Journal of Advancement in Computing (JAC)

A HYBRID FEATURE SELECTION APPROACH FOR ROMAN URDU TEXT CLASSIFICATION

Waqas Azeem[1], Chakir Aziza[2]
1#Faculity of Computer science University: Government college university Faisalabad
2#Faculty of Law, Economics and Social Sciences, Hassan II University, Casablanca, Maroc
Email: aziza1chakir@gmail.com
DOI: https://doi.org/10.36755/jac.v2i1.61

A B S T R A C T

Text classification is the task of assigning labels to unlabeled text data. Text classification has several applications like sentiment analysis, document classification, and fake news detection such as Machine learning (ML) methods have been used commonly in text classification in the last several years. The fundamental problem in ML is that these approaches heavily depend on feature selection methods. The models and feature selection methods used in this research. Several past types of research conclude that there is no uniform feature selection method that works well for all types of classifier tasks as well as Urdu is a resource-poor language. In this study, a proposed hybrid feature selection approach for Roman Urdu text not only reduces the dimension of the feature map but also increases the accuracy of ML models. Using 11000 and 20000 records have been used for Support Vector Classifier, Naive Base and Decision Tree which have given 80.81%, 72.94% and 76.78% respectively, among other tested methods. The best accuracy values achieved by each classifier and the hybrid features ChiSAE, CorrelationAE, and GainRAE. In future, text classification for better understanding of human being self-analysis as well as deep learning methods will be utilized for better authenticity.

A R T I C L E I N F O

Article History:
Received 12/09/2023
Accepted 30/01/2024
Available,online 31/01/2024

Keywords:
Hybrid Roman Urdu
Data Mining
Feature Extractions
Text Classification
Sentimental Analysis
.

1.INTRODUCTION

One of the most recent technological advances being made worldwide is in the field of artificial Text classification (TC) is the development of automatically classifying comment text into a set of labels. Text categorization is another term for text classification [1]. In-text categorization, the text represents a text document, news story, comment, sentence, or paragraph. Labels can also be referred to as classes or categories. TC is an important field of information retrieval and data mining. It has several important application in natural language processing like tweet analysis,

document organization, user sentiment analysis, spam or ham classification, automatic fake news detection, and abusive text detection [2].

The exponential growth of [3] users on e-commerce platforms, news blogs and channels, open-access digital libraries, and many social media websites are producing every day a large number of text comments, which is causing a large volume of unstructured text data. Unstructured data has several issues like poor quality, less reliability, waste of organizational resources, and little potential for making good decisions. There are two choices to organize the unstructured data either manually or automatically.

Manual organization of unstructured is almost impossible as it requires a lot of organizational resources like human labor and their effort, time, and budget. Humans perform poorly than automated systems to structure the unstructured text. The automatic method customs machine learning and natural language processing methods to structure the unstructured text data [4] . Machine learning (ML) methods have been widely used in automatic text classification methods in recent years. Artificial intelligence has an area called machine learning. Machine learning methods are a collection of algorithms that learn from a corpus of data during training and improve their performance automatically over test data. These algorithms were created to assist humans in making judgments about many parts of their lives. Machine learning algorithms have demonstrated substantial performance in a variety of disciplines, including image processing, natural language processing, and computer vision. The primary problem for these algorithms is selecting efficient features from large feature spaces, yet there is no uniform feature selection method that works well for all types of classifiers. For text classification in resource-rich languages like English, machine learning algorithms have been frequently deployed. Only a few text classification experiments employing ML have been done for resource-deprived languages like Urdu [5].

2. RECENT WORK

Text classification is a supervised learning task, so the process starts with text collection from different platforms like websites, blogs, and social media platforms and then annotating text (labelled it). Then dataset is divided into training and [2] testing subsets. The training subset is used to train the machine learning model while, after training, the testing subset is used to test the model performance. Before the text is given to a machine learning model, data is first preprocessed (tokenization, normalizing, cleansing, and removing stop words and rare words). Followed by preprocessing, feature selection is performed using feature selection methods like information gain (IG), gain ratio (GR), and n-grams. The objective of the feature selection is to reduce the vocabulary size, reduce feature space dimensions by extracting useful features from it, and reduce space and time requirements. On the extracted features [6] machine learning methods like (SVM), (KNN), and (NB) are applied to train and predict label of data. Feature selection methods can affect the performance of the classifier.

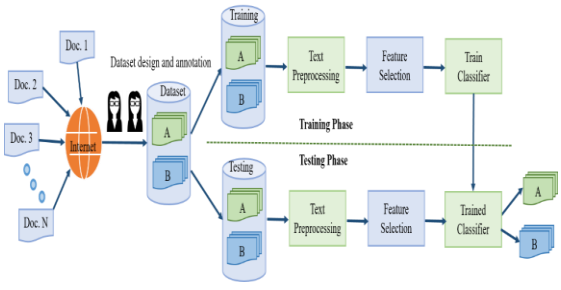


Fig.1 Automated Text Classification

Recent years have seen a substantial increase in the use of Roman Urdu datasets for natural language processing. The da-

JAC 2023, https://doi.org/10.36755/jac.v1i1.47

taset includes 15,000 remarks that were collected from various sources, including Twitter, Reddit, Urdu poetry, and social worker biographies. The main task is to do discourse-based sentiment analysis on Roman Urdu. To do this, data were first gathered, then they were transformed into standard expression using lexical normalization, and last they were examined to determine whether or not a discourse element was present. The average accuracy was 80%. The objective of this research was to develop a neural network approach for sentiment analysis [7] .

Fig.2 . Graphic Representation of the Hybrid Feature Process

Dataset [8] was collected from single platform and set a small size dataset (four thousand comments only, only one feature selection method was used (TFIDF). [9-12] performed on a single imbalanced dataset, best model SVM ML achieved low accuracy 64%, use of insufficient stop words list.

The proposed study is that a hybrid feature selection method to reduce high dimensional feature size and increase accuracy. Comparatively analyze the performance of the proposed hybrid feature selection method with individual feature selection methods. The data from Roman Urdu (vaccination recipients), including their comment text speech, as well as their impact and results, were analyzed in this work using statistical approaches and machine learning models.

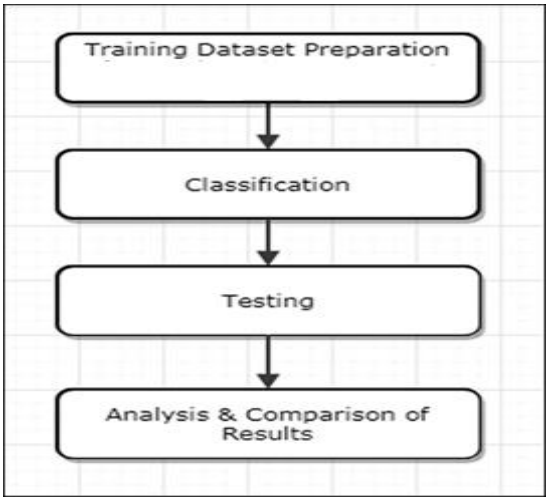


Fig.3: Research Framework

In Figure 3 training dataset has been utilized for the classification of roman Urdu text Positive, Negative and Neutral. Format, and reduce the data. After that, testing was applied along with the analysis with comparisons and results. In hybrid Roman Urdu employing a ML model that incorporates the stop words embedding CorrelationAE, ChiSAE, and GainRAE algorithms into an SVM architecture for Roman Urdu/Hindi hybrid classification. SVM, NB, and DT are known for being potential WEKA tools. Allow nearby input items to extract at hybrid classification on different

datasets 20000 and 11000 while datasets interact at single features so they may separately control how long dependence last. The best accuracy values achieved by each classifier and the hybrid features ChiSAE, CorrelationAE, and GainRAE [13-16].

