Social Media Fake News Classification Using Machine Learning Algorithm

Girma Yohannis Bade, Olga Kolesnikova , Grigori Sidorov,
José Luis Oropeza
Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC),
Mexico City, Mexico
Correspondence : girme2005@gmail.com

Abstract

The rise of social media has facilitated easier communication, information sharing, and current affairs updates. However, the prevalence of misleading and deceptive content, commonly referred to as fake news, poses a significant challenge. This paper focuses on the classification of fake news in Malayalam, a Dravidian language, utilizing natural language processing (NLP) techniques. To develop a model, we employed a random forest machine learning method on a dataset provided by a shared task(DravidianLangTech@EACL 2024)[1]. When evaluated by the separate test dataset, our developed model achieved a 0.71 macro F1 measure.

1 Introduction

The rise in usage of social media sites has made it easier for people to communicate with one another. Social media users can converse, share information, and keep up with current affairs. However, a lot of the current material that has surfaced on social media is misleading and, in certain situations, is an attempt to deceive users. This kind of stuff is frequently referred to as false news. Any incorrect or deceptive information that purports to be newsworthy is referred to as fake news (Subramanian et al., 2023; Yigezu et al., 2023e). Customers and retailers have both been impacted by fake reviews. Furthermore, in 2016, the issue of false news came to light, particularly in the wake of the previous US presidential election. Since both fake reviews and fake news involve producing and disseminating incorrect information or opinions, they are closely related phenomena (Ahmed et al., 2018).

The rapid expansion of internet news sources has made it exceedingly challenging to distinguish between fraudulent and true information (Bade, 2021; Yigezu et al., 2023d). Because of this, fake news is now widely spread and very difficult to evaluate and confirm. Discussing the subject case by case indeed presents a significant difficulty to both the public and the government (Yigezu et al., 2023b). For this reason, a system for fact-checking rumors and statements needs to be implemented, especially for those that receive thousands of views and likes before being disproved and disputed by reliable sources. Similarly, humans are incapable of identifying all of these false reports. Machine learning classifiers are therefore required to automatically identify these false news items (Ahmed et al., 2021). Even though a range of machine-learning methods have been employed to identify and categorize false information, these methods have limitations in terms of accuracy (Fayaz et al., 2022; Yigezu et al., 2023c). Several factors, including imbalanced datasets, ineffective parameter tuning, and bad feature selection, might be blamed for the low accuracy (Hakak et al., 2021).

Although numerous studies are in charge of taking fake news countermeasures in English, languages other than English are also taking advantage of natural language processing(NLP) to mitigate the growing challenge of their language aspect. In this regard, the Dravidian Language is gaining popularity in leveraging the NLP task including fake news classification, sentiment analysis, hate speech detection, and stress identification in general. In particular, this study focuses on the social media fake news classification in the Malayalam language which is one of the Dravidian languages. The golden standard dataset was offered to us by the task organizer as a shared task(DravidianLangTech@EACL 2024)(Subramanian et al., 2023). Via Codalab, we are given the datasets containing YouTube comments in the Malayalam language annotated for fake news detection. In this regard, we have used a random forest machine-learning model and developed a Malayalam language fake news classifier as intended in

[1] https://codalab.lisn.upsaclay.fr/competitions/16055

the shared task competition. The rest of the paper details the related works, system or methodology descriptions,the results generated, and the recommendations for future works.

2 Related Works

Several machine and deep learning algorithms have been used in different articles to detect and analyze bogus news on social media sites (Yigezu et al., 2023a). According to the article proposed in (Hakak et al., 2021), the ensemble classification had a higher accuracy in detecting fake news when compared to the state-of-the-art. Important features are collected from the fake news datasets by the suggested model, and these features are then classified using an ensemble model made up of three well-known machine learning models: Decision Tree, Random Forest, and Extra Tree Classifier.

A study (Kareem and Awan, 2019)was conducted to identify fake news in Pakistani media, as it is a challenging process to classify. The popular news website scrape was served as the source of the dataset for this investigation. 344 news stories that have been manually classified as True or Fake make up the generated corpus. It has employed seven distinct supervised Machine Learning (ML) classification methods for the result comparison, in addition to two feature extraction strategies (Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF). K-Nearest Neighbors (KNN) provided 70% accuracy, whereas logistic regression produced 69% accuracy, making it the highest-performing classifier.

To categorize and identify the bogus news on multiple social media sites, algorithms such as Naive Bayes, Support Vector Machines, Passive Aggressive Classifier, Random Forest, BERT, LSTM, and Logistic Regression were employed. The work is based on an ISOT dataset consisting of 44,898 news items that were collected from multiple sources and pre-processed using count vectorizer and TF-IDF. The SVM is deemed to be among the most accurate algorithms for spotting false information on social media (Kumar et al., 2023).

Deep learning's growth is greatly aided by its widespread use. Three types of neural network techniques: convolution filtering-based neural network approaches, sequential analysis-based neural network approaches, and attention mechanism-based neural network approaches can be distinguished from the model structure of the research work that now exist (Tash et al., 2023; Bade and Afaro, 2018). But nearly all of them were created with scenarios in a single language in mind, without taking into account mixed-lingual contexts. The article(Guo et al., 2023) proposed a unique false news detection model for the context of mixed languages through a multiscale transformer to completely capture the semantic information of the text, bridging this gap by extending the basic pretraining language processing model transformer into the multiscale format. Additionally, it improved accuracy in mixed language contexts by roughly 2%–10% as compared to baseline models that are frequently utilized.

The multi-modal transformer employing two-level visual characteristics (MTTV) was suggested as an alternative to text for the identification of false news(Wang et al., 2023). Firstly, news texts and photos are universally modeled as sequences that may be processed by a transformer. To enhance the use of news photographs, two-level visual features global feature and entity level feature are employed. Second, it creates a multi-modal transformer that expands the transformer paradigm for natural language processing, enabling complete interaction and semantic relationship capturing between multi-modal data. Furthermore, it suggested a scalable classifier to address the issue of class imbalance and enhance the classification balance of fine-grained false news detection. Comprehensive tests on two publicly available datasets show that our approach outperformed the state-of-the-art techniques by a significant margin.

Even though the majority of studies have been conducted on binary fake news classification tasks, the work (Shushkevich et al., 2023) addresses a more realistic scenario by assessing a corpus with unknown themes through multiclass classification, encompassing true, false, partially false, and other categories. Three BERT-based models: SBERT, RoBERTa, and mBERT are explored; artificial data generated by ChatGPT is used to improve results for class balance; and a two-step binary classification process is employed to improve outcomes. The testing results indicate that, while it is an optimal performance in comparison to past accomplishments, it still requires refinement to remain at the forefront of technology.

3 System Description

In this section, we offer thorough information regarding the dataset and the details of experimental tools. Moreover, it dives into the format of datasets, preprocessing, and experimental environmental tools to develop the proposed model.

3.1 Datasets

In the real world, the problems are always existing until the solutions are investigated. To investigate solutions for computational linguistic challenges, the availability of data is crucial (Bade, 2021; Bade and Afaro, 2018). The dataset for this particular task was provided on Codalab by the Shared_task (DravidianLangTech@EACL 2024) organizer (B et al., 2024). There are two subtasks provided to participate in this competition.Task_1 is to classify the given social media dataset that was prepared for this aim into fake or original and Task_2 is the Fake News Detection from Malayalam News(False, Half True, Mostly False, Partly False, and Mostly True). Among the offered tasks, we participated only in the first task(Task_1) based on our interests (Subramanian et al., 2023, 2024). The dataset is arranged in three different lists training, development, and test(without label) set. The training and development data sets are made available when we register for the competition on the Codalab and the test set was released when ten days left for the run submission deadline.

Table 1: The overview of dataset offer

No	Text	Label	Dataset	size
1	Masha Allah	Fake	Training	3257
2	à´¬à´¿à´œàµ	Original		
3	Well planned... China.	Fake	Development	815
4	à´ªà´°à´¨à´¾à´±à	Original		
5	Shame for entire Woman	—	Test	1019
6	à´ªàµà´°à´µà´¾à´¸à´¿	—		

The Table1 shows us three things:1) how the sample of the dataset and its class variable look like,2) how the code-mixed writing is taking place, and 3) how the test dataset is given without the class label. In the case of training and development datasets, the class variable or dependent variable is given with their features, however, in the case of test dates there is no class label given because it is expected that the model would predict its class label how it was in training data.

3.2 Preprocessing

Preprocessing is the process of preparing raw data for machine learning algorithms by cleaning, converting, and organizing the data and rendering it to the machine. It is the vital stage that fills in the gaps between raw data and useful insights because raw data is rarely in an ideal state (Bade and Seid, 2018). During the data preparation phase of machine learning tasks, there are typical or standard activities that we should use. The following are some among others.

Importing dependency libraries:- There are two libraries that we must always bring in. A library containing mathematical functions is called NumPy and the library used to import and manage the 'CSV' data sets is called Pandas.

