iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** XI **Month of publication:** November 2020

DOI: https://doi.org/10.22214/ijraset.2020.32258

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue XI Nov 2020- Available at www.ijraset.com

An Highly Efficient Algorithm to Improve the Performance of the Classifier using Persuasion and Sentimental Study

Reshma Israr[1], Pradeep Kumar Atulker[2], Rajendra Gupta[3]
[1, 2]*Research Scholar,* [3]*Associate Professor, Rabindranath Tagore University, Bhopal*

Abstract: ***The social media is not only a source that disseminates information to users, but it also allows users to communicate and share their thoughts and experience. The computer technology now-a-days have entered our imigination and it is impossible to imagine our lives without gadgets or the Internet. The part of data is subjective and includes opinions that can be analyzed to obtain the necessary data and to be used later for a variety of purposes for analysis and decision support. In this research study, the emotion related techniques are studied to derive opinions from tweets. To ensure efficient classification, it is important to implement an algorithm that performs well on this task. Therefore, the main goal of the research work is to investigate algorithms that can be applied to opinion estimation. To that extent, data preprocessing and multiple experiments are performed, that is, classifiers are trained and tested on two different datasets with two different classifiers i.e. Naive Bayes and Convolutional Neural Network.***

Keywords: Persuasion analysis, Sentimental analysis, Sentimental, Feature selection, Convolutional Neural Network

I. INTRODUCTION

Sentimental classification task is not a novel research area. However, the main focus of the research was on the analysis of large documents (reviews), but not on the microblogs that are sought today. Twitter is an example of a microblogging platform. A tweet is a short message (maximum 284 characters) that may contain opinions or express certain facts. This limit was doubled to 285 for all languages except Japanese, Chinese and Korean [2]. To classify a tweet is a difficult task as the tweet may contain irony, misspellings, emoticons, slang, abbreviations, and may contain only a few words. Numerous techniques exist that can be used for sentimental analysis work. The main approaches are machine learning [3], [5] and lexicon-based [2], [15], [10-12], [21]. The machine learning method uses a dataset for training classifiers that will be applied to further define the sense of a particular text. The lexicon-based methodology uses the Semantic orientation of words or phrases to define whether a text is positive or negative.

II. MACHINE LEARNING

The main work of this research study is an investigation of classification algorithms for extracting opinions from tweets and IMDB movie reviews. For this purpose, two methods are studied. First of all, Naïve Bayes algorithm that uses a bag-of-words representation for training classifier. Second is a convolutional neural network that converts words into word embeddings and then passes these embeddings through the layers to extract the polarity of tweets. As a result, the research work aims to perform experiments and investigate the performance of two different algorithms detecting positive and negative tweets/reviews. Furthermore, algorithm which gives better results has to be defined. Moreover, it is important to study how algorithms accuracy can be affected by data preprocessing, feature selection and data selection.

To apply machine learning algorithms, various steps needs to be applied:

1) *Data Collection:* Tweets to be analyzed have to be retrieved from Twitter as well as the dataset for training purpose has to be obtained.
2) *Preprocessing Data:* Tweets have to be pre-processed to remove the usernames, URLs, punctuation that do not contain any useful information. Moreover, words have to be lowercased.
3) *Training Process:* Data that was extracted as the training set is given to the classifier for learning.
4) *Data Classification:* When the training stage is complete the classifier can be used for analyzing the polarity of tweets or reviews. At first, the classifier is fed with the testing dataset to check the accuracy of the algorithm then real data can be given to the classifier to extract sentimentals from tweets.

After machine learning algorithms, the applied results are analyzed. The accuracy of algorithms and their performance time are analyzed.

III. EARLIER STUDY ON SENTIMENT ANALYSIS

The sentimental analysis is not a new task, it has been studied since 1990s. However, in 2000s attracted the interest of scientists due to its significance in different scientific areas, also SA had many unstudied research questions.

Moreover, the wide availability of opinionated data-pushed research in this area on a new stage. Since then SA became a rapidly developing area [11].

Sentimental analysis deals with the processing of opinion ated text to extract and categorize opinions from certain documents. The polarity of sentimental usually expressed in terms of positive or negative opinion (binary classification [13-14]). However, it can be multi-class classification [15-18], hence sentimental may have a neutral label or even broadened variation of labels like very positive, positive, neutral, negative, very negative, also labels can be associated with emotions like anger, sad, fearful, happy, etc.

Sentimental analysis is a developing area that arouses the interest of humans and especially organizations because sentimental analysis can be used for the decision making process.

Individuals are no longer limited to ask opinions from friends about a particular product or service, they can freely find such information on the Internet. It is significant to notice that sources that contain opinionated data are noisy sometimes, so it is important to extract the essential meaning from that information to use it further. Sentimental analysis uses different techniques and approaches to handling this challenging task [18] [22].

IV. MACHINE LEARNING PROCESS

The technique that can be used for sentimental analysis is machine learning that includes unsupervised and supervised machine learning methods that are explained below.

A. Unsupervised Machine Learning Process

The author in [27] uses an unsupervised machine learning approach for the review classification. Reviews are classified into recommended (thumbs-up) and not recommended (thumbs-down). The author retrieves phrases that consist of two words based on tags patterns. The patterns are designed in such a way that they have to capture sentimental phrases. Each phrase is a combination of adjective/adverb and verb/noun.

The Part-of-speech tagger (POS) is employed to the document to decide which phrases have to be retrieved. It is noticed taht a phrase is extracted if two words fall under one of the proposed patterns. Next step is a calculation of the semantic orientation of retrieved phrases from the review. The author applies the Pointwise Mutual Information (PMI) and Information Retrieval algorithm to find semantic orientation. PMI measures semantic similarity between two terms.

B. Supervised Machine Learning Methods

The most common and simple method that is used for text classification is supervised maching learning [39], [40-43]. The model is based on Bayes' theorem with the assumption that features are independent.

Naïve Bayes classifier defines the probability of the document belonging to a particular class. The advantages of the Bayes classifier are simplicity of the implementation, the learning process is quite fast, it also gives quite good results [39], [44]. However, the "naive" assumption may cause a problem because in the real world features are dependent.

C. Convolutional Neural Network (CNN)

Convolutional Neural Network CNN is organized by layers interleaving. Such network contains convolution, subsampling and fully-connected layers that can alternate in random order. Severyn and Moschitti [30] were working on Twitter sentimental analysis with deep CNN.

They proposed a one layer network that includes a convolutional layer that is passed through the non-linear activation function (ReLU) followed by max-pooling layer and further passed to soft-max classification layer. Neural language model was used for initializing word embedding out of an initial dataset of tweets [18]. Then word embeddings were refined using CNN on the distant supervised corpus. Authors claim that proposed system performs well.

Moreover, the author [19] was using CNN for sentence classification. He classified sentence into positive/negative as well into fine-grained classes, also he defined whether a sentence is subjective or objective and classified a question into 6 question categories. CNN includes convolutional, max-over-time-pooling and fully connected layers.

It was reported that model showed good results and "pre-trained vectors are 'universal' feature extractors that can be utilized for various classification tasks" [20].

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue XI Nov 2020- Available at www.ijraset.com

V. PROPOSED METHODOLOGY

This portion of work introduces the main steps that have to be performed for carrying out the sentimental classification, namely preprocessing and feature extraction. Moreover, two algorithms are used for the classification described in detail.

A. Preprocessing of Data

The first dataset is a dataset v1.0 introduced in Pang/Lee ACL 2005 that represents IMDB movie reviews. Dataset includes 10862 automatically labeled reviews, half of them are positive and another are negative. Dataset does not have split on training and testing data. Therefore, 80 per cent of data is taken as training data for creating a supervised learning model based on Naïve Bayes and neural networks, 20 per cent is taken as a test set for estimation of the accuracy of the classifiers.

The dataset of IMDB movie reviews is considered because such kind of reviews comprise a broad range of emotions and capture many adjectives suitable for sentimental classification. The second dataset is a dataset that contains automatically annotated tweets. This dataset was collected by the author, their approach based on usage of emoticons (":)", ":-)", ":)", ":D", "=)" mapped to positive emoticons and ":(", ":-(", ": (" mapped to negative). The total amount of tweets in the second dataset constitutes 1.6 million tweets, dataset evenly contains positive and negative tweets. The testing data includes 359 manually annotated tweets, which are labeled as positive and negative. The statistics of the datasets are presented in Table 1.

Table 1. The statistics of the datasets

Dataset	Type	Positive	Negative	Total number of tweets
IMDB movie reviews	Train	4788	4797	9594
	Test	414	534	1068
Tweets	Train	500000	500000	1200000
	Test	172	182	362

B. Feature Extraction

In the first experiment, the unigrams were selected as features for feeding the Naïve Bayes classifier. Sentence (IMDB movie review/ tweet) is split into words (unigrams) and represented as a set of words. Using unigrams end up in a large feature set that has to be reduced to eliminate uninformative features.

The Chi-square feature selection algorithms was investigated for the Naïve Bayes model. Chi-square is a statistical test that measures the independence between the class label and the feature itself. It estimates the importance of the terms by calculating their scores. In other words, it measures the correlation between terms and their classes.

The second experiment was conducted using a convolutional neural network. CNN uses filters (kernels) that play the role of feature detectors. Using initial dataset a vocabulary has to be formed, where each word is indexed. In this research work, two different datasets are used, the size of the IMDB movie reviews dictionary constitutes 19758 words and size of the tweets dictionary constitutes 204062 words.

The sentences of varied length normalized by padding them to the maximum length of the sentence. Overall, each sentence is converted to the vector representation and the whole input text is represented as a matrix. To feed the latter to the convolutional layer, it has to be further converted to the embeddings that are stored in a lookup table. In this research work, the word embeddings initialized randomly.

To select informative features from the initial dataset and move to higher-level perspective, convolution and pooling operations have to be employed.

VI. RESULTS AND ANALYSIS

This section discuss the results that were obtained after conducting the experiments using the Naïve Bayes algorithm and convolutional neural network. The experiment is performed on two different datasets. The first dataset contains the IMDB movie reviews, second contains the tweets. Both datasets are labeled.

To evaluate the quality of the classification algorithms three main metrics are used, namely precision, recall, and F_1 score. Moreover, during training and testing stages, computational time was measured that is also used in the analysis of algorithms' performance.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue XI Nov 2020- Available at www.ijraset.com

A. Evaluation Metrics of algorithms Measure

The effectiveness of the classification algorithms is usually estimated based on such metrics as precision, recall, F_1 score, and accuracy. Moreover, it is very important to take into account computational cost resources that algorithm needs for building the classifier and using it. Consider the metrics that were used for calculation of the precision, recall, F_1 score, accuracy (Table 2). The confusion matrix contains the estimated and actual distribution of labels. Each column corresponds to the actual label and each row corresponds to the estimated the label of the sentence.

Table 2. Confusion matrix for a binary classifier.

Estimated \ Actual	positive	negative
positive	TP	FP
negative	FN	TN

TP is the number of true positives: the sentence that is positive and was estimated as positive, *TN* is the number of true negatives: the sentence that is negative and was estimated as negative, *FP* is the number of false positives: the sentence that is negative but estimated as positive, *FN* is the number of false negatives: the sentence that is positive but estimated as negative.

Accuracy presents the proportion of the correct answers that are given by the classifier hence it can be estimated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision can be estimated using following formula:

$$Precision = \frac{TP}{TP + FP}$$

The precision shows positive answers that received from the classifier that are correct. The greater precision the less number of false hits. However, precision does not show whether all the correct answers are returned by the classifier. In order to take into account the latter recall is used:

$$Recall = \frac{TP}{TP + FN}$$

Whereas Recall shows the ability of the classifier to 'guess' as many positive answers as possible out of the expected.
The more precision and recall the better. However, simultaneous achievement of the high precision and recall is almost impossible in real life that is why the balance between two metrics has to be found. F_1 score is a harmonic mean of precision and recall:

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

B. Naïve Bayes Classifier

Naïve Bayes classifier was trained and tested on two datasets: IMDB movie reviews and tweets. For the Naïve Bayes classifier, all the experiment were conducted using the different amount of word for training the classifier, namely the *n* words that have the highest score were fed to the classifier.

This score was calculated using χ^2 test, for this purpose frequency distribution of all words in the dataset was found as well as the conditional frequency is defined to count how many times a word has occurred in the positive sentence and how many times in the negative.

The primary experiment involves the Naïve Bayes classifier which learned from IMDB movie reviews and evaluated on the IMDB movie reviews.

It is seen that the small dataset (up to 450 words) all demonstrated metrics have lower values compared to the usage of the larger amount of words for training.

The highest accuracy is reached when 4000 informative words are taken as features and it constitutes 82.50 per centage. Moreover, the classifier that is trained on 5500 of the best word also shows the highest values of recall and F_1 score. Recall equals to 92.74 percentage and F_1 score is 84.00 percentage. Nevertheless, the highest precision is gained when 4500 words are used for learning the classifier and makes up 82.6 percentage.

In case of sentimental classification, the precision is more important metric because the classifier has to be precise in detecting true positive answers. Hence, the usage of 4500 words is most favorable for training the classifier on IMDB movie review in order to get the optimal performance in recognizing the positive and negative tweets.

The next test is performed using the same classifier that is trained on IMDB movie reviews, but evaluation is done on tweets.

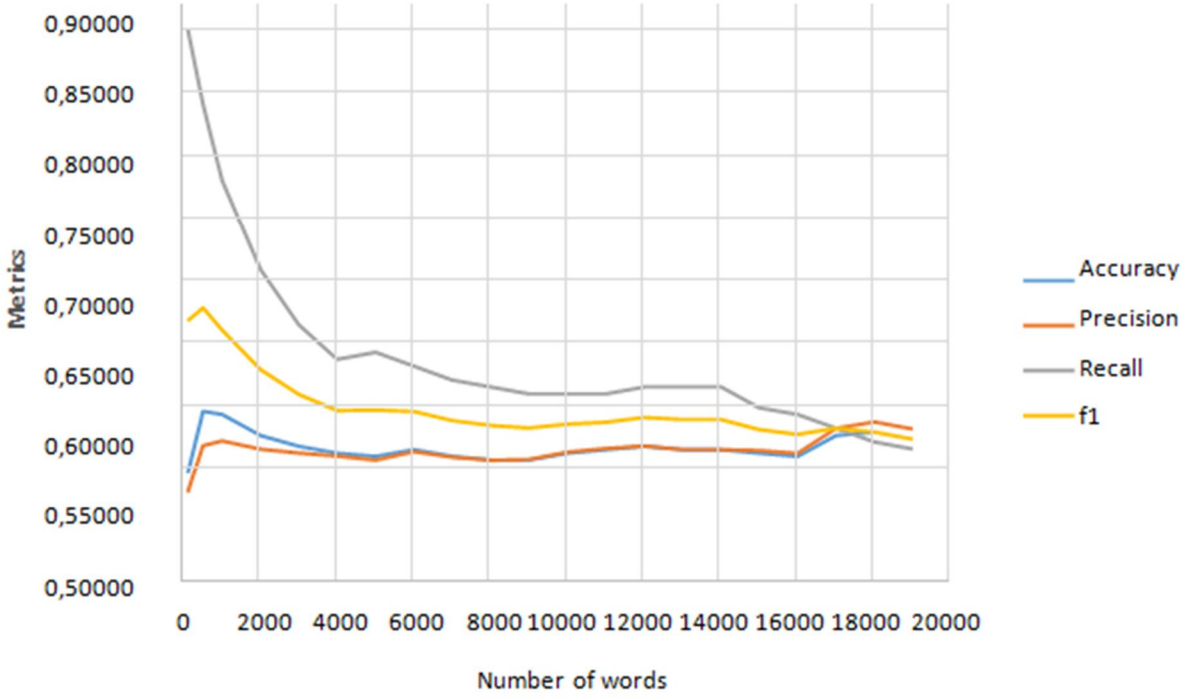


Figure 1 : Naïve Bayes classifier that was trained on the IMDB movie reviews

The above figure shows all the metrics got the lower values opposed to the previous case. The highest accuracy is reached when 6000 words are used for training the classifier and it equals to 68.20 percentage. Furthermore, F_1 score gets its optimal value of 68.22 percentage if 13576 words are used as features. However, the highest value of recall is gained when using only 100 words and it constitutes 84.00 percentage. On the other hand, the optimal precision is reached when the classifier is learned from the whole dataset. Such situation happens because different data is used for training and testing the system. The context of the data used for training has a huge impact on the performance of the algorithm. As mentioned above, tweets differ from the usual sentences, such as reviews due to its informal lexicon that classifier does not know.

The next experiment was conducted on the model that is trained on the larger dataset, which contains 1.6 Million tweets and tested on the tweets that were used for evaluation before. The result of the evaluation is depicted in Figure 2.

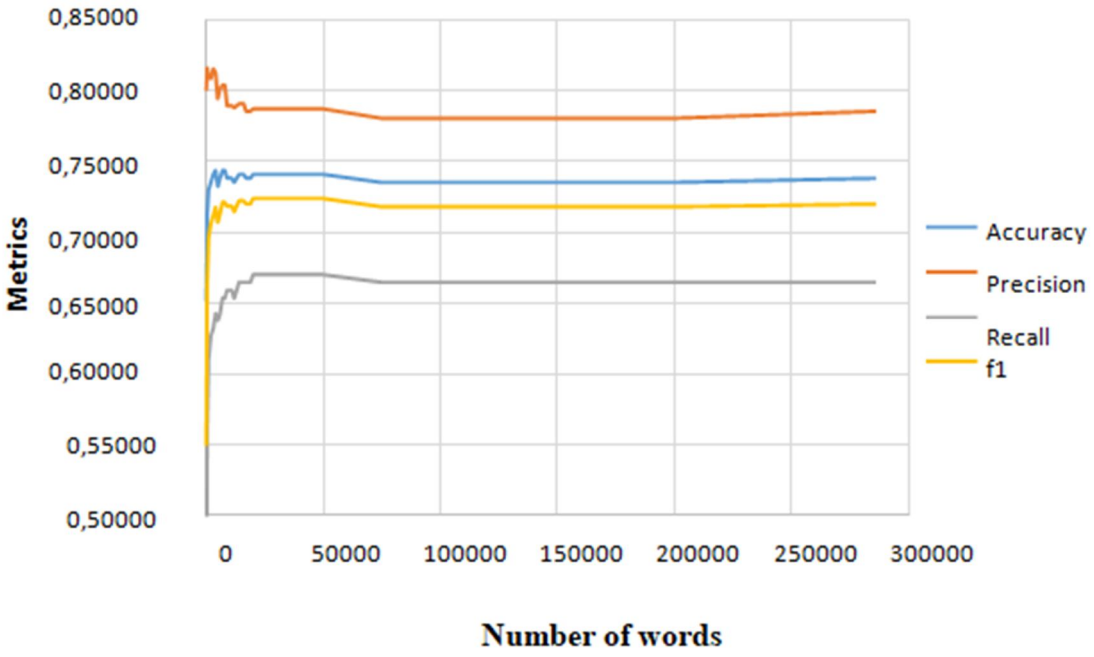


Figure 2 : Naïve Bayes classifier that was trained on the tweets

The classifier that is trained on the tweets classifies tweets better than the one that is trained on the IMDB movie reviews. The maximum of the accuracy is achieved when classifier takes 27152 words as features for learning and the accuracy constitutes 74.47 per cent. On the other hand, the highest values of the recall and F_1 score are reached when the number of features makes up 6788 words and equal to 69.33 per cent and 70.44 per cent respectively. The precision of the model that is trained and tested on tweets is 20.5 per cent higher than the precision of the one that is trained on IMDB movie reviews but tested on tweets and constitutes 78.90 percentage.

To sum up, when the classifier is trained and tested on the same type of data it shows better performance. Moreover, it has been found that the classification model that is based on the Naïve Bayes approach does not require huge training dataset, however, it needs the data samples from the same domain for training and testing the classifier.

Moreover, computational cost is estimated. It is clear that more specifically, during the training process that includes preprocessing and feature selection, the usage of virtual memory resource was evaluated. Figure 3 demonstrate the precision, recall and F-measure values while training the Naïve Bayes classifier on reviews.

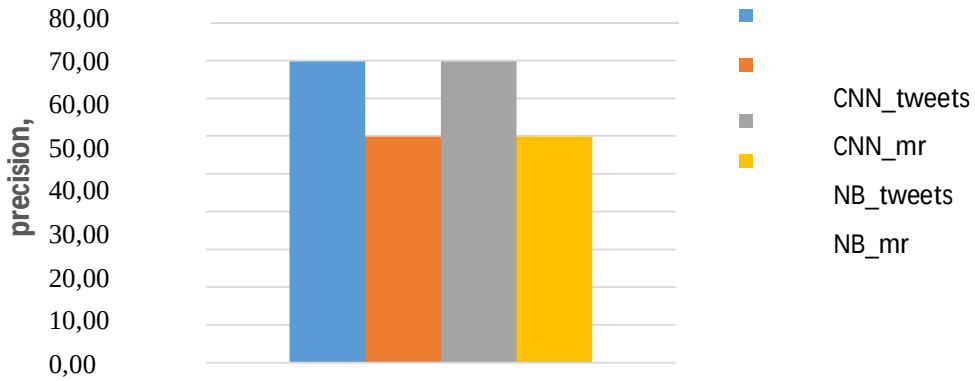


Figure 3. Precision of different classifiers

It is important here that CNN requires way more memory than NB. However, CNN classifier produces similar metrics as NB. Therefore, analysis of the results shows that investigated models may be further improved because metrics of the accuracy, precision, recall and F_1 score are not significant as they were expected, especially when employing CNN classifier.

In addition, it was investigated that the context of the dataset highly affects the performance of the classifier. If the task is to classify the data from whatever domain, then the classifier has to know samples that capture varied context. Hence, the quality of the dataset has an enormous impact on the effectiveness of the classification model.

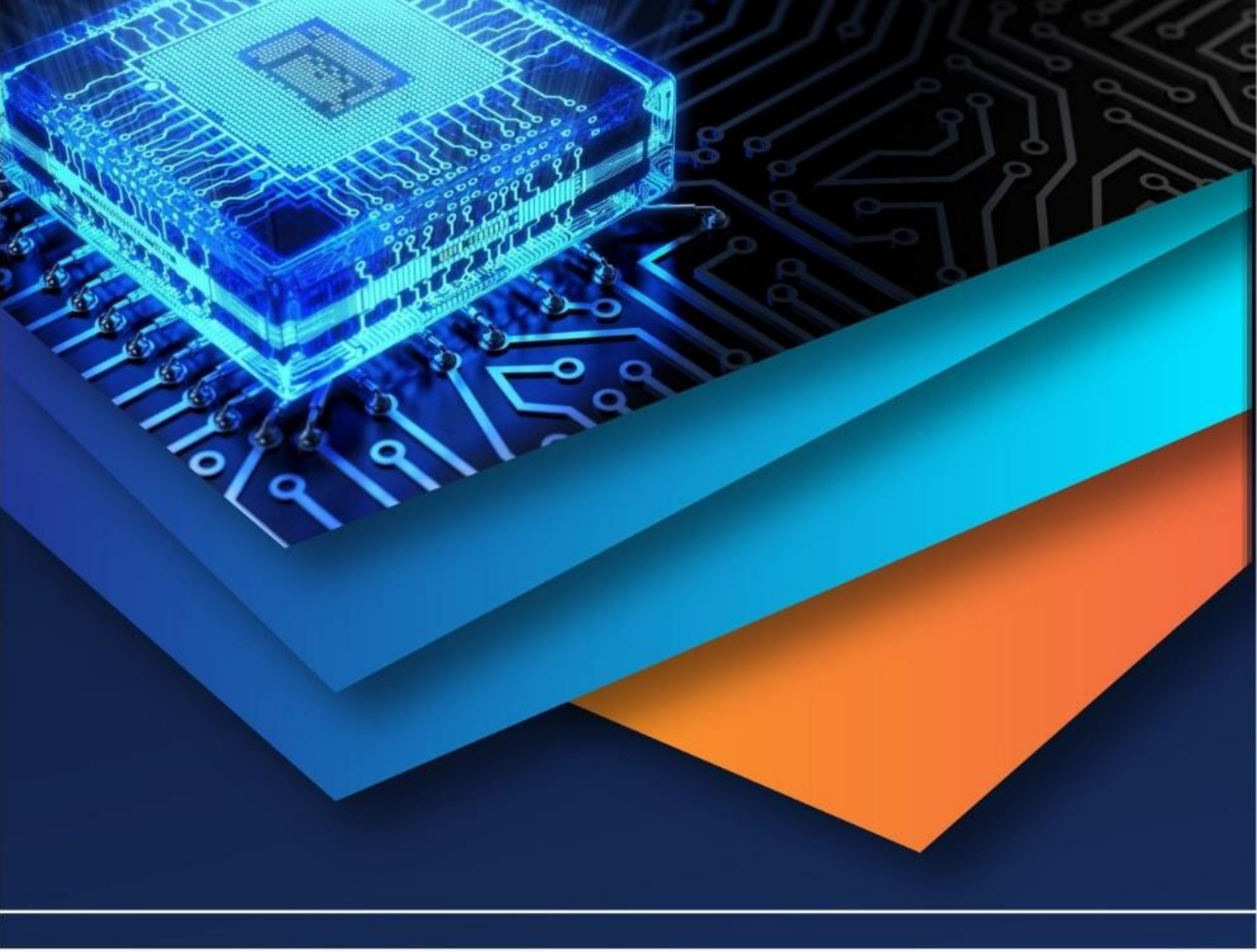
VII. CONCLUSION

In this research work binary classification is considered, namely, the tweet/review is assigned a positive or negative label according to the sentimental conveyed in it. Two different classifiers were investigated in order to estimate the sentimental. Classifiers performance is evaluated based on experiments. The first supervised method that was explored in this research is Naïve Bayes approach. As was expected it has shown sufficient results on the tweet classification. The best result of the precision that was achieved, made up 78.70 percentage when NB classifier was learned from the whole set of tweets. Another supervised approach that was studied for training the classifier is the one-layer convolutional neural network.

After evaluation of the CNN, the precision has slight growth and constituted 79.10 percentage. However, it was discovered that the CNN is extremely resource-demanding opposed to NB. In general, CNN performs better that Naive Bayes classifier, but it requires solid computational resources and large amount of training sample.

REFERENCES

[1] Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., & Perera, A. (2012, December). Persuasion and sentimental analysis on a twitter data stream. In Advances in ICT for emerging regions (ICTer), 2019 International Conference on (pp. 182-188). IEEE.
[2] Hallsmar, F., & Palm, J. "Multi-class sentimental classification on twitter using an emoji training heuristic", Vol. II, Issue-I (2018)
[3] Turney, P. D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews". In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics (2018)
[4] J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez- García, M. Á., & Valencia-García, R. Salas-Zárate, M. D. P., Medina-Moreira, "Sentimental Analysis on Tweets about Diabetes: An Aspect-Level Approach". Computational and mathematical methods in medicine, 2017.
[5] Chiavetta, F., Bosco, G. L., & Pilato, G. "A Lexicon-based Approach for Sentimental Classification of Amazon Books Reviews in Italian Language (2016).
[6] Hailong, Z., Wenyan, G., & Bo, J. (2015). Machine learning and lexicon based methods for sentimental classification: A survey. In Web Information System and Application Conference (WISA), 2014 11th (pp. 262-265). IEEE.
[7] Hu, M., & Liu, B. (2014). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177).
[8] Miller, G. A. (2005). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.
[9] Dong, Z., Dong, Q., & Hao, C. (2012). Hownet and its computation of meaning. In Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations (pp. 53-56).
[10] Musto, C., Semeraro, G., & Polignano, M. (2014). A comparison of lexicon-based approaches for sentimental analysis of microblog posts. Information Filtering and Retrieval, 59.
[11] Kim, S. M., & Hovy, E. (2014). Determining the sentimental of opinions. In Proceedings of the 20th international conference on Computational Linguistics (p. 1367). Association for Computational Linguistics.
[12] Park, S., & Kim, Y. (2016). Building thesaurus lexicon using dictionary-based approach for sentimental classification. In Software Engineering Research, Management and Applications (SERA), 2016 IEEE 14th International Conference on (pp. 39-44). IEEE.
[13] Severyn, A., & Moschitti, A. (2015). Twitter sentimental analysis with deep convolutional neural networks. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 959-962). ACM.
[14] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
[15] Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.
[16] Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188.
[17] Britz, D. (2015). Recurrent Neural Networks Tutorial, Part 1–Introduction to RNNs.
[18] Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101.
[19] Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In Mining text data (pp. 163-222). Springer US.
[20] Pang, B., Lee, L. (2005). IMDB movie Review Data. Sentence polarity dataset v1.0. https://www.cs.cornell.edu/people/pabo/IMDB movie-review-data/ (05.05.2017)
[21] Go, A., Huang, L., & Bhayani, R. (2009). Twitter sentimental dataset. http://cs.stanford.edu/people/alecmgo /trainingandtestdata.zip
[22] Liu, B. (2016). Sentimental analysis and Persuasion. Synresearch work lectures on human language technologies, 5(1), 1-167.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY