EAI Endorsed Transactions
on Scalable Information Systems Research Article EAI.EU

A Novel Ensemble Model for Complex Entities Identification in Low Resource Language

Preeti Vats[1,*], Nonita Sharma[2] and Deepak Kumar Sharma[3]

[1,2,3] Indira Gandhi Technical University for Women, Delhi, India

Abstract

The fundamental method for pre-processing speech or text data that enables computers to comprehend human language is known as natural language processing. Numerous models have been developed to date to pre-process data in the English language; however, the Hindi language does not support these models. India's national tongue is Hindi. In order to help the locals, the authors of this study used supervised learning methods like Linear Regression, SVM, and Naive Bayes algorithm to investigate a dataset of complicated terms in the Hindi language. Additionally, a sophisticated Hindi word classification model is suggested employing several methods based on the forecasts as well as collective learning strategies like Random Forest, Adaboost, and Decision Tree. Depending on how well the user's language is understood, the suggested model will assist in simplifying Hindi text. Authors attempt to classify the uncharted dataset using deep learning algorithms like Bi-LSTM and GRU approaches in further processing.

Keywords: NLP, Ensemble learning; Decision Tree; Hindi text identification

Received on 04 October 2023, accepted on 11 November 2023, published on 21 November 2023

doi: 10.4108/eetsis.4434

1. Introduction

The increased accessibility of electronic text documents from various web resources has led to an increased interest in supervised machine learning techniques in the real world. Documents must be categorized according to their relevant fields, which is crucial. Using specific qualities of the content, classification methods assist in grouping distinct items into categories. Machine learning makes it possible to categorize large-scale data handling, analysis, and assessment processes in this situation [1].

Text classification is an approach to text analysis that uses Natural Language Processing (NLP) to sort and classify data into various categories, forms, or any other specific pre-defined class. Text classification is a crucial NLP activity and a potent tool for identifying text and its context, making it simpler to access, find, and categorize [2]. Text classification is the automatic classification of texts into one or more predetermined classes based on labels. It involves combining many processes, such as retrieval, classification, and machine learning approaches, to categorize various patterns in each corpus. Fig 1.1 illustrates the text classification model for complexity detection.

In this study, authors suggested a difficult word classifier for Hindi text, a language that has received little attention in the literature on text simplification. Only a small amount of work has been done on Hindi's lexical simplification, despite the language having the third-highest number of speakers worldwide, after Mandarin Chinese and English [3]. Additionally, 43.63% of the population speaks Hindi, according to the most recent Indian census, making Hindi the official language of the government of India [1].

This research project is concerned with categorizing Hindi text into various groups. This study also compares various supervised machine learning techniques used to categorize text-balanced and text-unbalanced datasets. The authors investigated the variation between other evaluation criteria like precision, recall, and F1 score and accuracy.

[*] Corresponding author. Email: preeti017phdit22@igdtuw.ac.in

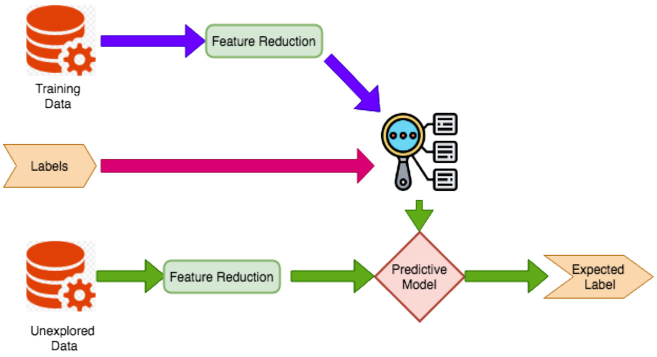

Figure 1. Text Classification model

This project aims to create a classifier model that, in addition to more accurately categorizing the dataset provided, also works with a distinct corpus that evaluates the terms using annotations from various annotators with different levels of experience in Hindi. The following queries are raised in this study

1) Which supervised learning approach produces the most favourable evaluation outcomes?

2) Can existing techniques be modified or new ones developed to enhance text classification?

3) How do these models stack up against different kinds of text corpora?

The following contributions are made to this research to investigate the answers to these questions:

•Using various well-known classifiers, well-known supervised learning models are applied to the balanced and imbalanced datasets.

•A new model is suggested and is based on three variations of current cutting-edge systems. The proposed weighting scheme achieves the overall maximum accuracy on the text corpus when compared to the current schemes and the proposed model. The dataset can be found.

Section 1 consists of a literature review of the given research work. Section 2 consists of various methodologies, their application, and the algorithm used to train the dataset. Section 3 explains the proposed model to classify and train the dataset more accurately. Section 4 describes the results and discussion. Section 5 concludes the research with a new proposed model.

2. Related Work

The details of associated research on various models developed for various corpora are included in this section. To uncover hidden patterns in the unstructured text data, "V Ehadi et. al."[1] discussed a categorization model that shows the context. This gave us a better comprehension of the meaning and themes of complicated words.

"G Venugopal et. al."[5] give a dataset of difficult Hindi words. They explain the apparent discrepancy in word difficulty ratings between native and non-native annotators. They adopted terms from a corpus of older Hindi works, such as stories and novels. 100 people, both native speakers and non-native speakers, annotated the different sentences. They purified the information, gathered it, and created a dataset that had feature values for the terms and their synonyms that were annotated. The corpus consists of a binary label with the values 1 and 0 indicating complex and simple, respectively. To simulate a real-world situation where readers with different vocabulary skills would consume the text, they have tried

to include readers conversant with a variety of languages other than Hindi. The classification of Hindi text is the focus of "M. Mehta et. al." [6] research. Research on the classification of the morphologically rich and low-resource languages such as Hindi which is written in Devanagari script has been hampered due to insufficient labelled corpus.

"HongChan li et. al." [15] talked about dimensionality reduction and feature selection. The many methods of dimensionality reduction were studied, and it was determined that the differential evolution approach, which assigns various excitation functions to several weak classifiers in order to train the ideal weight distribution, would work with the ensemble learning model.

"S.S.Samant et. al."[8] defines a supervised alternative to Tf-IDF and dimensionality reduction. When attempting to determine the significance of a word within a document as part of information retrieval, this approach performs admirably. Finding a word's relevance inside a category is the goal of term weighting methods, but the penalty factor IDF does not distinguish between documents from the positive category and the negative category.

"Venugopal et. al." [15] considers eight categorization models, five of which are ensemble models. Decision trees, support vector classifiers, nearest centroid classifiers, random forests, extra trees, Ada boost, gradient boosting, and XG boost were used to categorize the data. The classification was done utilizing k-fold cross-validation and five splits. Table 1 demonstrates other related work done in this field.

3. Tools and Techniques

Here, the authors used various classification algorithms to classify and train the data. Classification algorithms are generally those models that can identify the class of any data. Since supervised learning works with predetermined labels, these classification models are used to classify data in different classes or 'Labels'. The dataset is split into a training set and a testing set. Classifiers are used on text that has been labeled.

The proposed approach, called BagBoost, uses supervised learning to recognize tough phrases. An innovative strategy utilizing a soft voting classifier, a fundamental supervised learning model, was introduced by "Venugopal et al." [15]. Soft voting classifiers [15], which are made up of the predefined ensemble models bagging, boosting, and Random Forest, have an impact on BagBoost as well. Logistic regression is not employed since the characteristics are collinear.

The imbalanced data in this study is classified using six fundamental training models, including one ensemble learning model. Support Vector Machine, Logistic Regression, Decision Tree Using Cart Algorithm (Gini Index), K-Nearest Neighbour, and Naive Bayes algorithm were the models used to categorize the imbalanced data, and examples of each are shown below. The primary goal is to determine how the accuracy of supervised learning models varies depending on the evaluation criteria. Table 1 illustrates literature review for this research work.

