DFWC: Discriminative Feature Weight Correlation based Sentiment Analysis of Social Networks

Muddada Muralikrishna, G. Lavanya Devi

Abstract: ***Sentiment analysis is the buzz in current era of the social network-based opinion sharing about products, events, politics and many more.The lexicon analysis and machine learning are the major strategies to perform sentiment analysis. The phenomenal escalation of available data in social networks is indeed evincing the intensification of dimensions inportraying the opinion. Hence, the machine learning is the much effective than the lexicons analysis. However, the escalation of the dimensions in projection of the sentiment, indicating the need of the optimal learning methods enables to learn from the correlation between the features of vivid dimensions, which is the crucial requirement for available labeled data with multidimensional features. This manuscript portrayed a novel method that learns from the discriminative feature weight correlation to train the classifier in regard to identify the polarity of the sentiment to the given opinion about the target event. In this regard, the feature weights are used to identify the impact of the corresponding feature towards positive or negative polarity. The experimental study evincing the performance advantage of the proposal that compared to other contemporary models.***

Key words: sentiment analysis, query expansion ranking (QER), NLP, feature selection strategy.

1 INTRODUCTION

No boundary for the range of the information procured by texts & tweets and these short texts are often utilized for sharing the sentiments and opinions that people possesses regarding what is happening in the universe. Hence, task & corpus of twitter sentiment development is done to support research which will result to understand how the sentiment will be taken in texts & tweets. Contributing with the text genres which are informal presents the challenges aimed at processing the natural language due to utilization of language is informal with misspellings, novel words, punctuation, URLs, creative spelling and genre particular abbreviations & terminology. Another factor of social media information is messages of twitter that contains rich structured data regarding the people who are involved in conversation. In instance, twitter maintains the alive information of who is followed by whom and tags and re-tweets. Modeling the structured information is dynamic due to:

It will suggest to most precise tools aimed at bring out the associated information among the texts and

It offers means aimed at studying the social interactions properties practically

Revised Manuscript Received on March 26, 2019.

Muddada Muralikrishna,
Research scholar, Computer Science And Systems Engineering, AUCE (A), Andhra University, Visakhapatnam, Andhra Pradesh, India 1E.

Dr.G. Lavanya Devi,
Computer Science and Systems Engineering, AUCE (A), Andhra University, Visakhapatnam, Andhra Pradesh, India2.

The substantial quantity of study is conducted to examine the texts which are informal and categorize them with assistance of machine learning (ML) and lexicon methods. As no dictionary or lexicon is able to justify the quantity of piercing content within the texts, and in this research there were out of the scope [1]. The ML based methods are resourceful since they learn from text which is unstructured and form the feature vector by itself and will not depend on any set which is predefined of the features. The work [2] presents that these methods are resourceful in generalizing method on different datasets content. The extraction of feature is prominent task in the methods since they produce features vectors aimed at classifier for learning. Several algorithms are utilized for the classification and common amid them were SVM, Naïve Bayes (NB), & maximum entropy. And the contribution [3], [4] presents that SVM performs better when compared with other algorithms if there is imbalance of class and for this cause it is selected as base classifier.

One of the extensively agreed platforms of social media all over the world is routinely examined to assist the address problems originated from the several application fields. For instance, the researchers have utilized the twitter data for estimating the political result from all over the world [5], [6], [7]. The work [6] presents the complete survey of the contributions nevertheless proposes cautionary tale about the predictive power of the twitter data in this concern. The work [8], [9], [10] presents the other remarkable twitter data applications comprise of tracing & observing disease epidemics and natural disasters. The work [11] presents that twitter data is examined for uncovering the possible associated with the 311 civil criticisms. Contemporarily, the researchers of marketing have utilized the Twitter data for uncovering the opinions of the consumers of definite brand. Such kind of information can in turn assist marketers design novel and modify the present marketing schemes. Several such implementations require efficient & effective analysis of sentiment for huge amount of the tweets.

The analysis of sentiment is treated generally as the issue of classification where the reviews were usually classified into 4—irrelevant, negative positive & neutral on the basis of opinions stated in the review. Practically, the class of irrelevant is consolidated often with the neutral, hence converting into ternary classification issue. The work [12], [13], [14] presents that the extensive gathering of the classification algorithms like random forests and SVM has been implemented in this contribution. Contemporarily, the work [15], [16] presents that the occurrence of ensemble learners constructing for the classification of the twitter sentiment analysis. Combined with the ensemble learners, feature engineering presented to overbeat the individual learners.

II. Related Work

The work [17] presents that lexicon-based method is on the basis of unsupervised learning method where the sentiment lexicon will be built for determining the specified text polarity through predefined task of indicator of the negative & positive. The familiar lexicons contain AFINN Lexicon [18], Lexicon of NRC Emotion [19] and opinion Lexicon [20]. The work [21] presents that the average of polarity values derived from every text word in lexicon signifies in the form of sentiment score. This score reflects polarity orientation of text. The contemporary studies usually utilize former sentence as tie-breaker while the tested sentences classification is not derived from the function of scoring. Hence, the production of expressive lexicon on the basis of unsupervised tagging corpus is critical through lexicon based method [17]. The work [22] presents that since the lexicon creation is time taking procedure, and the lexicon which is pre-designed is not essentially applicable directly to diverse languages & domains, the methods based on conventional lexicon are dedicated for extracting specific words related to domain.

Wide-range of researches is made with lexicon based method. The work [23] presents that visual structure, the EmoWeb, suggested to examine dynamically the textual content sentiment depending on well-designed lexicon. Simulations are carried out by gathering the data from the websites for displaying the proposed strategy applicability. The work [24] presents that the technique NN (Neural Network) is utilized for overcoming the limitations of lexicon-based method, where some of words are eradicated by lexicon. The work [25] presents that "Portuguese context-sensitive lexicon" is constructed for the identification of sentiment polarity. The methodology LexReLi is employed for constructing the suggested lexicon, and effectiveness of diverse techniques combinations are also examined. The work [26] presents that "Arabic senti-lexicon" is formed for analyzing the sentiment for Arabic language, where the "Multi-domain Arabic Sentiment Corpus (MASC)" is constructed.

