A TWO LAYER FEATURE VECTOR GENERATION FOR OSIDL CLASSIFIER

[1]Salina Adinarayana, [2]E. Ilavarasan

[1] Research Scholar, [2] Professor

[1, 2] Department of Computer Science Engineering

[1, 2] Pondicherry Engineering College, Pondicherry, India

Abstract : In today's Internet era, massive generation of opinions in different forms are posted in Social Networking Sites (SNS) like Twitter and Movie review sites through smart gadgets and computing devices like laptop, tablet pc and desktop pcs. Extracting useful features from this massive collection of opinion posts is not an easy task. In this paper, we have proposed an efficient two layer approach for extracting features from twitter posts. These generated features are used further to efficiently classify the opinions through OSIDL classifier to generate opinion summary of the twitter posts.

Keywords: Classification, Opinion mining, Twitter Dataset, Machine learning, feature correlation

1. Introduction

SNS allows individuals to create a public/semi-public profile within a domain such that they can communicatively connect with other users within the network [1]. Social network has improved on the concept and technology of Web 2.0, by enabling the formation and exchange of User-Generated Content.

In social media posts, opinion extraction is a complex and tedious job. In this process there are many challenges to prepare ideal features set without having noise, missing values. One of the challenges to the machine learning researchers is to efficiently use datasets with noisy information for efficient opinion classification.

In Twitter, users wish to open up their sentiments. The sentiment analysis can be done effectively only with ideal and proper features set extracted .

As the opinions expressed in the twitter reviews are unstructured, we need to propose an efficient opinion selection approach to extract opinions from twitter dataset.

This paper has been structured as follows. Section 2 gives the recent work and research gaps in the feature extraction. The proposed opinion extraction approach for OSIDL classifier is presented in section 3. The performance evaluation criteria are presented in section 4. Results of the proposed work are analyzed with other bench mark algorithms and presented in section 5. The conclusion of the work with the future work is presented in section 6.

2. Related work

In this section, we have presented the recent contribution in opinion mining using different machine learning techniques. Wei Zhang et al., [2] have introduced a novel supervised machine learning approach to weigh the different areas of the Twitter message and the features of a world gazetteer. It creates a model that will prefer the correct gazetteer candidate to resolve the extracted expression. Jasmina Smailovic et al., [3] have proposed an active learning approach for sentiment analysis of tweet streams in the stock market domain. Support Vector Machine (SVM)is used for static Twitter data analysis problem to determine the best Twitter-specific text pre-processing setting for training and adapting the SVM classifier. It categorizes Twitter posts into three sentiment categories like positive, negative and neutral.

Hongmin Li et al., [4] have studied the labeled data usage from a prior source disaster, together with unlabeled data from the current target disaster to learn domain adaptation classifiers for the target using previous disasters tweets.

Feyza Gürbüz [5] have presented a comparison of data mining results before and after pre-processing with link analysis on a passenger satisfaction survey on a light rail public transport system.

In[6], authors proposed an algorithm to discover the useful opinions from imbalanced Tweets. The above recent related work shows that a novel approach for opinion mining of imbalance data corpus is required for efficient opinion mining. Based on the above

background, we have proposed an efficient approach for opinion mining from imbalanced twitter corpus. Adedoyin-Olowe et al., [7] have studied different data mining methods to analyze Social media. Bora et al., [8] have introduced a method for distant supervised learning which uses words with sentiment value as noisy label. Junseok Song et al., [9] have introduced an attribute weighting and feature selection approach using Naïve Bayes to handle the biased quantity of positive words and negative words of SNS data in calculating the weights, and eliminated irrelevant words during feature selection. Patricia and Andries [10] has proposed decision tree based categories for removing the irrelevant features. In[12] authors are presented feature extraction approach using supervised association rule mining-based approach .

[11]VSM is the most referred approach, for reducing dimensionality of features set and is easy elucidation because it achieves highly significant condensed document content information. In[13] authors have proposed approach for extracting highly relevant features for machine condition monitoring and related applications and the statistical distribution of the measurement values. Based on the literature survey, new feature extraction and selection is proposed to extract and select relevant features from Twitter review posts.

3. Twitter review dataset

3.1 Collecting Reviews

Reviews are collected from the Twitter. The portion of sample reviews from twitter and is obtained for analysis and is shown in Table-1. The experienced users will share the strengths and limitations of the product in the form of tweets. The product reviews help the new users to better understand the product before they purchase them. The individual reviews of users may mislead the opinion of the new users, so it is recommended to go through several reviews before summarizing the opinion on a product.

3.2 Opinion mining with twitter posts

In this work, we have considered Twitter posts as SNS dataset. This dataset considered for analysis is an imbalance dataset with 1155 opinions, in which 978 are positive opinions and 177 are negative opinions. The imbalance ratio (IR) of the considered dataset is 5.52. Particulars of these SNS datasets details are shown in Table-1.

Table 1. The Twitter datasets and their properties

S.No.	Dataset	Instances	Missing	Majority	Minority	IR
1	Twitter	1155	No	978	177	5.52

For analysis let us consider the twitter opinion mining dataset sample instances with features and class can be seen below, and extract features so that they can be used for opinion mining. Twitter posts are processed with feature extraction, balanced, classified and finally summarized the opinions. A sample review of twitter opinion posts is shown below:

Twitter Datasets:

```
@relation Twitter
@attribute Twitter numeric
@attribute body string
@attribute class {pos,neg}
@data
1229709107,'anyone feel motivated the fri afternoon prior to a holiday? Wanted to get lots done... but I want jammies and judge judy... \"SIR!\" &lt;3 her ',pos

1231217680,'I had the same issue with dominions site. Fixed it by using internet explorer ',neg
……
…….
…….
```

In most of the cases, the analysis of the dataset was done assuming it as a balance dataset. We have proposed to analyze an imbalance dataset, the reason is, almost all the real world datasets are in imbalance nature. The existing algorithms are not efficient in discovering the hidden knowledge from the imbalance twitter/movie review dataset. OSIDL classifier produces better results with this proposed feature extraction approach on imbalanced twitter corpus, so that we can summarize opinions efficiently and generate useful recommendations for twitter users. In the figure 1 we can observe the role of feature extraction in opinion mining; it is a two layer approach, detailed discussion is presented in

further sections. Opinion classification in figure-1 refers to OSIDL classifier to extract opinions efficiently and generate improved opinion polarity results. The performance our OSIDL classifier with this new Feature extraction is compared with C4.5 and REP algorithms and results are presented in subsequent sections.

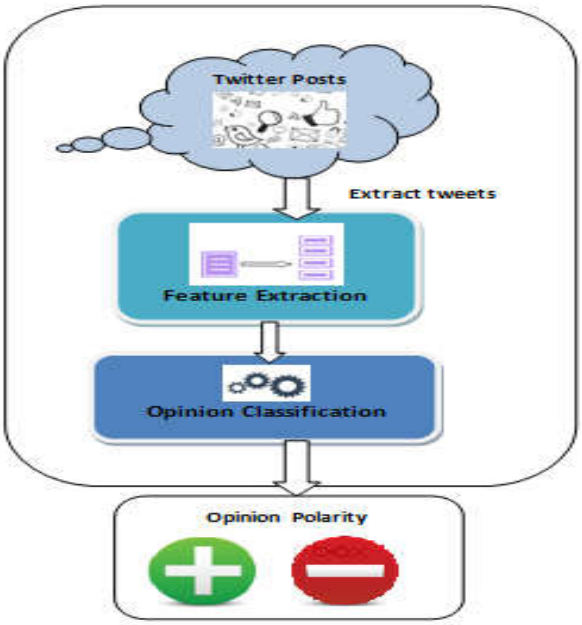

Figure 1. Role of feature Extraction in Opinion Mining

4. Proposed feature extraction

4.1 Preparing feature vector

The reviews collected are processed for removal of erroneous data. The reviews may be erroneously arranged in different categories and this may mislead the opinion of the new users. Proper categorization is to be validated for better product review summarization. The missing values, special characters and unnecessary blank spaces may be removed for improved performance of the product review algorithms.

We have taken the twitter review imbalance dataset here, so they must be converted into word vectors for better output in classification.

In this paper, a novel word2vect based approach is proposed for feature vector formation which groups the contents of twitter dataset and ranks them according to the relevancy of the given word by morphological analysis.

This is a two layered model in which first layer prepares the feature vectors from twitter tweets by applying CBOW model. CBOW model generated large number of pivot feature vectors. To apply semantic analysis on this large collection of feature vectors skip-gram model is applied to generate word vectors that will match with the context. This is a unique feature in this feature extraction approach. The advantage of this proposed approach is that it helps to retrieve user-expected features effectively. After this step the processed twitter posts are represented as shown below:

Preparing Twitter posts:

@relation Twitter

@attribute = numeric

@attribute About numeric

@attribute Agis numeric

@attribute Alt numeric

@attribute Although numeric

@attribute Amazing numeric

@attribute And numeric

@attribute Are numeric

@attribute As numeric

@attribute August numeric

@attribute BTW numeric

@attribute Beach numeric

@attribute Beatz numeric

@attribute Beautiful numeric

@attribute Been numeric

@attribute Behaviour numeric

@attribute Bella numeric

@attribute Best numeric

……

…….

…….

@attribute class {pos,neg}

@data

{0 1229709107,6 1,19 1,186 1,233 1,241 1,251 1,253 1,293 1,357 1,390 1,407 1,419 1,455 1,464 1,470 1,485 1,491 1,492 1,528 1,574 1,649 1,747 1,764 1,803 1,804 1}

{0 1231217680,114 1,294 1,443 1,483 1,675 1,698 1,747 1,792 1,834 1,875 1,917 1,921 1,941 1,942 1,992 neg}

{0 1229063765,233 1,269 1,402 1,483 1,521 1,605 1,656 1,745 1,747 1,764 1,877 1,889 1,896 1,897 1,905 1,944 1,950 1,985 1,987 1,992 neg}

……

…….

…….

The second layer in this approach is relevant features generation. The proposed two layer feature vector generation approach is shown in Figure-2.

4.2 Relevant feature generation

The number of features generated in the two layer feature extraction from the twitter dataset is very huge due to the large tokens used for opinion analysis. The unnecessary and redundant features are to be removed from the reviews collected. This can be done using an correlation method. In our work we have used CFS algorithm [11] for efficient

feature identification. The CFS algorithm follows a supervised approach, using feature to feature correlation and feature to class correlation. CFS is efficient enough to eliminate extraneous and superfluous features, and can discover relevant features as long as they do not sturdily depend on other features. It calculates the correlation between each attribute and the output variable and selects only those attributes that have a moderate-to-high positive or negative correlation and drop those attributes with a low correlation.

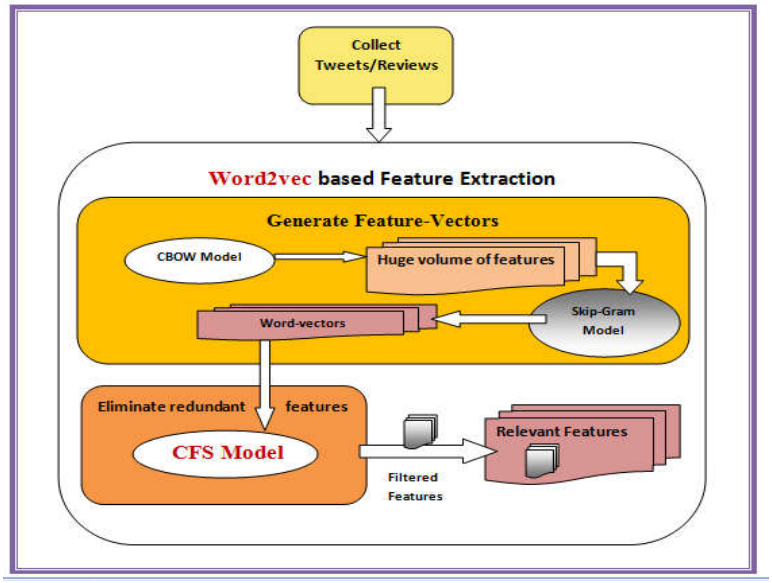

Figure-2. Two layer features vector generation approach

5. Classifying opinions

The Opinion classification through OSIDL takes the features set generated in previous step as input and then classify the opinions by generating opinion polarity. The numerical ratings that come with many of them enable us to classify them into finer-grained scales than just positive or negative categories [14]. This richer information makes it possible to rank items or quantitatively compare opinions of several reviewers, thus allowing more nuanced analyses to be carried out. Details of this OSIDL approach are already presented in our previous work [6] the extract of the approach is that it generates useful opinions from imbalanced twitter corpus. OSIDL performance with this new feature extraction is

greatly improved when compared with other bench mark classifiers c4.5 and REP. Details of the performance analysis is presented in further sections.

6. Performance evaluation

The formulas of all the evolution measures are given in the following equations. Performance of the approach is evaluated using accuracy, and AUC measures.

Accuracy (A) is the ratio of number of reviews predicted correctly to the total number of reviews in the corpus. The formula for calculating accuracy is given as:

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

The Area under Curve (AUC) is calculated using,

$$AUC = \frac{TP_{RATE} + TN_{RATE}}{2} \quad (2)$$

The True Positive Rate measures the proportion of positive opinions that are identified correctly from the review. The True Negative Rate measures the proportion of negative opinions that are correctly identified with in the review. A False Positive rate measure is the percentage of correctly received negative opinions from the review. Precision (Pr) is the ratio of number of positive reviews (TP) correctly predicted to the total number of reviews predicted as positive (TP+FP).

The Precision (Pr) is computed by,

$$\Pr = \frac{TP}{(TP) + (FP)} \quad (3)$$

Recall (Re) is the ratio of number of positive reviews (TP) correctly predicted to the actual number of positive reviews (TP+FN) in the corpus.

$$\mathrm{Re} = \frac{TP}{(TP) + (FN)} \quad (4)$$

The F-Score value is the harmonic mean of Pr and Re. It can have top value as 1 and least value as 0. The F-score is calculate using,

$$F\_Score = \frac{2 \times \Pr * \mathrm{Re}}{\Pr + \mathrm{Re}} \quad (5)$$

7. Result analysis

Opinion classification is the crucial phase in summarizing opinions of twitter corpus and will give better accuracy only with proper and ideal features set generation.

7.1 Result Analysis of Opinion classification

In this paper C4.5 and REP are used as bench mark algorithms to compare the performance of IOSDS_Classifier. The C4.5decision tree uses the recursive partition of the given space in the predefined classes. Another such approach is REP Tree algorithm for learning, which is a fast decision tree method to build a decision tree using information gain and prunes it using reduced-error pruning. The formula for calculating gain ratio is given below in equation (7),

$$Gain\ \text{Ratio}(a_i, S) = \frac{Inforamation\ \text{Gain}(a_i, S)}{Entropy(a_i, S)} \quad (6)$$

Information gain uses entropy as the impurity measure. The formula for calculating Information gain is given below,

$$Information\ \text{Gain}(a_i, S) = Entropy(y, S) - \sum_{v_{i,j} \in dom(a_i)} \frac{|\sigma_{ai=v_{i,j}} S|}{|S|} \times Entropy(y, \sigma_{ai=v_{i,j}} S) \quad (7)$$

In this result analysis of opinion miming problem, we have considered the imbalanced twitter dataset for classification. We used Weka [8] for implementation of the OSIDL algorithm with proposed Feature selection. By using our proposed approach, we could systematically classify their sentiments. In this section, the results of the OSIDL classifier using proposed feature selection with twitter imbalance dataset is compared and discussed with bench mark algorithm C4.5 and REP.

7.2 Results Analysis for IOSDS_Classifier

The table 2 presents the summary of Classification results on imbalance twitter datasets with validation measures accuracy, AUC. For the proposed algorithm to be a best

performing algorithm, the value of accuracy, AUC, precision, F-score, TP Rate and TN Rate should increase and the value of FP Rate and FN Rate should decrease.

The classification performance result of our proposed work is compared with C4.5 and REP classifiers and is represented in the figure 3. The overall accuracy of opinions classified correctly is 96.44 for IOSDS which is better than 95.99 for C4.5 and 93.88 for REP. For the AUC results, the proportion of opinions classified as correctly for both positive and negative instances is 0.982 for IOSDS whereas 0.935 for C4.5 and 0.881 for REP algorithm. The result of accuracy, AUC, TP Rate and FP Rate are considerably improved for IOSDS_Classifier on both the twitter and movie review datasets when compared with C4.5 and REP. The reason behind the improvement of results for IOSDS_Classifier over C4.5 and REP algorithms is due to the unique framework followed for efficient knowledge discovery from the imbalance twitter dataset. Classification results of proposed work are shown in Table-II for Twitter dataset. Graphical analysis of Accuracy and AUC results using C4.5, REP and IOSDS_Classifier are shown in Figure 3.

Table 2. Performance of IOSDS_Classifier on Twitter dataset

MEASURE	C4.5	REP	OSIDL_CLASSIFIER
ACCURACY	95.99●	93.80	96.44
AUC	0.935 ●	0.881●	0.982

●Dot indicates the dominance of IOSDS_Classifier over C4.5 and REP algorithm

Summarized results of Accuracy and AUC of the proposed approach in comparison with other approaches are shown in Fig-3.

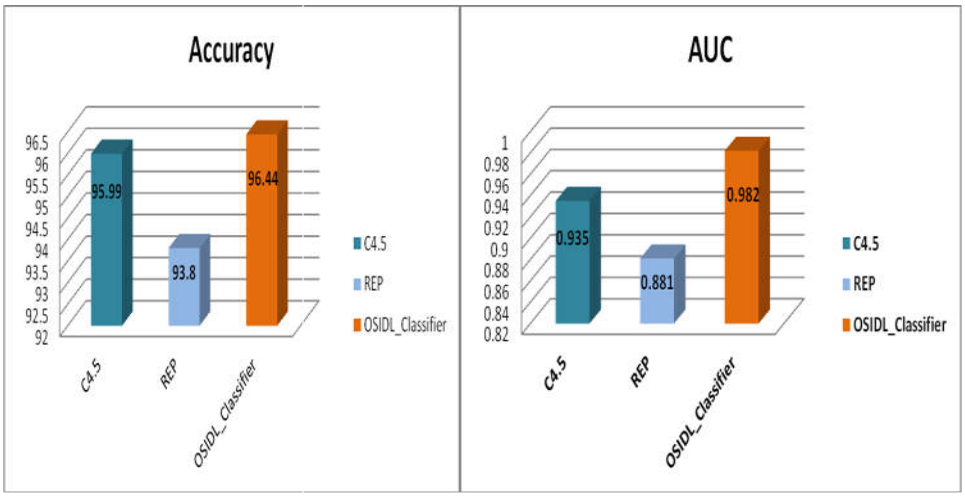

Figure 3.Summary of Accuracy and AUC results for Twitter dataset on IOSDS_Classifier

OSIDL with Twitter Dataset, the precision value of the proposed approach is improved from 0.964 to 0.966 when compared with C4.5 and increased from 0.956 to 0.966 when compared with REP. The F-Score value of the proposed approach is improved from 0.977 to 0.980 when compared with C4.5 and improved from 0.964 to 0.980 when compared with REP. The Recall value of the proposed approach is improved from 0.990 to 0.994 when compared with C4.5 and improved from 0.973 to 0.994 when compared with REP. Comparative analysis of C4.5 ,REP with our classifier performance is shown with another set of measures Precision, F-Score and Recall, are shown in Table- 3 and Figure 4.

Table 3: Performance evaluation with Precision-Score and recall measures on Twitter dataset

MEASURE	C4.5	REP	OSIDL_CLASSIFIER
PRECISION	0.964●	0.956●	0.966
F-SCORE	0.977 ●	0.964●	0.980
RECALL	0.990 ●	0.973●	0.994

● Dot indicates the dominance of OSIDL_Classifier over C4.5 and REP

The improved decision tree for Twitter dataset using the proposed approach is shown in the figure 5.

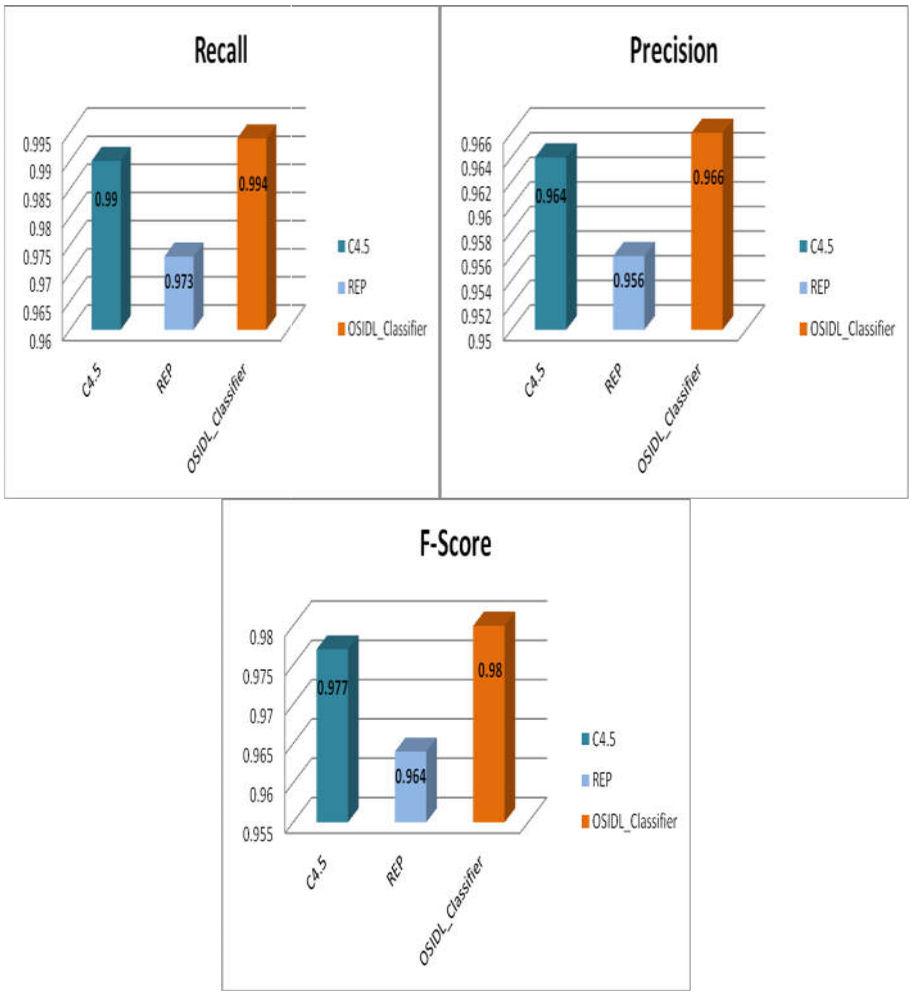

Figure 4. Precision and F-score and recall results for Twitter dataset on OSIDL classifier

The improved decision tree with the proposed approach for twitter dataset is shown in figure 5.

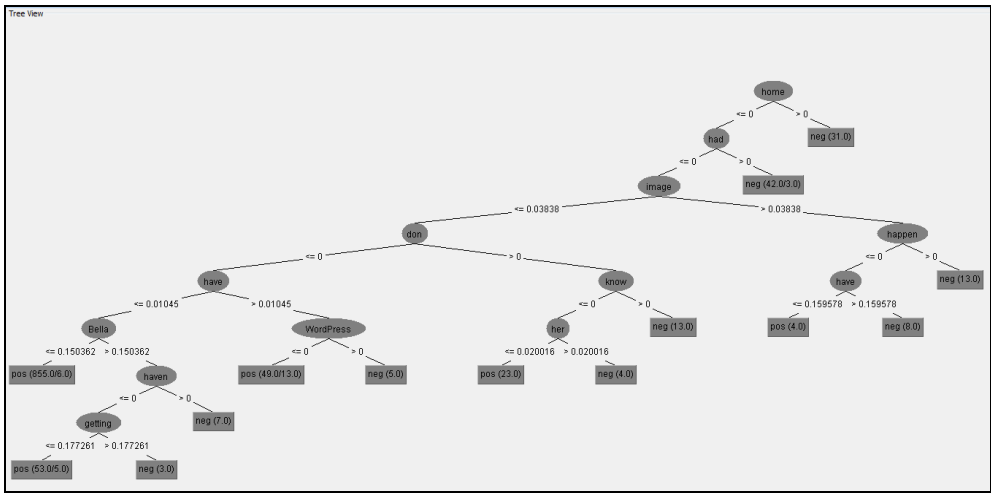

Figure 5. Decision Tree for improved Twitter dataset using OSIDL_Classifier

8. Conclusion

In future, we want to implement this approach for cross domain dataset with imbalance nature on different domains. Presented approach in this paper is aimed at different applications such as social network analysis, spam posts identification, product recommendation and movie review recommendation where additional clarity, efficiency, and ease of use is needed for human operators to be effective. After generating ideal features set classified the opinions efficiently and the efficiency OSIDL classifier is demonstrated by comparing the results with C4.5 and , REP and consistent improvements are observed. This proposed feature selection is an important stage for OSIDL classifier because it may have a substantial effect on its accuracy. It reduces the number of dimensions of the twitter posts.

REFERENCES

[1]. Chen, Z. S., Kalashnikov, D. V. and Mehrotra, S. Exploiting context analysis for combining multiple entity resolution systems. In Proceedings of the 2009 ACM International Conference on Management of Data (SIGMOD'09), 2009.

[2]. Zhang, W., T. Yoshida, and X. Tang, A comparative study of TF-IDF, LSI and multi-words for text classification. Expert Systems with Applications, 2011: p. 38(3): p. 2758-2765.

[3]. Shi, K., et al., Efficient text classification method based on improved term reduction and term weighting. The Journal of China Universities of Posts and Telecommunications, 2011. 18,Supplement 1(0): p. 131-135.

[4]. Hongmin Li, Nicolais Guevara, Nic Herndon DoinaCaragea, Kishore Neppalli, Cornelia Caragea, Anna Squicciarini, Andrea H. Tapia,"Twitter Mining for Disaster Response: A Domain Adaptation Approach",Short Paper – Social Media Studies Proceedings of the ISCRAM 2015 Conference - Kristiansand, May 24-27, Palen, Büscher, Comes & Hughes, eds.

[5]. Feyza GÜRBÜZ, Fatma TURNA," An application of data mining on tram faults for rule extraction", Journal of Engineering Technology (ISSN: 0747-9964), Volume 6, Issue 2, July, 2017, PP.514-526.

[6]. S. Adinarayana and E. Ilavarasan, "An efficient approach for opinion mining from skewed twitter corpus using over sampled imbalance data learning," 2017 International Conference on Intelligent Communication and Computational Techniques(ICCT),Jaipur,India,2017,pp.42-47.doi:10.1109/INTELCCT.2017.8324018.

[7]. Adedoyin-Olowe, Mariam, Mohamed Medhat Gaber, and Frederic Stahl. "A survey of data mining techniques for social media analysis." arXiv preprint arXiv:1312.4617 (2013).

[8]. Bora, Nibir Nayan. "Summarizing public opinions in tweets." International Journal of Computational Linguistics and Applications 3.1 (2012): 41-55.

[9]. Song, Junseok, et al. "A novel classification approach based on Naïve Bayes for Twitter sentiment analysis." KSII Transactions on Internet & Information Systems 11.6 (2017).

[10]. Patricia, E.N. Lutu, and Engelbrecht, A.P., A decision rule-based method for feature selection in predictive data mining, Expert Systems with Applications, 2010, 37,602–609.

[11]. M. A. Hall, Correlation-based feature selection for machine learning [Ph.D. thesis], University of Waikato, Hamilton, New Zealand, 1999.

[12]. HU, M. AND LIU, B. 2004b. Mining opinion features in customer reviews. In AAAI'04: Proceedings of the 19th national conference on Artifical intelligence. AAAI Press, 755–760.

[13]. T. Schneider, N. Helwig and A. Schütze, "Automatic feature extraction and selection for condition monitoring and related datasets," 2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Houston,TX,2018,pp.1-6 doi: 10.1109/I2MTC.2018.8409763.