DOI: http://dx.doi.org/10.26483/ijarcs.v8i9.5174

ISSN No. 0976-5697

Volume 8, No. 9, November-December 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

SENTIMENT POLARITY WITH SENTIWORDNET AND MACHINE LEARNING CLASSIFIERS

Akshaya R. Garje
Department of Computer Science & IT
Dr. Babasaheb Ambedkar Marathawada University
Aurangabad, India

Karbhari V. Kale
Department of Computer Science & IT
Dr. Babasaheb Ambedkar Marathawada University
Aurangabad, India

Abstract: Sentiment classification is concerned with using automated methods for predicting the orientation of subjective content on textual content documents, with applications on some of areas consisting of recommended and advertising and marketing systems, customer intelligence and information retrieval. SentiWordNet is not anything however an opinion lexicon derived from the WordNet database in which each term is related to numerical scores indicating their sentiments. This research offers the results of making use of the SentiWordNet lexical resource to the hassle of computerized sentiment classification on labelled dataset. Our method incorporates counting positive and negative scores to decide sentiment orientation, and an improvement is provided by means of constructing a information set of applicable features using SentiWordNet as supply, and additionally implemented to a machine learning classifier. We compared the accuracies results obtained with SentiWordNet and Machine Learning Classifiers.

Keywords: Sentiment Analysis, SentiWordNet, Naïve Bayes, Support Vector Machines.

I. INTRODUCTION

Sentiment Analysis (SA) or Opinion Mining (OM) is the computational have a look at of people's opinions, attitudes and feelings in the direction of an entity. The entity can represent people, events or topics that are maximum probable to be covered with the aid of evaluations. The two expressions Sentiment Analysis or Opinion Mining are interchangeable which expresses a mutual meaning. However, some researchers stated that Opinion Mining and Sentiment Analysis have slightly different notions. Opinion Mining analyzes human's reviews for an entity Sentiment Analysis the sentiment expressed is in a given text is recognized then analyzes it. Therefore, the target of SA is to find opinions, identify the sentiments they express, and then classify their polarity as shown in Fig. 1.

Sentiment Analysis is a classification process as shown in Fig. 1. There are three sentiment analysis levels of classification which are divided as Document, Sentence and Aspect based. First, Document based level of Sentiment Analysis which emphasizes on classification of an opinion document which gives a positive or negative sentiment. It considers the whole document a basic information unit. Second, Sentence-level Sentiment Analysis focuses to classify sentiment expressed in each sentence. At first, it is necessary to identify if the sentence is subjective or not. If the sentence is subjective, Sentence-level Sentiment Analysis will determine whether the sentence expresses positive or negative opinions. Wilson et al. have pointed out that sentiment expressions are not necessarily subjective in nature. However, document and sentence level classifications are not much different from each other as sentences are just meant to be short documents. Aspect level is needed because classification at document or sentence level, each property is not considered and it will not give more deep information about the entity. Third, Aspect-level Sentiment Analysis focuses to classify the sentiment with respect to the specific aspects of entities. The first step is to identify the entities and their aspects. The opinion holders can give different opinions for different aspects of the same entity. [1].

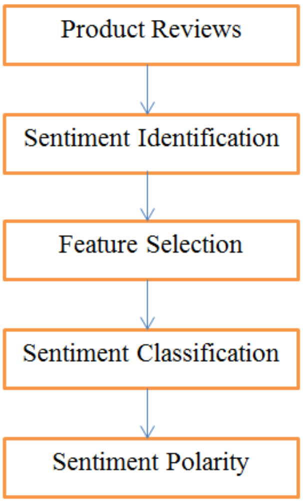

Figure 1. Sentiment Polarity Process.

Sentiment analysis influences users to classify whether the information about the product is satisfactory or not before they acquire it. Marketers and firms use this analysis to understand about their products or services in such a way that it can be offered as per the user's needs. There are two types of machine learning techniques which are generally used for sentiment analysis, one is unsupervised and the other is supervised. Unsupervised learning does not consist of a category and they do not provide with the correct targets at all and therefore conduct clustering. Supervised learning is based on labeled dataset and thus the labels are provided to the model during the process. These labeled dataset are trained to produce reasonable outputs when encountered during decision- making [2].

