Received December 4, 2019, accepted December 22, 2019, date of publication December 30, 2019, date of current version January 7, 2020.

*Digital Object Identifier 10.1109/ACCESS.2019.2963020*

# A Sentiment Polarity Categorization Technique for Online Product Reviews

**SAMINA KAUSAR[1,2], XU HUAHU[1], WAQAS AHMAD[2], MUHAMMAD YASIR SHABIR[2], AND WAQAS AHMAD[3]**

[1]School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
[2]Department of CS & IT, University of Kotli Azad Jammu and Kashmir, Kotli 11100, Pakistan
[3]College of Information Science and Technology, Beijing Normal University, Beijing 100875, China

Corresponding author: Muhammad Yasir Shabir (yasir.shabir14@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61572434 and Grant 91630206, and in part by the Shanghai Science and Technology Committee under Grant 16DZ2293600.

**ABSTRACT** Sentiment analysis is also known as opinion mining which shows the people's opinions and emotions about certain products or services. The main problem in sentiment analysis is the sentiment polarity categorization that determines whether a review is positive, negative or neutral. Previous studies proposed different techniques, but still there are some research gaps, i) some studies include only 3 sentiment classes: positive, neutral and negative, but none of them considered more than 3 classes ii) sentiment polarity features were considered on individual basis but none of them considered on both individual and on combined basis iii) No previous technique considered five sentiment classes with 3 sentiment polarity features such as a verb, adverb, adjective and their combinations. In this study, we propose a sentiment polarity categorization technique for a large data set of online reviews of Instant Videos. A comprehensive data set of five hundred thousand online reviews is used in our research. There are five classes (Strongly Negative, Negative, Neutral, Positive and Strongly Positive). We also consider three polarity features Verb, Adverb, Adjective and their combinations with their different senses in review-level categorization. Our experiments for review-level categorization show promising outcomes as the accuracy of our results is 81 percent which is 3 percent better than many previous techniques whose average accuracy is 78 percent.

**INDEX TERMS** Sentiment, opinion mining, social media, natural language processing.

## I. INTRODUCTION

The importance of users' sentiments has been realized by the business sector in the last decade. Since then social media platforms and other websites are used to extract users' opinions about products. Such phenomena is called sentiment analysis or opinion mining. Opinion mining is identifying, extracting and understanding the user's attitude or opinion by analyzing the text. This process usually involves natural language processing, statistical analysis and machine learning techniques for sentiment analysis. Various other names are also used including review mining, emotional analysis, opinion extraction, and subjectivity analysis [1]. Sentiment analysis has been defined by Smith [2] as, "Sentiment Analysis is the computational evaluation of documents to determine the fine-grained emotions that are expressed." Sentiment analysis is a study of people's opinions about a certain product, person, text, etc. It is their opinions that depict their mood for a specific entity whether we like it or not. It is a process of computationally identifying and categorizing the opinions provided in a review to determine whether it is positive, negative or neutral. Nowadays internet provides many different platforms for users to share their sentiments in textual form for different products. Many large organizations can increase revenues if they keep an eye on what people say about their products as people are the best judges. Based on reviews, large organization can enhance their products according to the needs of the customers. So, due to its utmost need it becomes the most important challenge in current era for NLP (Natural language processing). Hence for the extraction of subjective information from source material like product reviews, sentiment analysis techniques are widely used.

The associate editor coordinating the review of this manuscript and approving it for publication was Seok-Bum Ko.

### A. DIFFERENT LEVELS OF SENTIMENT ANALYSIS

Sentiment analysis is performed on three levels i.e., a) document level, b) sentence level and phrase level [3].The document level sentiment analysis focuses on classifying the entire document as positive or negative. In the document level classification, a single review of a single topic is considered. But in case of forums and blogs, comparative sentences may appear. There are two types of methods used in document level sentiment analysis, i) one is supervised learning and another is ii) unsupervised learning method. In supervised learning method, the traditional algorithms like naïve Bayesian and Support Vector Machine can be used to train the system. In order to train and testing the data, the review rating (1-5 stars) can be used. While in unsupervised learning method we extract just words inside the document. People compare one product with another similar product and hence document level sentiment analysis is not efficient in forums and blogs. The main issue is that not all the sentences in a document have relevance in expressing the opinion about an entity. Therefore, subjectivity and objectivity classification is very important in this type of classification [4].

In the sentence level sentiment analysis, the polarity of each sentence is calculated in [5], [6]. The same document level classification methods can be applied to the sentence level classification problem. It helps to find out the objective and subjective sentences. The subjective sentence contains the opinion words which help in determining the sentiments about an entity after which the polarity classification is done into positive and negative classes [7].

The phrase level sentiment classification is more specific approach for opinion mining. The phrases that contain opinion words are found out and phrase level classification is performed. This classification can have both advantages and disadvantages. In some cases, the exact opinion of an entity can be extracted correctly (advantage). In other cases where the contextual polarity matters, the result may not be accurate (disadvantage) [8].

### B. RESEARCH OBJECTIVES

Performing sentiment analysis on the product reviews, these reviews represent the user's opinions for specific products. Normal user of a product posts their reviews in the form of short text usually contains few sentences. These sentences are comprised of some important words. As we know that in English parts of speech, a word can have different meanings depending on the structure of the sentence. Identifying the parts of speech that can present true meaning of the sentence is a challenging task. The parts of speech are used to estimate the sentiments of the user comments. Adverbs are important part of any sentence and hence needed to be analyzed their role in determining the true sentiments of user. The different types of adverbs should be identified and analyzed for determining the sentiment of the sentence. So in this work we will identify and extract the different types of adverbs from the user review datasets and then estimate their importance in automatic classification of reviews in three sentiment classes i.e. positive, negative or neutral. For automatic classification different classifiers have been used in research community. It is yet to determine which classifier is the best to classify the reviews into the classes based on adverb features. The classifier works on some feature set. In this study we explore some very important features in the content (text) of reviews. These features are adverbs. We explore different types of verbs that can be used to classify the reviews into positive or negative classes.

In case of classifiers we are interested in determining the performance of different classifiers that are used by research community for classification. We investigate how these classifiers work on the extracted feature set and which of them achieve high performance.

### C. PROBLEM STATEMENT

Previous researchers proposed outstanding methods in order to determine the polarity of text. Usually they classify the text into three polarity classes i.e. Positive, Negative and Neutral where positive class contains those documents in which positive language is used, while negative class contains those documents where user has some bad experience with the product and finally neutral class presents those documents that are neither positive nor negative. In this study we introduced five polarity classes:Strongly negative, Negative, Neutral, Positive and Strongly Positive. Furthermore, there is also a need to investigate how parts of speech like adverbs can be used to assign polarity to the text. For such purpose we used product reviews instead of twitter tweets which are short in length.

### D. RESEARCH QUESTIONS

The following research questions have been identified during literature review;

**RQ1.** What is the impact of parts of speech (adverb) on sentiment analysis on product reviews?

**RQ2.** What is the impact of different combinations of adverbs on the classification?

**RQ3.** What is the best classifier for product review classification?

### E. RESEARCH METHODOLOGY

Our research methodology has been presented in Figure1. First we select sentiment analysis as research domain of this study. After selecting the research domain, we performed extensive literature review of the research domain. After literature review stage we identify some research questions that we consider need to be answered. Then we proposed methodology that answers our research questions. We implemented our methodology and performed experiments in order to answer our identified research questions during literature review. After completing our experiments, we have evaluated our results.

