facebook, tweets on Twitter and reviews on online shopping sites. Analysis of such vast amount of data to obtain opinions expressed is very important. Real time applications and websites generate big data which needs to be analysed and interpreted to obtain meaningful information. Sentiment analysis needs to be performed on big data collected from websites and social networking sites. A real world application of big data to perform sentiment analysis is in the Wimbledon championships [23]. Sentiment analysis is performed on tennis matches in real time. It analyses the sentiments of people based on the posts from social networks and provides the predictions in real time. Other application of big data in sentiment analysis is to provide information on opinions expressed on government policies by people on the blogs, social networking sites and other sources. Sentiment analysis is also performed by companies to analyse the sentiments of its customers based on the customer service interactions and provide improved customer support. The employee conversations can be analysed to know whether they are happy with the organization or facing certain difficulties. News and public data also provide various topics on which sentiment analysis can be performed to obtain useful information. All such sources have huge amount of data which requires application of big data techniques to perform sentiment analysis.

Retrieval Number D0926055419/2019©BEIESP
DOI: 10.35940/ijies.D0926.095619
Journal Website: www.ijies.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication
© Copyright: All rights reserved.

VI. CONCLUSION

This paper focuses on creation of a system for performing sentiment analysis in an efficient manner by modifying the existing system to improve the classification accuracy. The reviews will be classified into positive, negative or neutral classes and graphically represented using various visualization techniques to obtain useful information from the opinion expressed in the reviews. This system will be useful in classification of large amount of online reviews available in various online sale websites. It will be helpful for users shopping online as it will provide an overall opinion about a product or service.

REFERENCES

1. Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li, Dual Sentiment Analysis: Considering Two Sides of One Review, *IEEE Transactions on Knowledge and Data Engineering*, 2015
2. I. Councill, R. MaDonald, and L. Velikovich, What's Great and What's Not: Learning to Classify the Scope of Negation for Improved Sentiment Analysis, *Proceedings of the Workshop on negation and speculation in natural language processing*, pp. 51-59, 2010
3. D. Ikeda, H. Takamura, L. Ratinov, and M. Okumura, Learning to Shift the Polarity of Words for Sentiment Classification, *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2008
4. A. Kennedy and D. Inkpen, Sentiment classification of movie reviews using contextual valence shifters, *Computational Intelligence*, *vol. 22*, pp. 110–125, 2006
5. S. Li and C. Huang, Sentiment classification considering negation and contrast transition, *Proceedings of the Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 2009.
6. S. Li, S. Lee, Y. Chen, C. Huang and G. Zhou, Sentiment Classification and Polarity Shifting, *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2010.
7. T. Wilson, J. Wiebe, and P. Hoffmann, Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis, *Computational Linguistics, vol. 35, no. 3*, pp. 399-433, 2009.
8. Rui Xiaa, Feng Xu , Jianfei Yua, Yong Qi , Erik Cambria , Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis, *Information Processing and Management* 52(2016)
9. Nilam V. Kolekar, Sentiment Analysis and classification by considering negation polarity shifter and opinion summarization for product reviews, *International Journal in IT and Engineering, Vol.04 Issue-05, (May, 2016)*
10. Basant Agarwal and Namita Mittal, Prominent Feature Extraction for Sentiment Analysis, *Springer*, 2016
11. Rui Xia, Chengqing Zong, and Shoushan Li, Ensemble of feature sets and classification algorithms for sentiment classification, *Information Sciences*181.6 (2011): 1138-1152.
12. Farhan Hassan Khan, Usman Qamar, and Saba Bashir, A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet, *Knowledge and Information Systems* (2016): 1-22.
13. Fersini E., E. Messina, and F. A. Pozzi, Expressive signals in social media languages to improve polarity detection, *Information Processing & Management* 52.1 (2016): 20-35.
14. Jalilvand, Abbas and Naomie Salim, Feature unionization: A novel approach for dimension reduction, *Applied Soft Computing* (2016).
15. María del Pilar Salas-Zárate, et al., Feature-based opinion mining in financial news: An ontology-driven approach, *Journal of Information Science* (2016)
16. Onan, Aytuğ, Serdar Korukoğlu, and Hasan Bulut. A Multiobjective Weighted Voting Ensemble Classifier Based on Differential Evolution Algorithm for Text Sentiment Classification, *Expert Systems with Applications* (2016)
17. Perikos, Isidoros, and Ioannis Hatzilygeroudis, Recognizing emotions in text using ensemble of classifiers, *Engineering Applications of Artificial Intelligence* 51 (2016): 191-201.
18. Vishal Kharde and Sheetal Sonawane, Sentiment Analysis of Twitter Data: A Survey of Techniques, *International Journal of Computer Applications* (0975 – 8887) Volume 139 – No.11, April 2016.
19. Johannes V. Lochter, et al. ,Short text opinion detection using ensemble of classifiers and semantic indexing, *Expert Systems with Applications* 62 (2016): 243-249.
20. Wakade, Shruti, et al., Text mining for sentiment analysis of Twitter data, *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.
21. Nádia Félix Felipe da Silva, et al., Using unsupervised information to improve semi-supervised tweet sentiment classification, *Information Sciences* 355 (2016): 348-365.
22. [22] Yolanda Raquel Baca-Gomez, et al., Web Service SWePT: A Hybrid Opinion Mining Approach, *Journal of Universal Computer Science* 22.5 (2016):671-690.
23. [23] https://www.simplilearn.com/big-data-applications-in-industries-article accessed on 29-01-17,18:32.

Retrieval Number D0926055419/2019©BEIESP
DOI: 10.35940/ijies.D0926.095619
Journal Website: www.ijies.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication
© Copyright: All rights reserved.