II. LITERATURE SURVEY

Bruno Ohana, Brendan Tierney, assesses the use of SentiWordNet. The task of document level sentiment classification is accomplished the usage of the Polarity data set of movie reviews. SentiWordNet scores have been calculated as high-quality and bad terms on each document, and were used to determine sentiment orientation through assigning the document to the elegance with the highest score. This technique yielded a standard accuracy of 65.85% [3].

Bo Pang and Lillian Lee, Shivakumar Vaithyanathan, as in initial unigram results of their research, as an entire, the machine learning algorithms certainly exceeds the random-desire baseline of 50%. They additionally handily beat the two unigram baselines and carry out nicely in evaluation to the 69% baseline performed by restricted get entry to to the check-information information, although the development in the case of SVMs was now not so big [4].

Bhumika M. Jadav, Vimalkumar B. Vaghela, sentiment analysis was done for movie Review, Twitter and Gold dataset using optimized SVM. Here, Optimized Support Vector Machine towards Support Vector Machine and Naïve Bayes classifier were compared. Modifying hyper parameter value of RBF kernel SVM gives better result compare to Support Vector Machine and Naïve Bayes algorithm. Hyper parameters are soft margin constant C and Gamma γ. Proposed approach was found to be optimal value for hyper parameter which classified dataset with more accuracy than existing system [5].

Prof. (Dr.) N. S. Chandolikar, Based on the content material, textual content class is the approach of classifying textual content documents into distinctive classes. Methods for textual content which includes Support Vector Machine, Naïve Bayes, K-Nearest Neighborhood and Decision tree classifiers were used for training purpose. This research indented to cope with the sentiment categorized textual content type process and the classifiers were used on various overall performance measurement.

Milan Gaonkar, Prof. Amit Patil, the research categorizes a given tweet/paragraph whether it's far of Positive [True positive, False positive] or Negative [True negative, False negative] sentiment. In their paper a new method become proposed that makes use of lexicon database to assign every word in a textual content a price known as valence. The valence is how a each word is affecting the entire sentence wherein it is used. Every word in a sentence has its very own electricity and it tries to influence the overall semantic of the sentence. Higher the fee of valance of a phrase in the sentence, the extra influential it is. The approach proposed used lexicon based technique in addition to machine primarily based mastering. They used AFINN lexicon database to assign valance to words and Support Vector Machine (SVM), Naïve Bayes classifier (NB) machine studying algorithms for education and checking out the model [7].

Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, Sweta Tiwari, focused on sentiment focused web crawling framework to facilitate the quick discovery of sentimental contents of movie reviews and hotel reviews and analysis of the same. They used statistical strategies to capture elements of subjective fashion and the sentence polarity. Their studies elaborately discusses two supervised device gaining knowledge of algorithms: K-Nearest Neighbor (K-NN) and Naïve Bayes' and compares their typical accuracy, precisions as well as don't forget values. It become visible that in case of film evaluations Naïve Bayes' gave a ways better results than K-NN however for hotel evaluations those algorithms gave lesser, almost equal accuracies [8]

