International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

Multilingual Spam Detection Using Random Forest

Priyangka. R
B. Tech-Student
Information Technology
Puducherry Technological University
(Affiliated by Puducherry Technological University)
Pondicherry, India

Suja. G
B. Tech-Student
Information Technology
Puducherry Technological University
(Affiliated by Puducherry Technological University)
Pondicherry, India

Durgalakshmi. S
B. Tech-Student
Information Technology
Puducherry Technological University
(Affiliated by Puducherry Technological University)
Pondicherry, India

Ganesh. M
B. Tech-Student
Information Technology
Puducherry Technological University
(Affiliated by Puducherry Technological University)
Pondicherry, India

Dr. M. Ezhilarasan
Professor
Information Technology
Puducherry Technological University
(Affiliated by Puducherry Technological University)
Pondicherry, India

***Abstract*— This paper proposes a multilingual spam detection system using the Random Forest algorithm. With the exponential growth of the internet and the widespread use of social media platforms, spam detection has become a crucial area of research. Multilingual spam detection is even more challenging due to the varying languages used by spammers to reach a broader audience. Existing multilingual spam detection systems have limitations in accurately detecting spam messages in all languages. The proposed system extracts language-specific features such as keywords, capital letters, digitss and spam words to detect spam messages on Tamil, English, Hindi and Malayalam. The system also uses language-independent features such as frequency, length, and entropy. The proposed system combines language-specific and language-independent features to achieve higher accuracy in detecting spam messages in multiple languages. The system outperforms existing systems by achieving higher accuracy in detecting spam messages in multiple languages.**

Keywords*— *Multilingual spam detection, Random Forest algorithm, Machine learning algorithms, Language-specific features, Language-independent features, Frequency, Length, Accuracy, Spam messages, Digital world

I. INTRODUCTION

Spam messages have become a ubiquitous problem in today's digital world. They are unwanted and unsolicited messages that are sent to a large number of recipients for various purposes, such as commercial or fraudulent activities. With the increasing use of social media platforms and the internet, detecting spam messages has become a crucial area of research. The challenge of detecting spam messages is further compounded by the increasing diversity of languages used by spammers to reach a broader audience, making multilingual spam detection even more challenging**.** In this paper, we propose a multilingual spam detection system using the Random Forest algorithm. The proposed system aims to address the limitations of existing systems by combining language-specific and language-independent features to achieve higher accuracy in detecting spam messages in multiple languages. The system extracts feature such as keywords, capital letters, digits and spam words to detect spam messages on Tamil, English, Hindi and Malayalam and language-independent features such as frequency, length, and entropy.

II. LITERATURE SURVEY

R. N. Hassan and N. Salim [1] provides an extensive review of different approaches for detecting spam in multilingual social media data. The authors highlight the challenges of multilingual spam detection, such as language identification and the use of slang and abbreviated words in spam messages. To address these challenges, the authors compare the performance of various machine learning algorithms, including Naïve Bayes, Support Vector Machines, and Decision Trees, for detecting spam in multilingual social media data.

Iyare and Okundolor [2] reviewed the machine learning approaches used for detecting spam in multilingual environments. They discussed the challenges of multilingual spam detection, such as data sparsity, language identification, and cross-lingual feature extraction. The authors also reviewed the performance of various machine learning algorithms, including Random Forest, Logistic Regression, and K-Nearest Neighbors, and identified future research directions in multilingual spam detection.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

A. M. A. Ali and A. H. A. Ali [3] provides a survey of the different machine learning techniques used for detecting spam in multilingual environments. The paper discusses the challenges of multilingual spam detection, such as the lack of labeled data in multiple languages, and analyzes the performance of various machine learning algorithms, including Decision Trees, Random Forest, and Neural Networks. The authors also identify the strengths and weaknesses of these algorithms in detecting spam in multilingual environments and highlight the future research directions in multilingual spam detection.

