Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 2): pp. 2659-2665
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.02.282

News Text Classification Using Machine Learning Algorithms

[1]Jyoti Agarwal, [2] Piyush Agarwal, [3]Aditya Pai H, [4]Navadha Bhatt

[1]Department of CSE Graphic Era Deemed to be University, Dehradun
itsjyotiagarwal1@gmail.com

[2]Department of CSE Graphic Era Deemed to be University, Dehradun
piyush221292@gmail.com

[3]Department of CSE Graphic Era Deemed to be University, Dehradun
adityapaih2007y@gmail.com

[4] Department of Chemistry, Graphic Era Hill University, Dehradun, India.

ABSTRACT

In current scenario, lot of online news is available for different topics on Internet from which textual data is increasing rapidly. Due to this, it becomes essential to organize them properly so that important news can be searched easily as well as to avoid data loss. One effective solution for this problem is to classify the news into different classes or to extract most important and useful information. This paper is an attempt to provide a solution for by classifying the news text into different classes. For this, two different machine leaning algorithms (Random Forest and Decision Tree) are used. Experiment is performed on an online dataset taken from Kaggle to analyze which algorithm can be used to provide better results.

Key words: News, text, machine, learning, Random Forest Decision Tree

1.INTRODUCTION

News text is very much essential for having timely information which is also required in future. Now a days, Internet is full of news content, and it became a challenging task to store this information efficiently so that it can be extracted easily in future on time. News text also contain so much of irrelevant information which one is of no use and user want to read only crisp point about the news article. For this purpose, news text classification concept is used where text is classified based on certain attributes like topic, language, author name, date etc. (Stein, 2020). The major application of this classification is that it becomes easy to search a particular category of news for which a user may be interested which saves time, effort, and money. Supervised machine learning algorithms can be used for this purpose where a model is trained using training data and results are obtained for a particular dataset. This paper is also an attempt for working in this

direction.

The primary focus of this paper is to extract useful information from news articles. This extraction can be done based on topics, sentiments, author etc. In this work, this classification is done based on some selected news categories like sports, entertainment, politics etc. For this purpose, two different ML algorithms (Random Forest and Decision Tree) are used, and experiments are performed on an online dataset. Results shows that Random Forest was performing well compared to decision tree algorithm in terms of accuracy as well as f1 score.

The major contribution of paper will be that it will be very much useful to store important information efficiently which will help to save searching time in future and data will not be lost.

This paper is dived into following parts: section 2 describes review of related work in area of hand gesture recognition, section 3 focuses on research methodology for describing all the necessary steps to be followed to provide results, results are given in section 4 and in section 5 conclusion and future scope of this work is provided.

2. RELATED WORK

In 2013, support vector machine algorithm was used by researchers to classify twitter news into 12 different groups for Sri Lanka (Dilrukshi, 2013). In the same year, a review is done for different ML algorithms to classify news text and discuss their pros and cons (VasfiSisi & Derakhshi, 2013). In the next year 2014, authors have discussed new classification process and reviewed some existing classifiers and their methodologies (Rana, Khalid & Akbar, 2014). A specific review is done by researchers for Indian language content where they have reviewed supervised as well as unsupervised algorithms for news classification and they have analyzed that supervised algorithm performs better than in compare of unsupervised algorithms (Kaur & Saini, 2015). In 2015, a novel approach (TESC) was also designed for text classification using semi-supervised learning approach. Two different data sets Reuters-21578 and TanCorp V1.0 were used for experimental work which shows that TESC performs better than SVM and BPNN and provide better scalability also (Zhang, Tang, & Yoshida, 2015). A boosting based a combined approach of semi supervised learning and Universum learning was proposed for text classification. To analyze the performance of proposed approach different data sets are used and it was analyzed that the proposed algorithm was able to provide better results in the absence of labeled data (Liu, et al., 2015). In 2016, neural network approach was also applied by researchers to check its performance for news classification, and it was found that Random Forest was giving accuracy of 73% while neural network was able to provide 99.285% of average accuracy with average precision rate of 0.76125(Kaur & Khiva, 2016). A different news text classification model based on Latent Dirichlet (LDA) was proposed in 2016, which was using topic model to select features if news text. This model was able to provide better results for reducing the feature dimensions of the news text and improved results were obtained for a real news dataset (Li, Shang & Yan, 2016). In 2017, CNN approach was applied by

authors to propose a text classification method. Results were generated on two different datasets and more than 96% of accuracy was achieved for both the datasets (Li et al., 2017). For better accuracy of text classification major challenge is preprocessing of data. If data is not preprocessed, then there is major chance of incorrect output. Researchers have studies various evolutionary preprocessing tools for English text classification (Kadhim, 2018). A text classification study is also done for Indonesian news article where authors have explored various ML algorithms and analyzed that SVM was able to provide maximum f1 score of 93% (Londo et al., 2019). A comparison is also done between Multinomial and Bernoulli Naïve Bayes algorithm based on sentiments of newspaper article and concluded that Multinomial Naïve Bayes was able to provide slightly better results in compared to other one for the given dataset (Singh, et al., 2019). In the same year 10 different classification algorithms are used to classify Arabic text from news article where accuracy was between 87-97% (Qadi, et al., 2019). In 2020 text classification was also done for Azerbaijani language using ML algorithms and sentiment analysis is also done for product reviews (Suleymanov, 2020). Vox media dataset of 2017 was used by researchers to classify some important information like author name and topic of article in 2020 and analyzed that neural network was giving better results compared to tradition ML classification algorithms (Stein, 2020).

3. RESEARCH METHODOLOGY

The proposed research methodology is shown in Fig 1. Firstly, data is gathered from an online source (https://www.kaggle.com/datasets/rmisra/news-category-dataset) and its preprocessing is done. After that text tokenization is done where text is divided into tin parts. In next step, stemming is done which is word normalization approach used in Natural Language Processing (NLP). This process removes suffix from word. After selecting the features training and testing of the model is done. For this 80% of data is sued for training and rest 20% is used for testing. At the end final prediction is done to classify the news text which is shown in Fig 2.

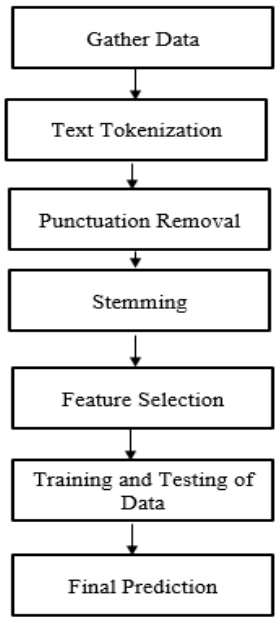

Fig. 1. Proposed Research Methodology

Fig 2. Results obtained from Random Forest Classifier

4. RESULTS ANALYSIS

Results are obtained from Random Forest and Decision Tree classification algorithms which was giving accuracy of 91.94% and 79.19% respectively which shows that Random Forest was able to provide higher accuracy in comparison to Decision Tree method. This comparison is shown in Table 1. Results obtained from the proposed research work are shown in Fig.3 and Fig. 4.

Table 1. Accuracy level of Algorithm

Algorithm	Accuracy (%)
Random Forest	91.94
Decision Tree	79.19

```
In [86]: from sklearn.ensemble import RandomForestClassifier
         model  = RandomForestClassifier(n_estimators = 500)
         model.fit(features_train, labels_train)
         model_predictions = model.predict(features_test)
         print('Accuracy: ', accuracy_score(labels_test, model_predictions))
         print(classification_report(labels_test, model_predictions))

Accuracy:  0.9194630872483222
              precision    recall  f1-score   support

           0       0.94      0.88      0.91        76
           1       0.93      0.89      0.91        47
           2       0.89      0.87      0.88        55
           3       0.93      0.97      0.95        65
           4       0.90      0.98      0.94        55

    accuracy                           0.92       298
   macro avg       0.92      0.92      0.92       298
weighted avg       0.92      0.92      0.92       298
```

Fig 3. Results obtained from Random Forest Classifier

```
In [99]: from sklearn.tree import DecisionTreeClassifier
         model = DecisionTreeClassifier(random_state=1)
         model.fit(features_train, labels_train)
         model_predictions = model.predict(features_test)
         print('Accuracy: ', accuracy_score(labels_test, model_predictions))
         print(classification_report(labels_test, model_predictions))

Accuracy:  0.7919463087248322
              precision    recall  f1-score   support

           0       0.74      0.75      0.75        76
           1       0.84      0.89      0.87        47
           2       0.78      0.64      0.70        55
           3       0.79      0.92      0.85        65
           4       0.84      0.76      0.80        55

    accuracy                           0.79       298
   macro avg       0.80      0.79      0.79       298
weighted avg       0.79      0.79      0.79       298
```

Fig 4. Results obtained from Decision Tree Method

5. CONCLUSION AND FUTURE SCOPE

This research work is about classification of news text using two different machine learning algorithms. For this purpose, Random Forest and Decision Tree algorithms were used and it was identified that Random Forest was giving better results in compared to decision tree algorithm.

This work will be useful for researchers and academicians to forward their research in this direction using more efficient algorithms. In future accuracy can be achieved by increasing training level of machine learning model. More classification algorithms can also be compared on different data sets to compare their efficiency level.

REFERENCES

[1] Dilrukshi, I., De Zoysa, K., & Caldera, A. (2013, April). Twitter news classification using SVM. In 2013 8th International Conference on Computer Science & Education (pp. 287-291). IEEE.

[2] VasfiSisi, N., & Derakhshi, M. R. F. (2013). Text classification with machine learning algorithms. Journal of Basic and Applied Scientific Research, 3(1), 31-35.

[3] Rana, M. I., Khalid, S., & Akbar, M. U. (2014, December). News classification based on their headlines: A review. In 17th IEEE International Multi Topic Conference 2014 (pp. 211-216). IEEE.

[4] Kaur, J., & Saini, J. R. (2015). A study of text classification natural language processing algorithms for Indian languages. VNSGU J Sci Technol, 4(1), 162-167.

[5] Zhang, W., Tang, X., & Yoshida, T. (2015). TESC: An approach to TExt classification using Semi-supervised Clustering. Knowledge-Based Systems, 75, 152-160.

[6] Liu, C. L., Hsaio, W. H., Lee, C. H., Chang, T. H., & Kuo, T. H. (2015). Semi-supervised text classification with universum learning. IEEE transactions on cybernetics, 46(2), 462-473.

[7] Kaur, S., & Khiva, N. K. (2016). Online news classification using deep learning technique. International Research Journal of Engineering and Technology (IRJET), 3(10), 558-563.

[8] Li, Z., Shang, W., & Yan, M. (2016, June). News text classification model based on topic model. In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) (pp. 1-5). IEEE.

[9] Li, L., Xiao, L., Wang, N., Yang, G., & Zhang, J. (2017, December). Text classification method based on convolution neural network. In 2017 3rd IEEE International Conference on Computer and Communications (ICCC) (pp. 1985-1989). IEEE.

[10] Kadhim, A. I. (2018). An evaluation of preprocessing techniques for text classification. International Journal of Computer Science and Information Security (IJCSIS), 16(6), 22-32.

[11] Londo, G. L. Y., Kartawijaya, D. H., Ivariyani, H. T., WP, Y. S. P., Rafi, A. P. M., & Ariyandi, D. (2019, March). A Study of Text Classification for Indonesian News Article. In 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT) (pp. 205-208). IEEE.

[12] Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019, April). Comparison between multinomial and Bernoulli naïve Bayes for text classification. In 2019 International Conference on Automation, Computational and Technology Management (ICACTM) (pp. 593-596). IEEE.

[13] Al Qadi, L., El Rifai, H., Obaid, S., & Elnagar, A. (2019, October). Arabic text classification of news articles using classical supervised classifiers. In 2019 2nd International conference on new trends in computing sciences (ICTCS) (pp. 1-6). IEEE.

[14] Stein, A. J., Weerasinghe, J., Mancoridis, S., & Greenstadt, R. (2020). News article text classification and summary for authors and topics. Comput. Sci. Inf. Technol.(CS & IT), 10, 1-12.

[15] Suleymanov, U., Kalejahi, B. K., Amrahov, E., & Badirkhanli, R. (2020). Text Classification for Azerbaijani Language Using Machine Learning. Comput. Syst. Sci.

Eng., 35(6), 467-475.