Overview of the shared task on Fake News Detection from Social Media Text

Malliga Subramanian[1], **Bharathi Raja Chakravarthi**[2],
Kogilavani Shanmugavadivel[1], **Santhiya Pandiyan**[1],
Prasanna Kumar Kumaresan[2], **Balasubramanian Palani**[3], **Muskaan Singh**[4],
Sandhiya Raja[1], **Vanaja**[1], **Mithunajha S**[1]
[1]Kongu Engineering College, Tamil Nadu, India
[2]Insight SFI Research Centre for Data Analytics, School of Computer Science,
University of Galway, Ireland
[3]Indian Institute of Information Technology, Kottayam, India
[4]Ulster University, UK
mallinishanth72@gmail.com

Abstract

The rapid proliferation of fake news has emerged as a significant challenge to the credibility of online information, largely due to the swift dissemination of content on social media platforms. This article provides a concise summary of the findings of the shared task on "Fake News Detection in Dravidian Languages[1] - DravidianLangTech@RANLP 2023". The aim of this shared task is to categorize social media posts as either fake or original, specifically focusing on content in Malayalam. The shared task garnered participation from 8 teams who presented their systems. These systems encompassed a spectrum of methodologies including machine learning techniques and transformer-based models like MuRIL, XLMRoBERTa, and Indic BERT. Notably, the XLMRoBERTa-based model demonstrated exceptional performance, achieving a macro F1-score of 0.90.

1 Introduction

Online social network (OSN) platforms such as Twitter, Facebook, WhatsApp, and Instagram are extensively used by millions of users in this modern internet era to publish and spread the news about emergent events to many people more quickly and without any validation or verification. According to the (Pennycook, 2020) the core ideas in cascading news and sensitive information are ingrained in truth notions and communication accuracy theories. According to a recent statistics report given in statistica 2023[2], Facebook has 2.96 billion monthly active users, while the count for Twitter has reached 556 million monthly active users, whereas WhatsApp, as well as Instagram, have more than 2 billion active users 1. A vast amount of news is shared and propagated among socially connected network users without knowing the authenticity of the news during the election campaign, trending events, and pandemic emergencies. Misinformation creators are intentionally flooding falsified and unverified information for various political and commercial purposes. Hence, a significant amount of misinformation and false news has proliferated over OSNs, negatively influencing readers, and causing numerous negative consequences on the economy, politics, and social security. Therefore, fake news detection (FND) is demanding in the current scenario. In (Pennycook et al., 2020) the authors identified primary methods currently available to spot false news and how these methods might be used in various contexts by conducting a systematic literature review. A pertinent example, the difficulties, and the ideal setting in which to use a certain technique are all provided for some approaches.

In general, FND methods can be categorized into two types: social context-based and content-based methods as shown in Figure 1. The former is more concerned with the user engagement data such as comments, reposts, and ratings, while the latter is associated with the article's news content like title, text, image, and video (Shu et al., 2017). The social context-based methods can be further divided into two categories: propagation structure-based and post-based methods. The propagation structure-based methods concentrate on propagation patterns or trends of fake news on social networks, while post-based methods examine the opinions or emotions expressed by the users in their posts. Due to the unstructured nature of the data, these two types of social-context techniques face

[1]https://codalab.lisn.upsaclay.fr/competitions/11176
[2]https://shorturl.at/dloDF

Third Workshop on Speech and Language Technologies for Dravidian Languages,
pages 59–63, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_008

the following challenges: data collection and analysis, noisy data, and missing data. Hence, the focus of this shared task is on content-based strategy. The Content-based methods are more straightforward and convenient to detect fake news, particularly at an early stage.

Transfer learning-based FND system is introduced in (Palani and Elango, 2023a) According to their perspective, BERT is a bidirectional language model since it considers the context of both a word's left and right sides. In contrast, GPT and ELMo are only trained in the right-to-left context and the left-to-right context, respectively. The local contextual features over space and the global semantic relation features over time are then extracted in the feature representation layer using multichannel CNN and stacked BiLSTM. The model may learn many characteristics from several viewpoints using a multichannel CNN. The model's various channels each extract features from the same input in their own unique ways, producing a more reliable representation(Shanmugavadivel et al., 2022).