Several learning algorithms is adopted for machine learning containing Decision Tree (DT), NN, Naïve-Bayes (NB), Support Vector Machine (SVM). The SVM proved as robust algorithm of classification, specifically in processing huge amount of the features [27], [28], [29]. Nevertheless, weakness of the less interpretability & computational overhead constraints its utilization. Another well-known algorithm is DT, where features & resulting classes were reflected as structure of tree. It will be implicit and is employed easily for classifying the textual-data. Nevertheless, exploring features will be confined owing to inelastic structure of tree, and over-fitting of feature might reduce the execution [30], [31], [32]. The NB will be commonly utilized model for the sentiment analysis because of its benefits simplicity & effectiveness. The work [21], [33], [34],[35] presents that it is utilized for selecting the class of maximum probability for unlabeled text on the basis of probability concept.

The procedure of feature weighting is to assign suitable weight to single features in accord with their significance to specified domains. The work [36] presents that normal thought of generalizing the feature selection is where existence of feature acts as criteria for the extraction. The work [21] presents that every feature is depicted in the form of binary vector where zero indicated the feature absence and one indicates its presence or existence. The work [37] presents several feature weighting models are introduced, &DF will be the easiest one where word single appearance is identical to manifold appearances. It calculates the amount of the documents possessing word and utilizes the values for depicting the resulting document. Another criterion is the TF, which explores weighting of feature in diverse direction. The work [38] presents that on the basis of intuition the term of manifold appearances will be more significant than single appearance.

The TF is proved as an effective, yet its performance will be reduced with manifold domain dataset due to its easy words measure. The "Term Frequency and Inverse Document Frequency (TF-IDF)" will be "state-of-the-art" model for weighting of feature [39] that measures comprehensively the frequency appearance and words distribution in allocating the weights.

Another new feature weighting method enhancing the "Twitter sentiment analysis" is the "Part of Speech-based Weighting (PSW)". It groups the categories unique words into diverse subsets on the basis of words feature called POS.

Our earlier contribution Frequency and Distribution Diversity based Optimal Feature Selection for Opinion Mining [40] is intended to project the word tokens as features, which is done by their distribution diversity when associated to positive and negative sentiment lexicons. This significant to identify the opinion polarity. However, the method is least significant, if reviews are influenced by other dimension of opinion representation.

The contemporary model proposed by Parlar et al., [41] is a novel feature selection model called "query expansion ranking (QER)" for sentiment analysis, which intended for attaining consistency in estimation with less false alarming. Nevertheless, this model is lessening the features dimensions, yet confined to the word tokens in the form of features only and is not dealing with manifold dimensions of "opinion representative features". Therefore, false alarming ratio in the label estimation is deliberately high when intended for learning from projected reviews in the social network.

In order to deal the constraints of the feature optimization in contemporary models, this manuscript endorsed a novel feature selection strategy called "Discriminative Feature Weight Correlation based Sentiment Analysis of Social Networks" that enables to deal the divergent dimensions of the opinion representative features such emojies, emoticons, and slang. The significance of the proposal is scaled by comparing the results obtained for performance metrics with counterpart method QER [41].

III. Methods and Materials

This section intended to briefs the method of (i) preprocessing, (ii)description of features adapted, (ii) feature optimization, (iv)

classifier used, (v) method of training applied on decision tree as classifier, (vi)method of sentiment polarity detection for the given opinion in social media format. The contribution of this article is considering the token of different types used toportray the opinion in social media. The diversified feature types used in proposal are,

The general and unique tendency of the social network users to portray their opinion is not so particular about to follow the NLP compatible format. The emojies, emoticons, slang, are two considerable options to present the opinion in social media. The other typical version that often notices in the opinions of negative label is sarcasm that often appears in the reviews and opinions about the target event, are product. The emojies are quite traditional, which are the symbols of special characters that intended to use to reflect the polarity of the sentiment of the given opinion. Similarly, the emoticons are quite recent strategy using in similar context. The emoticons are picture graphs of the character, objectives, face of emotions like weeping, fun mood, happy state, smiling, and disgusting or hating. The other dimension of opinion representation in social networks are the usage of slang. Slang is the shorthand representation of the statements often using in social media. As an example, the LoL is one of the slang used in social media, which indicates that "laugh out loud". The use of emojies, emoticons, and slang are wide spread and significantly frequent in social media, which is since, these methods defuse the number keys strokes while writing on social media.

A. Preprocessing

The given corpus of reviews and opinions from social media, which reflects the user's sentiment on target event or product. Each record of this corpus labeled either as positive or negative and these records are mix of text, emojies, emoticons, and slang. The records of this corpus with positive label indicates that the sentiment polarity of the corresponding record is positive and supporting the target event or product. The label negative indicates the sentiment polarity of the corresponding record is negative that trolling the target event or product.

The initial phase of sentiment analysis is to perform preprocessing on given corpus of labeled data. This phase intends to extract words as tokens, emojies, emoticons, and slang observed in each of the records labeled as positive or negative.

The words extracted from each record of the given corpus, the words other than stop words will be represented as a vector of words, which retains the label that assigned to the corresponding source record.

Similarly, the emojies, emoticons, and slang appeared in the record are separated as different vectors and each vector retains the label given to the corresponding source record.

Let the notation r^+ is the record having positive label, the vector vr_t^+ denotes the vector of word tokens, such that each entry of this vector represents a word, which is not the stop word and stemmed the "ing" and "ed" forms. Similarly, the vectors rv_t^+, rv_s^+, $vr_{et}^+, and\ vr_{ej}^+$ are representing the emoji, emoticons, and slang appeared in the corresponding record r^+.

The resultant vectors of the preprocessing applied on each labeled record are further used as input to optimize the features, such that the resultant features can be used to train the classifier adapted.

B. Feature Selection

This section, details the approach that adapted to identify the weights of each dimension of the features, record level, and corpus level. The each token of feature in regard to a dimension such as term, slang, emoji, and emoticons reflects their impact to decide the given record is positive or negative to sentiment polarity. The weight of all the tokens of vivid dimensions of the features can further used to denote the impact of the record to fall in one of the two labels of the sentiment polarity. Such that the impact of the records in regard to a specific label are used further to indicates the weight threshold of the corpus of the corresponding label. The methods of assessing the weight of the tokens related to vivid dimensions of the features, weight of the records fall under a label of the given two labels of the given training corpus, and the impact threshold of the corpus corresponding to a label are described in following.