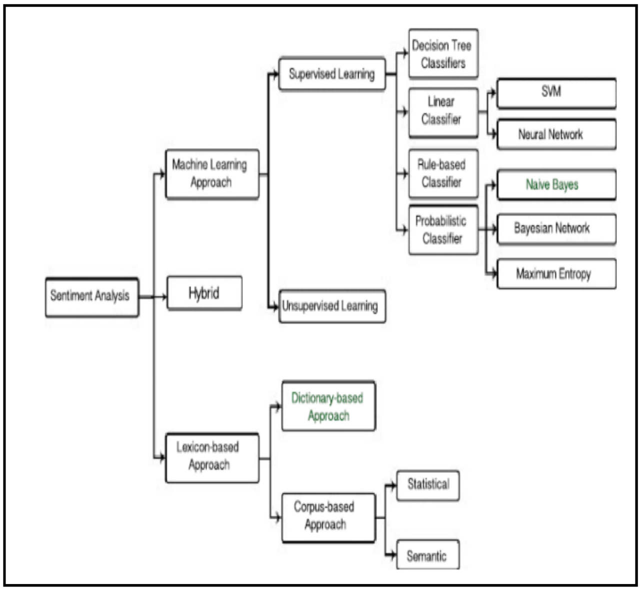

**FIGURE 1. Sentiment analysis approaches.**

## II. LITERATURE REVIEW

The techniques available for carrying out sentiment analysis can be classified into three main categories. Such as (a) Knowledge-based approaches, (b) Statistical based techniques and (c) Hybrid methods, the hybrid techniques or methods are the combination of the previous two approaches. Lexical Knowledge-based approaches normally focus on univocal words like happy, sad, afraid, etc., while statistical methods use the automated techniques to judge the sentiments based on machine learning analysis and hybrid approaches use both techniques collectively to analyze the results on reviews which are not clearly stated but have some link with the product.

Some studies are very much linked with our approach such as Fang and Zhan [9], proposed a process which is used to categorize the polarity based on parts of speech (POS). Another approach presented by Hu and Liu [10] provided a list of different words (i.e. both Positive and Negative words).The proposed list of words consisted of 2006 positive and 4783 negative words respectively. These words are based on online reviews which are used to extract the subjective information for this research. Moreover, in a proposed text categorization technique, Pang and Lee [11], proposed how to remove objective sentences by extracting the subjective ones as mainly we should focus on subjective contents and should not waste time for irrelevant material.

In another technique proposed by Gann and Day [12], the authors applied token based approach on twitter data as they assigned certain sentimental scores to every token which is being used to analyze that if a certain opinion is positive, negative or neutral. Some other techniques are also useful like topic modeling [13] in which the author proposed a process of automatically identifying the features or aspects of a product. Narrowing down the opinion, in the research community several approaches have been proposed on the sentiment analysis of micro blogging sites like twitter. Das and Chen [14], presented an approach for extracting the sentiment from stock message board where the authors suggested that market activities can have an effect on the sentiments of median and small range investors. Another study conducted by Nasukawa and Yi [15], focuses on subject centric aspect of sentiment analysis. The proposed technique develops a mechanism that determines the polarity scores (i.e. negative and positive sentiment scores) associated with a specific subject instead of determining or calculating the sentiment for complete documents. The proposed techniques have been evaluated using datasets of different domains, such as news and other web pages. The proposed technique achieves an overall precision score of 75 to 76 percent, depending on the different types of datasets used in the evaluation. Natural language processing techniques have been employed for carrying out the sentiment analysis. Conventionally, sentiment analysis can be performed using three main types of approaches, these approaches are a) Machine learning approach, b) Lexicon based approach and c) Hybrid approach [16].

R. Xia. *et al.* [17], developed a hybrid technique for sentiment analysis. The proposed technique combines both lexicon and machine learning based approaches for sentiment analysis. POS along with their associated and word-related features are selected from lexicon and then machine learning classifiers (i.e. Naiv Bayes, ME and SVM) are applied to determine the sentiment of words. In order to achieve better classification results, experiments were performed on the dataset using different combinations, such as fixed weighted, meta classifiers and ensemble combination techniques. Couple of variations of Naive Bayes were presented by Gamallo *et al.* [18], various variations of Naive Bayes classifier were applied to classify opinions into different classes. Features like Valence Shifters, Polarity Lexicon, Lemmas and Multiword were used in the experimentation. Nandi and Agrawal [19], presented a layered hybrid technique for sentiment analysis. The proposed approach has two layers; the first layer is a lexicon based approach while the second layer is machine learning approach. Machine learning classifiers are used to classify the sentiment of opinions into different classes, such as positive, negative and neutral classes.Rajganesh *et al.* [20], presented a hybrid approach for sentiment analysis. The approach is a feedback based recommendation system that uses sentiment analysis.

## III. PROPOSED METHODOLOGY

### *A. DATA COLLECTION*

The dataset used in this research is being crawled using python crawler. The dataset crawler fetch reviews of two products which are distinct in nature. Therefore, the dataset which has been crawled contains reviews of two products.

First is office product which includes Microsoft Word, Microsoft PowerPoint, Microsoft Excel and Microsoft Access Database. The other product is musical DVDs which contains two main albums that are pop tracks and slow tracks.

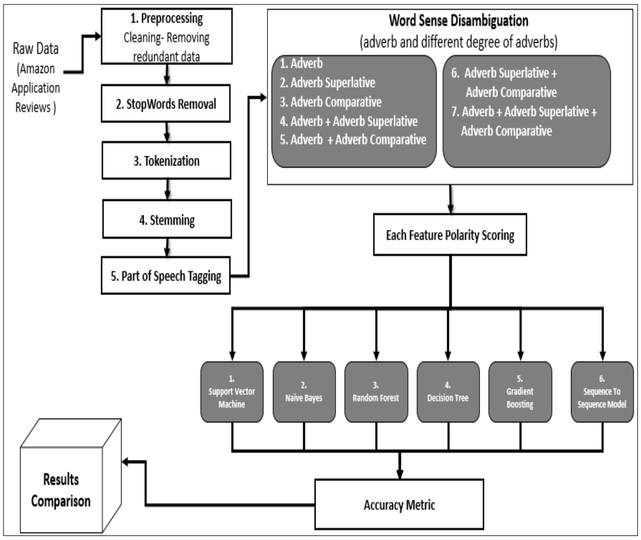
Raw Data (Amazon Application Reviews )
1. Preprocessing Cleaning- Removing redundant data
2. StopWords Removal
3. Tokenization
4. Stemming
5. Part of Speech Tagging
Word Sense Disambiguation
(adverb and different degree of adverbs)
1. Adverb
2. Adverb Superlative
3. Adverb Comparative
4. Adverb + Adverb Superlative
5. Adverb + Adverb Comparative
6. Adverb Superlative + Adverb Comparative
7. Adverb + Adverb Superlative + Adverb Comparative
Each Feature Polarity Scoring
1. Support Vector Machine
2. Naive Bayes
3. Random Forest
4. Decision Tree
5. Gradient Boosting
6. Sequence To Sequence Model
Accuracy Metric
Results Comparison

**FIGURE 2.** **Proposed methodology.**

### B. PRE PROCESSING

In the pre-processing step, in the first phase the boundary of sentence is to be determined and after verifying the sentence boundary the next phase is to tokenize the text into single words. Pre-processing step also includes removal of stop words, white spaces, new line tags, html tags, emotions and special symbols.

### C. REMOVAL OF STOP WORDS

Stop words are usually extra words which are not needed for sentiment polarity categorization. We remove all stop words in our data set which is beneficial for better accuracy.

### D. TOKENIZATION

We assign every word with a token and based on that token we get the score of the word from Senti Word Net library.

### E. STEMMING

We perform stemming of the complete data set in order to make sure that identical words in a review should be removed as this does not cause the repetition of identical words.

### F. PART OF SPEECH TAGGING (POS TAGGING)

The reviews consist of different parts of speech such as noun, adjective; verb and adverb are tagged using Natural language tool kit (NLTK). The main interest in this research is with adverbs and its forms so all forms of adverbs are extracted from the reviews. NLTK tagged some adverbs such as:

*Superlative Adverbs (RRS):* which modify general adverb with superior for e.g. best, longest and easiest etc.

*Comparative Adverbs (RBR):* which modify verbs along another adverb with comparison e.g. more, less and few etc.

*Adverb (RB):* which modifies verb using another adverb e.g. very, silently, much etc.

### G. REVIEW

I was smacked to realize that the new office is renewable **annually** and as I **only** require basic office and would **not** gain from the upgraded office programs. I looked **around** and found the 2013 which will do me for as long as my computer is alive and come to think of me **as well.** Compare but be sensible for home use - do you need the additional features and are you willing to pay **annually** for them. I wasn't and I am thrilled. I **already** have to make annual payments on other software I need such **as** protection but the annual cost **soon** mount up. Use your common sense. I do**n't** think many users of office know **how** to get the **most** out of it as home users unless they are studying or earning a living from the program or using it **professionally.**

In this review, the respective adverbs which appear in a review are underlined and bold but the problem is to understand how these adverbs narrate the story of any user and for sentiment how it will be classified. The different forms of adverbs such as annually, only, as-well, already, professionally are some general adverbs (RB) and most is general superlative adverb (RBS).

### H. SCORING FEATURES

Senti Word Net 3.0, lexical resource explicitly devised for supporting sentiment classification and opinion mining applications. Senti Word Net 3.0 is an improved version of Senti Word Net 1.0, a lexical resource publically available for research purposes. Senti Word Net is one of these lexicons that assigns to each synset of Word Net, three sentiment numerical scores, positivity, negativity and objectivity. Therefore, it is knowledge base which can be used for assigning scores. The total positive words present in it are 3,076,708 and negative words are 151,044. Every feature which is present in any document, review or text is assigned with some positive and negative scores(Salehan, M., & Kim, D. J. (2016)) [21].

### I. SENTENCE SCORING

The score of the sentence is calculated by the score of individual words present in that specific sentence.

$$\mathbf{Sen\ Score(s)\ 1/n} = \sum \mathbf{(i = 0)^\wedge n\ Pi} \qquad (1)$$

where,

- Sen Score(s) are scores for a sentence in a document or review.
- n is the total number of words present in a sentence.
- (Pi) polarity words present in sentence where i is the limit of words

Let us consider an example for calculating the sentence level scores.

*Sentence 1:* " The Microsoft version 2013 office is **very** good, and many things are enhanced **especially** the new style.

*Explanation:* The word "very" and especially are general adverbs. Now these two distinct adverbs will get the scores from Senti Word Net library and average is calculated for this sentence.

*Sentence Score:* The sentence score is positive because both adverbs have positive polarity score retuned by polarity lexicon "Senti Word Net".

Let us consider another example where negation occurs.

*Sentence 2:* "the Access is **not** that good **as** compared to SQL but others like Excel, Word is **much better** than before"

*Explanation:* This sentence contains "not"&"as" are a general adverb and "much" &" better" are superlative adverbs. Now these adverbs will get scores and to find the polarity of this sentence in which negation occurs firstly negativity is calculated by the formula as

$$\textbf{Neg Score} = \mathbf{1} - (\textbf{positive Scores} + \textbf{negative Scores}$$

Then, the total calculation will be constructed to understand the sentiments of a sentence.

*Sentence Score:* Thus, all the sentences are scored and finally take average for scoring the review of a product.

### J. REVIEW SCORING

The review score is calculated by the scores of sentences present in a review.

$$\textbf{Rev Score(r) 1/n} = \sum \textbf{(i = 0)\^{}n Si} \tag{2}$$

where,

- Rev Score (r) are scores of a document or review.
- n is the total number of sentences in a review.
- (Si) sentence present in a review where i is the limit of sentences.

For classifying the review using adverbs and its different forms, the respective review is tagged. After tagging the review different forms of adverbs are extracted. After extracting these forms, they are combined together for scores using Senti Word Net. Firstly, at sentence level and then at review level scores are assigned, the final scores of reviews are obtained and will be classified with 5 star rating class (Hu, Y.H., Chen, K., & Lee, P.J. (2017) [22]

### K. STAR RATINGS

For every review there is always a star rating which is assigned by a user on the basis of his/her experience for a particular product. Thus, Amazon also contains star ratings whenever customer shares opinions. To evaluate 5-starratings of the review, the first step is to find out the range which is from the highest to the lowest ratings. To calculate these star rates, range from 0 to 1,different researchers contributed such as Pappas & Popescu-Belis [23], Lak &Turetken [24], Boon *et al.* [25] and Lee and Pang [26] which indicates the highly positive and highly negative range i.e. −1 to 1 respectively. The Table 1 demonstrates the star ratings along with polarity values and classification as taken from the literature Kincl *et.al.* [27], Zhang *et.al.* [28], and Stieglitz and Dang-Xuan [29]

**TABLE 1.** Star ratings.

| Star Ratings | Polarity Values | Class |
|---|---|---|
| 1 | -1 to -0.5 | Strongly Negative |
| 2 | -0.5 to 0 | Negative |
| 3 | 0 | Neutral |
| 4 | 0 to 0.5 | Positive |
| 5 | 0.5 to 1 | Strongly Positive |

### L. CLASSIFICATION

Each review is a variable sequence of words and the sentiment of each review must be classified into above mentioned star rating classes [30], [31]. The Large Amazon Review Dataset contains 308,420 highly-polar reviews (good or bad) for training and testing. The problem is to determine whether a given review has a different sentiment depending on polarity of adverb features. Various methodologies have been practiced by different studies over the years starting from tree based classifier to neural network based approaches. We have chosen Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting and Sequence to Sequence Recurrent Neural Network.

#### 1) NAIVE BAYES CLASSIFIER

Naive Byes is probability-based classification algorithm widely used by the research community based on Bayes Theorem. Naive Byes classifier is based on assumption that the appearance of a specific attribute in a class is unrelated to the appearance of any other attribute. Naive Bayes model is useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Currently Google is using it, to mark an email as spam, or not spam. It is also used by some new agency in order to classify the news into different categories like technology, entertainment, politics and sports etc.

#### 2) DECISION TREE CLASSIFIER

Decision tree is classification algorithm that is widely used by research community for classification purpose. The decision tree is also used in classification of text into sentiment polarity. It falls under the machine learning category. As we discuss in related work chapter ,a lot of research utilize decision tree for classification of tweet into positive, negative and neutral tweets.

#### 3) RANDOM FOREST CLASSIFIER

Random Forest is a versatile machine learning method capable of performing both regression and classification tasks. It also undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps of data exploration, and does a fairly good job. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model. To classify a new object based on attributes, each tree gives a classification and we say the tree ''votes'' for that class. The forest chooses the classification having the most votes (over all the trees in the forest) and it takes the average of outputs by different trees.

#### 4) SUPPORT VECTOR MACHINE

Support Vector Machines are perhaps a standout amongst the most well-known and discussed machine learning algorithms. It remains in mainstream around the time they were created in the 1990s and keep on being the go-to technique for a high-performing algorithm with little tuning. It is a discriminative classifier, given labeled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new example. On the basis of this training, the algorithm is able to predict unknown input.

#### 5) GRADIENT BOOSTING CLASSIFIER

Gradient boosting technique is used by major search engine companies, i.e. Google, Bing, Yandex and Yahoo. They used it for web page ranking, but it is actually not limited to application domain and can be used for a variety of problems (Viola and Jones 2001) [32]. Gradient boosting classifiers are models made out of different weaker models that are trained individually and each model prediction is combined. This is an effective strategy and accordingly is extremely famous.

#### 6) SEQUENCE TO SEQUENCE MODEL

One of the powerful recurrent neural network is the long short-term model network or also called as LSTM. In our propose method, we will outline review into a real vector space, a mainstream method when working with text called word embedding. This is a procedure where terms are encoded as real valued vectors in a high dimensional space, where the likeness among terms describes the closeness in the vector space. Keras is an open source library which gives an advantageous approach to convert positive integer representations of words into a word embedding by an Embedding layer.

## IV. EVALUATION

Perform sentiment analysis on the product reviews,these reviews represent the user opinion on specific products. Normal user of a product posted their review in form of short text usually contain few sentences. These sentence are comprised of some important words. As we know that in English part of speech, a word can give different meaning depends on the structure of the sentence. Identifying the part of speech that can present true mean of the sentence is a challenging task. The part of speech are used to estimate the sentiment of user. Adverbs are important part of any sentence and hence needed to be analyzed their role in determining the true sentiment of user. The different types of adverbs should be identified and analyze for determining the sentiment of sentence. So in this work we will identify and extract different types of adverbs from the user review dataset and then estimate their importance in automatic classification of reviews in three classes of sentiment i.e. positive, negative or neutral. For automatic classification different classifiers has been used in research community. It is yet to determine that which classifier best classify the review in to classes based of adverb features. The classifier work on some feature set. In this study we explore some very important features in the content (text) of reviews. These features are adverb. We explore different types of verbs that can be used to classify the reviews in to positive or negative classes.

In case of classifiers we are interested in determining the performance of different classifiers that are used by research community for classification. We investigate how these classifiers work on the extracted feature set and which of them achieve high performance.

We compare the results with sentiment analysis using product review data in which authors (''Xing Fang &Justin Zhan 2015'') published in Springer open journal in 2015. The reason for comparing with aforementioned research work is that both the research (our and base paper) tackles the same problem but our approach towards solving the sentiment analysis problem is different as we are using polarity features (Adverb and its different forms)both on individual and combination basis because no previous techniques considered five sentiment classes plus three polarity features. Such difference is core novelty of our work. Three evaluation measures such as Precision, Recall, and F-Measures are used by using different classifiers firstly on using single features and later on binary features and find out which of them achieve high performance.

In order to evaluate or proposed methodology we used three evaluation measures. These include Precision, Recall and F-Measure (Tripathy, A., Agrawal, A., & Rath, S. K. 2016) [33]. These evaluation methods are presented in the following formula,

*Precision:* Precision is the ratio of correctly identified instance to the total instance in the data. This can be represented as,

$$\boldsymbol{Precision} = \mathbf{TP}/(\mathbf{TP}+\mathbf{FP})$$

*Recall:* Recall is the second evaluation measure we used to evaluate the performance of classifiers. This can be represented as,

$$\boldsymbol{Recall} = \mathbf{TP}/(\mathbf{TP}+\mathbf{TN})$$

*F-Measure:* Finally, we also calculated the F-Measure score of classifiers that can be shown as

$$\boldsymbol{F-Measure} = \mathbf{2} \times \mathbf{PrecisionRecall}/(\mathbf{Precision}+\mathbf{Recall})$$

### A. TOOLS AND TECHNIQUES

1) Natural Language Tool Kit for parts of speech tagging using specific tag set.
2) Microsoft Excel for pre-processing data and after processing data for results.
3) Python programming platform
4) Senti-Word Net is used for scoring the feature

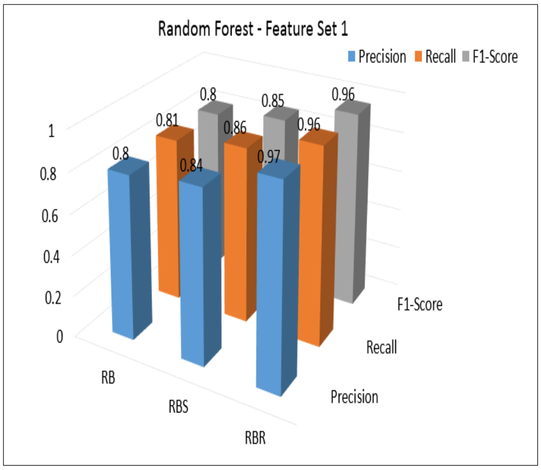

**FIGURE 3.** **Evaluation score of Precision, Recall and F-Measure for Random Forest Algorithm when using Single Feature Set.**

## V. RESULTS

### A. DATA SET ACQUISITION

The dataset was obtained by using developed crawler built in python programming language from Amazon. The proposed methodology uses a diverse dataset. The dataset contains reviews of two products.

- Office products
- Musical DVDs

These reviews hold product Id, product review, product rating stars and overall summary of a review along with some metadata. There are 30,842 reviews to thoroughly test the research effort. The collected dataset is diverse enough and belongs to different products for testing the research effort comprehensively.

### B. SINGLE FEATURE

This section discusses the results of classifiers when we only used single feature. This means when each feature was used by different classifier which classifier performs the best.

#### 1) RANDOM FOREST CLASSIFIER

Firstly we applied Random forest classifier on the feature set 1. It has been observed that F-measure score of RBR performed the best by securing the F-measure of 0.96. Similarly following forms were able to achieve the F- measure of more than or equal to 0.80: RB and RBS. However, general adverb obtains the lowest F- measure of 0.81 as compared to others.

#### 2) DECISION TREE CLASSIFIER

This time we applied Decision Tree classifier on the Feature set 1. It can be observed that RBR performed the best by securing the F-measure of 0.96. Similarly following forms were able to achieve the F- measure of more than or equal 0.78: RB and RBS. However, general adverb obtains the lowest F- measure of 0.78. This shows a decrease as compared to Random Forest Classifier.

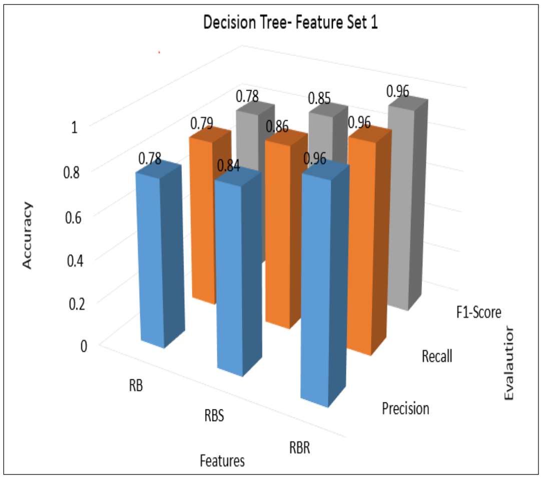

**FIGURE 4.** **Evaluation score of Precision, Recall and F-Measure for Decision Tree Algorithm when using Single Feature Set.**

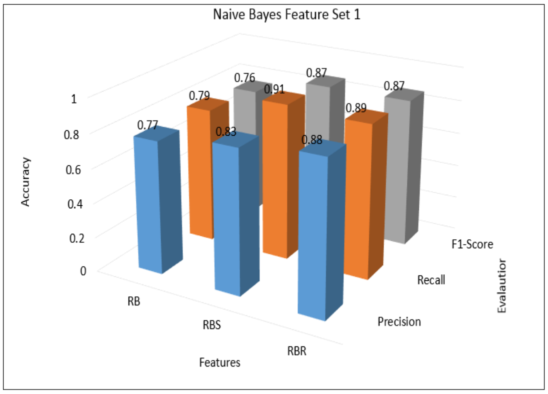

**FIGURE 5.** **Evaluation score of Precision, Recall and F-Measure for Naïve Bayes Algorithm when using Single Feature Set.**

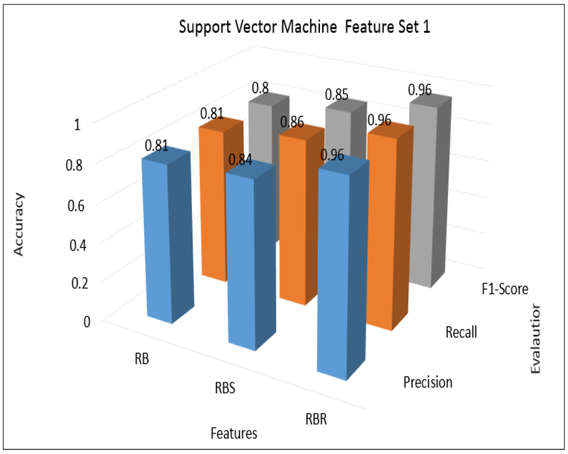

**FIGURE 6.** **Evaluation score of Precision, Recall and F-Measure for SVM Algorithm when using Single Feature Set.**

#### 3) NAÏVE BAYES CLASSIFIER

We applied Naive Bayes classifier on the Feature set 1. In terms of F-Measure score, RBR & RBS performed the best

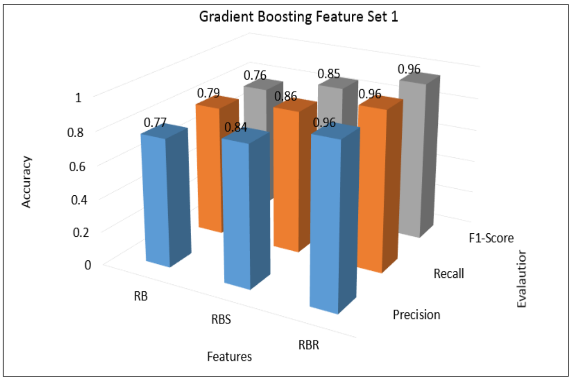

**FIGURE 7.** Evaluation score of Precision, Recall and F-Measure for Gradient Boosting Algorithm when using Single Feature Set.

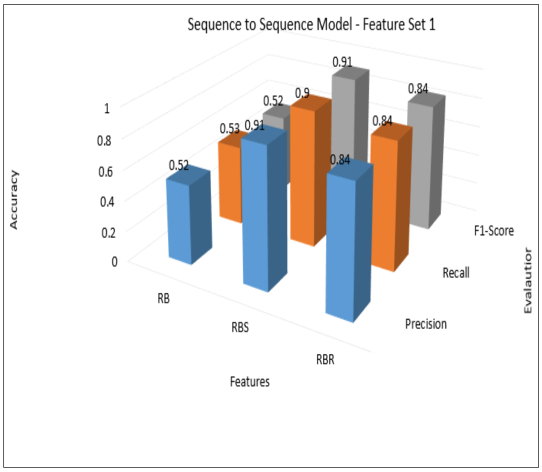

**FIGURE 8.** Evaluation score of Precision, Recall and F-Measure for Sequence to Sequence Model when using Single Feature Set.

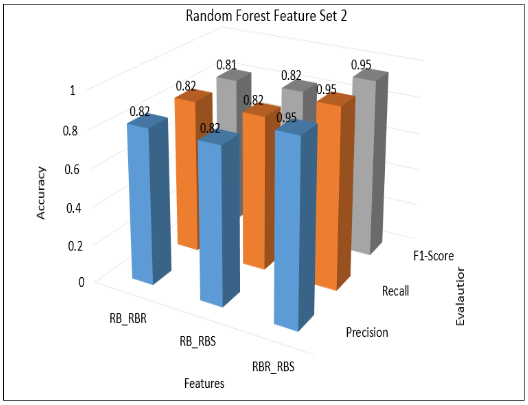

**FIGURE 9.** Evaluation score of Precision, Recall and F-Measure for Random Forest Algorithm when using Binary Feature Set.

by securing the F-measure of 0.87. Similarly following forms were able to achieve the F- measure of more than or equal to 0.76: RB. However, general adverb obtains the lowest F- measure of 0.76.

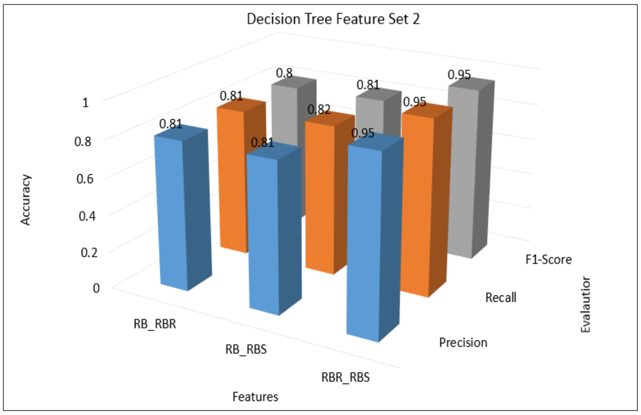

**FIGURE 10.** Evaluation score of Precision, Recall and F-Measure for Decision Tree Algorithm when using Binary Feature Set.

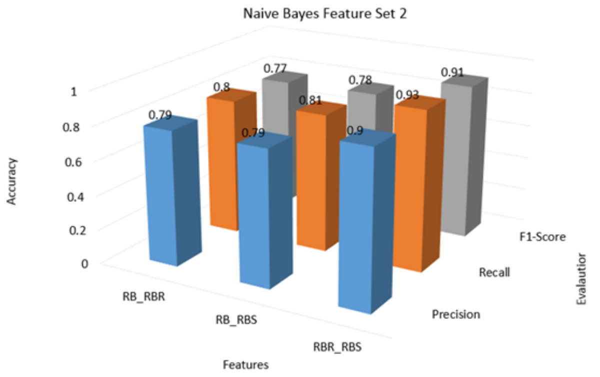

**FIGURE 11.** Evaluation score of Precision, Recall and F-Measure for Naive Bayes Algorithm when using Binary Feature Set.

### 4) SUPPORT VECTOR MACHINE

We applied Support Vector Machine classifier on the Feature set 1. If we observe the F-measure closely, it performed equally to the Random Forest Classifier. RBR performed the best by securing the F-measure of 0.96. Similarly following forms were able to achieve the F- measure of more than or equal to 0.85: RB and RBS.

### 5) GRADIENT BOOSTING CLASSIFIER

We applied Gradient Boosting classifier on the Feature set 1. It performed equally to the Random Forest Classifier and Support Vector Machine. RBR performed the best by securing the F-measure of 0.96. Similarly following forms were able to achieve the F- measure of more than or equal to 0.76: RB and RBS. However, general adverb obtained the lowest F- measure of 0.76.

### 6) SEQUENCE TO SEQUENCE MODEL

We applied Sequence to Sequence model on the Feature set 1. It performed the best for the RBS feature among all classifiers. RBS performed the best by securing the F-measure of 0.91. RB feature does not perform well and achieved the lowest score of 0.53. However, general adverb obtained the lowest F- measure of 0.84.

### *C. BINARY FEATURE*

This section will discuss the analysis of Bi Feature (those combinations which consist of combinations of two features) of three distinct forms of adverbs. When three adverbs forms are combined. There exist total of 7 combinations. In this combination, three Bi Features exist. We applied different classifiers & obtained results.

#### 1) RANDOM FOREST CLASSIFIER

Firstly, we applied Random forest classifier on the Feature set 2. Comparative adverbs (RBR) + Superlative adverbs (RRS) performed the best by securing the F-measure of 0.95. Similarly following forms were able to achieve the F- measure of more than or equal to 0.81: Adverb (RB) + Comparative adverbs (RBR) and Adverb (RB) + Superlative adverbs (RRS). However, general Adverb (RB) + Comparative adverbs (RBR) obtained the lowest F- measure of 0.81.

#### 2) DECISION TREE CLASSIFIER

This time we applied Decision Tree classifier on the Feature set 2. If we observe the F-measure closely Comparative adverbs (RBR) + Superlative adverbs (RRS) performed the best by securing the F-measure of 0.95. Similarly following forms were able to achieve the F- measure of more than or equal 0.78: Adverb (RB) + Comparative adverbs (RBR) and Adverb (RB) + Superlative adverbs (RRS).

#### 3) NAIVE BAYES CLASSIFIER

We applied Naive Bayes classifier on the Feature set 2. If we observe the F-measure closely. Comparative adverbs (RBR) + Superlative adverbs (RRS) performed the best by securing the F-measure of 0.91. Similarly following forms were able to achieve the F- measure of more than or equal to 0.77: Adverb (RB) + Comparative adverbs (RBR) and Adverb (RB) + Superlative adverbs (RRS). However, Adverb (RB) + Comparative adverbs (RBR) obtained the lowest F- measure of 0.77.

#### 4) SUPPORT VECTOR MACHINE

We applied Support Vector Machine classifier on the Feature set 2. It performed equally to the Random Forest Classifier. Comparative adverbs (RBR) + Superlative adverbs (RRS) performed the best by securing the F-measure of 0.94. Similarly other forms were able to achieve the F- measure of more than or equal to 0.82: Adverb (RB) + Comparative adverbs (RBR) and Adverb (RB) + Superlative adverbs (RRS)

#### 5) GRADIENT BOOSTING CLASSIFIER

We applied Gradient Boosting Classifier on the Feature set 2. It performed equally to the Random Forest Classifier. Comparative adverbs (RBR) + Superlative adverbs (RRS) performed the best by securing the F-measure of 0.95. Similarly other forms were able to achieve the F- measure of more than or equal 0.79: Adverb (RB) + Comparative adverbs (RBR) and Adverb (RB) + Superlative adverbs (RRS).

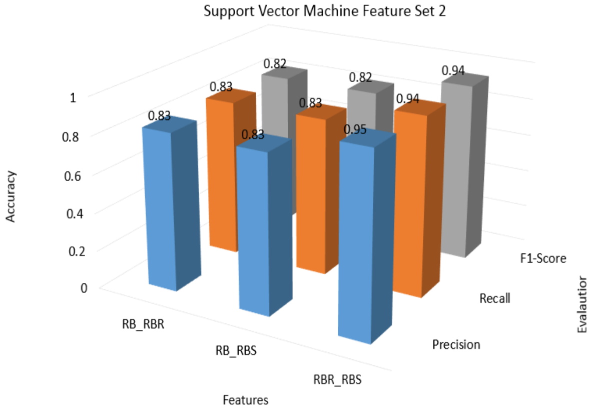

**FIGURE 12.** Evaluation score of Precision, Recall and F-Measure for SVM Algorithm when using Binary Feature Set.

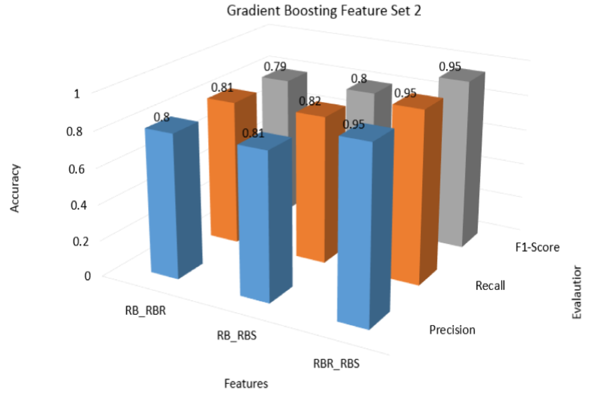

**FIGURE 13.** Evaluation score of Precision, Recall and F-Measure for Gradient Boosting Algorithm when using Binary Feature Set.

#### 6) SEQUENCE TO SEQUENCE MODEL

We applied Sequence to Sequence model on the Feature set 2. It performed the best for the Comparative adverbs (RBR) + Superlative adverbs (RRS) feature among all classifiers. Comparative adverbs (RBR) + Superlative adverbs (RRS) performed the best by securing the F-measure of 0.91. Adverb (RB) + Comparative adverbs (RBR) and Adverb (RB) + Superlative adverbs (RRS) feature does not performed well and achieved the lowest score of 0.53.

### *D. DISCUSSION*

We evaluated the performance of 6 classifiers, using three combinations of features. These features are consisted of three types of adverbs. At first run of each classifier we use singe types of adverbs, then in second run we combine two types of adverbs and total of 6 binary attributes have been tested while in third run we test all of them together.

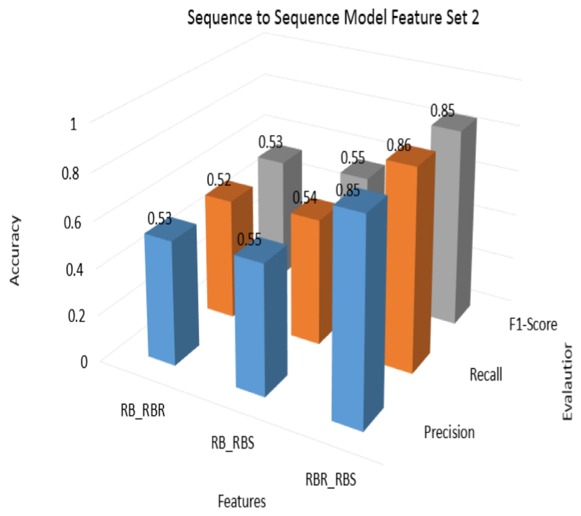

**FIGURE 14.** Evaluation score of Precision, Recall and F-Measure for Sequence to Sequence Model when using Binary Feature Set.

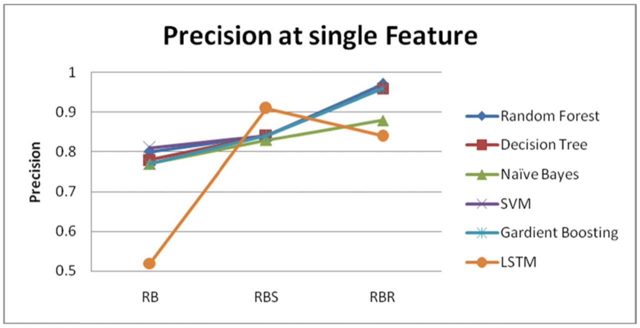

**FIGURE 15.** Precision score of all 6 classifiers at Single Feature.

### 1) PRECISION AT SINGLE FEATURE

As it can be shown in the above figure that LSTM performs very badly in terms of precision when using RB and RBR features. Whereas LSTM have good precision score when using RBS feature. On the other hand Naive Bayes is the second worst performance algorithm using all three types of adverbs. Random Forest, Decision tree, SVM and Gradient Boosting achieve high precision at RBR feature.

### 2) RECALL AT SINGLE FEATURE

Again LSTM (neural networks) algorithm failed to achieve high recall at RB and RBR attributes. While RBR attribute proved to be efficient for all classifiers except Naïve Bayes and LSTM.

### 3) F –MEASURE AT SINGLE FEATURE

LSTM achieved high F-score when using RBS feature, while have low recall when using RB and RBR feature. Random Forest, Decision Tree, SVM and Gradient Boosting algorithm achieve high F-measure score while Random Forest achieves high F-measure score at RB. But overall RBR have high F-measure score than all the other

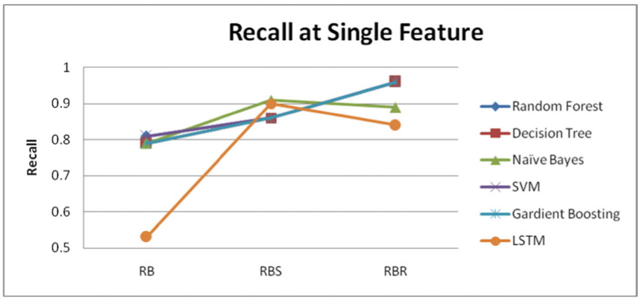

**FIGURE 16.** Precision score of all 6 classifiers at Binary Feature.

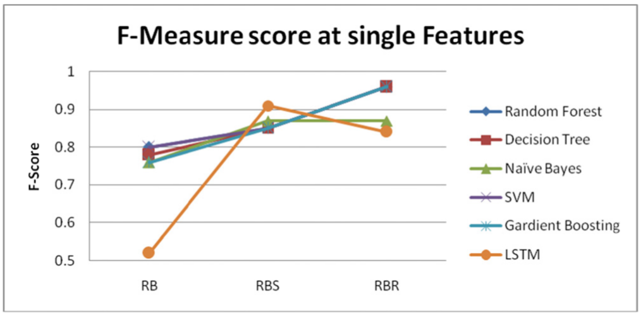

**FIGURE 17.** F-Score score of all 6 classifiers at Single Feature.

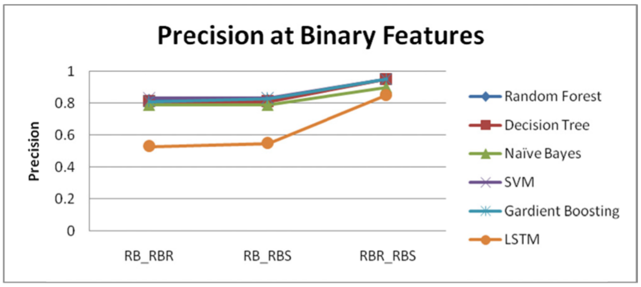

**FIGURE 18.** Precision score of all 6 classifiers at Binary Feature.

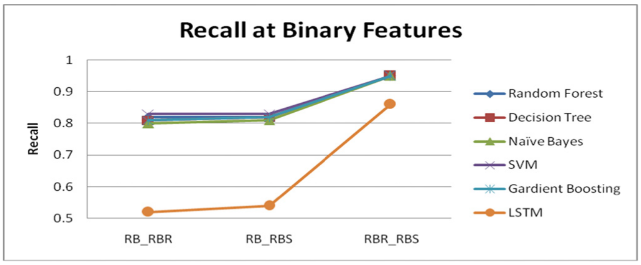

**FIGURE 19.** Recall score of all 6 classifiers at Binary Feature.

### 4) PRECISION AT BINARY FEATURE

LSTM algorithm performs poor in terms of precision when we combine the three types of adverbs. RBR and RBS combination are proved to be efficient as they achieve high precision on all the classifiers expect the LSTM algorithm. The other two combinations of RB-RBR and RB-RBS have low precision for all the classifiers presented in figure IV.

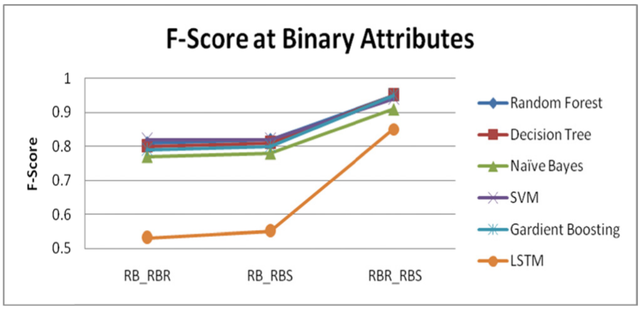

**FIGURE 20.** F-Measure score of all 6 classifiers at Binary Feature.

### 5) RECALL AT BINARY FEATURE

Same behavior as we have seen in case of Precision where LSTM performs poor while all the other classifier achieves good recall using RBR-RBS adverb combination.

### 6) F-MEASURE AT BINARY FEATURE

Figure 20 shows a comparison between F-Measure score for each classifier and it can be shown that LSTM and Naive Bayes has the least F-score as compare to other classifiers. Again combination of RBR-RBS proved to be more efficient.

## E. LIMITATION AND FUTURE RESEARCH

Like other research work, this research work has limitations. Automated sentiment analysis is helpful for analyzing big textual information, it still has limitation. The software we have used in this research work have the capability to process different types of textual information. But it has a drawback of processing different styles such as sarcasm. On contrary there are further areas for improvement in the field of natural language processing. Future research work can provide better insight regarding the information contained in online review using more advanced technology. Future research can also be look at how explanation of different aspect of customer reviews on product quality, marketing strategy influence in the field of data mining.

## ENDNOTES

Even though there are papers talking about spam on Amazon.com, we still contend that it is a relatively spam –free website in terms of reviews because of the enforcement of its review inspection process.

The product review data used for this research work can be downloaded at: http://www.ilabsite.org /? Page-id=1091.

## REFERENCES

[1] B. Liu, "Sentiment analysis and opinion mining," *Synth. Lectures Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.

[2] P. Smith, "Sentiment analysis: Beyond polarity thesis proposal," School Comput. Sci. Univ. Birmingham, Birmingham, U.K., Tech. Rep., Oct. 2011, pp. 1–42.

[3] R. S. Jagdale, V. S. Shirsat, and S. N. Deshmukh, "Sentiment analysis on product reviews using machine learning techniques," in *Cognitive Informatics and Soft Computing*. Singapore: Springer, 2018, pp. 639–647, doi: 10.1007/978-981-13-0617-4_61.

[4] M. Devika, C. Sunitha, and A. Ganesh, "Sentiment analysis: A comparative study on different approaches," *Procedia Comput. Sci.*, vol. 87, pp. 44–49, Jan. 2016.

[5] K. Schouten and F. Frasincar, "Survey on aspect–level sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 813–830, Mar. 2016.

[6] R. Arulmurugan, K. R. Sabarmathi, and H. Anandakumar, "Classification of sentence level sentiment analysis using cloud machine learning techniques," *Cluster Comput*, vol. 22, no. S1, pp. 1199–1209, Jan. 2019, doi: 10.1007/s10586-017-1200-1.

[7] P. Burnap, R. Gibson, L. Sloan, R. Southern, and M. Williams, "140 characters to victory?: Using Twitter to predict the UK 2015 general election," *Electoral Stud.*, vol. 41, pp. 230–233, Mar. 2016, doi: 10.1016/j.electstud.2015.11.017.

[8] A. S. Manek, P. D. Shenoy, M. C. Mohan, and V. K. R, "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier," *World Wide Web*, vol. 20, no. 2, pp. 135–154, Mar. 2017, doi: 10.1007/s11280-015-0381-x.

[9] X. Fang and J. Zhan, "Sentiment analysis using product review data," *J. Big Data*, vol. 2, no. 1, p. 5, Dec. 2015, doi: 10.1186/s40537-015-0015-2.

[10] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2004.

[11] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 1, nos. 1–2, pp. 1–135, 2008, doi: 10.1561/1500000011.

[12] W.-J. K. Gann, J. Day, and S. Zhou, "Twitter analytics for insider trading fraud detection system," in *Proc. 2nd ASE Int. Conf. Big Data*, 2014.

[13] Y. Liu, "Social media tools as a learning resource," *J. Educ. Technol. Develop. Exchange*, vol. 3, no. 1, pp. 101–114, Mar. 2017.

[14] S. R. Das and M. Y. Chen, "Yahoo! For Amazon: Sentiment parsing from small talk on the Web," Inst. Oper. Res. Manage. Sci., Catonsville, MD, USA, Tech. Rep., Sep. 2007, vol. 53, no. 9, pp. 1–16.

[15] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proc. Int. Conf. Knowl. Capture (K-CAP)*, 2003, pp. 70–77.

[16] A. Cambero, "A comparative study of Twitter sentiment analysis methods for live applications," B. Thomas Golisano College Comput. Inf. Sci., Rochester Inst. Technol., Rochester, NY, USA, Tech. Rep. 8, 2016.

[17] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Inf. Sci.*, vol. 181, no. 6, pp. 1138–1152, Mar. 2011, doi: 10.1016/j.ins.2010.11.023.

[18] P. Gamallo and M. Garcia, "Citius: A naive–Bayes strategy for sentiment analysis on english tweets," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 171–175.

[19] V. Nandi and S. Agrawal, "Political sentiment analysis using hybrid approach," *Int. Res. J. Eng. Technol.*, vol. 3, no. 5, pp. 1621–1627, 2016.

[20] N. Rajganesh, C. Asha, A. T. Keerthana, and K. Suriya, "A hybrid feedback based book recommendation system using sentiment analysis," *Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol.*, vol. 3, no. 3, pp. 2456–3307, 2018.

[21] M. Salehan and D. J. Kim, "Predicting the performance of online consumer reviews: A sentiment mining approach," in *Proc. ICIS*, 2014.

[22] Y.-H. Hu, K. Chen, and P.-J. Lee, "The effect of user-controllable filters on the prediction of online hotel reviews," *Inf. Manage.*, vol. 54, no. 6, pp. 728–744, Sep. 2017.

[23] N. Pappas and A. Popescu-Belis, "Explicit document modeling through weighted multiple–instance learning," *J. Artif. Intell. Res.*, vol. 58, pp. 591–626, Jul. 2018.

[24] P. Lak and O. Turetken, "Star ratings versus sentiment analysis—A comparison of explicit and implicit measures of opinions," in *Proc. 47th Hawaii Int. Conf. Syst. Sci.*, Jan. 2014, pp. 796–805, doi: 10.1109/HICSS.2014.106.

[25] C. Boon, C. Hawkins, K. Bisht, S. J. Coombes, B. Bakrania, K.-H. Wagner, and A. Bulmer, "The king's speech metalanguage of nation, man and class in anecdotes about George III," *Boon*, vol. 16, no. 2, pp. 281–299, Jul. 2012.

[26] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2005, pp. 115–124.

[27] T. Kincl, M. Novák, and P. Štrach, "A cross-cultural study of online marketing in international higher education—A keyword analysis," *New Educ. Rev.*, vol. 32, no. 2, pp. 49–65, 2013.

[28] J. Q. Zhang, G. Craciun, and D. Shin, "When does electronic word-of-mouth matter? A study of consumer product reviews," *J. Bus. Res.*, vol. 63, no. 12, pp. 1336–1341, Dec. 2010.

[29] S. Stieglitz and L. Dang-Xuan, "Emotions and information diffusion in social media—Sentiment of microblogs and sharing behavior," *J. Manage. Inf. Syst.*, vol. 29, no. 4, pp. 217–248, Apr. 2013.

[30] F. Ali, D. Kwak, P. Khan, S. R. Islam, K. H. Kim, and K. Kwak, "Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling," *Transp. Res. C, Emerg. Technol.*, vol. 77, pp. 33–48, Apr. 2017.

[31] F. Ali, K.-S. Kwak, and Y.-G. Kim, "Opinion mining based on fuzzy domain ontology and support vector machine: A proposal to automate online review classification," *Appl. Soft Comput.*, vol. 47, pp. 235–250, Oct. 2016.

[32] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Aug. 2005, pp. 511–518.

[33] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst. Appl.*, vol. 57, pp. 117–126, Sep. 2016, doi: 10.1016/j.eswa.2016.03.028.

**SAMINA KAUSAR** received the M.S. degree in computer science from the International Islamic University Islamabad, Pakistan, in 2007. She is currently a Researcher and also a Ph.D. Scholar with the School of Computer Engineering and Science, Shanghai University, China. She has been an Assistant Professor with the University of Kotli Azad Jammu and Kashmir, Pakistan. Her research interests are in the fields of big data, bioinformatics, computer networks, cloud computing, data mining, and machine learning algorithms.

**XU HUAHU** is currently a Doctoral Supervisor and a Professor with the School of Computer Engineering and Science, Shanghai University, where he is also the Director of Information Office. He is the Chairman of the Shanghai Security and Technology Association. His research interests include multimedia technology, CIMS, and computer network technology.

**WAQAS AHMAD** received the B.S.I.T. degree from the University of Azad Jammu and Kashmir. He is currently pursuing the M.S.C.S. degree with the University of Kotli Azad Jammu and Kashmir, Pakistan. His research interests include data mining, the Internet of Things, and computer networks.

**MUHAMMAD YASIR SHABIR** was born in Kotli, Azad Jammu and Kashmir, Pakistan. He received the B.S.I.T. degree from the University of Azad Jammu and Kashmir, and the M.S. degree in computer sciences from International Islamic University Islamabad, Pakistan. He is currently a Lecturer with the University of Kotli Azad Jammu and Kashmir, AJ&K. His major research areas are computer networks and security, cloud computing, big data, machine learning algorithms, and the IoT.

**WAQAS AHMAD** received the M.S. degree in computer science from International Islamic University at Islamabad, Islamabad, Pakistan, in 2012. He is currently pursuing the Ph.D. degree with Beijing Normal University, Beijing, China. From 2013 to 2015, he was a Visiting Faculty Member in different institutions, Pakistan. His areas of interest are game theory, mechanism design, crowd sourcing, privacy preservation in mobile cloud computing, data mining, and mobile crowd sensing.