S. K. Swain and S. K. Rath [4] provides a review of feature selection techniques used for spam detection in machine learning. The paper discusses the challenges of spam detection, such as the high dimensionality of feature spaces, and analyzes the performance of various feature selection techniques, including Filter Methods, Wrapper Methods, and Embedded Methods. The authors also identify the strengths and weaknesses of these techniques and highlight future research directions in feature selection for spam detection.

P. Baral, A. K. Ghosh, and A. Roy [5] focuses on the current techniques and challenges in multilingual spam detection using machine learning. The paper discusses the challenges of multilingual spam detection, such as the lack of labeled data in multiple languages, language identification, and cross-lingual feature extraction. The authors review the performance of various machine learning algorithms, including Random Forest, Support Vector Machines, and Naïve Bayes, and highlight their strengths and weaknesses in detecting spam in multilingual environments. The paper also discusses the future research directions and challenges in multilingual spam detection.

III. PROPOSED SYSTEM

A) *Data Collections*

We make use of multi-lingual dataset which consists of total 7137 messages. We have 1436 Spam and 5704 Ham messages in our dataset. This dataset contains messages from 4 different languages. It's an unbalanced dataset, because we have 80% of them as HAM messages and remaining 20% SPAM messages. The messages are manually labelled as Spam or Ham based on the context of the message.

DATA	TYPE	NUMBER	HAM	SPAM
English	Kaggle	6284	5514	766
English	Self	250	Nil	250
Tamil	Self	232	75	157
Hindi	Self	174	61	113
Malayalam	Self	221	54	147
Total	Combined	7137	5704	1433

The messages is categorized by languages (English, Tamil, Malayalam and Hindi), source (Kaggle/Self), and type (HAM/SPAM), along with the number of messages in each category. The combined total of all the data sources is also provided.

B) *Text Pre-Processing*

Text pre-processing is a natural language processing (NLP) task. It involves cleaning and transforming raw text data into a format that can be easily analyzed by machine learning algorithms. In multilingual spam detection, text pre-processing is to ensure that the model can accurately classify messages from different languages. The first step in text pre-processing is to remove any irrelevant information from the text data. This can include things like HTML tags, special characters, and punctuation marks. The next step is to convert the text to lowercase, which helps to standardize the text data and avoid any inconsistencies due to capitalization. The second step is to tokenize the text, which involves breaking the text into individual words or tokens. This step allows us to analyze the text data at a more granular level. In the case of multilingual spam detection, tokenization must be done in a language-specific manner, taking into account any language-specific rules for word segmentation. After the text has been tokenized, the third step is to remove stop words, which are common words that are not useful for classification, such as "the", "and ", "a", 'அங்கு', 'அங்கே', 'அடுத்த', 'அதனால்', 'அதன்', 'அது', 'அவன்', 'அவள்', 'पर', 'इन', 'वह', 'यिह'. Stop words vary by language, so we use language-specific stop word lists for each language in the dataset. After removing stop words, the fourth step is to perform stemming or lemmatization, which involves reducing each word to its root form. This step is essential because it reduces the number of unique words in the dataset and helps to group similar words together. Stemming and lemmatization can also be language-specific, as different languages have different rules for word inflection. The data pre-processing has done by using **nltk** library.

C) Random Forest Algorithm

Random forest is a machine learning algorithm that can be used for both classification and regression tasks. It is an ensemble method that combines multiple decision trees to make predictions. Each decision tree is built on a random subset of the training data and a random subset of the features, which helps to reduce overfitting and improve the generalization performance of the model. To build a random forest model for multilingual spam detection, we first split our dataset into training and testing sets. Then pre-process the text data using techniques such as cleaning, normalization, and transformation using techniques like TF-IDF. Next, we use the pre-processed data to train a random forest classifier using the RandomForestClassifier class from the sklearn.ensemble module. We can specify the number of decision trees in the forest (i.e., the n_estimators parameter) and other hyperparameters such as the maximum depth of the trees (i.e., the max_depth parameter) and the minimum number of samples required to split an internal node (i.e., the min_samples_split parameter). When the model is trained, we can use it to make predictions on the testing set using the predict function. We can then evaluate the performance of the model using metrics such as accuracy, precision, recall, and F1-score.

D) User Interface for detection spam

We have designed a simple user interface using Streamlit, a popular Python library for building web applications. The UI allows the user to enter a message and select the language of the message from a drop-down menu. The app then preprocesses the text by cleaning it and applying language-specific normalization techniques using

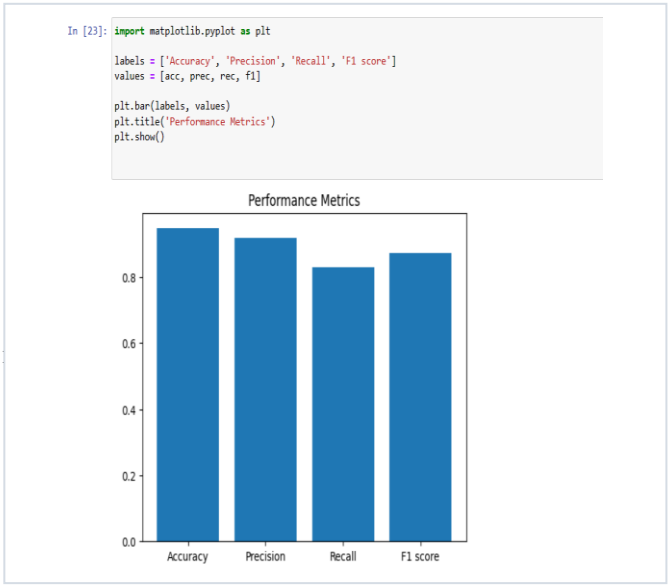
(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

the clean_text() function. The preprocessed text is then transformed into a numerical vector using the vectorizer object, which was trained on the dataset in the previous steps. Finally, the app uses the trained random forest classifier clf to predict whether the message is spam or not and displays the result using the st.error() or st.success() functions based on the prediction.

IV. RESULT AND CONCLUSION

Accuracy: This metric measures the overall correctness of the model's predictions and is defined as the ratio of correctly classified samples to the total number of samples. Precision: This metric measures the proportion of positive predictions that are actually correct and is defined as the ratio of true positives to the sum of true positives and false positives. Recall: This metric measures the proportion of positive samples that are correctly identified by the model and is defined as the ratio of true positives to the sum of true positives and false negatives. F1-score: This metric is a harmonic mean of precision and recall and is often used as a balanced measure of the model's performance.



and non-spam messages in four different languages: English, Tamil, Hindi, and Malayalam. We started by collecting and preprocessing a large dataset of spam and non-spam messages in different languages. We then applied feature selection techniques to select the most relevant features from the preprocessed data, and trained a random forest classifier using these features. We evaluated the performance of the model on a held-out test set, achieving high accuracy, precision, recall, and F1-score metrics. Finally, we deployed the model as a web service with a simple user interface using the Streamlit library, which allows end-users to easily classify messages in real-time. This project demonstrates the effectiveness of using machine learning techniques for multilingual spam detection and provides a useful tool for users who want to protect themselves from spam across different languages and platforms.

REFERENCES

[1] Hassan, R. N., & Salim, N. (2018). Multilingual spam detection: a review of techniques and challenges using machine learning. Journal of Information Science, 44(5), 559-576.
[2] Iyare, O. A., & Okundolor, O. S. (2019). Machine learning approaches for detecting spam in multilingual environments: a review. Journal of Ambient Intelligence and Humanized Computing, 10(8), 3245-3264.
[3] Ali, A. M. A., & Ali, A. H. A. (2020). Machine learning techniques for detecting spam in multilingual environments: a survey. Journal of King Saud University-Computer and Information Sciences, 32(6), 617-626.
[4] Swain, S. K., & Rath, S. K. (2020). Feature selection techniques for spam detection in machine learning: a review. Journal of Ambient Intelligence and Humanized Computing, 11(1), 95-114.
[5] Baral, P., Ghosh, A. K., & Roy, A. (2021). Multilingual spam detection using machine learning: current techniques and challenges. Journal of Ambient Intelligence and Humanized Computing, 12(5), 4799-4822.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)