C. Feature levelWeight

List all unique terms, unique Slang tokens, unique Emojies, and unique Emoticons of the positive label as set T^+, set Sl^+, set Ej^+, and set Et^+ in respective order which is as follows

Let the sets $T^+, Sl^+, Et^+, and\ Ej^+$ which are empty at their empty state.

$\forall_{i=1}^{|R_+|}\left\{r_i^+ \exists r_i^+ \in R_+\right\}$ Begin

$T^+ = T^+ \bigcup r_i v_t^+$

$Sl^+ = Sl^+ \bigcup r_i v_s^+$

$Ej^+ = Ej^+ \bigcup r_i v_{ej}^+$

$Et^+ = Et^+ \bigcup r_i v_{et}^+$

End

Find the term weight of each term exists in set T^+, which is the coverage frequency of the corresponding term in records of the positive label.

$\forall_{i=1}^{|T^+|}\left\{t_i \exists t_i \in T^+\right\}$ Begin

$$isc(t_i) = \frac{\sum_{j=1}^{|R^+|}\{1 \exists t_i \in r_j \wedge r_j \in R^+\}}{|R^+|}$$

End

Find the slang weight of each token exists in set Sl^+ as follows, which is the coverage frequency of the corresponding token in records of the positive label.

$\forall_{i=1}^{|Sl^+|}\{s_i \exists s_i \in S^+\}$ Begin

$$isc(s_i) = \frac{\sum_{j=1}^{|R^+|}\{1 \exists s_i \in r_j \wedge r_j \in R^+\}}{|R^+|}$$

End

Find the emoticon weight of each term exists in set Ej^+ as follows, which is the coverage frequency of the corresponding term in records of the positive label.

$\forall_{i=1}^{|Et^+|}\{et_i \exists et_i \in Et^+\}$ Begin

$$isc(et_i) = \frac{\sum_{j=1}^{|R^+|}\{1 \exists et_i \in r_j \wedge r_j \in R^+\}}{|R^+|}$$

End

Find the emoji weight of each token exists in set Et^+ as follows, which is the coverage frequency of the corresponding token in records of the positive label.

$\forall_{i=1}^{|Ej^+|}\{ej_i \exists ej_i \in Ej^+\}$ Begin

$$isc(ej_i) = \frac{\sum_{j=1}^{|R^+|}\{1 \exists ej_i \in r_j \wedge r_j \in R^+\}}{|R^+|}$$

End

Impact Threshold of the feature dimensions

The impact threshold $ist_T^+, ist_{Sl}^+, ist_{Ej}^+, or\ ist_{Et}^+$ of the each dimension in respective order of term, slang, emoji, and emoticon of the features is estimated further, which is the absolute difference of the average of the weight of all tokens of the corresponding dimension of the features, and their root mean square distance.

Impact threshold of the feature dimension terms is

$$\langle T^+ \rangle = \frac{\sum_{i=1}^{|T^+|}\{isc(t_i) \exists t_i \in T^+\}}{|T^+|}$$

$$eT^+ = \frac{\sum_{i=1}^{|T^+|}\left\{\sqrt{\left(\langle T^+ \rangle - isc(t_i)\right)^2} \exists t_i \in T^+\right\}}{|T^+|}$$ // assessing root mean square error

$$ist(T^+) = \sqrt{\left(\langle T^+ \rangle - eT^+\right)^2}$$

Impact threshold of the feature dimension slang is

$$\langle Sl^+ \rangle = \frac{\sum_{i=1}^{|Sl^+|}\{isc(s_i) \exists s_i \in Sl^+\}}{|Sl^+|}$$

$$eSl^+ = \frac{\sum_{i=1}^{|Sl^+|}\left\{\sqrt{\left(\langle Sl^+ \rangle - isc(s_i)\right)^2} \exists s_i \in Sl^+\right\}}{|Sl^+|}$$ // assessing root mean square error

$$ist(Sl^+) = \sqrt{\left(\langle Sl^+ \rangle - eSl^+\right)^2}$$

Impact threshold of the feature dimension Emojies is

$$\langle Ej^+ \rangle = \frac{\sum_{i=1}^{|Ej^+|}\{isc(ej_i) \exists ej_i \in Ej^+\}}{|Ej^+|}$$

$$eEj^+ = \frac{\sum_{i=1}^{|Ej^+|}\left\{\sqrt{\left(\langle Ej^+ \rangle - isc(ej_i)\right)^2} \exists ej_i \in Ej^+\right\}}{|Ej^+|}$$ // assessing root mean square error

$$ist(Ej^+) = \sqrt{\left(\langle Ej^+ \rangle - eEj^+\right)^2}$$

Impact threshold of the feature dimension Emoticons is

$$\langle Et^+ \rangle = \frac{\sum_{i=1}^{|Et^+|}\{isc(et_i) \exists et_i \in Et^+\}}{|Et^+|}$$

$$eEt^+ = \frac{\sum_{i=1}^{|Ej^+|}\left\{\sqrt{\left(\langle Et^+ \rangle - isc(et_i)\right)^2} \exists et_i \in Et^+\right\}}{|Et^+|}$$ // assessing root mean square error

$$ist(Et^{+}) = \sqrt{\left(\left\langle Et^{+}\right\rangle - eEt^{+}\right)^{2}}$$

Similarly, the further process finds the unique tokens of the each feature dimension in different sets $T^{-}, Sl^{-}, Ej^{-}, and\ Et^{-}$ in respective order of feature dimensions terms, slang, emojies, and emoticons. Then, the weight of eachtoken of each feature dimensions of the corpus having records labeled as negative.

Later the process, finds the weight thresholds $ist(T^{-}), ist(Sl^{-}), ist(Ej^{-}), and\ ist(Et^{-})$ of each dimension of the features in regard to the records labeled as negative.

