IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Using Lexicon and Random Forest Classifier for Twitter Sentiment Analysis

M. Thenmozhi[1*], R. Indira[2], R. Dharani[3]

[1,2,3]Dept of Computer Science and Engineering, Pondicherry Engineering College, Puducherry, India

[*]*Corresponding Author: thenmozhi@pec.edu, Tel.: 9500893708*

DOI: https://doi.org/10.26438/ijcse/v7i6.591594 | Available online at: www.ijcseonline.org

Accepted: 14/Jun/2019, Published: 30/Jun/2019

***Abstract*—**Today users prefer blogs and review sites to purchase products online. Thus, user reviews are considered as an important source of information in sentiment analysis applications for decision making. Machine Learning and Lexicon based sentiment analysis are the two popular methods available in the literature. The Machine Learning based classifiers does not work for unlabelled dataset such as tweets. On the other hand existing Lexicon based sentiment analysis approaches are becoming less efficient due to data sparseness, low accuracy and non-consideration n-gram words. N-grams can improve the accuracy of sentiment classification. Following these limitations the proposed work provides a combination of Lexicon and Machine learning based approach to perform sentiment analysis on Twitter datasets.

***Keywords*—**Sentiment Analysis, Sentiment Classification, Lexicon based Analysis, Sentiment Score

I. INTRODUCTION

Users share their views and experiences about a particular product in the social media services which are increasingly being used by online communities with the rapid increase in social networks and blogs [1]. The writing of these user reviews to promote products is growing in number due to the economic importance of these reviews. Therefore user reviews are considered as an important source of information in the sentiment analysis application in decision making. Sentiment analysis, which is also referred as opinion mining is an approach to natural language processing that can identify the sentiment or emotion in the given text [2]. This is a popular way for organizations to determine and categorize opinions about a product, service or idea [3]. The sentiment analysis has grown to be one of the most active research areas which coincide well with those of the social media on the Web. It has a huge volume of opinionated data recorded in digital forms and also widely studied in data mining, Web mining, text mining, and information retrieval.

Sentiment analysis involves the use of data mining, machine learning (ML) and artificial intelligence (AI) to mine text for sentiment and subjective information [4]. These algorithms classify the reviews into positive, negative or neutral. The presence of noisy text, emoticons, modifiers, unigram or bigram words and domain specific words in the reviews lead to incorrect classification. Though the existing supervised approach gives better performance, it is not always preferred due to lack of labelled training datasets [5]. In case of unlabelled datasets Lexicon based methods are preferred [6, 7, 8, 9]. The existing Lexicon based methods is semi manual and if automatic procedures are used, then they are domain specific. The objective of the proposed approach is to label the unlabelled tweets into positive, negative and neutral using Lexicon based method and then feed the labelled tweets to allow the classifier such as Random Forest to learn so as to improve the classification accuracy. The proposed work uses different dictionaries such as emoticons, modifiers, Senti-n-gram, as well as domain specific terms to analyze the tweets.

Section I provides the introduction of the area of research, Section II contains the related work of lexicon based approaches for sentiment analysis, Section III presents the proposed work, Section IV describes the results and discussion and Section V concludes research work with future directions.

II. RELATED WORK

In [9] author presents a hybrid approach to create health-related domain specific lexicon using bootstrapping modal, health reviews dataset and domain-specific corpus. The domain specific lexicon is then used for classification and scoring of health-related user sentiments. In [10] they proposed a framework to generate sentiment lexicons using domain-specific corpus and small sets of seed words. Their framework could induce and release historical sentiment lexicons and community-specific sentiment lexicons large online communities. In [11] the proposed work is a semi-supervised method for sentiment analysis. It incorporates lexicon-based methodology with the machine learning algorithm to revise the sentiment scores defined in SentiWordNet, they have applied mathematical models such

as information gain and cosine similarity. In [12] authors proposed a sentiment analysis method for Chinese micro-blog text based on the sentiment dictionary. Here the sentiment dictionary is extended by using various related dictionaries and the sentiment value of a micro-blog text is obtained which is then classified as positive, negative and neutral. In this paper [13] the authors proposed a rule-based framework for sentiment analysis, which uses a SentiWordNet based classifier and domain specific classifier. The work mainly concentrates on emoticon classification, and identification of modifiers and negations. In their work [14] the authors have presented a survey of existing sentiment analysis approaches. In [15], the authors have proposed a spark framework for analyzing the sentiment for Twitter streaming data.