Loading the data set:- In most cases, data sets are offered in a csv format. Tabular data is stored in plain text in a CSV file. In a file, every line represents a data record. To read a local CSV file as a data frame, the pandas library's (read_csv) function was utilized.

Handling Missing Data:- In real-world datasets, handling missing data is a prevalent difficulty. Preprocessing methods like imputation and the removal of missing data or null values ensure that the model is fed accurate and comprehensive data. For a variety of reasons, data may be missing, and it must be handled to prevent our machine-learning model from performing worse. In addition, we used "raw['category'].fillna(0, inplace=True)" to handle empty strings of class labels.

Data Cleaning:- is finding and fixing inaccuracies or flaws in the data.

Handling Outliers:- Anomalies that drastically depart from the average might cause distortions in the process of learning. Preprocessing techniques such as transformation or scaling lessen the negative effects of outliers on model performance.

Data Encoding:- Since machine learning algorithms usually operate on numerical data, it is necessary to properly encode our text inputs in numerical equivalent. To do so we have specifically used the TF-IDF text vectorization technique. It preserves the semantics and instance positions in addition to converting the provided text into a numeric representation. However, in the case of converting 'class label', we used the "to_numeric()" function known as "raw['category'] = pd.to_numeric(raw['category'], errors='coerce')"

3.3 Model Selection and Experimentation

The selected machine learning model for this study is a random forest. This is because several decision trees are combined in a random forest, an ensemble learning technique to produce predictions that are more reliable and accurate (Tonja et al., 2022). In a random forest, every decision tree is trained using a random subset of features and a random subset of the data (bootstrap samples). The diversity among the individual trees is increased and overfitting is lessened by this randomization (Yigezu et al., 2023b). During prediction, the ultimate result is established by combining all of the trees' predictions, either by average (for regression) or by majority voting (for classification). The capacity to manage complicated datasets, high-dimensional data, and non-linear interactions is a well-known feature of random forests. They are also frequently utilized in machine learning applications and are less prone to overfitting than a single decision tree.

Experimental setup:- This section discusses the details of the developmental tool and the dependency libraries we used. For this research, we used Jupyter Notebook3 which is the Integrated Development Environment(IDE) of Python. After the tool setup was finished, we imported the four basic dependency libraries known as pandas, TfidfVectorizer, RandomForest, Joblib. Among those, the first three(pandas, TfidfVectorizer, RandomForest) are found in the Sklearn module. Pandas is used to read CSV files from the local drive to a Python-run environment, TfidfVectorizer is for converting text data inputs into a numerical representation, and Random-Forest is the principal algorithm to train the input data based on the predefined class. Finally, joblib is a standalone module for saving the trained model for later use.

4 Result and Discussion

The Random Forest based developed model classified the test dataset into two classes as they present in training data.

Table 2: Class label test data overview of manually or by annotator classified and machine or our model classified classification distribution.

Class	Manually classified	**Machine classified**
Original	512	**628**
Fake	507	**391**
Total	1019	**1019**

From the Table 2 can understand that 116 instances of the class 'Fake' were incorrectly classified into the 'Original ' class category. Here 'manually classified' in column_2 refers to the answer key that has been released after the competition is over, whereas the 'machine classified' in column _3 refers to our model. The test output was sent to the organizers to test the performance of the model in macro F1 metrics. According to the result published, our model has achieved 0.71 macro F1. It is also a promising result to the other code mixed social media posts.

5 Conclusion

In this particular task, we have developed a classifier to classify social media posts into two binary classes, fake and original. The model has used the Random Forest algorithm method. The numeric features are extracted using TF-IDF techniques. The newly developed model has been evaluated with the new unseen test dataset and the results also promising for other code mixed languages.

6 Future work

Since social media posts that classify fake news are very critical, the jobs ought to be transferred to other various languages. Furthermore, by offering additional algorithms for the languages utilized here and expanding the number of dataset sizes, the performance of the suggested model in this study should be enhanced.

Acknowledgements

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20232138, 20231567, and 20232080 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

Limitation and Ethics Statement

Finding words outside of one's lexicon or linguistic occurrences that were not taken into consideration during preprocessing are limitations. Code-mixing can bring linguistic variances that the current language processing algorithms may not be able to handle well enough, which could result in incorrect classifications. Future studies could improve the model's performance and generalization capacities by addressing these linguistic issues. Notably, out of all the participating systems, our method achieved the 12th rank in the shared job. Our model performs well in classifying Fake News comments in code-mixed text, even in the face of competition from other participants and obstacles in the competition. Furthermore, our work obeyed the computational ethics[2].

References

Alim Al Ayub Ahmed, Ayman Aljabouh, Praveen Kumar Donepudi, and Myung Suh Choi. 2021. Detecting Fake News using Machine Learning: A systematic literature review. *arXiv preprint arXiv:2102.04458*.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9.

Premjth B, Bharathi Raja, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Prashanth Karnati, Sai Rishith Reddy Mangamuru, and Janakiram Chandu. 2024. Findings of the Shared Task on Hate and Offensive Language Detection in Telugu Codemixed Text (HOLD-Telugu). In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Girma Yohannis Bade. 2021. Natural Language Processing and Its Challenges on Omotic Language Group of Ethiopia. *Journal of Computer Science Research*, 3(4):26–30.

Girma Yohannis Bade and Akalu Assefa Afaro. 2018. Object Oriented Software Development for Artificial Intelligence. *American Journal of Software Engineering and Applications*, 7(2):22–24.

Girma Yohannis Bade and Hussien Seid. 2018. Development of Longest-Match Based Stemmer for Texts of Wolaita Language. *vol*, 4:79–83.

Muhammad Fayaz, Atif Khan, Muhammad Bilal, and Sana Ullah Khan. 2022. Machine learning for fake news classification with optimal feature selection. *Soft Computing*, 26(16):7763–7771.

Zhiwei Guo, Qin Zhang, Feng Ding, Xiaogang Zhu, and Keping Yu. 2023. A Novel Fake News Detection Model for Context of Mixed Languages Through Multiscale Transformer. *IEEE Transactions on Computational Social Systems*, pages 1–11.

Saqib Hakak, Mamoun Alazab, Suleman Khan, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, and Wazir Zada Khan. 2021. An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems*, 117:47–58.

Irfan Kareem and Shahid Mahmood Awan. 2019. Pakistani Media Fake News Classification using Machine Learning Classifiers. In *2019 International Conference on Innovative Computing (ICIC)*, pages 1–6.

Ashish Kumar, M Izharul Hasan Ansari, and Kshatrapal Singh. 2023. A Fake News Classification and Identification Model Based on Machine Learning Approach. In *Information and Communication Technology for Competitive Strategies (ICTCS 2022) Intelligent Strategies for ICT*, pages 473–484. Springer.

Elena Shushkevich, Mikhail Alexandrov, and John Cardiff. 2023. Improving Multiclass Classification of Fake News Using BERT-Based Models and ChatGPT-Augmented Data. *Inventions*, 8(5):112.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Premjith B, Sandhiya Raja, Vanaja, Mithunajha S, Devika K, Hariprasath S.B, Haripriya B, and Vigneshwar E. 2024. Overview of the Second Shared Task on Fake News Detection in Dravidian Languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the Shared Task on Fake News Detection from Social Media Text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Moein Tash, Jesus Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. LIDOMA@ DravidianLangTech: Convolutional Neural Networks for Studying Correlation Between Lexical Features and Sentiment Polarity in Tamil and Tulu Languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 180–185.

Atnafu Lambebo Tonja, Mesay Gemeda Yigezu, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbuk. 2022. Transformer-based model for word level language identification in code-mixed kannada-english texts. *arXiv preprint arXiv:2211.14459*.

[2] https://www.aclweb.org/portal/content/acl-code-ethics

Bin Wang, Yong Feng, Xian-cai Xiong, Yong-heng Wang, and Bao-hua Qiang. 2023. Multi-modal transformer using two-level visual features for fake news detection. *Applied Intelligence*, 53(9):10429–10443.

Mesay Gemeda Yigezu, Girma Yohannis Bade, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023a. Multilingual Hope Speech Detection using Machine Learning.

Mesay Gemeda Yigezu, Selam Kanta, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023b. Habesha@ DravidianLangTech: Abusive Comment Detection using Deep Learning Approach. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 244–249.

Mesay Gemeda Yigezu, Tadesse Kebede, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023c. Habesha@ DravidianLangTech: Utilizing Deep and Transfer Learning Approaches for Sentiment Analysis. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 239–243.

Mesay Gemeda Yigezu, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023d. Transformer-Based Hate Speech Detection for Multi-Class and Multi-Label Classification.

Mesay Gemeda Yigezu, Moges Ahmed Mehamed, Olga Kolesnikova, Tadesse Kebede Guge, Alexander Gelbukh, and Grigori Sidorov. 2023e. Evaluating the Effectiveness of Hybrid Features in Fake News Detection on Social Media. In *2023 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 171–175. IEEE.