Proceedings of 4th RIT Post Graduates Conference (RIT PG Con-18)
NOVATEUR PUBLICATIONS
JournalNX- A Multidisciplinary Peer Reviewed Journal (ISSN No:2581-4230)
April, 13th, 2018

HIERARCHICAL TEXT CLASSIFICATION USING DICTIONARY BASED APPROACH AND LONG-SHORT TERM MEMORY

Karishma D. Shaikh
M. Tech. Student
Department of Computer Science and Engg.
Rajarambapu Institute of Technology, Sakharale, India.
shaikhkarishma2110@gmail.com

Amol C. Adamuthe
Assistant Professor, Dept. of Information Technology,
Rajarambapu Institute of Technology, Sakharale, India.
amol.admuthe@gmail.com

***Abstract*:**
The text classification process has been well studied, but there are still many improvements in the classification and feature preparation, which can optimize the performance of classification for specific applications. In the paper we implemented dictionary based approach and long-short term memory approach. In the first approach, dictionaries will be padded based on field's specific input and use automation technology to expand. The second approach, long short term memory used word2vec technique. This will help us in getting a comprehensive pipeline of end-to-end implementations. This is useful for many applications, such as sorting emails which are spam or ham, classifying news as political or sports-related news, etc.

Keywords—dictionary based approach; long-short term memory approach; preprocessing; text classification

I. INTRODUCTION

The rapid development of technologies has risen in the count of electronic documents in world wild [1]. The World Wide Web urgently needs efficient and effective classification algorithms to help people navigate and browse online documents quickly [2]. This case increases the significance of text categorization for the purpose of classifying the text into the appropriate categories based on the textual content. Text categorization is useful in a variety of areas such as spam filtering, sentiment analysis, topic detection, author identification, and language identification [1].

Since the birth of digital documents, automatic text categorization has always been an essential application and a demanding research issue. Text classification is vital sub-area of text mining that allots documents to already defined categories or classes. Different forms of text collection are digital libraries, web pages and news articles are significant information. Therefore, text classification is main research challenge of information science, computer science and library science. The application contains opinion mining (sentiment analysis), organization, news filtering, search, document organization, and spam filtering [3].

Text classification problem can be achieved several settings of the program. The basic text classification idea, like distinct pattern recognition problems, contains the text preprocessing and categorization stage. Due to the nature of the problem, text preprocessing mechanisms need to extract digital content which is in the form of text file. Then classifier is classified text document by predicting the class of the document [4].

Text preprocessing is an important part of any natural language processing because the words, characters, and sentences recognize and then basic units proceed to further processing stages. It is a set of activities that preprocess text documents because textual data typically consist special formats such as date formats, number formats, and very commonly used or unwanted content can be removed.

Neural network models have proven to achieve superior accuracy in document and sentence modeling. Deep learning approach CNN and RNN are the main forms of this modeling task, which take a completely different approach to understanding natural language. In this work, a novel unified model called LSTM is proposed for text categorization.

II. LITERATURE REVIEW

This section covers review of different research carried out by the researcher for text classification using text preprocessing and various machine learning (ML) based approaches and various lexicon approaches (LA).

A. ML based approach

ML is the ability to learn things automatically and that uses computers to predict emotions. It is the application of artificial intelligence and not require any manual intervention to perform the task. Two major categories of ML are supervised or unsupervised.

In the paper [1] Uysal, Alper Kursat *et al.* proposed an IGFSS, in which the end of feature selection scheme. IGFSS improves the performance of GFS. For this reason, IGFSS used LFS methods to differentiate the classes and generating feature sets.

In paper [7] Zulfany Erlisa Rasjid *et al.* focus on data classification using two out of the six approaches of data classification, which is k- NN (k-Nearest Neighbors) and Naïve Bayes. The Corpus used three thousand text documents and over twenty types of classifications. Out of

the twenty types of classifications, six are chosen with the most number of text documents. The accuracy of K-NN is better than naïve bayes. By using information retrieval, it is possible to obtain an unstructured information and automatic summary, classification and clustering.

In the paper [8] Parlak *et al.* proposed the system which uses two datasets for analyzing the impact of the feature selection for classification. Two pattern classifiers were used such as gini index and the recognition feature selector.

In paper [10] Kamran Kowsari *et al*. discussed traditional supervised classifiers. It has degraded as the number of documents has increased. Kamran Kowsari et al. approached this problem differently from current document classification methods that view the problem as multi-class classification. Instead, we perform hierarchical classification using an approach we call Hierarchical Deep Learning.

