Journal of Web Engineering, Vol. 16, No.1&2 (2017) 161-176
© Rinton Press

MIXED-SENTIMENT CLASSIFICATION OF WEB FORUM POSTS USING LEXICAL AND NON-LEXICAL FEATURES

HIKMAT ULLAH KHAN

Department of Computer Science,

COMSATS Institute of Information Technology, Wah, Pakistan
Hikmat.Ullah@ciitwah.edu.pk

Received August 11, 2016
Revised December 12, 2016

The sentiment detection of the content has become an active research domain in recent years due to the increased availability of public views and opinions in the social web forums. Earlier works detect the sentiment arousal and valence using a lexicon or a dictionary. This paper aims to classify a post content in the social web forums by identifying the mixed-sentiment views and targets to find such posts in which users' views have both positive and negative emotions. Identification of the mixed-sentiment content has several potential applications such as monitoring public views, making products related business decisions and predicting users' behaviors. I propose a non-lexical feature set and compare with the conventional lexicon-based sentiment feature set. The four state of the art classification algorithms applied on the large dataset of public forum verify that the proposed non-lexical features are helpful to find the mixed-sentiment in online forums. The main contribution is the proposal and validation of such features which do not need a lexicon. In addition, a comprehensive analysis of the dataset has been carried out using the power law analysis. The features have been ranked according to their significance in the classification model to identify mixed-sentiment content in the social web forum.

Key words: Opinion Mining, Web Forum, Supervised Learning, Mixed-Sentiment, Feature analysis.
Communicated by: D. Schwabe & F. Vitali

1 Introduction

Social web media such as users' comments, reviews, blogs, debates and discussions contain information on diverse topics. Users express their views and opinions from anywhere in the world and discuss various topics of interest. A large number of users discuss social, religious, political and technical issues. An important aspect of public discussions is sentiment and usually all such topics to have conversations where people share their arguments in favor of their point of view or disagreeing with others' views. Opinion mining aims at analyzing the users' opinions and emotions towards different products, issues and people [1]. Various opinion mining studies analyze social web content and help us in having a deep understanding of the human behavior [2], [3], [4], and [5]. In opinion mining, subjectivity mining classifies the content into subjective or objective [6] and emotion polarity detection categorizes the opinions into positive or negative. Identification of mixed-sentiments having positive emotions and negative emotions is one of the domains of opinion mining. Few related works focus to find contentions from discussions and debates [7], to mine contrasting opinion on political texts [8], to discover sentiment based contradictions [9]. The details are provided in related work Section.

The basic aim of the proposed research is to find such content in online forums in which users express mixed-sentiments. Identification of such discussions has various potential applications such as monitoring public views, making products related business decisions and predicting users' behaviors. Sentiment mixture is helpful to find the social and political issues about which the people express bifurcated sentiments, to find out certain products which receive reviews of diverse emotional valence from users and to foresee reaction about certain forthcoming government policies. To explain mixed opinion for better understanding, we may say a review of a product or a song can be referred as mixed opnion, if a number of people like it and also dislikes it. Similarly, a mixed review contains both good and bad opinions. For instance[a], "the music got mixed reviews because some people thought it was wonderful and others disliked it" is the example of a mixed review. Our aim is to identify whether the comments contain mixed-sentiment or not.

In this research work, non-lexical feature set is proposed and the feature set is compared with lexicon based features to find mixed-sentiment content. This feature-centric approach uses a large real world dataset of Web forum containing thousands of threads on diverse topics. The non-lexical features do not need any lexicon for computation and are computable irrespective of the content language. Sentiment mining imposes many challenges such as natural language processing, co-reference resolution and relation extraction. Social web presents challenges to detect for opinion mining due to the content-generation facility. Users' comments are usually grammatically incorrect due to informal an writing style, spelling mistakes, abbreviations, hashtags [10]. Social web content suffers from lack of contextual information and contains sarcasm and irony which leads to disorientation [10], [11]. To cater to such difficulties, I propose a non-lexical feature set which does not need a lexicon for computation and is helpful to identify mixed-sentiment views in web discussions.

The rest of the paper is organized as follows: The next section reviews the relevant literature. Section 3 formulates the problem. Section 4 proposes the post features, provides an algorithm and framework of research methodology. Section 5 describes the experimental setup. Section 6 discusses the results before concluding the paper.

2 Related Work

Opinion mining is an active field of research. A number of works of opinion mining focus on the social web. Let us review few recent work of opinion mining, sentiment mining in online forums and relevant studies of mixed-opinions.