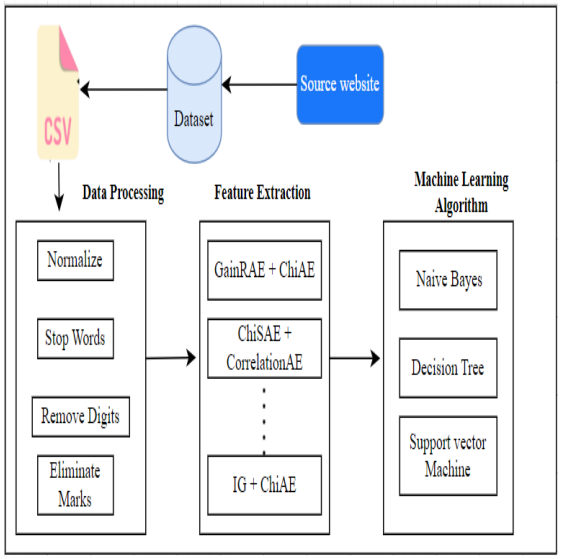


Fig.4: Proposed Model

3. RESULTS AND DISCUSSIONS

The results from the dataset and Decision Tree Algorithm, Multilayer Perception and Navies then applied the feature classification using WEKA tool. In addition, the last series contains dataset of respondents who were used for their classification. The data was analyzed using descriptive statistics and classification. This study gets publicly in worldwide. Secondary data were also collected and examined to determine the feasibility and relevance of the study. This collection contains 11000 and 20000 two data set with different categories and specifications. In this study, the first phase of was checked sample data correction with WEKA tool, while the second phase was removed and unremoved stop words data set accuracy result with hybrid classification and to resolve the issue of class imbalance for Roman Urdu text, to provide a list of stop words of Roman Urdu to remove frequent words in the text. However, the proposed study is that a hybrid feature selection method to reduce high dimensional feature size and increase accuracy.

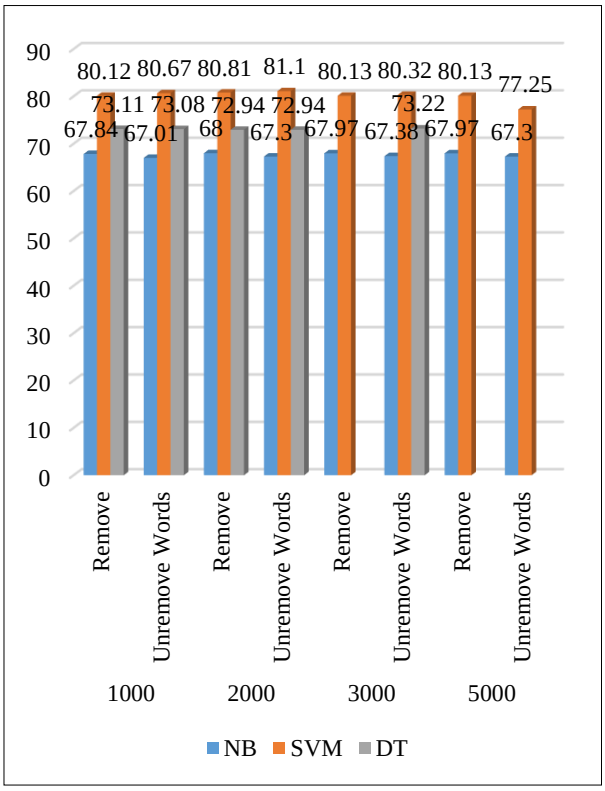
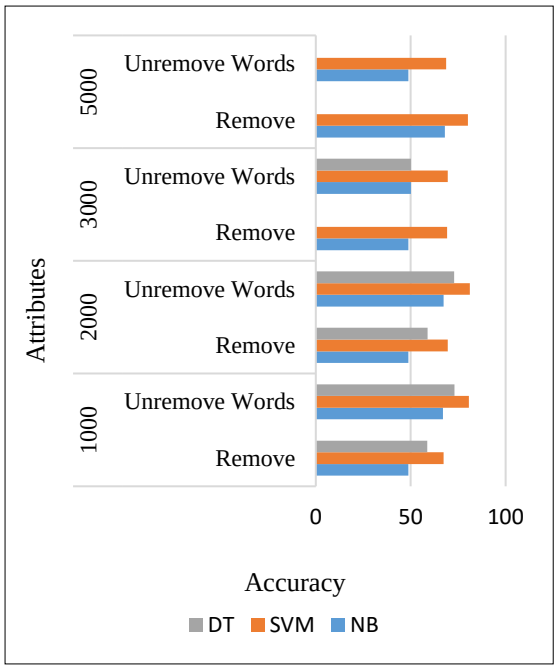


Fig.5 Comparison Accuracy of Remove and Unremoved Stop Words in 11000 Datasets

Fig 5 shows that removed datasets accuracy is better than unremoved datasets while datasets 2000 is better performed than other datasets. The accuracy of SVM 80.12 is dominated than other ML NB and DT.Fig.6 Comparison Accuracy of Remove and Unremoved Words in 20000 Datasets

The proposed Methodology for hybrid feature selection for Roman Urdu text comparing with 11000 and 20000 datasets remove and unremoved are discussed in this chapter as a research framework. The proposed method evaluates classification text in Chi SE, Info gain and using statistics analysis and ML algorithms. The chapter also discusses to classification of positive and negative comments text hybrid classification. The proposed study is that a hybrid feature selection method to reduce high di-

JAC 2023, https://doi.org/10.36755/jac.v1i1.47



mensional feature size and increase accuracy. Comparatively analyze the performance of the proposed hybrid feature selection method with individual feature selection methods.

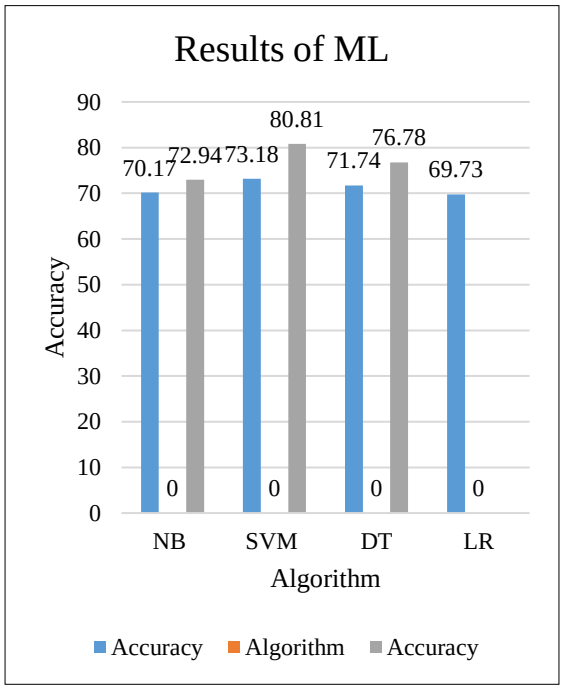


Fig.7: Previous Study Performance and proposed study results of ML

Fig 7 shows that SVM is better performed than other classifiers while proposed study is compared with previous SVM is showed good results 80.2% with accuracy

4. CONCLUSION

The effort done to establish a categorization system for roman Urdu comments text is presented in this publication. The method we used is stepwise; in the first stage, the roman data is tokenized, and then stop words are deleted from the tokenized data and then restored using several methods. By utilizing three distinct deep learning techniques, the experimental development of ten different news classes demonstrates high accuracy. In this study, Roman Urdu news is categorized using the Deep Learning models NB, SVM, and DT. It's because just one or two characters in the Romanized Urdu and Urdu script designed significant words. Commonly used stop words, these words. SVM classifier performs 80.12 % with 11000 features stop words removed and highest performance is 80.81% with single feature (2000) using SVM classifier and similarly, performance DT is 73.94% and NB has 68%. Results show that the approach performs rather well when tested with certain epochs and batch sizes; accuracy results are compared to those of NB, SVM, and DT methods with various parameters. However, several aspects still require development, such as the need for a robust stop words list and more precise stemming classifier algorithms while mining roman data.In comparison of stop remove words and unremoved, removed stop words datasets accuracy is better than unremoved stop datasets while datasets 20000 is better performed than other datasets. The accuracy of SVM 80.12 is dominated other classifiers ML NB and DT. The proposed study is that a hybrid feature selection method to reduce high dimensional feature size and increase accuracy. Comparatively analyze the performance of the proposed hybrid feature selection method with individual feature selection methods.

REFERENCES

[1] Kadhim, A. I. J. A. I. R. (2019). Survey on supervised machine learning techniques for automatic text classification. ICRAIE, 52(1), 273-292.

[2] Zhou, X., Gururajan, R., Li, Y., Venkataraman, R., Tao, X., Bargshady, G., Kondalsamy Chennakesavan, S. (2020). A survey on text classification and its applications. Web Intelligence, 18(3), 205-216.

[3] Rafae, A., Qayyum, A., M M., Karim, A., Sajjad, H., & Kamiran, F. (2015). An unsupervised method for discovering lexical variations in Roman Urdu informal text. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 12th September, (pp. 823–828).

[4] Esmaeilzadeh, A., & Taghva, K. (2022). Text classification using neural network language model (nnlm) and bert: An empirical comparison. In Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference,07 August, USA, (pp. 175-189). Springer

[5] Ameer, I., Sidorov, G., Gomez-Adorno, H., & Nawab, R. M. A. J. I. A. (2022). Multi-Label Emotion Classification on Code-Mixed Text: Data and Methods. 10(3), 8779-8789

[6] Luo, X. J. A. E. J. (2021). Efficient english text classification using selected machine learning techniques. 60(3), 3401-3409.

[7] Sharf, Z., & Rahman, S. U. (2018). Performing natural language processing on roman urdu datasets. International Journal of Computer Science and Network Security, 18(1), 141-148.

[8] Qutab, I., Malik, K. I., Arooj (2022). Sentiment Classification Using Multinomial Logistic Regression on Roman Urdu Text. 4(2), 223-335.

[9] Tehreem, T. J. a. p. a. (2021). Sentiment analysis for youtube comments in roman urdu.

[1[10] Ullah, A., Khan, S. N., & Nawi, N. M. (2023). Review on sentiment analysis for text classification techniques from 2010 to 2021. Multimedia Tools and Applications, 82(6), 8137-8193.

[11] Sebai, D., & Shah, A. U. (2023). Semantic-oriented learning-based image compression by Only-Train-Once quantized autoencoders. Signal, Image and Video Processing, 17(1), 285-293.

[12] Mustafa, R., Rai, S., Ullah, U., & Naz, M. S. (2023). Summary in General Summary of an Overview of Opinion Mining. Journal of Advancement in Computing, 1(1), 9-13.

[13] Alam, T., Gupta, R., Qamar, S., & shah A. (2022). Recent applications of Artificial Intelligence for Sustainable Development in smart cities. In Recent Innovations in Artificial Intelligence and Smart Applications (pp. 135-154). Cham: Springer International Publishing.

[14] Aznaoui, H., Raghay, S., Ullah, A., & Khan, M. H. (2021). Energy efficient strategy for WSN technology using modified HGAF technique. iJOE, 17(06), 5.

[15] Ouhame, S., & Hadi, Y. (2020). A Hybrid Grey Wolf Optimizer and Artificial Bee Colony Algorithm Used for Improvement in Resource Allocation System for Cloud Technology. International Journal of Online & Biomedical Engineering, 16(14).

[16] Branch, S. R., & Rey, S. (2018). Providing a load balancing method based on dragonfly optimization algorithm for resource allocation in cloud computing. International Journal of Networked and Distributed Computing, 6(1), 35-42.

JAC 2023, https://doi.org/10.36755/jac.v1i1.47