The record level weight of each dimension of the features of the label positive will be assessed in further phase, which is as follows,

$\underset{i=1}{\overset{|R^{+}|}{\forall}}\left\{r_i \exists r_i \in R^{+}\right\}$ Begin

$$isc_{+}\left(r_i^{t}\right) = \frac{\sum_{j=1}^{|T^{+}|}\left\{isc(t_j)\exists t_j \in T^{+} \wedge t_j \in r_i\right\}}{\left|T^{+}\right|}$$

// record level weight of the corresponding dimension called as terms, which is the average of weight observed for all features of the dimension term that are existing the corresponding record r_i^{+}

$$isc_{+}\left(r_i^{s}\right) = \frac{\sum_{j=1}^{|Sl^{+}|}\left\{isc(s_j)\exists s_j \in Sl^{+} \wedge s_j \in r_i\right\}}{\left|Sl^{+}\right|}$$

// record level weight of the corresponding dimension called as slang, which is the average of weight observed for all features of the dimension slang that are existing the corresponding record r_i^{+}

$$isc_{+}\left(r_i^{ej}\right) = \frac{\sum_{j=1}^{|Ej^{+}|}\left\{isc(ej_j)\exists ej_j \in Ej^{+} \wedge ej_j \in r_i\right\}}{\left|Ej^{+}\right|}$$

// record level weight of the corresponding dimension called as emojies, which is the average of weight observed for all features of the dimension emojies that are existing the corresponding record r_i^{+}

$$isc_{+}\left(r_i^{et}\right) = \frac{\sum_{j=1}^{|Et^{+}|}\left\{isc(et_j)\exists et_j \in Et^{+} \wedge et_j \in r_i\right\}}{\left|Et^{+}\right|}$$

// record level weight of the corresponding dimension called as emoticons, which is the average of weight observed for all features of the dimension emoticons that are existing the corresponding record r_i^{+}

End

The adaptation of the aforesaid process enables to find the record level weights divergent dimensions of the features of the records labeled as negative, which are further denoted as notations $isc_{-}\left(r_i^{t}\right), isc_{-}\left(r_i^{s}\right), isc_{-}\left(r_i^{ej}\right), and\ isc_{-}\left(r_i^{et}\right)$ representing for each record r_i of the label negative.

Corpus Level Impact Thresholds

This section defines the corpus level impact thresholds of each feature dimension in regard to each label, which is as follows:

Corpus level Impact threshold of the terms of the records labeled as positive

$$\left\langle R_t^{+}\right\rangle = \frac{\sum_{i=1}^{|R^{+}|}\left\{isc_{+}(r_i^{t})\exists r_i \in R^{+}\right\}}{\left|R^{+}\right|}$$

// finding the average of the term level weight of the records labeled as positive

$$eR_t^{+} = \frac{\sum_{i=1}^{|R^{+}|}\left\{\sqrt{\left(\left\langle R_t^{+}\right\rangle - isc_{+}(r_i^{t})\exists r_i \in R^{+}\right)^{2}}\right\}}{\left|R^{+}\right|}$$

// assessing root mean square error

$$ist\left(R_t^{+}\right) = \sqrt{\left(\left\langle R_t^{+}\right\rangle - eR_t^{+}\right)^{2}}$$

Corpus level Impact threshold of the slang of the records labeled as positive

$$\left\langle R_s^{+}\right\rangle = \frac{\sum_{i=1}^{|R^{+}|}\left\{isc_{+}(r_i^{s})\exists r_i \in R^{+}\right\}}{\left|R^{+}\right|}$$

// finding the average of the slang level weight of the records labeled as positive

$$eR_s^{+} = \frac{\sum_{i=1}^{|R^{+}|}\left\{\sqrt{\left(\left\langle R_s^{+}\right\rangle - isc_{+}(r_i^{s})\exists r_i \in R^{+}\right)^{2}}\right\}}{\left|R^{+}\right|}$$

// assessing root mean square error

$$ist\left(R_s^{+}\right) = \sqrt{\left(\left\langle R_s^{+}\right\rangle - eR_s^{+}\right)^{2}}$$

Corpus level Impact threshold of the emojies of the records labeled as positive

$$\left\langle R_{ej}^{+}\right\rangle = \frac{\sum_{i=1}^{|R^+|}\left\{isc_+(r_i^{ej})\exists r_i \in R^+\right\}}{\left|R^+\right|}$$ // finding the average of the emojies level weight of the records labeled as positive

$$eR_{ej}^{+} = \frac{\sum_{i=1}^{|R^+|}\left\{\sqrt{\left(\left\langle R_{ej}^{+}\right\rangle - isc_+(r_i^{ej})\exists r_i \in R^+\right)^2}\right\}}{\left|R^+\right|}$$ // assessing root mean square error

$$ist\left(R_{ej}^{+}\right) = \sqrt{\left(\left\langle R_{ej}^{+}\right\rangle - eR_{ej}^{+}\right)^2}$$

Corpus level Impact threshold of the emoticons of the records labeled as positive

$$\left\langle R_{et}^{+}\right\rangle = \frac{\sum_{i=1}^{|R^+|}\left\{isc_+(r_i^{et})\exists r_i \in R^+\right\}}{\left|R^+\right|}$$ // finding the average of the emoticons level weight of the records labeled as positive

$$eR_{et}^{+} = \frac{\sum_{i=1}^{|R^+|}\left\{\sqrt{\left(\left\langle R_{et}^{+}\right\rangle - isc_+(r_i^{et})\exists r_i \in R^+\right)^2}\right\}}{\left|R^+\right|}$$ // assessing root mean square error

$$ist\left(R_{et}^{+}\right) = \sqrt{\left(\left\langle R_{et}^{+}\right\rangle - eR_{et}^{+}\right)^2}$$

The similar process on negative labeled records denotes the corpus level impact thresholds $ist\left(R_t^-\right), ist\left(R_s^-\right), ist\left(R_{ej}^-\right), and\ ist\left(R_{et}^-\right)$ in regard to feature dimensions terms, slang, emojies, and emoticons of the records representing the negative sentiment polarity.

The impact scales and thresholds attained from the given training set are further used to assign label to the given test records that explored in following section.