Opinion mining is used to separate subjective and objective views. It also aims to recognize the positive, neutral or negative polarity of the opinions [12] [13]. A number of works summarize the product reviews of online review sites [14] [15]. A review usually has a single role, of depicting user's feedback about a certain product. In contrast, thread structure comprises a sequence of posts from multiple users and these posts serve multiple roles including feedback, junk and question [16]. Hai et

[a] http://www.englishbaby.com/vocab/word/4084/mixed-reviews. Accessed on Spetember 19, 2016.

al., [17] propose a new approach for inter-corpus feature extraction from online review corpora by introducing feature filtering criterion for opinion identification. Lavaniya and Varthini [18] proposed a feature based sentiment classification approach for web opinion documents. Gangemi et.al., [19] built a model using cognitively inspired frames for the detection of holder, topic and sub-topic of opinion.

Now focusing to research works related to online forums, a study by Hassan et al., [3] examines the thread structure to categorize the users' attitude towards other users. This sentence level classification facilitates the identification of attitudinal sentences, interaction dynamics of discussions and formation as well as break up of users' groups in online forums. Other promising work includes identification of evaluative and non-evaluative sentences from opinions in online posts [20]. The dialog structure of discussions is analyzed from debate perspective [21] and disagreement among the posts [22]. Weninger et al., [4] explore the notion of discussion threads for social news site of Reddit Community by evaluating the comment thread on the basis of the least common ancestor of hLDA clusters. It shares that the depth of comments in a discussion increases with the passage of time. Duan and Zhai carry out the study of smoothing schemes to improvise natural forum thread structure. They proposed new smoothing schemes for the natural language model. Their scheme is twofold: consists of model expansion and count expansion. Further weighting functions such as distance and content similarity are incorporated to improve accuracy of posts retrieval [23]. Biyani et al. [24] [28] conducted a study for the subjective analysis of online forum threads. They used structural features of threads for identification of thread subjectivity orientation. The main aim of the proposed approach focuses on finding mixed-sentiment posts and discussions in online forums using sentiment, dialog and thread-structure features.

The concept of analyzing mixed-sentiment is found in various studies. One of the first works, analyzes the negative and positive sentiments in web [25]. Andre Bizau et al., [26] describes a natural language method to find opinion diversity expressed in text. They capture positive and negative reviews from the online forum. Another work detects opinionated claims in online discussions using machine learning techniques in LiveJournal and Wikipedia data [27]. It uses lexical features for sentiment analysis for identification of users' arguments and claims. Fang Yi et al., [8] use political text to analyze the contrasting opinions of the users on politics and to quantify their difference. A similar work [7] states that contentions are the important feature of forums which discuss political, social and religious issues and tries to discover the agreement and contention indicator expressions at post and discussions level in web data. A number of these works use lexical features to mine the sentiment.

3 Problem Formulation and Problem Statement

A user can initiate a new topic by creating a new thread in an online forum. Th main aim is to identify only those posts which have positive and negative emotions in it and to find out those discussions having where positive and negative emotional views. During discussion, topic can drift in the threads and this assumption may not hold right always but such exceptional cases are out of the scope of the paper.

Formally, a forum post p is a sequence of words in a Vocabulary set V, a forum thread t is a sequence of posts i.e., $t = \{p_1, p_2, \dots, p_L\}$ where p_i is the i^th^ post in the thread and forum f to be a collection of threads $f = \{t_1, t_2$,…,$t_m\}$ where t_j is a thread. A post having high positive and negative sentiment values is regarded as mixed-sentiment post, denoted as p_{ms} and denoted as p_{nms} otherwise.

Given an online forum f and the set of thread t having a number of posts, my aim is to classify each post p_i into one of the two given classes: Mixed-sentiment posts (denoted by p_{ms}) or otherwise (p_{nms}) in case of posts

4 Research Methodology and Features Engineering

The proposed research is based on introduction of feature sets. I present the feature sets, the framework of the research methodology and the algorithm to compute the both the feature sets.

I propose several thread-structure features in addition to various proposed dialogs and sentiment based features. The main purpose is to investigate the effects of all the proposed features in identification of mixed-sentiment posts and threads. The table 1 provides the list of symbols used to calculate the proposed features and the post features are presented in table 2.

Table 1: List of Symbols used in the paper

Symbols	Description
T	Set of Threads
P	Set of Posts
U	Set of Users
t	$t \in T$
p	$p \in P$
u	$u \in U$
N_{un}^{p}	Number of Username mentioned in post
N_{q}^{p}	Number of textQuoted in post
N_{ur}^{p}	Number of URL in post
N_{c}^{p}	Number of Capital words in post
N_{s}^{p}	Number of sentiment words in post (using SentiWordNet)
N_{w}^{p}	Number of Words in post
$S_{p'}^{p}$	Positive score of the post (using SentiWordNet)
S_{n}^{p}	Negative score of the post (using SentiWordNet)

N^{p}_{pw}	Number of positive words in post. (A word is a positive word if its positive value > 0 and negative value = 0) (using SentiWordNet)
N^{p}_{nw}	Number of negative words in post. (A word is a positive word if its negative > 0 and positive value = 0) (using SentiWordNet)
N^{p}_{mw}	Number of Mixed-sentiment words in post. (A word is a Mixed-word if its negative > 0 and positive value > 0) (using SentiWordNet)
S^{p}_{mw}	Mixed-sentiment Score of the post based on Mixed-words.

The post features set consists of two types lexical based sentiment features and non-lexical based dialog act features set

4.1 Lexical Features

These features take into account the users' sentiment. Mixed-sentiment content has high positive and negative sentiment valence. The features are computed using various resources such as sentiment lexicon (like SentiWordNet and WordNet-Affect) and sentiment analysis tool (such as SentiStrength [29] [30] and LIWC [31]). I applied SentiWordNet [32] which is a widely used lexicon to calculate the positive and negative scores of a post content. The positive and negative sentiment scores of a post are denoted as pPositiveScore and pNegativeScore respectively and their absolute difference shows how close are both sentiment score (pSentiScore). Similarly the sentiment word feature (pSentiWordsScore) is the difference of the number of positive and negative words. The positive and negative words have been counted with the help of SentiWordNet. A positive word has positive value greater than zero and negative value is equal to zero. Similarly, a negative word has negative value greater than zero while positive value is equal to zero. The mixed-word has both positive and negative values greater than zero. An opinionative post is usually lengthier than an informative one (pPostLength) so it is considered as a feature. [33].

4.2 Non-Lexical Features

Mixed-sentiment topics have a higher chance of dialog and thus the discussions about mixed-sentiment topics have more chance of conversations among the users. Let me posit that dialog features help to detect mixed-sentiment posts. The first feature is the existence of URLs (boolURL) as the user share links to the web pages or other posts within the forum as an argument or evidence about their point of views. A user mentions some other user's name who already posted his comments to seek his/her attention for replying his question or sharing point of view. It is assumed that such user mentions are more common in an emotional conversation than in technical discussion so user mention is taken as the dialog feature (boolUsername). Likewise, a user copies text of an earlier post (boolQuotedText). The content in upper case depicts shouting, showing negative emotion or low valence. So the relevant feature (boolCapital) may be helpful to recognize dialog. Boolean features exhibit the existence of certain characteristics.

Table 2: Post Feature Sets

Symbol	Feature Name	Description	Calculating Formula
	Lexical Features		
S_s^p	pSentiScore	Sentiment Score of the post	$S_s^p = S_{p'}^p - S_n^p$
S_w^p	pSentiWords	Sentiment Score based on Sentiment Words	$S_w^p = N_{pw}^p - N_{nw}^p$
S_d^p	pMixWordsScore	Mixed Words Score based on Mixed-sentiment Words	$S_{mw}^p = \frac{N_{mw}^p}{N_w^p}$
N_l^p	pPostLength	Post length	
	Non-Lexical Features		
b_{un}^p	boolUsername	Existence of Username mentioned in the post	$b_{un}^p = \begin{cases} 1, & if\ N_{un}^p > 0 \\ 0, & otherwise \end{cases}$
b_q^p	boolQuotedText	Existence of earlier thread posts quoted in the post	$b_q^p = \begin{cases} 1, & if\ N_q^p > 0 \\ 0, & otherwise \end{cases}$
b_{ur}^p	boolURL	Existence of URL in post	$b_{ur}^p = \begin{cases} 1, & if\ N_{ur}^p > 0 \\ 0, & otherwise \end{cases}$
b_c^p	boolCapital	Existence of Capital Case words in the post	$b_c^p = \begin{cases} 1, & if\ N_c^p > 0 \\ 0, & otherwise \end{cases}$

4.3 Algorithm and Proposed Framework

The proposed framework for the research is presented in the Figure 1 and the algorithm to compute the feature mentioned above are given in the algorithm.

ALGORITHM: Mixed-Sentiment Classification of Posts in Web Forum

Input: Data of Posts in Web Forum

Output: Post classified as Mixed-Sentiment or non-Mixed-Sentiment

```
Initialize $S_s^p$, $S_s^p$, $S_s^p$, $N_l^p$, $b_{ur}^p$, $N_c^p$
FOR each $t \in T$
    FOR each $p \in P$
        $N_w^p =$ CountWords($p$)
        $S_{p'}^p$ = CalculatePostiveSWNScore ($p$)
```

```
            $S_n^p$ = CalculateNegativeSWNScore($p$)
            $N_{pw}^p$ = CountPostitiveSWNWords($p$)
            $N_{nw}^p$ = CountPostitiveSWNWords($p$)
            $N_{mw}^p$ = ComputeMixedOpinionSWNWords($p$)
▷ Computation of Sentiment Feature Set ($F_S^p$)
            $S_s^p = S_{p'}^p - S_n^p$
            $S_w^p = N_{pw}^p - N_{nw}^p$
            $S_{mw}^p = \frac{N_{mw}^p}{N_w^p}$
            $N_l^p = ComputePostLength(p)$
            $F_S^p = [F_S^p ; S_s^p, S_s^p, S_s^p, N_l^p]$
▷ Computation of Post Dialog Feature Set ($F_D^p$)
            IF $p$ contains URL THEN
                    $b_{ur}^p = 1$
               End IF
            IF $p$ contains Capital Word THEN
                    $N_c^p = N_c^p + 1$
               End IF
            IF $p$ contains Username of Earlier posts in the thread $t$ THEN
                    $b_{un}^p$ = 1
            End IF
            IF $p$ contains Quoted text from Earlier posts in the thread $t$ THEN
                    $b_q^p = 1$
               End IF
            $F_D^p = [N_{ur}^p, N_c^p, b_{un}^p, b_q^p]$
      End FOR
End FOR
Class = Classifier($F_S^p$,$F_D^p$)
IF Class = 1 then
    $p' = p_{ms}$
Else
    $p' = p_{nms}$
End IF
STOP ▷End of the Algorithm
```