In the paper [12] Pratiksha Y. Pawar *et al.* done the comparison of different types of text categorization methods. The author studied supervised unsupervised and semi-supervised approaches.

In the paper [13] Dawu *et al.* done document classification using semantic matching method for. First, several heuristic selection rules are defined to quickly select relevant concepts for documents in the Wikipedia semantic space. Second, based on the semantic representation of each text document. Finally, the evaluation experiment proves the effectiveness of the proposed method, that is, it can improve the classification efficiency of Wikipedia matching without compromising the accuracy of the classification.

In paper [14] Muhammad Nabeel Asim *et al.* compares the performance of nine popular feature metrics on six datasets using naive Bayes, SVM classifiers. The nine metrics include chi-squared, odds ratio, information gain, Gini index, Poisson ratio, normalized difference measure, balanced accuracy measure, distinguishing feature selector, and binomial separation.

In paper [15] Abdal Raouf Hassan *et al*. proposed CNN and bidirectional RNN over pre-trained word vectors. It utilizes bidirectional layers as a substitute for pooling layers in CNN.

In paper [16] Lihao ge *et al.* developed a method using word2vec. For classification accuracy author combine the proposed method with mutual information and chi-square.

In paper [18] Piotr Semberecki *et al.* proposed LSTM. The author tested different feature vector approach and after that, he used word2vec approach for the better result.

In paper [19] Muhammad Diaphan Nizam Arusada *et al.* shows the strategy to make optimal training data by using customer's complaint data from Twitter. The author used both NB and SVM as classifiers. The experimental result shows that our strategy of training data optimization can give good performance for multi-class text classification model.

In paper [21] Neha Sharma *et al.* modified model for naive Bayes classifier for multinomial text classification has been proposed by modifying the conventional bag of words model. The experimental results over benchmark datasets prove its superior performance than original naive Bayes multinomial model.

B. Lexicon Based Approach

In paper [5] Gelbukh *et al.* proposed hierarchical dictionary approach. Using this technique author classify the main topic of the document. The dictionary consists keyword related to topics and hierarchy of that topics.

In paper [22] Reshma Bhonde *et al*. proposed dictionary approach. In that manually create a dictionary of positive and negative sentiment words. Using this dictionary classify the document into positive and negative classes.

In paper [23] Geli Fei *et al.* proposed a dictionary-based approach. This approach gives the better result than traditional supervised approaches.

In paper [24] Santanu Mandal *et al.* proposed dictionary algorithm. It uses the comparisons of positive, superlative and comparative degrees on the word; for every negative and positive sentiment words.

In paper [6] Alper Kursat Uysal *et al*. gives information about the impact of preprocessing. Impact in terms of classification accuracy, text language, dimension reduction and text domain.

In paper [9] Dharmendra Sharma *et al.* done summarizing the effect of stemming and stop word onto feature selection. Most of the researchers in text categorization are focusing more on the development of algorithms for optimization of preprocessing technique for text categorization.

In paper [11] Bali *et al.* proposed feature selection methods for decreasing the size of the data by eliminating unnecessary content which is not related to classification.

In paper [17] Suchi Vora *et al.* addresses the issues of how does a feature selection method affect the performance of a given classification algorithm by presenting an open source software system that integrates eleven feature selection algorithms and five common classifiers; and systematically comparing and evaluating the selected features and their impact over these five classifiers using five datasets. The five classifiers are SVM, random forests, naïve Bayes, KNN and c4.5 decision tree. gini index or Kruskal-Wallis together with SVM often produces classification performance that is comparable to or better than using all the original features.

In paper [20] Kinga Glinka *et al*. examined the, combining problem transformation methods with different

approaches to feature selection techniques including the hybrid ones. Author checked the performance of the considered methods by experiments conducted on the dataset of free medical text reports. There are considered cases of the different number of labels and data instances. The obtained results are evaluated and compared by using two metrics: classification accuracy and hamming loss.

III. PROBLEM FORMULATION

Dictionary-based approach and neural network approach which is LSTM addresses the issues occurs in many companies. Many Companies are facing problem for searching there required text document. It is very headache for searching where that required text or document is available and it required lots of time. Because of that problem, many organizations are lost their valuable time for finding required text document such as from where we get that document. This type of issues is solved in our proposed system. Our proposed systems are providing the classification of that text so a user can easily check their text document. We used a lexical approach which is dictionary based approach and machine learning approach which is the long-term memory.

IV. PROPOSED METHODOLOGY AND DISCUSSION

The idea in this paper is to use a simple dictionary based approach and to use standard LSTM. LSTM overcome the problem which is vanishing gradients descent and that problem occurs in RNN. For creating a baseline implementation, we will be using a dictionary-based approach. The dictionary will be populated based on domain-specific inputs. The dictionary will be expanded using automated techniques. This will help us in getting an end-to-end implementation of full pipeline. In subsequent phases, we will be expanding our approach with LSTM.

A. Dictionary-Based Approach

Baseline DA classifying text document into class labels. It takes the input as token and output in the form of class name/label.

Step 1: *Take input as a text file.*
Step 2: *Break the sentences of the file into words as token using word tokenization and store it into the file.*
Step 3: *Delete the stop words from that file.*
Step 4: *Apply stemming from that file using Porter Stemmer algorithm.*
Step 5: *Create manual dictionaries of classes. And apply step 2, 3, 4.*
Step 6: *Create global word list of classes.*
Step 7: *Compare global word list with different classes and store it into an array list.*
If (word (global list) ∈ (class file))
Set 1;
Otherwise Set 0;
Step 8: *Do step 7 until finish for all classes.*
Step 9: *Compare global word list with input file (i.e. preprocessed performed file) store it into an array list.*
If (word (global list) ∈ (input file))
Set 1;
Otherwise Set 0;
Step 10: *Performed dot product operation on the entire class array list with input array list.*
Step 11: *Check the max (dot product) and assign that number as the class label.*

Figure 1 Pseudo code of dictionary-based approach.

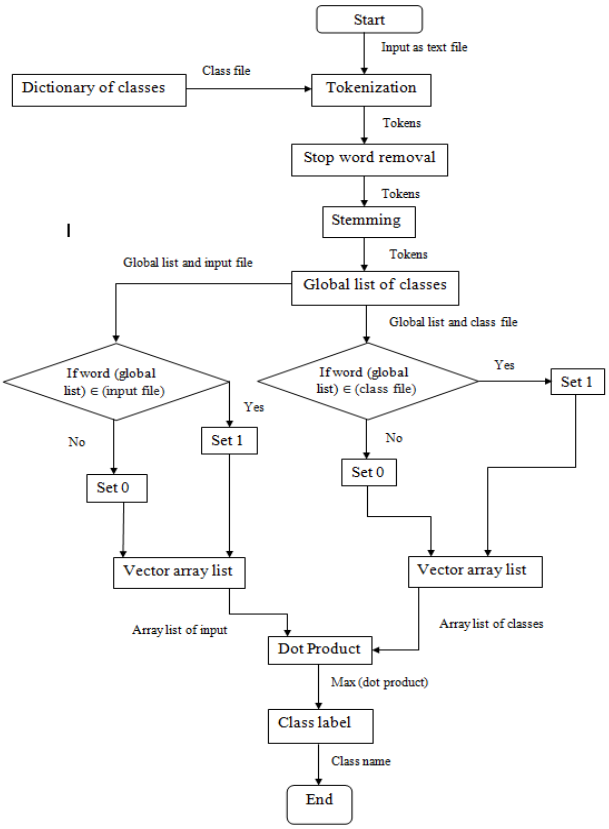

Figure 2 Flowchart demonstrating dictionary-based approach for text classification

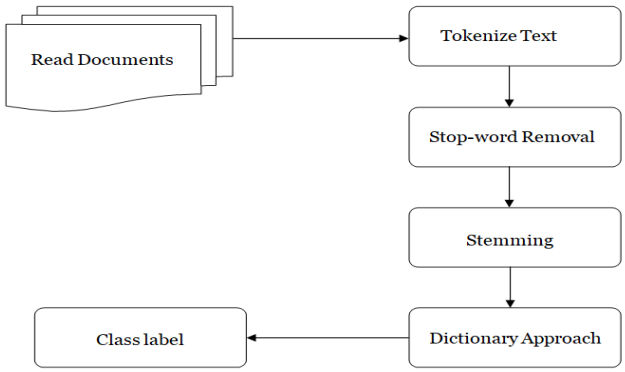

Figure 3 Text classification using Dictionary-Based Approach

The proposed system mainly contains preprocessing, classification technique. In this system row data uses as input which is in the form of text. This row data is going for preprocessing. Text preprocessing system consists of activities like tokenization, stemming, stop word removal and vector formation. Tokenization is used for breaking the text into symbols, phrases, words, or token. Next step is to stop word elimination. It is used to remove unwanted tokens and those tokens that are used frequently in English and these words are useless for text mining. Stop words are language specific words which carry no information. The most commonly used stop words in English are e.g. is, a, the etc. Stemming is another important preprocessing step. For example, the word walks, walking, walked all can be stemmed into the root word"WALK". Now the text data

is ready for mining process and this data is classified using supervised classification technique. Here supervised classification technique is used for classification module. We are using both labeled and unlabeled data for doing classification.

B. *LSTM Approach*

LSTM also used to classify the text document into class name/label. It takes the input of word and class name/label as output.

LSTM having one memory cell, three gates which are forgotten gate, input gate, and the output gate. Input gate is used to control the flow of input, output gate is used to control the flow of output and forget gate decide the what information going to store in the memory cell and what information going to throw away from the cell. Input and output gate usually used Tanh function. Forget gate always used the sigmoid function. Sigmoid and Tanh are the activation function.

The formula of sigmoid activation function for input x:

$$\text{Sig(x)} = 1/\ (1+e^{-x})$$

Sigmoid function satisfies all properties of the good activation function. The output of this function is always in between 0 to 1 i. e. always positive.

The formula of Tanh activation function for input:

$$\text{Tanh (input)}=(e^{input} - e^{-input})/(e^{input} + e^{-input})$$

Activation function Tanh is faster and gradient computation is less expensive. The output of this function always increases and decreases and in between -1 to 1 i.e. always positive or negative.

Equations of forget gate, input gate and output gate:

$$F_t=\text{sig}\ (W_f*X_t+B_f)$$

Where,

F_t is the output of forget gate layer,

W_f is the weight of the forget gate layer,

X_f is the input of that cell,

B_f is the bias of the forget gate layer.

$$I_t=\tanh\ (W_i*X_t+B_i)$$

Where,

I_t is the output of input gate layer,

W_i is the weight of the input gate layer,

X_t is the input of that cell,

B_i is the bias of the input gate layer.

$$O_t=\tanh\ (W_o*X_t+B_o)$$

Where,

O_t is the output of output gate layer,

W_o s the weight of the output gate layer,

X_t is the input of that cell,

B_o is the bias of individual output gate layer.

Calculating the partial derivatives of all the activation function of input, output and forget gate to minimize the loss:

$$F_t^{'} = F_t\ (1\text{-}(\ F_t\ )^2)$$

$$I_t^{'} = I_t(1\text{-}\ (I_t)^2)$$

$$O_t^{'} = O_t\ (1\text{-}(O_t)^2)$$

Where,

$F_t^{'}, I_t^{'}, O_t^{'}$ are the partial derivatives of forget gate, input gate, output gate respectively.

Step 1: *Decide the data parameters.*

Step 2: *Model those parameters.*

Step 3: *Load and save the data.*

Performed stop word elimination and clean data (remove special characters).

Step 4: *Save parameters to file.*

Step 5: *Performed cross-validation of data (used k-fold cross-validation).*

Step 6: *Train data for the classifier.*

Step 7: *Used LSTM classifier.*

Step 8: *Repeat step 5 to 7 until the end of training data.*

Figure 4 Pseudo code of LSTM classifier.

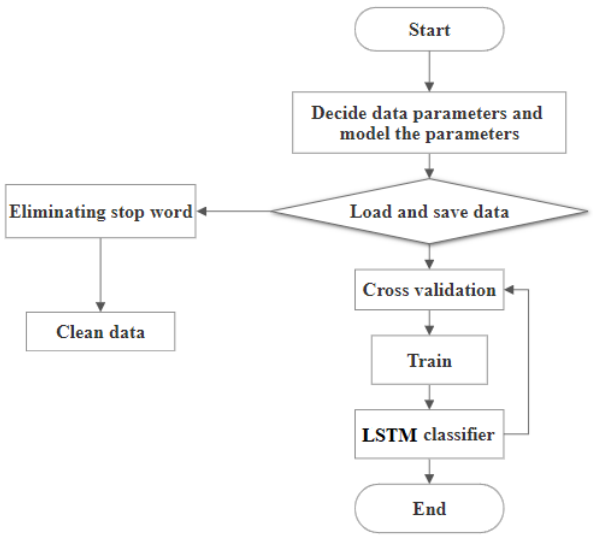

Figure 5 Flowchart of LSTM classifier.

V. RESULT AND DISCUSSION

The result of the dictionary was obtained using text preprocessing and dot product, for LSTM was obtained using 2 layers, hidden size was 128, dropout probability was 0.5, learning rate 0.001, batch size was 32, a number of

the epoch was 50, evaluate module after 100 steps and used k-fold cross-validation.

```
tokens tokenizer working the file hieghted a an
working worked [ platform ]
shutdown reboot suspend tokenization FOR baller run
lotus temple god tulip worship basket
```

Figure 6 Input file

Figure 6 shows the sample text file which we want to classify.

```
football cricket racket pitch ball bat
basket ground cricket bat man ball run
```

Figure 7Dictionary of sports class

Figure 7 shows the dictionary of sports class which contains the keywords related to sports.

```
religion temple god prayer flower
worship religion temple god prayer
flower worship religion temple
```

Figure 8 Dictionary of temple class

Figure 8 shows the dictionary of temple class which contains the keywords related to temple.

```
lily tulip lotus sunflower rose
flower Aster Amaranth Azalea
```

Figure 9 Dictionary of flower class

Figure 9 shows the dictionary of flower class which contains the keywords related to flower.

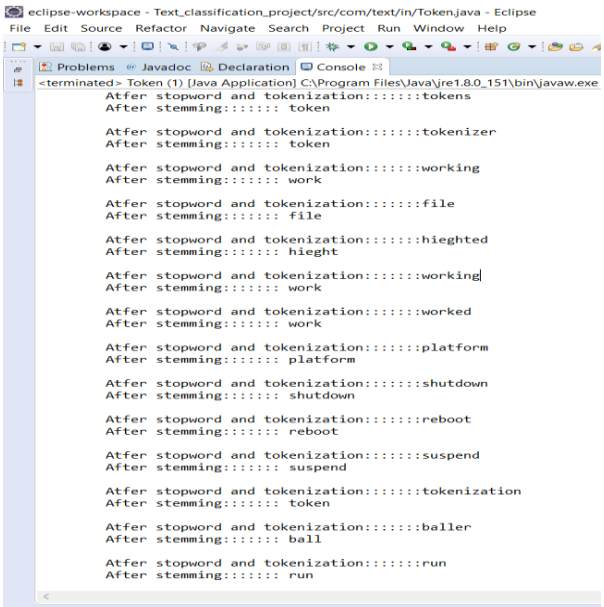

Figure 10 Text preprocessing on the input file.

Figure 10 shows the text preprocessing steps on the input file. Above operation is performed on all class dictionaries. Using those preprocessed files performed classification steps which were discussed in section algorithm. After performing all the operation will get the accurate text classification.

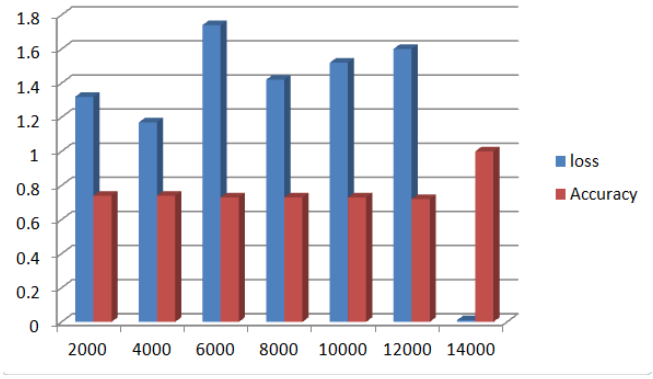

Figure 11 Result of LSTM after calculation of accuracy and loss.

Figure 11 shows the 1 accuracy and 0.1 loss after 15000 iterations. After every 100 steps we calculated cross-validation for 50 epoch and 32 batch size tried to decrease the loss. In the paper, we took dataset which having content and label. After trained dataset, we performed LSTM classifier on that data and apply 15000 iterations and calculate accuracy and loss.

VII. CONCLUSION

Text classification is used to classify the document into different categories. In the paper, we implemented dictionary-based approach and LSTM for text classification. The dictionary-based approach is simple approach and in that, we used text preprocessing and dot product. Using dictionary-based approach we understood that how classification is done. In LSTM we used the word as input from a text file which we want to classify and after that, we used LSTM algorithm classification. It is complex to implement but it gave high accuracy for text classification.

REFERENCE

[1] Uysal, Alper Kursat."An improved global feature selection scheme for text classification." *Expert systems with Applications* 43 (2016): 82-92.

[2] Onan, Aytug, Serdar Korukoglu, and Hasan Bulut."Ensemble of keyword extraction methods and classifiers in text classification." *Expert Systems with Applications* 57 (2016): 232-247.

[3] Gelbukh, Alexander, Grigori Sidorov, and Adolfo Guzmán-Arenas."Text categorization using a hierarchical topic dictionary." *Proc. Text Mining workshop at 16th International Joint Conference on Artificial Intelligence (IJCAI'99), Stockholm, Sweden*. 1999.

[4] Uysal, Alper Kursat, and Serkan Gunal."The impact of preprocessing on text classification." *Information Processing & Management* 50.1 (2014): 104-112.

[5] Rasjid, Zulfany Erlisa, and Reina Setiawan."Performance Comparison and Optimization of Text Document Classification using k-NN and Naive Bayes Classification Techniques." *Procedia Computer Science* 116 (2017): 107-112.

[6] Wang, Yong, Julia Hodges, and Bo Tang."Classification of web documents using a naive bayes method." *Tools with Artificial*

Intelligence. Proceedings. 15th IEEE International Conference on. IEEE, 2003.

[7] Wang, Yong, Julia Hodges, and Bo Tang."Classification of Web Documents Using a Naive Bayes Method." *Tools with Artificial Intelligence, Proceedings.15th IEEE International Conference on.* IEEE, 2003.

[8] Kowsari, Kamran, et al."Hdltex: Hierarchical deep learning for text classification." *arXiv preprint arXiv:1709.08267* (2017).

[9] Bali, Monica, and Deipali Gore."A survey on text classification with different types of classification methods." *International Journal of Innovative Research in Computer and Communication Engineering* 3 (2015): 4888-4894.

[10] Pawar, Pratiksha Y., and S. H. Gawande."A comparative study on different types of approaches to text categorization." *International Journal of Machine Learning and Computing* 2.4 (2012): 423.

[11] Feng, Guozhong, et al."A Bayesian feature selection paradigm for text classification." *Information Processing & Management* 48.2 (2012): 283-302.

[12] Yao, Zhao, and Chen Zhi-Min."An optimized nbc approach in text classification." *Physics Procedia* 24 (2012): 1910-1914.

[13] Lai, Siwei, et al."Recurrent Convolutional Neural Networks for Text Classification." *AAAI*. Vol. 333. 2015.

[14] Ge, Lihao."Improving Text Classification with Word Embedding." (2017).

[15] Vora, Suchi."A Comprehensive Study of Eleven Feature Selection Algorithms and Their Impact on Text Classification."

[16] Semberecki, Piotr, and Henryk Maciejewski."Deep learning methods for subject text classification of articles." *Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on*. IEEE, 2017.

[17] Mohasseb, Alaa, et al."Domain specific syntax based approach for text classification in machine learning context." *The 16th International Conference on Machine Learning and Cybernetics (ICMLC)*. IEEE, 2017.

[18] Glinka, Kinga, Rafał Wozniak, and Danuta Zakrzewska."Improving Multi-Label Medical Text Classification by Feature Selection." *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2017 IEEE 26th International Conference on*. IEEE, 2017.

[19] Sharma, Neha, and Manoj Singh."Modifying Naive Bayes classifier for multinomial text classification." *Recent Advances and Innovations in Engineering (ICRAIE), 2016 International Conference on*. IEEE, 2016.

[20] Kannan, S., and Vairaprakash Gurusamy."Preprocessing Techniques for Text Mining." (2014).

[21] Fei, Geli. "A dictionary-based approach to identifying aspects implied by adjectives for opinion mining." *Proceedings of COLING: Posters* (2012): 309-318.

[22] Mandal, Santanu, and Sumit Gupta."A novel dictionary-based classification algorithm for opinion mining." *Research in Computational Intelligence and Communication Networks (ICRCICN), Second International Conference on*. IEEE, 2016.

[23] Hailong, Zhang, Gan Wenyan, and Jiang Bo."Machine learning and lexicon based methods for sentiment classification: A survey." *Web Information System and Application Conference (WISA), 11th*. IEEE, 2014.