D. Classifying by DFWC

This section portrays the method of using these weights and corresponding thresholds to label the given record that representing the opinion on the target event/product.

Each record r given unlabeled records ^{t}R will be underwent the preprocessing that delivers the features of the vivid dimensions term, slang, emojies, and emoticons as different vectors $tv, sv, ejv, and\ etv$ in respective order.

Further, the weight of each entry in the vector ^{tv} that represents the features of the text tokens will be used to identify their weight towards the both labels positive and negative, which is as follows

$$isc(r_t^+) = \frac{\sum_{i=1}^{|tv|}\left\{isc_+(t_i)\exists t_i \in tv \wedge t_i \in T^+\right\}}{|T^+|}$$ //term level weight of the record towards positive label

$$isc(r_t^-) = \frac{\sum_{i=1}^{|tv|}\left\{isc_-(t_i)\exists t_i \in tv \wedge t_i \in T^-\right\}}{|T^-|}$$ //term level weight of the record towards negative label

$$isc(r_s^+) = \frac{\sum_{i=1}^{|sv|}\left\{isc_+(s_i)\exists s_i \in sv \wedge s_i \in Sl^+\right\}}{|s^+|}$$ //slang level weight of the record towards positive label

$$isc(r_t^-) = \frac{\sum_{i=1}^{|sv|}\left\{isc_-(s_i)\exists s_i \in sv \wedge s_i \in Sl^-\right\}}{|Sl^-|}$$ //slang level weight of the record towards negative label

$$isc(r_{ej}^+) = \frac{\sum_{i=1}^{|ejv|}\left\{isc_+(ej_i)\exists ej_i \in ejv \wedge ej_i \in Ej^+\right\}}{|Ej^+|}$$ //emojies level weight of the record towards positive label

$$isc(r_{ej}^-) = \frac{\sum_{i=1}^{|ejv|}\left\{isc_-(ej_i)\exists ej_i \in ejv \wedge ej_i \in Ej^-\right\}}{|Ej^-|}$$ //emojies level weight of the record towards negative label

$$isc(r_t^+) = \frac{\sum_{i=1}^{|tv|}\left\{isc_+(t_i)\exists t_i \in tv \wedge t_i \in T^+\right\}}{|T^+|}$$ //emoticons level weight of the record towards positive label

$$isc(r_t^-) = \frac{\sum_{i=1}^{|tv|}\left\{isc_-(t_i)\exists t_i \in tv \wedge t_i \in T^-\right\}}{|T^-|}$$ //emoticons level weight of the record towards negative label

Further, the record scope towards positive and negative labels will be assessed as follows:

$$isc_+(r)+= \begin{Bmatrix} 1 & if\ \left(isc(r_t^+) > ist(R_t^+)\right) \\ 0 & else \end{Bmatrix} \quad isc_-(r)+= \begin{Bmatrix} 1 & if\ \left(isc(r_t^-) > ist(R_t^-)\right) \\ 0 & else \end{Bmatrix}$$

$$isc_+(r)+= \begin{Bmatrix} 1 & if\ \left(isc(r_s^+) > ist(R_s^+)\right) \\ 0 & else \end{Bmatrix} \quad isc_-(r)+= \begin{Bmatrix} 1 & if\ \left(isc(r_s^-) > ist(R_s^-)\right) \\ 0 & else \end{Bmatrix}$$

$$isc_+(r)+= \begin{Bmatrix} 1 & if\ \left(isc(r_{ej}^+) > ist(R_{ej}^+)\right) \\ 0 & else \end{Bmatrix} \quad isc_-(r)+= \begin{Bmatrix} 1 & if\ \left(isc(r_{ej}^-) > ist(R_{ej}^-)\right) \\ 0 & else \end{Bmatrix}$$

$$isc_{+}(r)+=\begin{cases}1 & if\ \left(isc(r_{et}^{+})>ist(R_{et}^{+})\right)\\ 0 & else\end{cases}\qquad isc_{-}(r)+=\begin{cases}1 & if\ \left(isc(r_{et}^{-})>ist(R_{et}^{-})\right)\\ 0 & else\end{cases}$$

$if\ \left(isc_{+}(r)>isc_{-}(r)\right)$, then the label of the given record will be concluded as positive, else the label will be concluded as negative.

IV. EXPERIMENTAL STUDY

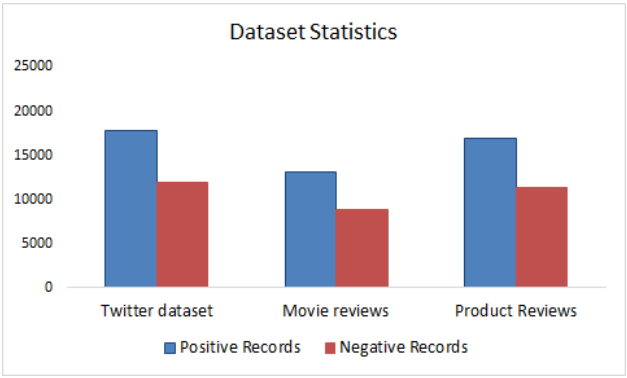
An experimental study conducted on three divergent datasets of target audience opinion over the movies, twitter trends and products. The key criteria for selection of datasets are the quantity, diversity and review presentation. Dataset statics pertaining to input corpus and the training set are explored in section datasets and statistics. Performance analysis based on experimental results is identified with classification process adapted for selective test based on Adaboost classifier in section performance analysis.

A. Datasets and Statistics

Product review datasets [42], Twitter dataset sentiment analysis [43], movie reviews dataset [44] are the major sets chosen for analysis. 27886 records are part of movie dataset wherein 21993 are considered for experiments and the rest discarded for complexity noticed for pre-processing step. Original twitter dataset [43] with tweets marked as positive and negative. Various records wherein 29700 records were retained post the pre-processing stage. Other dataset product reviews of Amazon Instant video from the Amazon product that is constructed in the form of USCD [42], which are considered for experiments. 28310 reviews were chosen for pre-processing. Majority of the datasets were pre-processed and numerous instances that were available were depicted in Table 1. Data sets of Movie Reviews and Product Review contain uniform distribution, whereas Twitter datasets are skewed.

Table 1: Data Statistics for Each Dataset

Dataset Statistics					
Dataset	**Total # of records**	**Positive Records**	**Negative Records**	**Records used for**	**Records used for**
				training	**testing**
Twitter dataset	29700	17820	11880	19800	9900
Movie reviews	21933	13160	8773	14622	7311



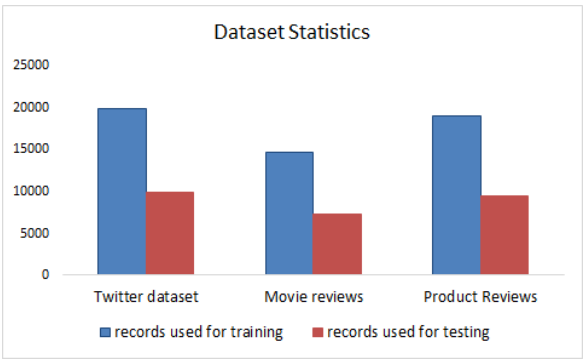


Figure 1: The statistics of the transactions depicted in datasets

For every dataset, overall positive and negative labeled records are classified as 3:3 ratio for training and the testing process categorically. Statistics pertaining to training set are detailed in Table 2 Wherein the positive and negative records are adapted from every dataset and are depicted with sentiment lexicons for each of the corresponding datasets.

Table 2: Training Set Statistics

	Records with Positive Sentiment Labels	Records with Negative Sentiment Labels	Positive Sentiment Lexicons	Negative Sentiment Lexicons
Twitter Trends	10692	7128	23	16
Movie Reviews	7896	5264	22	15
Product Reviews	10192	6794	13	8

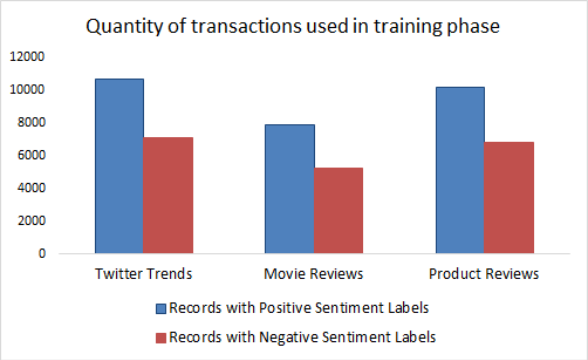


Figure 2: The statistics of the transactions depicted in training set

B. Performance Analysis

In this section, the results attained from the classification process based on Adaboost classifier, wherein the optimal features chosen from proposed feature selection strategy DFWC and results attained from classification process based on same classifier which trains the optimal features chosen from the novel model ofQER. But the optimal feature selection on the basis of both models that carry divergent datasets used in section datasets and statistics using the same statistics. Critical argument of the paper related to contemporary models and the solution for assessing the performance analysis section.

Precision, sensitivity, accuracy are some of the metrics of statistical assessment [45], which are assessed for results for both proposed and the contemporary feature selection models. Diversity and robustness for the intrinsic elements considered for predictive accuracy when compared to contemporary model.

Test statistics and the values attained for performance metrics are discussed in Table 3 and Table 4 for the DFWC. Label prediction stats that are attained for DFWC and QERare seen in Figure 4 and Figure 5 respectively.

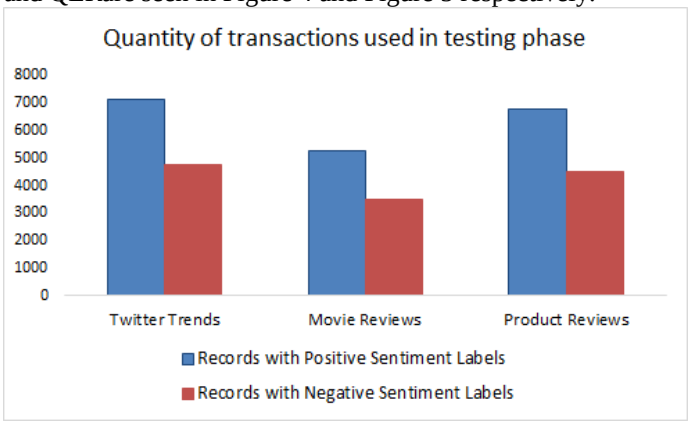


Figure 3: Statistics of Test sets used in experiments

Table 3: Test Set and performance statistics of DFWC

	Records with Positive Sentiment Labels	Records with Negative Sentiment Labels	True Positives	False Positives	True Negatives	False Negatives	Precision	Specificity	Sensitivity	Accuracy
Twitter Trends	7128	4752	6658	456	4296	470	0.94590104	0.91404	0.944063	0.922054
Movie Reviews	5264	3509	4950	315	3194	314	0.95017094	0.920231	0.95035	0.928303
Product Reviews	6794	4530	6366	426	4104	428	0.947279152	0.91596	0.947003	0.924585

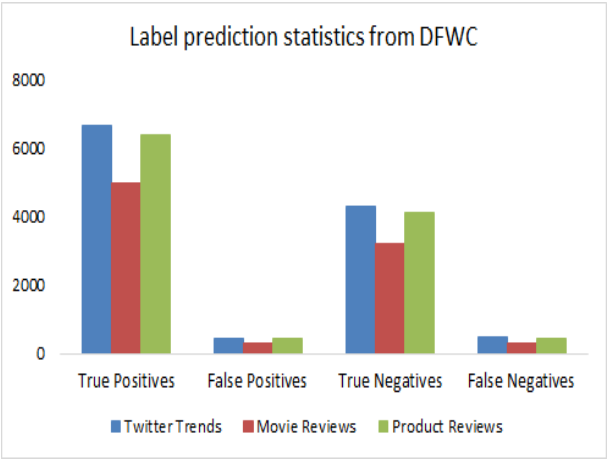


Figure 4: Label prediction statistics observed from DFWC

Table 4: Test Set and Performance Statistics of QER

	Records with Positive Sentiment Labels	Records with Negative Sentiment Labels	True Positives	False Positives	True Negatives	False Negatives	Precision	Specificity	Sensitivity	Accuracy
Twitter Trends	7128	4752	6016	931	3821	1112	0.865985317	0.804082	0.843996	0.82803
Movie Reviews	5264	3509	4266	631	2878	998	0.871145599	0.820177	0.81041	0.814317
Product Reviews	6794	4530	5530	879	3651	1264	0.862849118	0.80596	0.813953	0.810756