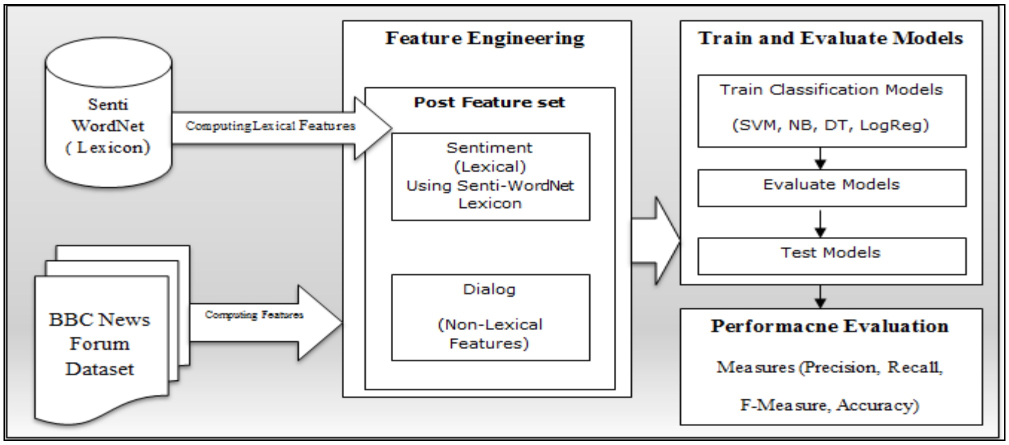

Figure 1: A Framework of Proposed Research Methodology

5 Experimental Setup

Let me here briefly describe the classification algorithms applied, the BBC forum dataset used and the performance evaluation measures used to analyze the results.

5.1 Classification Algorithms

Oracle Data Miner, a tool for data mining and analysis, is used. The classification algorithms applied include Naïve Bayes, Support Vector Machine, Decision Tree and Logistic Regression. The use of supervised learning techniques is a well-known approach which produced good results in subjectivity and opinion mining [34], [12], [35], [24].

5.2 Dataset

The choice of dataset is significant as it should cover diverse topics from factual to opinionative and a large number of users from all over the world share their views. Dataset of BBC Forum [36], a public discussion forum, provides positive and negative emotions for each post. The dataset has been taken from the CyberEmotions[b] team which provided the dataset free of cose for the research purposes. It contains discussions about topics of news, social issues, political and religious views for the period from July 2005 to June 2009. The statistics of the dataset are given in the table 3 as follows:

Table 3: BBC Forum Dataset statistics

Threads	97,946
Posts	2,474,781
Users	18,045
Average Posts in a Thread	10
Average Users in a Thread	8

[b] http://www.cyberemotions.eu/data.html. Accessed on September 07,2016.

Average Thread Life	112
Average Thread Length	331

Oracle Data Miner [37] uses k-fold cross validation techniques, the value of k is set as 10 which is usually considered as its standard value. The positive and negative emotional values for each post are between 1 to 5. The higher the value, the stronger is the emotion. The low emotion value is between 1 and 2 while strong emotional value is 3, 4 or 5. A few posts have 5 as emotion value. To evaluate the high positive and negative emotion, a post is categorized as Mixed-sentiment having both positive emotion and negative emotion greater than 2. There are total 1,92,158 posts, out of those only 92, 159 posts are Mixed-sentiment posts.

5.3 Performance Evaluation Measures

For the evaluation purposes, the standard performance evaluation measures of Accuracy, Precision, Recall and F-Measure have calculated using the following:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F - Measure = 2. Precision. \frac{Recall}{Precision+Recall} \quad (4)$$

where TP, TN, FP, FN represents True Positive, True Negative, False Positive and False Negative respectively.

6 Results and Discussion

Here, let me discuss the results in three steps. First, the BBC dataset is described by carrying out its statistical and power law analysis. Then the results obtained by the classification algorithms are elaborated and lastly, the features have been ranked accordin to their significance in the proposed method.

6.1 Dataset Analysis

A relationship exists between the properties of a dataset, which are represented in the form of variables. A relationship can either be linear or non-linear. If a value of a quantity alters as a power of second quantity value, then such a relationship is expressed using power law. The power law is used to study the probability distributions which is done for data analysis . The distributions of a large data of physical, biological, and human related fields of life follow the power law distributions. For instance, the frequencies of words in most languages, the ranges of earth quakes etc., observe power law form. For the analysis of the number of users and their participation in the social web forum, power law analysis has been used as well. The number of posts or messages in threads varies following power law as well. It has been followed using the various research works [24,28,35]. It has been observed that the empirical power-law distributions hold for a limited range and a number of such values observe the law but others fir the power law in the tail.

In order to represent the uncertainty in the experimental values, a deviation term ε is added in the power law $O(x^k)$ or simply $y = ax^k + \varepsilon$ for observing the value of a deviation from the main function of the power law. The power law distribution is given as follows:

$$p(x) = cx^{-\propto} \qquad \textit{For } x > x_{min} \qquad (5)$$

If the observed value of $\propto$ is found to be higher than unity, then the tail spread across an infinite region. It is required that the minimum value (xmin) should be as less as possible. The constant value of co-efficient C is the scaling factor that helps to make sure the total are is 1, which is the requirement of the probability distribution.

To study the overall data distribution in the data of BBC over all the five years of the dataset, the statistical analysis is carried out on user's posts quantities. The objective of the analysis is to observe the relationship between the user behavior and the distribution of posts overs the years. Different users have different behavior in threads of online forum threads. An active user may share more comments as compared to other inactive users and this behavior is usually common in all the social web forums. The analysis and sentiment level have also been done in [38,39,40].

Let us consider a user activity, referred as <ai> , defined as the count of the posts of a user i and this is represented as a. The maximum number of the posts by a user in BBC dataset is, i.e., amax = 18,289. It represents that the single user has shared more than eighteen thousand messages. On the other hand, the average posts per user or the average activity is <aavg> = 136, and the median value of user activity is amed = 3, which is very less and suggest that the majority of the users remain very less active in the forum. The observed statistics are given in Table 4 which shows the results for the BBC dataset [36]. The similar trend is observed for the other two recent datasets as well.

Table 4: A Comparision of Dataset Characteristics

Dataset Features	***BBC***
Users Count	18,045
Maximum number of messages by a single user $<a_{max}>$	18,289
Average posts in threads $< a_{avg} >$	136
Median $< a_{med} >$	3
Value of β	2.04
Value of Best Fit R^2	0.950

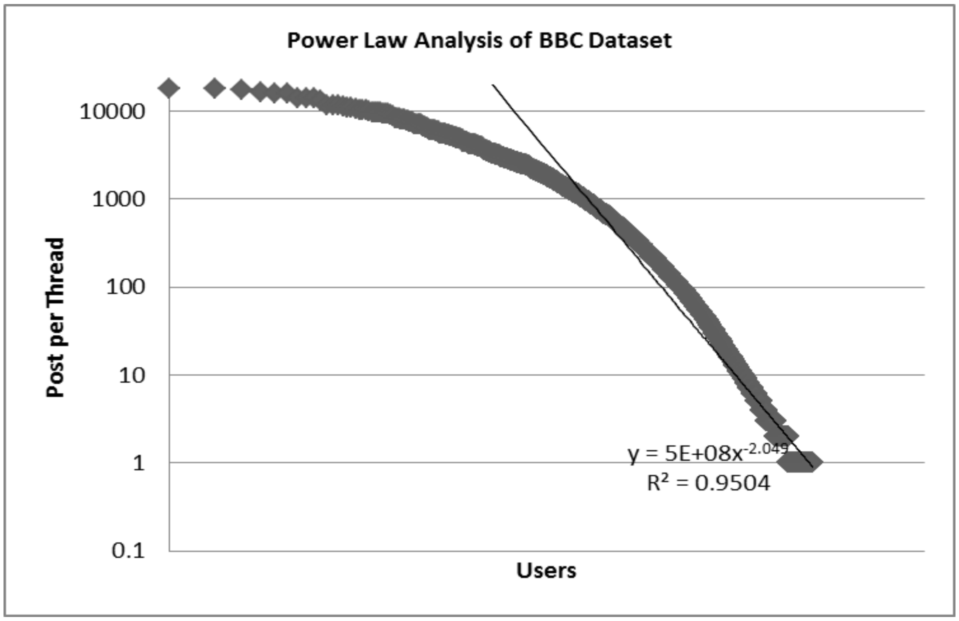

Figure 1: Power Law Analysis of BBC dataset using Log Scale Chart

The power law analysis of the BBC dataset in Figure 1 shows that the dataset follows the normal distributions as majority of the datasets. The value of R2, which is known as the “co-efficient of Regression”, represents the accuracy of the curve with respect to the data. It ranges from 0 to 1. The higher values depict the accurate result or the fitness of the curve. Its value for BBC dataset is 0.95 which reveal the very high level of accuracy or “best fit” with respect to its data. Similarly, the value of β shows high number of active users of the data. Its value is relatively higher which reveal that the users activity remains similar in recent and BBC dataset as on average more number of users are active in the forum of BBC news.

6.2 Classification Results

First, the results are elaborated by comparing the baseline as well as proposed features by applying the classification algorithm. The table 5 shows that both conventional sentiment features and proposed dialog features contribute to find mixed-sentiment posts and their combined feature set shows better results as evidenced from the Recall results. Considering the performance evaluation measures, higher Recall and F-measure are better as compared to Accuracy and Precision.

To compare the classification algorithms, decision tree outperforms other classification algorithms showing the best overall results. Logistic regression performs better using sentiment features. Similar results are evident using thread-structure features also in which decision tree shows better results comparing accuracy, recall and F-measure based results with those of other classification methods. Logistics regression show optimal results in precision. Naïve Bayes and SVM show relatively similar results as well. The maximum results obtained using both feature sets and their combination is presented in Figure 2. An analysis of the figure reveals that the lexical features do have their importance and have better results as compared to dialog results, but the combination of non-lexical results along with lexical results outperform the conventional lexical results. The behavior of the feature sets is consistent using various performance evaluation measures.

Table 5: Results to find Mixed-sentiment Post

Feature Set	**Classification Algorithm**	**Accuracy**	**Precision**	**Recall**	**F-Measure**
Lexical	***NB***	0.667	0.635	0.712	0.671
	DT	**0.685**	0.649	**0.743**	**0.693**
	SVM	0.670	0.632	0.742	0.683
	LogReg	0.650	**0.691**	0.487	0.571
Non-Lexical	***NB***	0.642	0.636	0.590	0.612
	DT	**0.642**	0.636	0.590	0.612
	SVM	0.513	0.489	0.400	0.440
	LogReg	0.642	**0.636**	**0.590**	**0.612**
Lexical+Non-Lexical	***NB***	0.682	0.649	0.731	0.687
	DT	**0.692**	0.636	**0.833**	**0.721**
	SVM	0.602	0.555	**0.836**	0.667
	LogReg	0.679	**0.683**	0.609	0.644

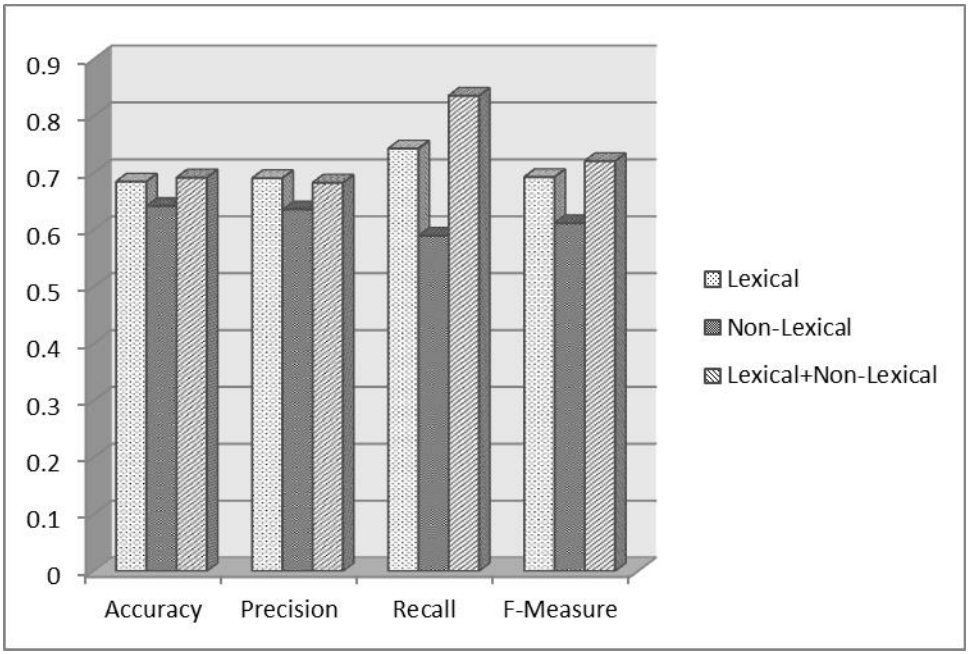

Figure 2: A Comparison of Lexical and Non-Lexical Feature Sets.

6.3 Features Analysis

In addition to providing the classification results, the Oracle data miner compute Wald chi-square statistic to evaluate the statistical significance of each coefficient (feature f in this case) in the model.

$$W_j = \frac{f_j}{SE_{b_j}} \tag{5}$$

where W is the Wald's statistic having normal distribution, f is the feature (coefficient) and SE is the standard error. The value of Wald's statistic is squared to yield a Wald statistic with a chi-square distribution.

$$SE_j = sqrt(diag(H^{-1})_j) \tag{6}$$

where

$$H = [X^T WX] \tag{7}$$

The ranking of each feature within feature set is shown in table 6. The top-ranked post length reveals that users write lengthy content in mixed-sentiment discussions. The feature of mixed-sentiment words (pMixWordsScore) is helpful to find mixed-sentiment posts. The sentiment score is more significant as compared to the sentiment words feature, which is understandable as the former shows the strength of the sentiment valence while the former counts the number of sentimental words. In dialog features, the top-ranked capital content verifies that it shows strong hate or negative emotion which is a common feature in controversial topic discussions. The feature of mentioning the user's name is also significant. It is understandable that quoted text may be copied in the reply-post in a question-answer thread or a user quote above text to add argument in favor of the earlier post. In other words, the direct mentioning the user is more important than quoting the text of the earlier post. The feature of provision of URL does not enjoy high rank because a reference may be given in an informative post to provide links to a detail of a fact or it may be given in reply to question and does not necessary be an important characteristic of an opinionative post.

Table 6: Top Post features ranked by chi-square values

Sentiment Features	**Dialog Feature**
numCharPost	boolCapital
pMixWordsScore	boolUsername
pSentiWords	boolQuotedText
	boolURL

7 Conclusions and Future Work

In this paper, non-lexical feature set is compared with the conventional lexical feature set to identify the mixed-sentiment posts in web forums. Mixed-sentiment content contain high positive and the high negative sentiments and depict the high diversity in emotional valence. The analysis on real-world large dataset is carried out and the four state of the art classification algorithms have been applied that use feature sets for binary classification of web posts. The introduction of the non-lexical features for classification of online forum posts is the main contribution in this work. The proposed feature set decreases the complexity of the learning model in comparison to the existing lexicon based models without compromising the performance. The proposed model may help to find the products having positive as well negative feedback, the social issues about which people have a strong difference of views. This is helpful to identify controversial topics or to find certain issues or policies about which the public has contradictory or bifurcated opinions. The potential future work is to classify the threads and identify the mixed-opinion discussions in online threads. In addition, the future plan is to do topic-sensitive classification of the online forum threads using non-lexical features or to use a different dataset such as web blog data.

Acknowledgements

The author wishes to acknowledge the Cyberemotions team for providing the BBC dataset for research purpose.

References

1. Pang B, Lee L. Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retr. 2008. 2(1). 1-135.
2. Chung J E, Mustafaraj E. Can Collective Sentiment Expressed on Twitter Predict Political Elections?. Proceedings of AAAI Conference on Artificial Intelligence, 2011.
3. Hassan A, Qazvinian V, Radev D. What's with the Attitude?: Identifying Sentences with Attitude in Online Discussions. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 2010.
4. Weninger T, Zhu XA, Han J. An Exploration of Discussion Threads in Social News Sites: A Case Study of the Reddit Community. Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, New York, NY, USA, 2013.
5. Petz GE. Opinion Mining on the Web 2.0. Characteristics of User Generated Content and Their Impacts. Proceedings of Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, A. Holzinger and G. Pasi, Eds., Springer, Berlin Heidelberg, 2013. p. 35-46.
6. Tsytsarau M, Palpanas T.Survey on mining subjective data on the web. Data Mining and Knowledge discovery. 2012. 24(3). 478-514.
7. Mukherjee A, Liu B. Mining Contentions from Disussions and Debates. Proceedings of Knowledge discovery and data mining, Beijing, China, 2012.
8. Fang Y, Si L, Somasundaram N, Yu Z. Mining Contrastive Opinions on Political Texts using Cross-Perspective Topic Model. Proceedings of Web search and data mining, 2012.
9. Tsytsarau M, Palpanas T, Denecke K., Scalable Discovery of Contradictions on the Web. Proceedings of World Wide Web, Raleigh, NC, USA, April 26-30, 2010.
10. Maynard KB, Rout D. Challenges in developing opinion mining tools for social media. Proceedings of @NLP can u tag user generated content? Workshop at LREC, 2012.
11. Azam BS. Opinion Mining: Issues and Challenges (A survey). International Journal of Computer Applications. 2012. 49(9).

12. Cambria E, Schuller B, Xia Y, Havasi C. New Avenues in Opinion Mining and Sentiment Analysis. IEEE Intelligent Systems. March 2013. 28(2). 15-21.
13. Liu B. Sentiment analysis and subjectivity. Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca, 2010.
14. Hu M, Liu B. Mining and Summarizing Customer Reviews. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2004.
15. Fabbrizio GD, Aker A, Gaizauskas R. Summarizing Online Reviews Using Aspect Rating Distributions and Language Modeling. IEEE Intelligent Systems. May 2013. 28(3). 28-37.
16. Sumit B, Prakhar B, Prasenjit M. Classifying User Messages For Managing Web Forum Data. Proceedings of 15th International Workshop on the Web and Databases, p. 13-18, 2012.
17. Hai Z, Chang K, Kim JJ, Yang C. Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance. IEEE Transactions on Knowledge and Data Engineering. March 2014. 26(3).623-634.
18. S. Lavanya, B. Varthini, Sentiment classification of web opinion documents, in Electronics and Communication Systems (ICECS), 2014 International Conference on, 2014.
19. Gangemi A, Presutti V, Reforgiato RD. Frame-Based Detection of Opinion Holders and Topics: A Model and a Tool. IEEE Computational Intelligence Magazine. 2014. 9(1). 20-30.
20. Zhongwu Z, Liu B, Zhang L, Xu H, Jia P. Identifying Evaluative Sentences in Online Discussions. Proceedings of AAAI, 2011.
21. Walker MA, Anand P, Abbott R, Tree JF, Martell C, King J, That is Your Evidence?: Classifying Stance in Online Political Debate. Decis. Support Syst. November 2012. 53(4). 719-729.
22. Abbott R, Walker M, Anand P, Fox JE, Bowmani R, King J. How Can You Say Such Things: Recognizing Disagreement in Informal Political Argument. Proceedings of the Workshop on Languages in Social Media, Stroudsburg, PA, USA, 2011.
23. Duan H, Zhai C. Exploiting Thread Structures to Improve Smoothing of Language Models for Forum Post Retrieval. Proceedings of the 33rd European Conference on Advances in Information Retrieval, Berlin, Heidelberg, 2011.
24. Biyani P, Bhatia S, Caragea C, Mitra P. Using non-lexical features for identifying factual and opinionative threads in online forums. Knowledge-Based Systems. 2014. 69. 170-178.
25. Mei Q, Ling X, Wondra M, Su H, Zhai C. Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. Proceedings of World Wide World, New York, USA, 2007.
26. Bizau A, Rusu D, Mladenic D. Expressing Opinion Diversity. Proceedings of DiverseWed-2011, Knowledge diversity on the Web, 2011.
27. Rosentinal S, McKeown K. Detecting Opinionated Claims in Online Discussions. Proceedings of IEEE 6th International Conference on Semantic Computing, 2012.
28. Biyani P, Bhatia S, Caragea C and Mitra P. Thread Specific Features are Helpful for Identifying Subjectivity Orientation of Online Forum Threads. Proceedings of COLING, 2012.
29. Thelwall M, Buckley K, Platoglou G, Kappas A. Sentiment Strength detection in short informal text. Journal of the American Society for Information Science and Technology. 2010. 61(12). 2544-2558.
30. Thelwall M, Buckley K, Platoglou G. Sentiment Strength detection for the social web, Journal of the American Society for Information Science and Technology. 2011. 63(1). 163-173.
31. Pennebaker WJ, Mehl RM, Niederhoffer GK. Psychological Aspects of Natural Language Use: Our Words, Our Selves. Annual Review of Psychology. 2003. 54(1). 547-577.
32. Esuli A, Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentimet Analysis and Opinion Mining. Proceedings of Seventh Conference on International Language Resource and Evaluation, 2010.

33. Chemiei A, Sobkowicz P, Sienkiewicz J, Paltogios P, Buckley K, Thelwall M and Holyst JA. Negative emotions boost user activity at BBC forum. Physica A: Statistical Mechanics and its Applications. 2011. 390(16). 2936-2944.
34. Pang B, Lillian L, Shivakunar V. Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of Conference on Empirical Methods in Language Processing (EMNLP), Philadelphia, July 2002.
35. Biyani P, Caragea C, Singh A, Mitra P. I Want What I Need!: Analyzing Subjectivity of Online Forum Threads. Proceedings of the 21st ACM International Conference on Information and Knowledge Management, New York, NY, USA, 2012.
36. Paltoglou G, Thelwall M. Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media, ACM Transactions on Intelligent Systems and Technologies. September 2012. 3(4). 1-19.
37. Tamayo P, Berger C, Campos M, Yarmus J, Milenova B, Mozes A, Taft M, Hornick R, Krishnan R, Thomas S, Kelly M, Mukhin D, Haberstroh B, Stephens S and Myczkowski J. Oracle Data Mining. Data Mining and Knowledge Discovery Handbook. 2005. p. 1315-1329.
38. Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance N, Cost-effective outbreak detection in networks, Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. P. 420-429, 2007.
39. Severyn A, Moschitti A, Uryupina O, Plank B, Filippova K, Multi-lingual opinion mining on YouTube, Image Processing & Management, 2016. 52(1). 46-60.
40. Zhou g, Zhu Z, He T, H X, Cross-lingual sentiment classification with stacked utoencoders,Knowledge and Information Systems, 2016. 47(1). 27-44.