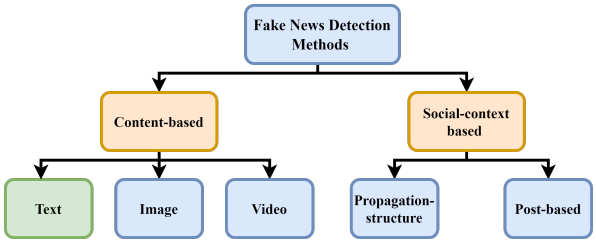

Figure 1: Categorization of FND methods

In (Ahmad et al., 2020), the authors applied machine learning-based ensemble methods with the help of textual properties to distinguish fake news from the original one.

2 Related Work

The researchers used pre-trained language models such as BERT, and RoBERTa for contextual word embedding and then used DL-based models to detect fake news. In (Palani and Elango, 2023b), authors present the DL-based FND framework in which RoBERTa and FFN are used to extract contextual dependent features and to detect fake news respectively. Similar to FND there are numerous works published such as Hope speech detection (Chakravarthi et al., 2022a) and Homophobia, Transphobia Detection (Chakravarthi et al., 2022b) in social media posts. The author in (Chakravarthi, 2022) employs a DL-based hope speech detection model in which T5-sentence and Indic-BERT are used for word embedding to capture the contextual relationship among words. Then the contextual features are sent as input to CNN to detect the hope speech comments(Subramanian et al., 2022). The proposed model's performance is evaluated on a multilingual dataset named HopeEDI which is introduced in the shared task 2021 (Chakravarthi, 2020).

Dhivya (Chinnappa, 2021) proposed a two-stage hope detection process in which the language detector identifies the language of the model, and the hope detector classifies the text into hope speech, non-hope speech, or not lang. The various pre-trained language models and DL-based models are proposed, and their performance is evaluated on three hope speech datasets in English, Tamil, and Malayalam.

3 Task Description

The task aims to identify fake news from the posts or comments in Dravidian Languages such as Malayalam which is collected from the YouTube OSN. Each comment/post is annotated at the comment/post level and assigned the class labels fake or real.

4 Dataset Description

The dataset is balanced since the number of samples in fake and real classes is almost nearer. The dataset contains a total of 5,091 comments of which 2,512 are fake news and 2,579 are real news. The dataset is split into training, validation, and testing. The detail of the dataset is shown in Table 1.

Table 1: Summary of the dataset

Dataset	Fake News	Real News	Total
Training	1,599	1,658	3,257
Validation	406	409	815
Testing	507	512	1,019
Total	2,512	2,579	5,091

5 Methodology

In this shared task, there are eight teams actively participated and implemented their models. They evaluated their model's performance on our fake news dataset.

DeepBlueAI (Luo and Wang, 2023): Team DeepBlueAI used a pre-trained language model

such as XLM-RoBERTa to identify fake news. The authors employed the XLM-RoBERTa model to extract the context-aware features from the textual news. Then the contextual feature vector is sent as input to the fully connected layer with softmax to classify fake or real news. Their model achieves an F1-score of 0.90 in this task.

AbhiPaw (Bala and Krishnamurthy, 2023): Team AbhiPaw_ ABHi presented a Multilingual Representations for Indian Languages (MuRIL) for FND. The F1-score of 0.87 is achieved with their proposed model.

NITK-IT-NLP (R L and Kumar M, 2023): Team NITK-IT-NLP used a multilingual version of the transformer-based MuRIL model for developing the FND system. They also introduced focal loss as the loss function while training the model. The model achieves the F1-score of 0.87 for FND.

NLPT (Raja et al., 2023): Team NLPT_Malayalam employed a pre-trained language model called XLM-RoBERTa for FND. The proposed model achieves an F1-score of 0.87 which is better than the ML-based models. The reasons for the improvement are a self-attention mechanism of transformers, a byte-level BPE, and a dynamic masking pattern during training.

MUCS (Sharal Coelho and Shashirekha, 2023): Team MUCS proposed TF-IDF to convert words into vectors based on the occurrence of the words. The extracted features of TF-IDF are passed as input to the different ML-based classifiers to predict the given news as fake or real. The model achieves the F1-score of 0.83 for FND.

ML_AI_IIITRanchi (Kumari et al., 2023): Team ML_AI_IIITRanchi proposed an ensemble ML-based FND system which uses Bag-of-words and Indic BERT for the textual features extraction. Then, ensemble ML-based classifiers such as Random Forest (RF) and AdaBoost are employed to detect the fake news. The F1-score of 0.78 is achieved with their proposed model for FND. DLRG_RR: Team DLRG_RR presents the ML-based FND system in which TF-IDF is used to transform the words into vectors and Passive Aggressive Classifier (PAC) is adopted for FND. Their model achieves the F1-score of 0.73.

NLP_SSN_CSE (Balaji et al., 2023): Team NLP_SSN_CSE employed various pre-trained transformer-based language models, such as BERT, ALBERT, and XLNET for FND. These models are effective in extracting the contextual relationships within the text which lead to improved accuracy in FND tasks for the Malayalam language. Self-attention mechanism captures the most relevant features from the text to detect fake news. The precision, recall, and F1 measure are around 0.75, indicating a balanced performance in identifying both real and fake news. The accuracy of 0.75 suggests the model's ability to make correct predictions overall.

6 Results and Discussions

The performance assessment of the proposed models by the participating teams was conducted using the macro F1 score metric, which is a widely used measure for evaluating classification models. The results of this evaluation are presented in Table 2 below, showcasing the ranking of the teams that took part in the collaborative task. In total, eight teams submitted their respective solutions for evaluation.

Securing the top position, the team "DeepBlueAI" which achieved the first rank demonstrated an impressive macro F1 score of 0.90. Their accomplishment was attributed to the adept utilization of the XLM-RoBERTa model. This pre-trained transformer model was fine-tuned by the team's authors to effectively discern fake comments. This achievement underscores the efficacy of leveraging powerful transformer-based architectures for addressing the task.

Moving on, teams ranked 2nd, 3rd, and 4th garnered identical macro F1 scores of 0.87, highlighting their consistent performance. Among these, the team "AbhiPaw," positioned at the 2nd rank, strategically employed the Multilingual Representations for Indian Languages (MuRIL) model. While MuRIL was primarily designed as a multilingual language model for Indian languages, the team harnessed its potential for the classification task. This innovative approach signifies the adaptability of pre-trained models across diverse downstream applications.

In the lower rankings, the last two teams, securing the 7th and 8th positions, achieved a macro F1 score of 0.73. Their methodology involved the implementation of the Passive Aggressive Classifier (PAC) coupled with the feature weighting technique known as Term Frequency-Inverse Document Frequency (TF-IDF). The teams meticulously experimented with various configurations of the maximum document frequency parameter in the

Table 2: Rank list for Malayalam task

S.No.	Team Name	Macro F1	Rank
1	DeepBlueAI(Luo and Wang, 2023)	0.90	1
2	AbhiPaw(Bala and Krishnamurthy, 2023)	0.87	2
3	NITK-iIT-NLP(R L and Kumar M, 2023)	0.87	2
4	NLPT(Raja et al., 2023)	0.87	2
5	MUCS(Sharal Coelho and Shashirekha, 2023)	0.83	3
6	ML_AI_IIITRanch(Kumari et al., 2023)	0.78	4
7	DLRG_RR	0.73	5
8	NLP_SSN_CSE(Balaji et al., 2023)	0.73	5

Tfidfvectorizer, leading to their placement at the 7th rank. Similarly, the team "NLP_SSN_CSE," positioned at rank 8, employed a similar approach, utilizing an array of pre-trained transformer models like BERT, ALBERT, XLNet, and mBERT. Despite their diverse model ensemble, their performance closely aligned with the team ranked 7th.

Overall, the ranking table provides a comprehensive overview of the distinct strategies and models adopted by each participating team. This evaluation sheds light on the varying degrees of success achieved by exploiting transformer-based models, language-specific architectures, and traditional classification techniques, all contributing to the advancement of fake news detection.

7 Conclusion

This paper presents an overview of the fake news detection shared task conducted at DravidianLangTech-RANLP 2023, specifically focusing on the Malayalam language. The task garnered participation from eight teams, each submitting predictions for evaluation. The methods employed by these teams varied, spanning from traditional TF-IDF vectorizers with machine learning to contemporary pre-trained transformer models for data representation. An analysis of the methodologies revealed a consistent trend: transformer-based methods outperformed other techniques, as indicated by evaluation metrics such as classification accuracy and confusion matrices. This suggests the potency of transformer models in effectively capturing fake news detection performance. In summary, the paper summarizes the DravidianLangTech 2023 fake news detection shared task for Malayalam, highlighting diverse strategies, and underscoring the prevalence of transformer-based methods for improved performance.

Acknowledgments

The author Bharathi Raja Chakravarthi was supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2).

References

Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. 2020. Fake news detection using machine learning ensemble methods. *Complexity*, 2020:1–11.

Abhinaba Bala and Parameswari Krishnamurthy. 2023. Abhipaw @ fake news detection in dravidian languages-dravidianlangtech@ranlp 2023. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Varsha Balaji, Shahul Hameed T, and Bharathi B. 2023. Nlp_ssn_cse@dravidianlangtech-ranlp 2023: Fake news detection in dravidian languages using transformer models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Bharathi Raja Chakravarthi. 2020. Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53.

Bharathi Raja Chakravarthi. 2022. Hope speech detection in youtube comments. *Social Network Analysis and Mining*, 12(1):75.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Cn, John Philip McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumaresan, Rahul Ponnusamy, et al. 2022a. Overview of the shared task on hope speech detection for equality,

diversity, and inclusion. In *Proceedings of the second workshop on language technology for equality, diversity and inclusion*, pages 378–388.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Philip McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022b. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377.

Dhivya Chinnappa. 2021. dhivya-hope-detection@ ltedi-eacl2021: multilingual hope speech detection for code-mixed and transliterated texts. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 73–78.

Kirti Kumari, Shirish Shekhar Jha, Zarikunte Kunal Dayanand, and Praneesh Sharma. 2023. Ml&ai_iiitranchi@dravidianlangtech-ranlp 2023:leveraging transfer learning for the discernment of fake news within the linguistic domain of dravidian language. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Zhipeng Luo and Jiahui Wang. 2023. Deepblueai@dravidianlangtech-ranlp 2023. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Balasubramanian Palani and Sivasankar Elango. 2023a. Bbc-fnd: An ensemble of deep learning framework for textual fake news detection. *Computers and Electrical Engineering*, 110:108866.

Balasubramanian Palani and Sivasankar Elango. 2023b. Ctrl-fnd: content-based transfer learning approach for fake news detection on social media. *International Journal of System Assurance Engineering and Management*, 14(3):903–918.

Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, 31(7):770–780.

McPhetres J. Zhang Y. Lu J. G. Rand D. G. Pennycook, G. 2020. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7):770–780.

Hariharan R L and Anand Kumar M. 2023. Nitk-it-nlp@dravidianlangtech-ranlp 2023: Impact of focal loss on malayalam fake news detection using transformers. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Eduri Raja, Badal Soni, and Sami Kumar Borgohain. 2023. nlpt malayalm@dravidianlangtech : Fake news detection in malayalam using optimized xlm-roberta model. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. An analysis of machine learning models for sentiment analysis of tamil code-mixed data. *Computer Speech & Language*, 76:101407.

Kavya G Sharal Coelho, Asha Hegde and Hosahalli Lakshmaiah Shashirekha. 2023. Mucs@dravidianlangtech2023: Malayalam fake news detection using machine learning approach. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Malliga Subramanian, Ramya Chinnasamy, Prasanna Kumar Kumaresan, Vasanth Palanikumar, Madhoora Mohan, and Kogilavani Shanmugavadivel. 2022. Development of multi-lingual models for detecting hope speech texts from social media comments. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 209–219. Springer.