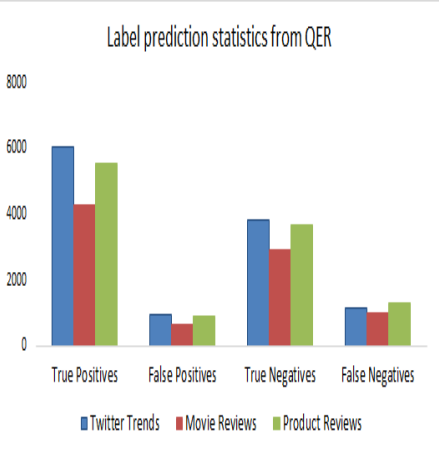


Figure 5: Label prediction statistics observed from QER

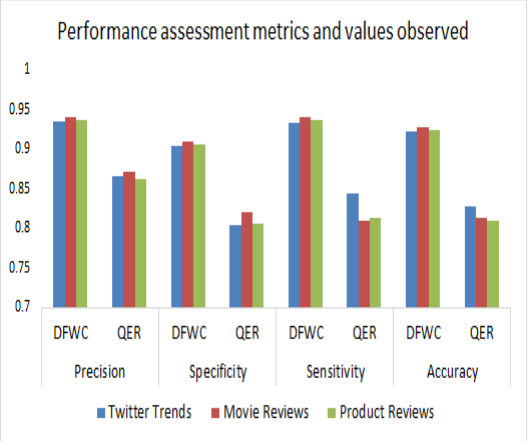


Figure 6: The results obtained for performance metrics from both DFWC and QER

It is imperative that the statistics discussed in Table 3 and Table 4 for the proposed model reflects that outperformed the contemporary model for predictive accuracy. Such metric used in the form of sensitivity reflects the ability in terms of predicting the sentiments that are positive, which are envisaged for 90% from the optimal features chosen for DFWC. However, the sensitivity observed for the novel model is around 82% which is comparatively lower than the model proposed (results represented in Figure 6). In the situation of negative sentiment prediction the presentation of models were 90% and 81% respectively for the proposed and the benchmark model.

Table 5: Average Predictive accuracy at divergent sizes of the input datasets

	10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
DFWC	0.929	0.929	0.928	0.928	0.927	0.926	0.926	0.926	0.926	0.925
QER	0.918	0.918	0.918	0.917	0.827	0.826	0.825	0.825	0.824	0.818

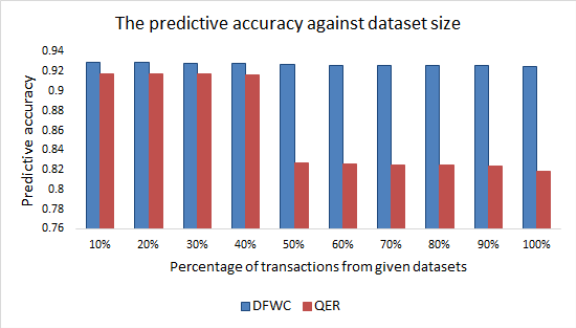


Figure 7: Average predictive Accuracy observed against divergent count of transactions from the given input corpuses

Average predictive accuracy observed is DFWC and another contemporary model levelQER, wherein the three input corpses are discussed in Table 5 and the figurative representation is depicted in Figure 7. Pertaining to the experiments that are carried out iteratively for the all the chosen datasets, the first set performed 10% of the transactions based on every dataset and is incremented based on 10% as the iteration is carried out in further.

Predictive accuracy attained from the DFWC over the corpus tuples reflect that DFWC is sturdy and retains the stability in terms of predictive accuracy, irrespective of dataset size. In contrast, even the predictive accuracy targeted for the contemporary model focus on QERusing the same experimental conditions that are not stable and is inversely proportionate for the size of the corpus size. A statistical model of t-test that is applied reflect that the predictive accuracy observed in DFWC is much stable and distinct when compared to the other compared models. Results of t-test reflect that there is more accuracy by t-value being 4.27245, that has positive outcome. Hence, it can be stated that DFWC is much accurate with degree of probability as 0.000458.

V. CONCLUSION

The contribution of this manuscript endeavored to define a discriminative features weight correlation as fitness scale to classify the given records of opinion in regard to sentiment polarity. Unlike the traditional methods of the contemporary literature, the proposal is considering the multidimensions of the features that are equally significant in regard to the opinions expressed in social web. The experimental study denoting that, the classification accuracy of the proposal is significant and scaled far optimal that compared to the contemporary method found in recent literature. The future version of this work can incorporate the evolutionary computation techniques that uses discriminative features weight correlation as fitness objective.