III. METHODOLOGY

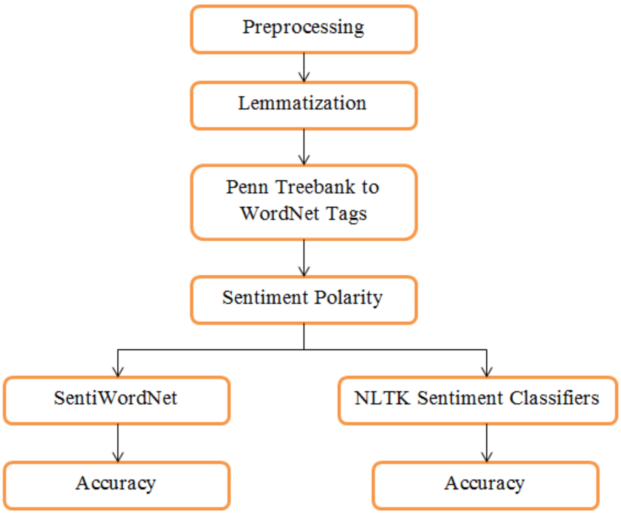

Figure 2. Proposed Methodology

A. *Preprocessing*

First, Pre-processing of the data is the process of cleaning and preparing the text for classification. Online texts contain usually abundant noise and uninformative parts such as HTML tags, scripts and advertisements. In addition of on words level, many words in the text do not have an impact on the general orientation of it.

Keeping those words makes the dimensionality of the problem high and hence the classification more difficult since each word in the text was treated as one dimension. Here is the hypothesis of having the data properly pre-processed that was to reduce the noise in the text should help improve the performance of the classifier and speed up the classification process, thus aiding in real time sentiment analysis [9].

B. *Lemmatizer and PennTreebank tags to simple Wordnet tags conversion*

Lemmatizers can also be used in text categorization to treat different variation of the same root words as one for statistical counting; for instance, to bring verbs to the infinitive form, and nouns to the singular and masculine form [10].

After lemmatizing, the dataset has to convert Penn Treebank tags to Simple WordNet tags.

C. *SentiWordNet Polarity*

Sentiment polarity is calculated the use of SentiWordnet. SENTIWORDNET 3.0, a lexical resource explicitly devised for assisting sentiment classification and opinion mining programs. SENTIWORDNET 3.0 is an progressed version of SENTIWORDNET 1.0, a lexical aid publicly available for studies functions, now currently licensed to greater than 300 studies businesses and used in a ramification of research projects international. Both SENTIWORDNET 1.0 and 3.0 are the result of routinely annotating all WORDNET synsets according to their levels of positivity, negativity, and neutrality. SENTIWORDNET 1.0 and 3.0 vary in the variations of WORDNET which they annotate (WORDNET 2.0 and 3.0, respectively), in the set of rules used for automatically annotating WORDNET, which now consists of (moreover to the preceding semi-supervised gaining knowledge of step) a random-walk step for refining the rankings [11].

D. Machine learning Classifiers.

Supervised machine mastering is the search for algorithms that cause from externally provided instances to provide widespread hypotheses, which then make predictions approximately further instances. In other phrases, the goal of supervised gaining knowledge of is to build a concise version of the distribution of class labels in terms of predictor capabilities. We have used two classifiers here that are Naïve Bayes and Support Vector Machine.

*1) Naive Bayesian networks (NB):*Naive Bayesian networks (NB) are very simple Bayesian networks that are composed of directed acyclic graphs with handiest one determine and several kids with a sturdy assumption of independence among toddler nodes inside the context of their parent (Good, 1950).Thus, the independence model (Naive Bayes) is primarily based on estimating (Nilsson, 1965). Comparing those two probabilities, the larger opportunity shows that the magnificence label value that is more likely to be the real label (if R>1: predict i else are expecting j). Cestnik et al (1987) first used the Naive Bayes in ML community. Since the Bayes class algorithm makes use of a product operation to compute the chances P(X, i), it's far specifically vulnerable to being unduly impacted via probabilities of 0. This may be prevented through the use of Laplace estimator or m-esimate, by means of adding one to all numerators and adding the variety of added ones to the denominator (Cestnik, 1990).

2) Support Vector Machines (SVM): Support Vector Machines (SVMs) are the newest supervised machine learning technique (Vapnik, 1995). An excellent survey of SVMs can be found in (Burges, 1998), and a more recent book is by (Cristianini & Shawe-Taylor, 2000). Thus, in this take a look at aside from a short description of SVMs we can seek advice from a few greater latest works and the landmark that were published before these works. SVMs revolve around the belief of a “margin”—either side of a hyperplane that separates information classes. Maximizing the margin and thereby growing the most important possible distance between the setting apart hyperplane and the instances on either aspect of it's been established to lessen an upper certain on the anticipated generalization mistakes [12].

IV. EXPERIMENT AND RESULTS

A. Data Description

Here, in this research, sentiment labelled dataset is used. Sentences labelled reviews with positive or negative sentiment are present in the dataset . It is in format of sentence and its score. Score is either 1 (for positive) or 0 (for negative) .The sentences come from three different websites/fields: imdb.com, amazon.com and yelp.com. For each website, there exist 500 positive and 500 negative sentences. Those were selected randomly for larger datasets of reviews.

We attempted to select sentences that have a clearly positive or negative connotation; the goal was for no neutral sentences to be selected.

B. Proposed methodology

Here, we have 3 website dataset, which are labelled sentences. All the three are at first preprocessed. In this preprocessing, all the data is cleaned by removing HTML tags and all other unwanted things. Then these datasets are lemmatized by WordNetLemmatizer ( ).The dataset is POS (Part of Speech) tagged. Here, there is a conversion between Penn Treebank tags to Simple WordNet tags. This conversion is performed on the respective data. For sentiment polarity; comparison of accuracies is performed on these dataset for both SentiWordNet (SWN) and NLTK classifiers. Two classifiers i.e. Naïve Bayes as well as Support Vector Machine are used as Machine Learning technique. One new addition is using bigrams. Bigrams are pairs of consecutive words. The accuracies results are shown in table below:

Table I. Accuracies of SentiWordnet, SVM and Naïve Bayes on different datasets.

Dataset	***Amazon***	***IMDB***	***Yelp***
Methods			
SWN	68.0%	68.6%	70.5%
SVM (Linear SVC)	79.0%	73.6%	82.5%
NB (Multinomial)	80.0%	75.1%	77.5%

V. CONCLUSION

In this research, three different website data reviews are used for training and testing the best between SentiWordNet and machine learning classifier methods. Using bigrams instead of unigrams is a trick for improving performance in text classification.

As the table 1 shows us, the accuracies obtained using SentiWordNet are less than as compared to both machine learning classifiers. Hence, it can be said that machine learning classifiers with approximately 80-83% accuracies..

VI. ACKNOWLEDGMENT

The Authors acknowledge and extend our gratitude to UGC who have supported for development of UGC SAP (II) DRS Phase-I F.No.-3-42/2009 to Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. Thanks to research supervisor for technical discussion and support at every stage of implementation.

VII. REFERENCES

[1] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams Engineering Journal 5.4 (2014): 1093-1113.

[2] Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." Contemporary computing (IC3), 2014 seventh international conference on. IEEE, 2014.

[3] Ohana, Bruno, and Brendan Tierney. "Sentiment classification of reviews using SentiWordNet." 9th. it & t conference. 2009.

[4] 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.

[5] Jadav, Bhumika M., and Vimalkumar B. Vaghela. "Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis." International Journal of Computer Applications 146.13 (2016).

[6] Prof. (Dr.) N. S. Chandolikar, ‘Performance evaluation of classifiers for sentiment labeled text Dataset’ International

Journal of Scientific Engineering and Applied Science (IJSEAS), Volume-1, Issue-7, October 2015, ISSN: 2395-3470.

[7] Milan Gaonkar1, Prof. Amit Patil,' Sentiment Classification Using Product Reviews', IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727 PP 69-73.

[8] Dey, Lopamudra, et al. "Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier." arXiv preprint arXiv:1610.09982 (2016).

[9] Haddi, Emma, Xiaohui Liu, and Yong Shi. "The role of text pre-processing in sentiment analysis." Procedia Computer Science 17 (2013): 26-32.

[10] Bonatti, Rogerio, et al. "Effect of Part-of-Speech and Lemmatization Filtering in Email Classification for Automatic Reply." AAAI Workshop: Knowledge Extraction from Text. 2016.

[11] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." LREC. Vol. 10. 2010.

[12] Kotsiantis, Sotiris B., I. Zaharakis and P. Pintelas. "Supervised machine learning: A review of classification techniques."(2007):3-24.