ISSN:0975-9646
G. Vaitheeswaran et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (1) , 2016, 306-311

Combining Lexicon and Machine Learning Method to Enhance the Accuracy of Sentiment Analysis on Big Data

[1]G. Vaitheeswaran, [2]Dr. L. Arockiam

[1]Research scholar, [2]Associate Professor,
[1,2]Department of Computer Science,
[1,2] St. Joseph's College (Autonomous), Tiruchirappalli—620 002.

Abstract : During the day-to-day life, our opinions and opinions of others toward a product or event play a vital role in the decision-making process. In recent times, the outburst of social media over the web has an abundant impact on an individual's and the organization's decision-making process about certain content. Twitter which is a leading micro blogging website allows the people to post their opinions, state of mind, or status toward products or any events such as politics and sports. As more number of people lively connected to the twitter, more data are generated by the people which lead to the big data. Sentiment analysis performs an essential role to extract the content of the tweets. This research work consists of two components: a lexicon builder and a sentiment classifier. In addition the contextual valence shifter was used to evaluate the context sentiments and to solve the context-dependent problem. The combination of these two components will produce better accuracy.

Keywords: Big data, sentiment analysis, hybrid approach.

1. INTRODUCTION

The rapid growth of the Internet and online activities (such as clickstreams, blogging and microblogging, social media communications, e-commerce, online transactions, ticket booking, surveillances, conferencing, chatting) has enabled the research and industrial community to extract, transform, load, and analyze very huge amount of structured and unstructured data, which is referred to Big Data. Data can be analyzed using a combination of data mining, text mining, web mining, and natural language processing techniques. The enormous amount of data related to customer opinions/reviews is quite difficult to analyze, and it needs extant approaches to get a generalized opinion summary. Various web resources, news reports, e-commerce web sites, social networks (such as YouTube, Facebook, Twitter, Pinterest), blogs and forums help to express opinions, which can be utilized to understand the opinions of the general public and consumers on social events, political movements, company strategies, marketing campaigns, product preferences, and monitoring reputations [15]. Research communities, academicians, and industrialists have been working thoroughly on sentiment analysis for the past three and a half decade to achieve these tasks.

Sentiment analysis (SA) is a computational study of opinions, sentiments, emotions, and attitude which are expressed in texts toward an entity [16]. Online media and social networking sites (SNS) are used to express and share public experiences in the form of product reviews, blogs, and discussion forums. Collectively, these media contain vastly unstructured data, the combination of data formats such as text, images, audios, videos, and animations that are useful in making public awareness for various issues. Online media affords the platform for broader sharing of ideas and boosting public for group discussions with open views. It delivers a better means to get a quick response and feedback on different Global issues and entities in the form of textual posts, news, images, and videos. Thus, it can be utilized to analyze peoples' opinions about learning the behaviors of consumer, patterns market, and trends of society [17]. Twitter has 320 million monthly active users and they posts 500 million tweets every day[1]; Facebook has 936 million daily and 1,440 million monthly active users[2] as of December, 2015. Thus, it helps to extract heterogeneous reviews posted by people from diverse societies for different purposes such as improvement of quality of products and services, and prediction of consumers' demand and needs. The sentiment found within critiques, feedback and comments, which provide fruitful information for many different purposes.

Twitter is a microblogging website which allows users to tweet not more than 140 characters. This short message contains a rich source for processing sentiment analysis. An usual tweet holds images, audios, videos, url, word variations, emoticons, hash tags, contextual texts, etc. This creates a problem to analyze the polarity of words. After preprocessing the tweets, the general sentiment analysis tasks will be performed using lexicon or machine learning-based approaches. In this work, we have used the lexicon-based approach to classify the tweets as positive or negative using the emoticon and the sentiment orientation of the words. We have proposed a new algorithm based on a hybrid approach to provide better accuracy. Measuring the depth of the sentiment, which mostly relies on the contextual word, is one of the major issues and challenges

[1]https://about.twitter.com/company
[2]http://www.socialbakers.com/statistics/facebook/

involved in the process of sentiment analysis. The contextual semantic orientation word will be used as a testing set in the classifier model to evaluate the performance of the proposed algorithm. The proposed algorithm has mainly focused on twitter conventions and contextual words. The twitter conventions have been explained in the section "proposed method".

The remaining article has been explained as follows: In this section, brief concepts on sentiment analysis approaches have been provided. In the literature review section, a review of some related work on sentiment analysis using hybrid approach has been presented briefly. The main objective of this proposed work has been provided in the Section 3. The methodological diagram and proposed algorithm for hybrid approach on sentiment analysis have been presented in the Section "proposed work." The article has been concluded in the Section "conclusion."