REFERENCES

1. Yang, Ang, et al. "Enhanced twitter sentiment analysis by using feature selection and combination." 2015 International Symposium on Security and Privacy in Social Networks and Big Data (SocialSec). IEEE, 2015.
2. Venugopalan, Manju, and Deepa Gupta. "Exploring sentiment analysis on twitter data." 2015 Eighth International Conference on Contemporary Computing (IC3). IEEE, 2015.
3. Kanakaraj, Monisha, and Ram Mohana Reddy Guddeti. "NLP based sentiment analysis on Twitter data using ensemble classifiers." 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN). IEEE, 2015.
4. Jain, Anurag P., and Vijay D. Katkar. "Sentiments analysis of Twitter data using data mining." 2015 International Conference on Information Processing (ICIP). IEEE, 2015.
5. Tumasjan, Andranik, et al. "Predicting elections with twitter: What 140 characters reveal about political sentiment." Fourth international AAAI conference on weblogs and social media. 2010.
6. Gayo-Avello, Daniel. "" I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"--A Balanced Survey on Election Prediction using Twitter Data." arXiv preprint arXiv:1204.6441 (2012).
7. Ceron, Andrea, et al. "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France." New Media & Society 16.2 (2014): 340-358.
8. Lamb, Alex, Michael J. Paul, and Mark Dredze. "Separating fact from fear: Tracking flu infections on twitter." Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013.
9. Lee, Kathy, Ankit Agrawal, and Alok Choudhary. "Real-time disease surveillance using twitter data: demonstration on flu and cancer." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013.
10. Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." Proceedings of the 19th international conference on World wide web. ACM, 2010.
11. Culotta, Aron, and Jennifer Cutler. "Mining brand perceptions from twitter social networks." Marketing science 35.3 (2016): 343-362.
12. Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams engineering journal 5.4 (2014): 1093-1113.
13. Gonçalves, Pollyanna, et al. "Comparing and combining sentiment analysis methods." Proceedings of the first ACM conference on Online social networks. ACM, 2013.
14. Dacres, Shana, Hamed Haddadi, and Matthew Purver. "Topic and sentiment analysis on OSNs: a case study of advertising strategies on twitter." arXiv preprint arXiv:1312.6635 (2013).
15. Da Silva, Nadia FF, Eduardo R. Hruschka, and Estevam R. Hruschka Jr. "Tweet sentiment analysis with classifier ensembles." Decision Support Systems 66 (2014): 170-179.
16. Wang, Gang, et al. "Sentiment classification: The contribution of ensemble learning." Decision support systems 57 (2014): 77-93.
17. Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and Trends® in Information Retrieval 2.1–2 (2008): 1-135.
18. Nielsen, Finn Årup. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs." arXiv preprint arXiv:1103.2903 (2011).
19. Mohammad, Saif M., Svetlana Kiritchenko, and Xiaodan Zhu. "Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets." arXiv preprint arXiv:1308.6242 (2013).
20. Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.
21. Kolchyna, Olga, et al. "Twitter sentiment analysis: Lexicon method, machine learning method and their combination." arXiv preprint arXiv:1507.00955 (2015).
22. Hailong, Zhang, Gan Wenyan, and Jiang Bo. "Machine learning and lexicon based methods for sentiment classification: A survey." 2014 11th Web Information System and Application Conference. IEEE, 2014.
23. de Diego, Isaac Martín, et al. "A visual framework for dynamic emotional web analysis." Knowledge-Based Systems 145 (2018): 264-273.
24. Cambria, Erik. "Affective computing and sentiment analysis." IEEE Intelligent Systems 31.2 (2016): 102-107.
25. Machado, Mateus Tarcinalli, Thiago AS

Pardo, and Evandro Eduardo Seron Ruiz. "Creating a Portuguese context sensitive lexicon for sentiment analysis." International Conference on Computational Processing of the Portuguese Language. Springer, Cham, 2018.
26. Al-Moslmi, Tareq, et al. "Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis." Journal of Information Science 44.3 (2018): 345-362.
27. Jaramillo, Francisco, et al. "On-line estimation of the aerobic phase length for partial nitrification processes in SBR based on features extraction and SVM classification." Chemical Engineering Journal 331 (2018): 114-123.
28. López, Julio, Sebastián Maldonado, and Miguel Carrasco. "Double regularization methods for robust feature selection and SVM classification via DC programming." Information Sciences 429 (2018): 377-389.
29. Maldonado, Sebastián, and Julio López. "Dealing with high-dimensional class-imbalanced datasets: embedded feature selection for SVM classification." Applied Soft Computing 67 (2018): 94-105.
30. Achlerkar, Pankaj D., S. R. Samantaray, and M. Sabarimalai Manikandan. "Variational mode decomposition and decision tree based detection and classification of power quality disturbances in grid-connected distributed generation system." IEEE Transactions on Smart Grid 9.4 (2018): 3122-3132.
31. Liu, Xiaoqian, et al. "Differentially private classification with decision tree ensemble." Applied Soft Computing 62 (2018): 807-816.
32. Li, Fenglian, et al. "Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets." Information Sciences 422 (2018): 242-256.
33. Lee, Chang-Hwan. "An information-theoretic filter approach for value weighted classification learning in naive Bayes." Data & Knowledge Engineering 113 (2018): 116-128.
34. Li, Tong, et al. "Differentially private Naive Bayes learning over multiple data sources." Information Sciences 444 (2018): 89-104.
35. Xu, Shuo. "Bayesian Naïve Bayes classifiers to text classification." Journal of Information Science 44.1 (2018): 48-59.
36. Xia, Zhihua, et al. "A Privacy-Preserving Handwritten Signature Verification Method Using Combinational Features and Secure KNN." IEEE Access 6 (2018): 46695-46705.
37. Zheng, Yuhui, et al. "Student's t-hidden Markov model for unsupervised learning using localized feature selection." IEEE Transactions on Circuits and Systems for Video Technology 28.10 (2018): 2586-2598.
38. Xu, Yan, and Lin Chen. "Term-frequency based feature selection methods for text categorization." 2010 Fourth International Conference on Genetic and Evolutionary Computing. IEEE, 2010.
39. Xia, Zhihua, et al. "A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data." IEEE transactions on parallel and distributed systems 27.2 (2016): 340-352.
40. Muddada Muralikrishna, et al. "Frequency and Distribution Diversity based Optimal Feature Selection for Opinion Mining." Jour of Adv Research in Dynamical & Control Systems, 10.2 (2018).
41. Parlar, Tuba, Selma Ayşe Özel, and Fei Song. "QER: a new feature selection method for sentiment analysis." Human-centric Computing and Information Sciences 8.1 (2018): 10.
42. http://jmcauley.ucsd.edu/data/amazon/.
43. http://thinknook.com/wp-content/uploads/2012/09/Sentiment-Analysis-Dataset.zip.
44. http://www.cs.cornell.edu/people/pabo/movie-review-data/.
45. Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011).

AUTHORS PROFILE

1Muddada Muralikrishna,
Research scholar, Computer Science And Systems Engineering, AUCE (A), Andhra University, Visakhapatnam, Andhra Pradesh, India
1E-mail: muralikrishna1926@gmail.com
2Dr.G. Lavanya Devi,
lavanyadevig@yahoo.co.in.
Computer Science and Systems Engineering, AUCE (A), Andhra University, Visakhapatnam, Andhra Pradesh, India2E-mail: