Review of Machine Learning Approaches for Multiclass Classification of Urdu Text

Noman Tahir[1]

1 Introduction

The advancement in almost every field of research is greater in one perspective that human life is getting easier. The other perspective is the data, which is increasing massively in terms of research data, news data, or other specific departments such as medical, meteorology, etc. This raises a challenge to the data analysis community when it comes to analysing huge amounts of data to get specific information.

There exist many methods to summarise and classify huge amounts of data but most of them are originally made for and trained on the English language. There are also some methods that are trained on some European languages that are spoken in many countries but for the Asian languages, the scenario is different. While dealing with Asian languages, specifically the Urdu language there are insufficient data resources and also rare approaches to analyze data written in Urdu text. Keeping fact, it is spoken in Pakistan mainly and also India with almost more than 200 million speakers combined(Ashraf, 2023).

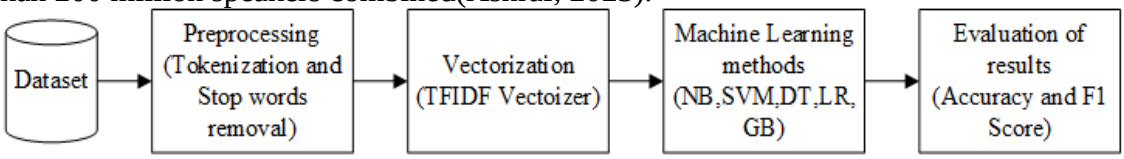

Figure 1: Research Methodology

This study aims to figure out the challenges to Urdu text analysis and best-performing Machine Learning (ML) methods (Gasparetto, Marcuzzo, Zangari, & Albarelli, 2022) from the state of the art that works well with the Urdu text. The aim is developed considering the usage of the best ML method for further enhancement or analysis of newly gathered data resources.

The methodology of this study, according to Figure 1, works in such a way that it collects a dataset from Kaggle[2] which contains Urdu news data collected from multiple news platforms. The data is further classified into three categories: *entertainment*, *crime*, and *cricket*. The data is then passed through pre-processing by tokenizing the words, and removing the stop words(Rahimi & Homayounpour, 2023). The lemmatization is not applied in the pre-processing because the model needs to be trained on the exact words.

The next step deals with the vectorization of tokens, the TFIDF vectorizer is used for this task. The next step is to split the data into training and testing, this study divides the dataset with 80% and 20% split. The training and testing data were then passed separately to each model. This study selected the Naïve Bayes, Decision Tree, Support Vector Machine (SVM), Linear Regression, and Gradient boosting as the most common ML approaches for text analysis (Khurana, Koli, Khatter, & Singh, 2023). The accuracy and F1 score were considered as

[1] student of the Doctoral degree program, Department of Computer Science and Engineering, field of study Natural Language Processing, e-mail: noman@kiv.zcu.cz

[2] https://www.kaggle.com/datasets/disisbig/urdu-news-dataset (last accessed 18-05-2023)

evaluation parameters and scores for both are calculated for each approach (Amini, Rahmani, & Technology, 2023).

The results in Table 1 shows that the SVM's performance is better as compared to the other ML methods. SVM scored 94 % score for accuracy and F1 score both. However, the Naïve Bayes was able to score 89% accuracy and 88% F1, the Decision tree was also near with 89% score for both parameters, the Linear Regression and Gradient boosting scored same with 91% accuracy and 90% F1 score. So overall, it is claimed that SVM is the best among all in order to deal with the Urdu text.

	Naïve Bayes	Decision Tree	SVM	Linear Regression	Gradient Boosting
Accuracy	89	89	94	91	91
F1 Score	88	89	94	90	90

Table 1: Results of all methods

2 Conclusion and Future work

While the identification of the best ML method was successful, but this study has some limitations. The approaches applied in this research use the parameters with default settings. On the one hand, it is good that all approaches use the same input data from the TFIDF vectorizer, but it could be an option to try any other vectorizer along with TFIDF to check the operability of methods. This article serves as the beginning of a venture toward the identified challenges in Urdu text analysis. The future work with this study is the implementation of different vectorizers with these methods to check the best combination with ML methods. Applying Deep Learning and Transformer methods with the same goals to get quality results.

Acknowledgment

I would like this opportunity to thank my mentor, Prof. Karel Ježek for putting his efforts into providing me the guidelines to make this possible.

References

Amini, M., Rahmani, A. J. I. J. o. S., & Technology, A. (2023). Machine learning process evaluating damage classification of composites. *International Journal of Science and Advanced Technology, 9*(2023), 240-250.

Ashraf, H. J. L. p. (2023). The ambivalent role of Urdu and English in multilingual Pakistan: a Bourdieusian study. *Language Policy*, *22*(1), 25-48.

Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. J. I. (2022). A survey on text classification algorithms: From text to predictions. *Information, 13*(2), 83.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications, 82*(3), 3713-3744. doi:10.1007/s11042-022-13428-4

Rahimi, Z., & Homayounpour, M. M. (2023). The impact of preprocessing on word embedding quality: a comparative study. *Language Resources and Evaluation, 57*(1), 257-291. doi:10.1007/s10579-022-09620-5