III. METHODOLOGY

The proposed approach has two modules, procedure for sentiment analysis and procedure for sentiment classification as shown in Figure1.While the first module construct the feature vector by using a lexicon based method. The second module applies random forest classifier to classify the reviews. The proposed approach automatically generates feature vector by utilizing the scores generated by an emoticon, modifiers [12], domain-specific terms [10] and Senti-N-Gram dictionary [6]. This reduces manual computation and also increases the classification accuracy. It calculates the scores of the n-grams by following a rule based approach using a random corpus of reviews.

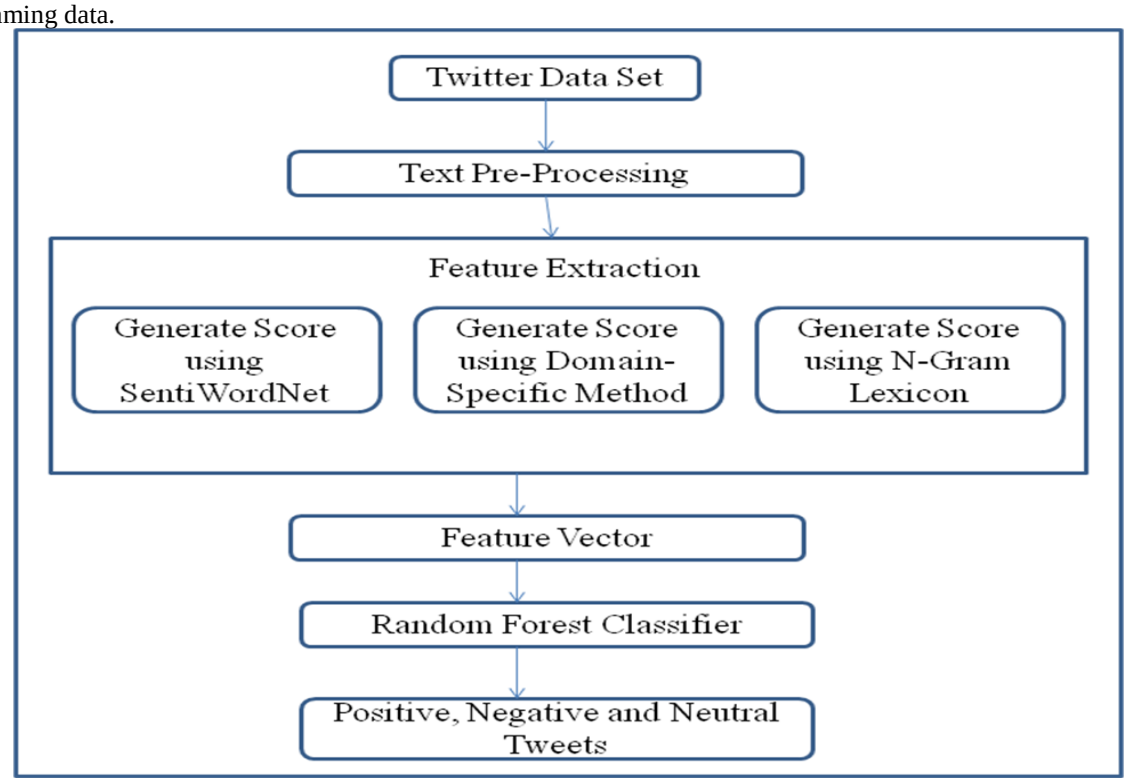

Figure 1. Proposed Senti-n-gram Approach

A. *Lexicon based Feature Vector Construction*

i. Text Pre-Processing

The Twitter data is published on online social media, hence text pre-processing is the first step for Sentiment Analysis. Pre-processing involves removing stop words, special symbols, URL's etc. and prepare the text for further processing.

ii. Generate Sentiment Score using SentiWordNet

To generate sentiment score for a particular the proposed work uses the SentiWordNet lexicon along with other repositories such as Emoticons, Modifiers and Negations. Emoticons represent the emotional state, mind, mood and feelings of the users online. Modifiers are those which either increase or decrease the polarity of a sentiment word. Words such as *few, very, lot, slightly* etc are some of the modifiers. Negation words such as *never, couldn't, don't* are likely to change the polarity sentiment of a word. The presence of a particular word in these repositories implies a score of '1' otherwise '0'.

iii. Generate Sentiment Score using Domain-specific Method

The word 'terrific' may seem to have an negative impact for sentiment score. But considering the example 'They played a terrific game', here in the sport domain the word seems to have a positive impact. Thus, to identify the exact sentiment of a particular word with respect to a domain the proposed approach utilizes a domain-specific lexicon such as SocialSent.

iv. Generate Sentiment Score using N-Gram Lexicon

Senti-n-gram is generated by considering unigrams along with bigrams and trigrams using an algorithm proposed in [10]. Polarity calculation of a sentence follows few steps,

a. Sentiment bearing n-grams are found from the pre-processed sentences
b. Score of n-grams are added
c. If the total score is zero, then the sentence is not considered for further process
d. Now, if the final score is greater than 0 then the polarity is defined as positive otherwise negative.

v. Ratio of Positive and Negative or Neutral Tweets

This finds a ratio between positive and negative sentences for a tweet. If the number of positive sentences is greater than the number of negative sentences, then the tweet is considered as positive, otherwise negative. If the number of positive sentences is equal to the number of negative sentences, then it is considered as neutral.

vi. Feature Vector

Each classified tweet is labelled as Class, where Class = {Positive, Negative}. Based on the final result of the previous step, we can conclude that whether the class of a tweet is positive or not.

B. Feature Vector Classification

The output of Lexicon based feature vector construction module is fed as input to the random forest classifier for training the classifier. It uses the Bagging approach in building classification models. Random forest builds multiple decision trees by taking a subset of features randomly. The class label will be assigned depending on the majority of votes. Once the classifier is trained it can then be used to classify the Twitter dataset. Following are the steps of the Random forest classifier algorithm:

Input: B = Number of Trees, N = Training Data, F = Total Features, f = Subset of Features
Output: Bagged class label for the input data.

a. For each tree in Forest B:
  1) Select a bootstrap sample S of size N from training data.
  2) Create the tree Tb by recursively repeating the following steps for each internal node of the tree.
    - Choose f at random from the F.
    - Select the best among f.
    - Split the node.

b. Once B Trees is created, Test instance, will be passed to each tree and class label will be assigned based on the majority of votes.

IV. RESULTS AND DISCUSSION

The performance of the proposed work is evaluated Twitter dataset. Precision, Recall, F-Measure and Accuracy metrics have been used to compare the proposed work with the baseline methods. Table 1 represents the experimental results. From the obtained results it is inferred that, the proposed approach with the combination of Lexicon and Machine learning classifier performs better compared to the baseline lexicon methods such as VADER and SO-CAL for Positive, Negative and Neutral sentiments. Figure 2 shows the comparison for the Accuracy metric. The accuracy of the proposed approach is high compared to the existing baseline methods.

Table 1 Experiment results

Sentiment	Metric	VADER	SO-CAL	Proposed
Positive	Precision	82.89	80.23	95.18
	Recall	93.33	92.21	96.34
	F-Measure	87.80	85.8	95.76
Negative	Precision	92.80	90.40	95.52
	Recall	81.69	76.35	94.12
	F-Measure	86.89	82.78	94.81
Neutral	Precision	69.35	71.31	73.17
	Recall	82.69	63.50	78.95
	F-Measure	75.44	67.18	75.95

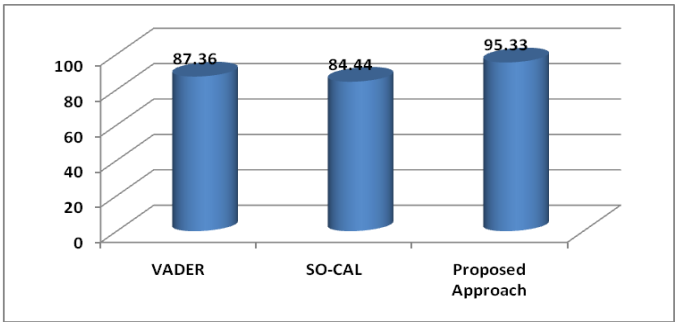

Figure 2. Results of Accuracy

V. CONCLUSION AND FUTURE SCOPE

The existing supervised approaches for sentiment analysis give better performance, but it are not always preferred due to lack of labelled training datasets. Lexicon based methods perform well for unlabelled datasets. But Lexicon such as SentiWordNet and VADER are created by human annotators.