Table 1. Literature Review of Proposed Model

Citation	Approach used	Novelty	Evaluating parameter
[1]	Structured Topic Modeling	Text Classification	Labels with 0-5
[2]	Multi-label Text Classification	Framework using Labelled powerset with SVM	Accuracy
[3]	Rule-based approach for sentence simplification in Hindi	Addressing sentence complexity in the context to NLP applications	Labels [0-3] given to sentences based on complexity
[4]	KNN, Naïve Bayes, Logistic regression, SVM	Hindi Text Categorization based web sources	Accuracy
[5]	LSTM, Bi-LSTM, CNN	IndicFT, FastText embeddings for Hindi Text	Model Accuracy
[7]	Naïve Bayes	Bagging Naïve Bayes and Fine-Tuned NB	Accuracy and Box plot
[8]	Term Weighting	New Term weighing Model ifn-modRF	F1 score
[9]	Sense-based normalization of features and analysis of performance classifier	Corpus for complex entities found in Hindi	AUC and F1 Score
[12]	Ensemble learning and Data mining	Text Identification	Accuracy
[13]	Machine Learning and Deep learning approach	Improved SVM classifier	Precision and Recall
[14]	Ensemble learning versus classification techniques	Soft Voting Classifier	Accuracy and precision
[15]	Ensemble Approach of identification of Text	Hindi and English Text Dataset	F1 score
[17]	Lazy Classifiers KNNand Logistic Regression	Classification of Hinditext using lazy classifiers	Accuracy
[18]	Feature selection and LDA	Ensemble Classifier stream text Data	Accuracy
[19]	Word embedding using Vanilla Transformers	Chat Bot for Natural Language processing	F1 score
[20]	Multi-Label Text Classification	Multi-Layer Perceptron	Transformer accuracy
[21]	Text Classification Alalgorithm	Pros and Cons of supervised learning algorithm	Accuracy, F1 score,AUC
[22]	Stacking Classifier, and feature selection	Two-layer Stack model	Accuracy, Recall and Precision
[23]	Support vector machine and KNN	overlap-sensitive margin (OSM) classifier	Accuracy and F1score

4. Methodology

4.1 Text mining and Information retrieval

Information retrieval is a vital part of any machine-learning model to handle the dataset. The retrieval and evaluation of textual information from document repositories is the subject of information retrieval (IR). Text mining is the process of extracting useful information from a text corpus and examining nontrivial patterns. The authors created a 2-D array of a given dataset using Tf-IDF vectorization techniques for information retrieval and text mining.

4.2 Data Balancing

After evaluating the metrics of different supervised learning models on the raw dataset, the authors have a task to handle the imbalanced dataset. Since we have 4363 easy words and 2956 complex words, balancing the dataset is important to get accurate results. Here authors use the upsampling method for dataset balancing.

4.3 Proposed Model

This section includes a description of the proposed model 'BagBoost' a supervised learning model used to identify complex words. Here, the authors proposed a new model using a soft voting classifier, which constitutes basic ensemble learning models that are bagging, boosting, and Random Forest. It collectively calculates results based on the voting of trained data. Figure 2 illustrates Voting Classifiers are aggregation techniques that can combine two or more models in the flowchart of the BagBoost model.

4.4 Evaluation Metrics

To assess the effectiveness of machine learning models, a few hyper-parameters, including Validation Accuracy, F1 score, precision, accuracy, recall, Confusion Matrix, and AUC curve, have been developed.

5. Results and Discussion

This section analyses the findings and discusses the various supervised machine-learning models used in the study. As noted in section 3, this study employs a Support Vector Classifier, Naive Bayes, Logistic Regression, KNN, and decision tree utilizing the Cart algorithm, as well as an ensemble learning model random forest on the raw dataset. Table 2 summarizes and compares all machine-learning models used in this study. Figure 3 plots the accuracy of each machine-learning model.

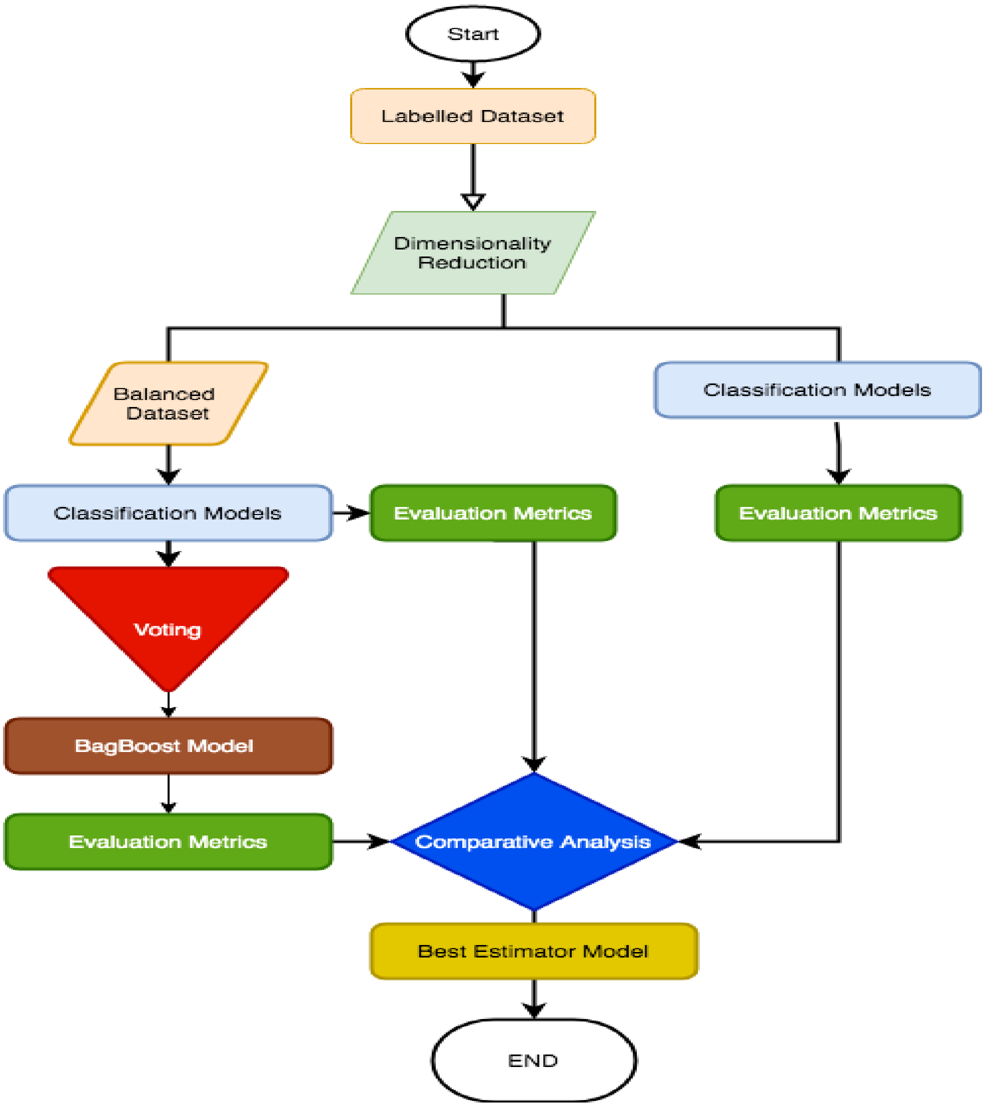

Figure 2. Flow Chart of BagBoost Model

Table 2. Comparison of different supervised algorithms (raw dataset)

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	60.89	0.72	0.61	0.64
SVM	61.50	0.60	0.62	0.58
Naive Bayes	61.43	0.78	0.61	0.67
Logistic Regression	61.50	0.77	0.62	0.66
Decision Tree	59.59	0.96	0.60	0.73
KNN	46.55	0.51	0.47	0.46

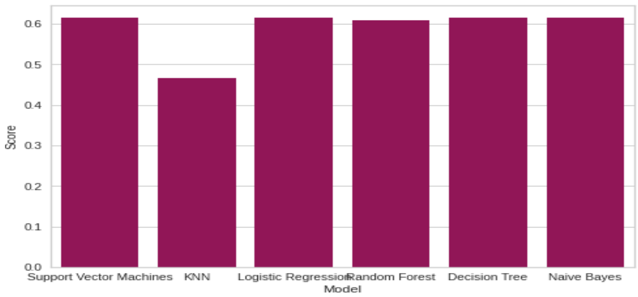

Figure 3. Graph for accuracy score of different learning algorithms (Raw Dataset)

Table 3. Comparison of different supervised algorithms (Balanced dataset)

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.67	0.66	0.67	0.65
SVM	0.65	0.66	0.66	0.65
Naive Bayes	0.64	0.64	0.67	0.65
Logistic Regression	0.64	0.64	0.65	0.64
Decision Tree	0.51	0.59	0.51	0.36
KNN	0.63	0.62	0.65	0.63

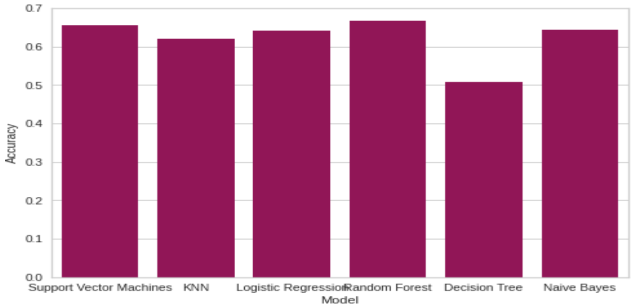

Figure 4. Graph for accuracy score of different learning algorithms (Balanced Dataset)

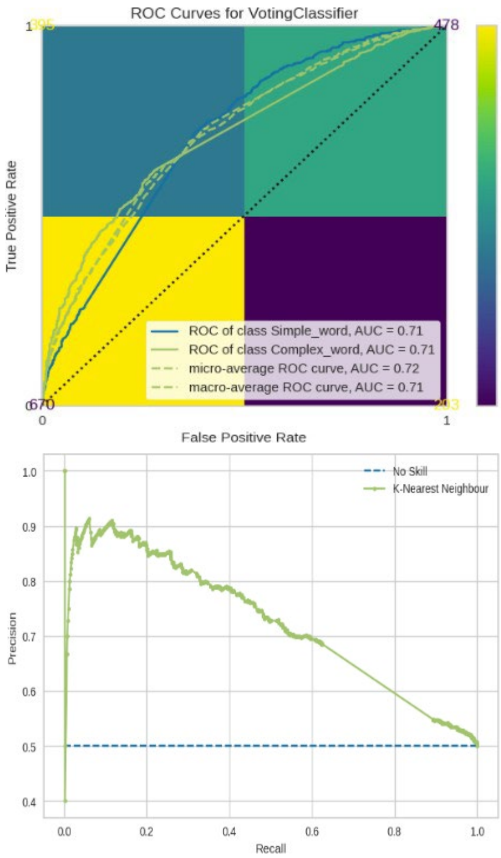

Figure 5. ROC and AUC curve for Proposed Model

After up-sampling the corpus, the authors applied these models to the balanced dataset once more. Table 3 shows the results of balancing the dataset. A graph of accuracy is shown in figure 4.

The proposed 'BagBoost' model provides an accuracy of 0.68, which is better than any former machine learning model. It also evaluated precision as 0.67 and recall as 0.66. It also calculated the best F1 score as 0.65. Hence figure 5 depicts AUC and ROC curve for the BagBoost model. This section follows the conclusion of the proposed model.

6. Conclusion

The proposed model 'Bagboost' performs well on the given dataset and provides a better understanding of the machine-learning model for low-resource languages. It also distinguishes entities found in Hindi text as simple and complex. This model further helps to create semantic graphs in low-resource languages that lead to the construction of conversational AI machines in the linguistic domain. In the future, authors perform this study for other low-resource languages. This study doesn't have any conflict of interest.

References

[1] Ebadi, A., Tremblay, S., Goutte, C., & Schiffauerova, A. (2020). Application of machine learning techniques to assess the trends and alignment of the funded research output. Journal of Informetrics, 14(2), 101018.

[2] Camponogara, E., Jia, D., Krogh, B. H., & Talukdar, S. (2002). Distributed model predictive control. IEEE Control Systems Magazine, 22(1), 44-52.

[3] Soni, A., Jain, S., & Sharma, D. M. (2013, October). Exploring verb frames for sentence simplification in Hindi. In Proceedings of the Sixth International Joint Conference on Natural Language Processing (pp. 1082-1086).

[4] Soni, V. K., & Selot, S. (2021, October). A Comprehensive Study for the Hindi Language to Implement Supervised Text Classification Techniques. In 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC) (pp. 539-544). IEEE.

[5] Mehta, M., Pandey, U., Chaudhary, Y., Sharma, R., Gill, I., Gupta, D., & Khanna,

[6] A. (2021, December). Hindi Text Classification: A Review. In 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) (pp. 839-843). IEEE.

[7] Joshi, R., Goel, P., & Joshi, R. (2020). Deep learning for Hindi text classification: A comparison. In Intelligent Human Computer Interaction: 11th International Conference, IHCI 2019, Allahabad, India, December 12–14, 2019, Proceedings 11 (pp. 94-101). Springer International Publishing.

[8] El Hindi, K., AlSalman, H., Qasem, S., & Al Ahmadi, S. (2018). Building an ensemble of fine-tuned naive Bayesian classifiers for text classification. Entropy, 20(11), 857.

[9] Samant, S. S., Murthy, N. B., & Malapati, A. (2019). Improving term weighting schemes for short text classification in vector space model. IEEE Access, 7, 166578-166592.

[10] Venugopal, G., Pramod, D., & Shekhar, R. (2022, June). CWID-hi: A Dataset for Complex Word Identification in Hindi Text. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 5627-5636).Zhou, Z.H. Ensemble Methods Foundations and Algorithms; CRS Press: Boca Raton, FL, USA, 2012.

[11] Rokach, L. (2010). Pattern classification using ensemble methods (Vol. 75). World Scientific.

[12] Zhang, Cha, and Yunqian Ma, eds. Ensemble machine learning: methods and applications. Springer Science & Business Media, 2012.

[13] Seni, G., & Elder, J. F. (2010). Ensemble methods in data mining: improving ac- curacy through combining predictions. Synthesis lectures on data mining and knowledge discovery, 2(1), 1-126.

[14] Quan, Z., & Pu, L. (2022). An improved accurate classification method for online education resources based on support vector machine (SVM): Algorithm and experiment. Education and Information Technologies, 1-15.

[15] Venugopal, G., Pramod, D., & Jatinderkuma, R. S. (2022). Revisiting the role of classical readability formulae parameters in complex word identification (Part 2). Computer Science Journal of Moldova, 88(1), 49-63.

[16] Roy, A., Kapil, P., Basak, K., & Ekbal, A. (2018, August). An ensemble approach for aggression identification in English and Hindi text. In Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018) (pp. 66-73).

[17] Bafna, P. B., & Saini, J. R. (2020, March). Hindi Verse Class Predictor using Concept Learning Algorithms. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 318-322). IEEE.

[18] Wang, Z., Liu, J., Sun, G., Zhao, J., Ding, Z., & Guan, X. (2020, June). An ensemble classification algorithm for text data stream based on feature selection and topic model. In 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 1377-1380). IEEE.

[19] Sergio, G. C., & Lee, M. (2021). Stacked DeBERT: All attention in incomplete data for text classification. Neural Networks, 136, 87-96.

[20] Yadav, S., & Sharma, N. (2018). Homogenous ensemble of time-series models for Indian stock market. In Big Data Analytics: 6th International Conference, BDA 2018, Warangal, India, December 18–21, 2018, Proceedings 6 (pp. 100-114). Springer International Publishing.

[21] Yadav, S., & Sharma, N. (2018). Homogenous ensemble of time-series models for indian stock market. In Big Data Analytics: 6th International Conference, BDA 2018, Warangal, India, December 18–21, 2018, Proceedings 6 (pp. 100-114). Springer International Publishing.

[22] Sharma, N. (2021). Jaiditya Dev, Monika Mangla, Vaishali Mehta Wadhwa, Sachi Nandan Mohanty, and Deepti Kakkar. A heterogeneous ensemble forecasting model for disease prediction. New Generation Computing, 39(3-4), 701-715.

[23] Sultana, N., Sharma, N., & Sharma, K. P. (2019, April). Ensemble model based on NNAR and SVR for predicting influenza incidences. In Proceedings of the Inter- national Conference on Advances in Electronics, Electrical & Computational Intelligence (ICAEEC).

[24] Kowsari, K. (2019). Jafari Meimandi, K. Heidarysafa, M.Mendu, S.Barnes, L.Brown, D.: Text Classification Algorithms: A Survey. Information, 10(4).

[25] Wahba, Y., Madhavji, N., & Steinbacher, J. (2022, March). Reducing Misclassification Due to Overlapping Classes in Text Classification via Stacking Classifiers on Different Feature Subsets. In Advances in Information and Communication: Proceedings of the 2022 Future of Information and Communication Conference (FICC), Volume 2 (pp. 406-419). Cham: Springer International Publishing.