2. RELATED WORKS

According to Lei Zhang et al. [1], the lexicon-based approach gives high precision but low recall. To improve recall, the Pearson's chi-square test was used to identify the opinionated word not in the lexicon. Then, the SVM–based classifier is trained to assign polarities to the entities in the newly identified tweets. Instead of being labeled manually, the examples are given by the lexicon-based approach. The test data were used as the result of the chi-square's resulted data. The LMS method produced better accuracy of 85.4% than the existing baseline method. Later [2] in some other work, the Naïve Bayes technique gives a better result than the maximum entropy and support vector machine for unigram model. The Naïve Bayes classifier produced 88.2% accuracy. The accuracy was again improved by using the semantic analysis WordNet and produced 89.9%.The emoticons and context-dependent word were not considered in this work.

Considering the unigrams and bigrams features together will produce better accuracy. Since the original lexicon was built with the unigrams pattern, but in such a case, it is difficult to find polarity information of the sentiment word with bigram patterns that has a negative word or an adverb that intensifies the meaning of the sentiment. According to Hanhoon Kang et al. [4], the unigrams patterns along with the bigrams pattern such as negative words and intensive adverbs were included in the lexicon which resulted the better accuracy.

The twitter convention languages were used as the features to evaluate the performance of the support vector machine classifier [5]. Another hybrid approach by Andrius Mudinas et al. [6] uses a sentiment lexicon constructed using public resources for initial sentiment detection. The sentiment words were used as features in the support vector machine classifier. The weight of such a feature is the sum of the sentiment value in the given review. The adjectives which are not present in the sentiment lexicon were calculated by its occurring frequencies. Cross-style setting method was used to compare the results. The software reviews are used as training set, and the performance was evaluated by applied the movie reviews as the test set or vice-versa. The hybrid approach achieved 82.30% accuracy.

Tweets are microblogging website, and the characters are limited to 140 characters. Nowadays, people are using emoticons or emoji to express their emotions, feelings, and state of mind. Instead of analyzing the sentiment words, analyzing the emoticons will produce good results. Most of the tweets contain emoticons to express sentiments. Geeta et al. [7] adjudged that it is not necessary to use external dictionaries or any other lexicons of polarized words to find the sentiment polarity in tweets. A training dataset was automatically generated by referring to the sentiment present in tweets containing emoticons. It is able to map all common expressions with new words, slangs, and errors.

Alistair Kennedy et al. [13] made a study on CVS about its characteristics and occurrence. The study has proposed two algorithm CVS and Term Counting and have compared the results of both. The comparison leads CVS with higher accuracy. Vo Ngoc Phu et al. [14] improved CVS algorithm and proposed a methodology for sentiment analysis which is a combination of CVS and Term Counting. By combining both processes, researchers have noticed an increase in the accuracy rate.

The author [3] presented a methodology, how to use the big data analysis method to process the sentiment analysis effectively. The author discussed the tools such as Hbase, Mahout, and Hadoop framework for sentiment analysis.

3. MOTIVATION AND OBJECTIVE

From the above literature review, most of the works have been carried out using the emoticon and context dependent separately to find the polarity of tweets. This motivated to build a new hybrid model based on combining both lexicon and machine learning methods to enhance the accuracy.

The objective of this work was to provide a novel hybrid-based approach using emoticon and contextual word identification to classify the sentiment words of tweets and also to establish a classifier model to produce better accuracy than the existing method.

4. PROPOSED METHOD

This section presents the proposed approach. Figure 1 shows an architectural overview of the proposed hybrid (Lexicon and Machine Learning)-based method. We will discuss the techniques in the following section before we discuss the proposed hybrid-based algorithm.

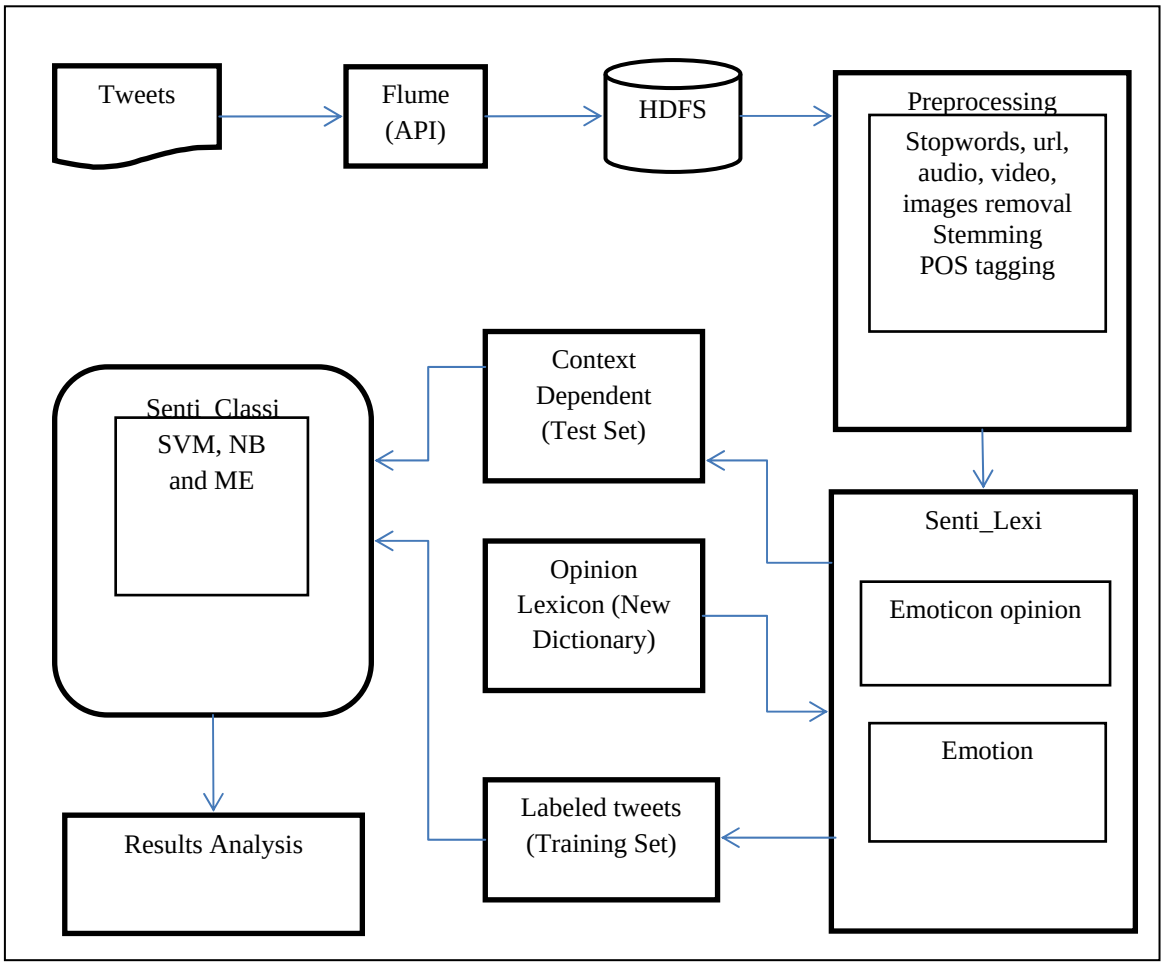

Figure 1. Methodological diagram for Lexicon- and Machine Learning-based approach

A. Tweets

The tweets have its own conventions. The following are some of the examples of tweets conventions.

1. Length of tweet—tweets are limited to 140 characters. This is different from other lexicons.
2. Hashtags (“#”) “#BigData” and “#Sentiment” are hash tags. These are used to organize tweets on particular topics.
3. URLs—the URLs are used to track the external sources. For example, “owl.li/6Mq1Q” is a URL shortened by the application Hootsuite.
4. Retweet (“RT”)—this is the easiest and most common way to share someone else’s content.
5. Emoticons (Smileys)—Smilies or emoticons are frequently used in tweets to express the user state of mind. For example ☺ (Happy), :-p, ☹(sad).
6. Colloquial expressions—most of the users express their state of mind in an abnormal way. For example, “This mobile awesommmeeeeeeeeeee”, “It is veryyyyyyyyy baddddddd”.
7. Reply to a user (@Username)—For example, have you registered for our world-class #**SuperPhoneLe1s**? Check out the details on @**Flipkart** http://www.flipkart.com/le-1s

B. Flume

Flume is a tool, which collects data from different sources to Hadoop’s HDFS. It ingests log data from multiple web servers into a centralized store (HDFS, HBase) in an efficient manner. It can be scaled horizontally. Flume acts as a channel to import huge volumes of event data produced by social networking sites such as Facebook and Twitter, and e-commerce websites such as Amazon and Flipkart into HDFS at a higher speed. Cloudera posted the Twitter–Flume configuration in GitHub repository [8].

C. HDFS

HDFS (Hadoop Distributed File System) is the primary storage system used by Hadoop applications. HDFS creates multiple replicas of data blocks and distributes them on compute nodes throughout a cluster to enable reliably, extremely rapid computations.

D. Preprocessing

Before processing sentiment analysis, preprocessing is the most important step to remove the noisy data. The most commonly used preprocessing techniques are stop words removal, url removal, lemmatization, and stemming. In

twitter, the noisy data such as retweets (duplicates which do not add any value) whose text starts with \RT" are to be eliminated. The punctuations are kept as people often express the sentiment with characters such as ":-)", ":-(".

E. Opinion lexicon

The lexicon-based approach depends on opinion or sentiment words that express positive, negative, and neutral sentiments. The most popularly used opinion lexicons in the sentiment analysis are Senticnet [10], SentiWordNet [9], NRC emotions [12], MPQA Subjectivity Lexicon [11]. Words that denote a desirable state (e.g., \happy" and \awesome") have a positive polarity, while words that denotes an undesirable state have a negative polarity (e.g., \sad" and \ugly"). Although opinion polarity normally applies to adjectives and adverbs, there are verb and noun opinion words as well. Researchers have compiled sets of opinion words and phrases for adjectives, adverbs, verbs, and nouns, respectively. Note that there are also many words whose polarities depend on the contexts in which they appear [1]. For example, "*The product is awesome but the price is unexpected*". The review contains the positive sentiment awesome, but the unexpected sentiment word toward the price denotes the negative polarity. The normal sentiment calculation will provide the polarity as neutral. To provide a solution to this problem, we included the Contextual Valence Shifter (CVS) [14] method to calculate such context opinion.

F. Proposed Senti_Lexi_Classi technique

After cleaning, we perform sentence segmentation, which separates the collected tweets into individual sentences. Afterward, we tokenize and perform, part of speech tagging (POS) for each sentence. Our proposed hybrid method contains two functions. There are as follows:

(i) Senti_Lexi() – to labeled the collected dataset as positive, negative, and neutral.
(ii) Senti_Classi() – to build a machine learning-based classifier.

For our work, we have to perform a comparative study between Naive Bayes, Support Vector Machine, and Maximum Entropy to find the outperformed classifier model. To find the polarity of negation and context-dependent text, we use ngrams (N<4) technique, as tweets contains 140 characters. To assign polarity value for the context-dependent text, we used the Contextual Valence Shifter (CVS) [] method. Unigrams are used to find the polarity of the emoticons. Later the labeled dataset obtained from the unigrams technique which is used as a training set to build the machine learning-based classifier. The labeled dataset obtained from the ngrams are used as test set to find the accuracy of the proposed model. Later the results of the training set and test set are to be compared. The results of the proposed model will be analyzed using the four parameters, namely accuracy, F-measure, precision, and recall which are discussed in the results analysis section.

Figure 2 represents the procedure of the hybrid model for sentiment classification. The procedure for labeling the preprocessed datasets is given in the Figure 3. Figure 4 represents the process of machine learning-based classifier to find the accuracy.

Procedure: Senti_Lexi_Classi /*Lexicon and Machine Learning*/
Input : Tweets
Output : Accuracy

1. Begin
2. Preprocess the collected tweets
    a. Stopwords removal
    b. Urls, audio, video, images removal
    c. Apply porter stemmer /*using Porter stemmer*/
    d. Perform part of speech (POS) /*using Stanford POS tagger*/
3. Senti_Lexi()
4. Print positive, negative and neutral tweets
5. Senti_Classi()
6. Print precision, recall, F-score and Accuracy
7. End

Figure 2. Procedure for Hybrid approach

Procedure: Senti_Lexi
Input : Tweets
Output : labeled tweets

1. Begin
2. Each tweet
3. Apply Unigrams /*Training Set*/
   Calculate positive, negative and neutral
4. Apply N-grams /*Test Set*/
   Calculate positive, negative and neutral
   Sum positive, negative and neutral value for each sentence
5. End
6. Count total positive, negative and neutral tweets
7. End

Figure 3. Procedure for Senti_Lexi

Procedure : Senti_Classi /* SVM classifier to measure accuracy*/
Input : Labeled positive, negative and neutral tweets
Output : Accuracy

1. Begin
2. Train model using classified tweets.
3. Build Clasifier Model
4. Test the Model using the newly identified opinionated tweets
5. Result analysis
6. End

Figure 4. Procedure for Senti_Classi

G. Results Analysis

In this section, we discuss the results obtained through Naïve Bayes, Maximum Entropy, and Support Vector Machine and compared their relative performances into four parameters, namely accuracy, F-measure, precision, and recall where,

Accuracy

Accuracy is computed by the following equation,

$$Accuracy = \frac{Tp + Tn}{Tp + Tn + Fp + Fn}$$

F-measure

F-measure is the harmonic mean of precision and recall.

$$F = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Precision

Precision is the fraction of the documents retrieved that are relevant to the user's information need.

$$Precision = \frac{tp}{tp + fp}$$

Recall

Recall is the fraction of the documents that are relevant to the query, which are successfully retrieved.

$$Recall = \frac{tp}{tp + fn}$$

True positive (TP) is correctly identified
True negative (TN) correctly rejected
False positive (FP) is incorrectly identified
False negative (FN) is incorrectly rejected

CONCLUSION AND FUTURE ENHANCEMENT

The methodological diagram discussed in this article delivers a big picture for processing sentiment analysis of short text data. This article has presented a novel hybrid-based model for analyzing sentiment analysis on big data. A new algorithm Senti_Lexi_Classi has been proposed to provide improved accuracy. Analyzing the emoticon and contextual text along with the machine learning classifier will produce better accuracy than the existing works. The less amount of research has been carried out in emoticon- and contextual-based analysis, which will bring more issues and challenges to the industrialists and academicians. Improving the emoticon and contextual word dictionaries will produce better results.

REFERENCES

1. Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu, *"Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis"*, HP Laboraties, 2011.
2. Geetika Gautam and Divakar yadav, *"Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis"*, Contemporary Computing (IC3), Seventh International Conference on August 2014, Pages 437-442.
3. Kalyankumar B Waddar, and K Srinivasa, *"Opinion Mining in Product Review System using Big Data Technology Hadoop"*, International Journal of Advanced Computational Engineering and Networking, Volume 2, Issue 9, 2014.
4. Hanhoon Kang, Seong Joon Yoo, and Dongil Han, *"Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews"*, Expert Systems with Applications, Elsevier, Volume 39, Issue 5, April 2012, Pages 6000–6010.

5. Prerna Chikersal, Soujanya Poria, and Erik Cambria, *"SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning"*, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 647-651.
6. A. Mudinas, D. Zhang, and M. Levene, *"Combining lexicon and learning based approaches for concept level sentiment analysis"*, Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, New York, NY, USA, Article 5, 2012, Pages. 1-8.
7. Geeta.G.Dayalani, Dr.Seema, and Prof.B.K.Patil, *"Emoticon-based unsupervised sentiment classifier for polarity analysis in tweets"*, International Journal Of Engineering Research and General Science, Volume 2, Issue 6,October-November, 2014.
8. Internet source as on 24-01-2016, https://github.com/cloudera/cdh-twitter-example/blob/master/flume-sources/flume.conf
9. Available from http://sentiwordnet.isti.cnr.it/
10. Available from http://sentic.net/sentire/
11. Available from http://mpqa.cs.pitt.edu/
12. Available from http://www.saifmohammad.com/WebPages/ResearchInterests.html
13. Alistair Kennedy and Diana Inkpen, *"Sentiment Classification of Movie Reviews Using Contextual Valence Shifter"*, Computational Intelligence, Volume 22, Issue 2, 2006, Pages 110–125
14. Vo Ngoc Phu and Phan Thi Tuoi, *"Sentiment Classification using Enhanced Contextual Valence Shifters"*, International Conference on Asian Language Processing, IEEE, 2014, Pages 224 – 229
15. M.R. Saleh, M.T. Martin-Valdivia, A. Montejo-Raez and L.A. Urena-Lopez, *"Experiments with SVM to classify opinions in different domains"*, Expert Systems with Applications, Volume 38, Issue 12, November–December 2011, Pages 14799–14804.
16. W. Medhat, Ahmed Hassan and Hoda Korashy, *"Sentiment analysis algorithms and applications: A survey"*, Ain Shams Engineering Journal, Volume 5, Issue 4, December 2014, Pages 1093–1113.
17. O. Popescu, and C. Strapparava, *"Time corpora: Epochs, opinions and changes"*, Knowledge Based Systems, Issue 3, Vol-13, 2014, Pages 3-13.