In order to overcome the limitations of these two approaches the proposed work is a combination of Lexicon and Machine learning based Sentiment analysis. The proposed work is fully automatic and domain independent. This also eliminates human annotators which also reduce cost and computational load for n-gram based sentimental analysis.

REFERENCES

[1] Chen, Yubo, Scott Fay, and Qi Wang. "*The role of marketing in social media: How online consumer reviews evolve*." Journal of interactive marketing, Vol.25, No. 2, pp. 85-94, 2011.

[2] Pang, Bo, and Lillian Lee. "*Opinion mining and sentiment analysis*." Foundations and Trends® in Information Retrieval, Vol.2, No. 1–2, pp. 1-135, 2008.

[3] Fang, Xing, and Justin Zhan. "*Sentiment analysis using product review data*." Journal of Big Data, Vol.2, no. 1, pp. 5, 2015.

[4] Vohra, S. M., and J. B. Teraiya. "*A comparative study of sentiment analysis techniques*." Journal JIKRCE, Vol.2, no. 2, pp.313-317, 2013.

[5] Agarwal, Basant, and Namita Mittal. "*Machine learning approach for sentiment analysis.*" In Prominent feature extraction for sentiment analysis, pp. 21-45. Springer, Cham, 2016.

[6] Dey, Atanu, Mamata Jenamani, and Jitesh J. Thakkar. "*Senti-N-Gram: An n-gram lexicon for sentiment analysis.*" Expert Systems with Applications, Vol. 103, pp.92-105, 2018.

[7] Al-Ayyoub, Mahmoud, Safa Bani Essa, and Izzat Alsmadi. "*Lexicon-based sentiment analysis of Arabic tweets*." IJSNM, Vol.2, no. 2, pp. 101-114, 2015.

[8] Trinh, Son, Luu Nguyen, Minh Vo, and Phuc Do. "*Lexicon-based sentiment analysis of Facebook comments in Vietnamese language*." In Recent developments in intelligent information and database systems, pp. 263-276. Springer, Cham, 2016.

[9] Asghar, Muhammad Zubair, Shakeel Ahmad, Maria Qasim, Syeda Rabail Zahra, and Fazal Masud Kundi. "*SentiHealth: creating health-related sentiment lexicon using hybrid approach.*" SpringerPlus 5, no. 1, pp. 1139, 2016.

[10] Hamilton, William L., Kevin Clark, Jure Leskovec, and Dan Jurafsky. "*Inducing domain-specific sentiment lexicons from unlabeled corpora*." In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, vol. 2016, p. 595. NIH Public Access, 2016.

[11] Khan, Farhan Hassan, Usman Qamar, and Saba Bashir. "*A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet.*" Knowledge and information Systems, Vol. 51, no. 3, pp. 851-872, 2017.

[12] Zhang, Shunxiang, Zhongliang Wei, Yin Wang, and Tao Liao. "*Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary*." Future Generation Computer Systems, Vol. 8, pp. 395-403, 2018.

[13] Asghar, Muhammad Zubair, Aurangzeb Khan, Shakeel Ahmad, Maria Qasim, and Imran Ali Khan. "*Lexicon-enhanced sentiment analysis framework using rule-based classification scheme*." PloS one, Vol. 12, no. 2, e0171649, 2017.

[14] C. Nanda, M. Dua, *"A Survey on Sentiment Analysis"* International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.2, pp.67-70, 2017.

[15] Amit Palve, Rohini D.Sonawane, Amol D. Potgantwar, *"Sentiment Analysis of Twitter Streaming Data for Recommendation using, Apache Spark"*, International Journal of Scientific Research in Network Security and Communication, Vol.5 , Issue.3 , pp.99-103, Jun-2017.

Authors Profile

M.Thenmozhi received her B.Tech in Computer Science and Engineering from Pondicherry University in 2001 and M.E in Computer Science and Engineering from Anna University in 2006. She obtained her Ph.D from Pondicherry University in the year 2015. She has been working as Assistant Professor in Department of Computer Science and Engineering, Pondicherry Engineering College for the past 12 years. She has published more than 24 papers in International Journal and Conferences. Her research interest includes Data warehousing, Data Modeling, Data Mining, Data Analytics and Ontology.

