Multilingual Sentiment Analysis: A Machine Learning Approach with a Focus on Malayalam

Shubham Ojha
Student / VIT Chennai
shubhamojha2109@gmail.com

Rajalakshmi R
Prof / SCOPE, VIT Chennai
rajalakshmi.r@vit.ac.in

Abstract

In this study, Multilingual Sentiment Analysis is a sophisticated computational technology used in this work to identify and categorise a range of emotions present in texts written in different languages. The Malayalam language is given special attention, which helps to illuminate the complex emotional background present in this linguistic area.

This research analyses and discerns between real and fake feelings in social media postings and comments using machine learning algorithms.

It also applies its analytical skills to the classification of news articles, carefully classifying them into false, mainly false, true, half true, and mostly true truthfulness categories. This experiment is evidence of how machine learning can transcend linguistic borders and deepen our comprehension of international digital communications.

1. Introduction

Multilingual sentiment analysis plays a crucial role in computational linguistics by enabling the interpretation of emotions across different languages. This technique is vital for analysing public opinions and social media trends globally. Our project focuses on two Malayalam datasets: one categorizing YouTube comments as original or fake, and another classifying news content into five veracity levels.[11] We explored various machine learning and deep learning approaches, including traditional classifiers (SVM, RF, Logistic Regression, Naive Bayes) and advanced models using BERT classifiers and the Indic BERT tokenizer.[12] This approach aims to enhance the accuracy of sentiment analysis in Malayalam, demonstrating the potential of combining traditional and cutting-edge methods for effective multilingual sentiment analysis.

2. Literature Review

2.1. Salvador Contreras Hernández, María Patricia Tzili Cruz, and José Martín Espínola explored sentiment analysis of COVID-19-related tweets in Mexico using BERT-based models. Their research, focusing on semi-supervised learning with Spanish language models, demonstrated superior precision over multilingual BERT and traditional classifiers. This underscores the effectiveness of language-specific models in capturing public sentiment, offering significant implications for public health decision-making during the pandemic. [1]

2.2. George Manias, Argyro Mavrogiorgou, and Athanasios Kiourtis investigated multilingual sentiment analysis on Twitter, emphasizing the importance of language- and domain-agnostic approaches. Their study assessed

the efficacy of four BERT-based classifiers against a zero-shot classification method. The findings suggest that while BERT-based classifiers are highly effective, zero-shot classification stands out as an innovative and scalable strategy, even though it may not reach the fine-tuned accuracy of its counterparts. [2]

2.3. Amina Amara, Mohamed Ali Hadj Taieb, and Mohamed Ben Aouicha spearheaded a study on COVID-19 trend analysis through the lens of Facebook data across seven languages using Latent Dirichlet Allocation (LDA). The research stands out by leveraging an underexplored platform for multilingual topic modelling, using graph visualization to trace the progression of public interest in the pandemic. The outcomes present unique insights into global sentiment and conversational trends on Facebook, mapping the chronological growth of discourse surrounding COVID-19. [3]

2.4. Rami Mohawesh, Sumbal Maqsood, and Qutaibah Althebyan's research offers a novel semantic solution to multilingual fake news detection by utilizing capsule neural networks. Their framework incorporates word embeddings and n-gram features, significantly enhancing fake news detection across languages. The results show a marked improvement over existing methods, highlighting the capability of capsule neural networks to adeptly manage the intricacies of multilingual text analysis. [4]

2.5. Anjum and Rahul Katarya have developed HateDetector, an advanced technique tailored for detecting hate speech on social media in various languages. Incorporating an improved seagull optimization algorithm and a hybrid diagonal-gated recurrent neural network, their method shows notable enhancements in accuracy, precision, recall, and F-measure. These results position HateDetector as a promising tool for effectively monitoring and curbing hate speech across multiple languages and social media platforms. [5]

2.6. Caio Mello, Gullal S. Cheema, and Gaurish Thakkar examine the construction of Olympic legacy narratives through multilingual sentiment analysis of news articles on the London 2012 and Rio 2016 Olympics. Their methodology intertwines four sentiment analysis (SA) algorithms with explainable AI techniques to scrutinize methodological constraints. While recognizing SA's value in content analysis, the research reveals complexities inherent in multilingual and specialized domains. A blend of leading classifiers paired with clear AI practices promises improvements, and an intriguing utopian versus dystopian narrative dichotomy in Olympic legacy portrayal is disclosed. [6]

2.7. Purbani Kar and Swapan Debbarma tackle the challenge of detecting hate speech and analyzing sentiments in multilingual code-mixed social media texts. They present an enhanced seagull optimization

algorithm coupled with a novel hybrid diagonal gated recurrent neural network. The proposed methodology has shown considerable improvement in precision, recall, and F-measure over traditional approaches, asserting its effectiveness as a powerful tool for hate speech and sentiment analysis in diverse linguistic settings. [7]

2.8. Simran Sidhu, Surinder S. Khurana, Munish Kumar, and Parvinder Singh provide a thorough review of sentiment analysis techniques in Hindi, focusing on negation handling and the development of Hindi SentiWordNet. They explore a range of methodologies, including both semantic and machine learning approaches, and assess tools such as lexicons, stemmers, and morphological analyzers. The paper concludes with a call for more advanced sentiment analysis research for Hindi, highlighting its vast native-speaking community and expanding online footprint, while also pointing out potential areas for future investigation. [8]

2.9. Christian E. Lopez and Caleb Gallemore introduce a sizable multilingual Twitter dataset designed to support research on COVID-19 social discourse. With over 2.2 billion tweets enriched by sentiment analysis and named entity recognition, this resource permits an in-depth examination of discussions surrounding the pandemic. Their conclusion affirms the dataset's importance as a tool for tracking the progression of public sentiment and conversational patterns about COVID-19, enabling diverse analyses of social media data. [9]

2.10. Siroos Rahmani Zardak, Amir Hossein Rasekh, and Mohammad Sadegh Bashkari address the gap in sentiment analysis tools for Persian text by customizing the BERT algorithm for this context. Their work benchmarks the BERT algorithm's performance against former approaches across various datasets, concluding that BERT excels in analysing Persian sentiment, evidenced by superior accuracy and F1 scores. This breakthrough underscores BERT's adaptability and potency in processing Persian language datasets for sentiment analysis. [10]

3. Research Objective

- To implement and evaluate traditional machine learning algorithms (SVM, RF, Logistic Regression, Naive Bayes) for sentiment classification on Malayalam datasets.
- To adapt and fine-tune BERT-based deep learning models for advanced sentiment analysis of Malayalam YouTube comments and news articles.
- To compare the effectiveness of direct BERT classification against models using translated text (via Helsinki-NLP pipeline) and Indic BERT tokenization for Malayalam language processing.
- To determine the accuracy, precision, recall, and F1-score of each model and establish which provides the most reliable sentiment analysis on validation datasets.

- To investigate the efficacy of machine learning versus deep learning approaches in detecting fake comments and classifying news articles into nuanced truthfulness categories.
- To contribute to the body of knowledge in multilingual sentiment analysis by addressing the challenges of language-specific sentiment assessment and providing comparative insights into different algorithmic approaches.

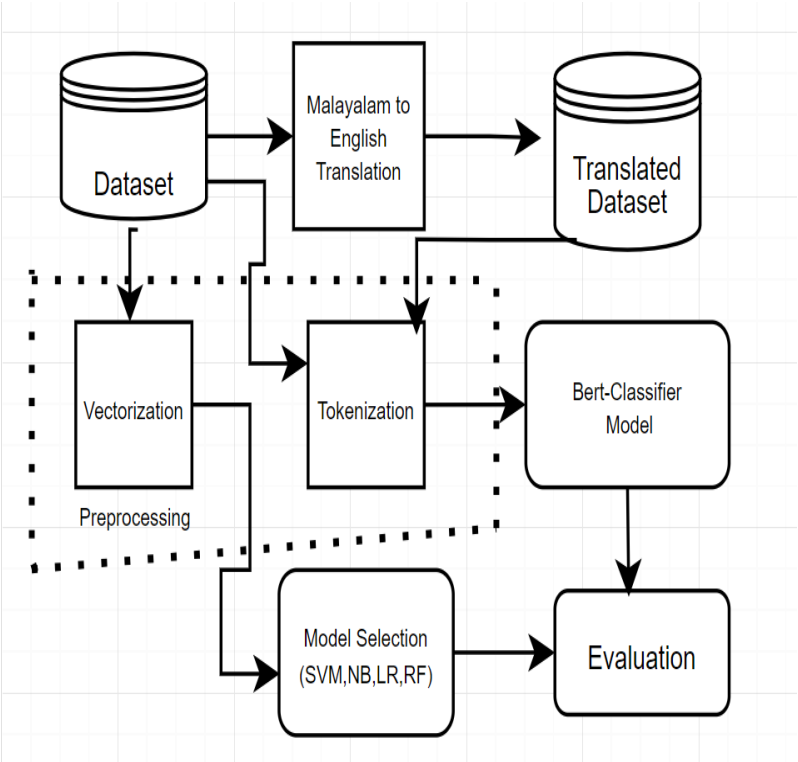
4. Proposed Work

Our project embarks on advancing multilingual sentiment analysis with a concentrated focus on the Malayalam language, engaging with a pair of distinct datasets sourced from the CodaLab Fake News Detection in Dravidian Languages competition (Dravidian-LangTech@EACL 2024). [11] The first dataset comprises YouTube comments, each meticulously categorized as original or fake, serving as a testament to the intricacies of digital discourse.[11] The second dataset encompasses a diverse collection of news items, each painstakingly classified into one of five truthfulness categories: false, mainly false, true, half true, and mostly true.[11] This granular classification scheme presents a nuanced spectrum of information authenticity, critical for the discerning algorithms we deploy.

The analytical journey of the project begins with the application of four classical machine learning classifiers—Support Vector Machine (SVM), Random Forest (RF), Logistic Regression, and Naive Bayes—executed on both datasets. The objective is to compare and contrast their validation performance rigorously, thereby determining the most efficacious model. The model that prevails in validation is then subjected to the crucible of testing within the datasets to ascertain its generalizability and robustness.

Venturing into the realm of deep learning, the project explores three distinct methodologies. Initially, we implement a BERT classifier to delve into the sentiment analysis directly. Subsequently, we enhance our linguistic reach by employing the Helsinki-NLP pipeline to translate the Malayalam text into English, thereafter applying the BERT classifier to this translated corpus.[12] The final stride in our deep learning endeavour employs the Indic BERT tokenizer, specially designed to improve tokenization of the Malayalam script, thus tailoring the BERT classifier to the linguistic nuances of Dravidian syntax and semantics. Through these varied approaches, we seek to construct a comprehensive analysis mechanism that stands at the vanguard of sentiment analysis for Malayalam language data, aiming for the highest echelons of accuracy and interpretability.

5. Flowchart



5.1. Traditional ML Models Steps

- Data Preparation: Import the necessary libraries, including scikit-learn and pandas. Load the training dataset for model training.
- Vectorization of Text: To convert the textual data into numerical vectors, use the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer.
- The Logistic Regression Model: To find the ideal regularisation parameter (C), use crossvalidation and Logistic Regression using a hyperparameter grid search. Utilize the determined ideal hyperparameters to train the classifier.
- The Random Forest Model: To determine the optimal set of hyperparameters, such as the number of estimators, maximum depth, minimum samples split, and minimum samples leaf, use Random Forest in conjunction with a grid search. Utilize the optimized hyperparameters to train the Random Forest model.
- The Support Vector Model (SVM): Use the Support Vector Machine (SVM) technique for categorization. Utilizing the TF-IDF vectorized training data, train the SVM model.
- The Naive Bayes Model: Multinomial Naive Bayes should be used for categorization. Utilizing the TF-IDF vectorized training data, train the Naive Bayes model.
- Verification Assessment of the dataset: To evaluate the model, load the validation dataset. Take care of missing values and vectorize the text column. Evaluate each classifier's accuracy score on the validation set.
- Ideal Model Choice: Determine which classifier on the validation set has the best accuracy. Declare this classifier to be the best model to test further. Evaluation of Testing Dataset.
- Testing: Load the testing dataset. Utilizing the TF-IDF vectorizer, vectorize the text column of testing data. To predict labels for the test dataset, use the best classifier available.
- Conclusion: Conclude the study by summarizing the chosen classifier's performance on the testing dataset.

5.2. Deep Learning Steps

- Data Preparation: Begin by importing essential libraries such as transformers for the BERT model, torch for the deep learning framework, and pandas for data manipulation. Load the training dataset into a DataFrame for further processing.
- Tokenization: Utilize the BERT tokenizer to convert the text into tokens that are understandable by the model. This step will include converting the Malayalam sentences into a format with token IDs, attention masks, and segment IDs suitable for BERT.
- Model Configuration: Choose a pre-trained BERT model that is

optimized for the Malayalam language or the multilingual version that includes Malayalam. Configure the BERT model with appropriate parameters, paying attention to the number of epochs, learning rate, and batch size for training.

- Fine-Tuning: Using the tokenized text data, fine-tune the BERT model on the Malayalam sentiment analysis task. This involves training the model on the dataset, adjusting weights, and ensuring the model learns the context of the dataset effectively.
- Validation Assessment: After the model has been fine-tuned, evaluate its performance on a separate validation set to gauge its effectiveness. Process the validation dataset similarly to the training set, with tokenization followed by the creation of DataLoader objects for the BERT model.
- Testing: Load the testing dataset and process it through the BERT models using the same tokenization and DataLoader creation steps as the validation set. Apply the selected model to obtain predictions for sentiment classification.
- Conclusion: Conclude the process by summarizing the performance of the BERT model on the testing dataset. Discuss the effectiveness of the model, its ability to generalize to unseen data, and any observations regarding its performance on multilingual sentiment analysis.

6. Result

The evaluation matrix used here are accuracy, training time, bar graph and Confusion matrices.

YouTube Comment Originality Detection

Model	Accuracy (%)	Training Time
Naive Bayes	78.8	0.51 sec
SVM	50.1	0.27 sec
Random Forest	75.5	12.61 sec
Logistic Regression	78.6	1.57 sec
BERT	75.2	1 hr 30 min 22 sec
EnglishTranslation+BERT	74.0	3 hr 56 min 27 sec
IndicBERT	75.1	1 hr 29 min 27 sec

Table 1- YouTube Comment Originality Detection Accuracy

Fake News Detection

Model	Accuracy (%)	Training Time
Naive Bayes	63.5	0.62 sec
SVM	65.8	0.37 sec
Random Forest	68.6	20.76 sec
Logistic Regression	61.7	12.07 sec
BERT	60.52	1 hr 56 min 12 sec
EnglishTranslation+BERT	65.7	4 hr 22 min 40 sec
IndicBERT	62.31	1 hr 50 min 18 sec

Table 2- Fake News Detection Accuracy

Table 1 presents the evaluation results for various machine learning models on the YouTube Comment Originality Detection task. The models compared include Naive Bayes, SVM (Support Vector Machine), Random Forest, Logistic Regression, BERT (Bidirectional Encoder Representations from Transformers), English Translation + BERT, and IndicBERT. These models were assessed based on their accuracy and training time. Naive Bayes demonstrated the highest accuracy (78.8%) with the shortest training time (0.51 sec), making it the most

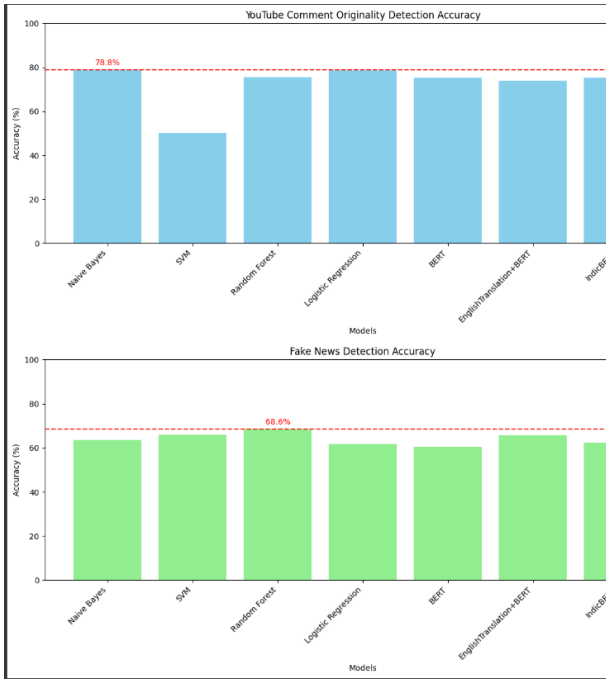


Fig 1- Bar Graph Showing accuracy of all models

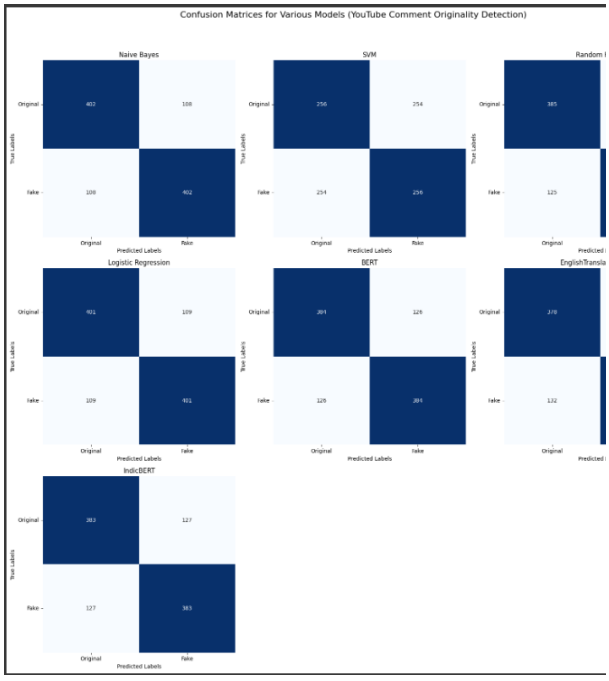


Fig 2- Confusion Matrix for all the models for Malayalam Comment Detection

efficient model for this task. In contrast, SVM had a significantly lower accuracy (50.1%) and a fast-training time (0.27 sec). Logistic Regression showed comparable accuracy to Naive Bayes (78.6%) but with a slightly longer training time (1.57 sec). The BERT-based models, while powerful

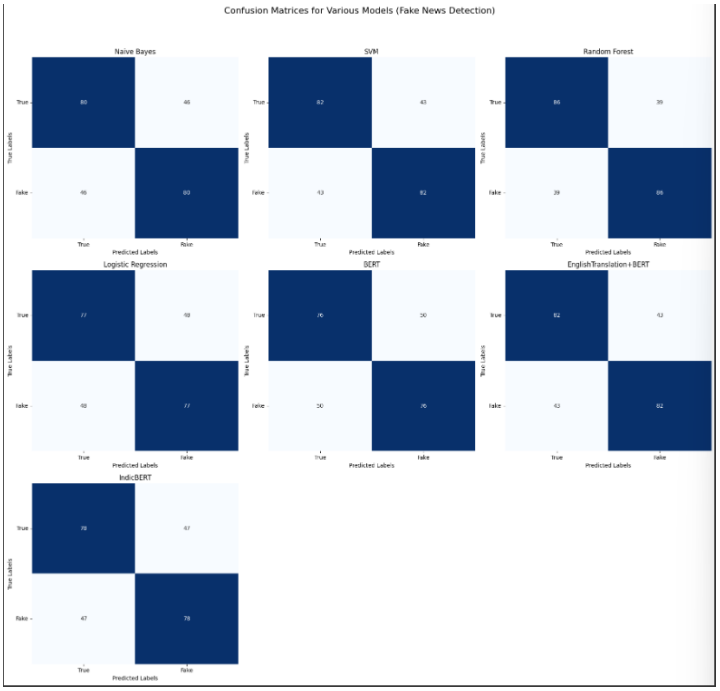


Fig 3- Confusion Matrix for all the models for Malayalam News Classification

in terms of linguistic understanding, required much longer training times, ranging from 1 hour 29 min 27 sec for IndicBERT to 3 hours 56 min 27 sec for English Translation + BERT.

Table 2 outlines the results for the Fake News Detection task. Similar models were evaluated, and Random Forest achieved the highest accuracy (68.6%) with a moderate training time (20.76 sec). SVM showed a balanced performance with an accuracy of 65.8% and a quick training time (0.37 sec). Naive Bayes and Logistic Regression offered lower accuracy rates (63.5% and 61.7%, respectively) with relatively short training times. The BERT model had the lowest accuracy (60.52%) with a training time just under 2 hours. English Translation + BERT improved the accuracy slightly to 65.7% but required the longest training time at over 4 hours. IndicBERT's accuracy was 62.31% with a training time of 1 hour 50 min 18 sec, suggesting that while the BERT-based models were powerful, they were not the most efficient in terms of training time versus performance gain for this specific task.

7. Conclusion

The comparative analysis of sentiment analysis models on Malayalam datasets underscores the dominance of Naive Bayes for YouTube comment classification and Random Forest for fake news detection in terms of accuracy. Despite longer training times, BERT-based models offer deep linguistic insights, albeit not surpassing traditional models in accuracy.

A critical factor influencing performance disparities is dataset size. Deep learning models like BERT thrive on vast data, whereas the modest datasets of 3,200 comments for YouTube and 1,700 articles for news may limit their potential. Traditional models exhibit notable efficiency, possibly due to their compatibility with smaller datasets and less complex feature spaces.

The restricted dataset size emerges as a pivotal variable, potentially constraining deep learning algorithms' effectiveness. Expanding datasets or employing data augmentation strategies could enhance deep learning models' performance, leveraging their ability to grasp complex language patterns.

This research illuminates current sentiment analysis model capabilities for Malayalam texts while emphasizing the necessity for larger datasets to maximize deep learning techniques' potential. Future endeavours should focus on dataset expansion or data augmentation strategies to enhance deep learning models' performance, striking a balance between model choice, dataset size, and computational efficiency in multilingual sentiment analysis research.

References

[1] Contreras Hernández, S., Tzili Cruz, M. P., Espínola Sánchez, J. M., & Pérez Tzili, A. (2023). Deep learning model for covid-19 sentiment analysis on twitter. New Generation Computing, 41(2), 189-212.

[2] Manias, G., Mavrogiorgou, A., Kiourtis, A., Symvoulidis, C., & Kyriazis, D. (2023). Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. Neural Computing and Applications, 35(29), 21415-21431.

[3] Amara, A., Hadj Taieb, M. A., & Ben Aouicha, M. (2021). Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis. Applied Intelligence, 51, 3052-3073.

[4] Mohawesh, R., Maqsood, S., & Althebyan, Q. (2023). Multilingual deep learning framework for fake news detection using capsule neural network. Journal of Intelligent Information Systems, 60(3), 655-671.

[5] Anjum, & Katarya, R. (2023). HateDetector: Multilingual technique for the analysis and detection of online hate speech in social networks. Multimedia Tools and Applications, 1-28.

[6] Mello, C., Cheema, G. S., & Thakkar, G. (2023). Combining sentiment analysis classifiers to explore multilingual news articles covering London 2012 and Rio 2016 Olympics. International Journal of Digital Humanities, 5(2), 131-157.

[7] Kar, P., & Debbarma, S. (2023). Multilingual hate speech detection sentimental analysis on social media platforms using optimal feature extraction and hybrid diagonal gated

recurrent neural network. The Journal of Supercomputing, 79(17), 19515-19546.

[8] Sidhu, S., Khurana, S. S., Kumar, M., Singh, P., & Bamber, S. S. (2023). Sentiment analysis of Hindi language text: a critical review. Multimedia Tools and Applications, 1-30.

[9] Lopez, C. E., & Gallemore, C. (2021). An augmented multilingual Twitter dataset for studying the COVID-19 infodemic. Social Network Analysis and Mining, 11(1), 102.

[10] Zardak, S. R., Rasekh, A. H., & Bashkari, M. S. (2023). Persian Text Sentiment Analysis Based on BERT and Neural Networks. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 47(4), 1623-1634.

[11] Subramanian, M., Chakravarthi, B. R., Shanmugavadivel, K., Pandiyan, S., Kumaresan, P. K., Palani, B., Singh, M., Raja, S., Vanaja, & S, Mithunajha. (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.

[12] Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020, November). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 4948-4961).