Automatic Classification of Gender Stereotypes in Social Media Post

Gersome Shimi[1], Jerin Mahibha[1,*,†] and Durairaj Thenmozhi[3,†]

[1]Madras Christian College,Chennai, India
[2]Meenakshi Sundararajan Engineering College, Chennai, India
[3]Sri Sivasubramaniya Nadar College of Engineering, Chennai, India

Abstract

Every day, millions of information are shared on the internet through social media. The contents of the social media posts are based on the person's wishes, emotional expressions, ambitions, passions, and achievements. Among these posts there are possibilities of hurtful messages such as sexist contents, getting embedded. It may sometimes be intentional or unintentional, but also may disturb the mental well-being of the recipient. So automatic identification of these sexist languages and terms in social media posts has to be taken into immediate consideration. EXIST (sEXism Identification in Social Media Network) 2024, a shared task has addressed this issue. This shared task addresses binary classification(Task1), multiclass classification(Task2) and multilabel classification(Task3). We contributed Language Agnostic BERT Sentence Embeddings(LaBSE) based MultiLayer Perceptron (MLP) classifier, eXtreme Gradient Boosting (XGBoost) Classifier, and ensemble Convolutional Neural Network (CNN) model for Task1 and LABSE with MLP classifier and XGBoost Classifier for Task2.

Keywords

ensemble CNN, LaBSE, MLP, XGBoost, Classifier, sexism,

1. Introduction

Social media platforms have become a basic amenity for communication in the modern world[1]. It is an effective tool for posting content from diverse fields like sports, politics, religion, race, or culture. According to data reportal global media statistics, the world spends approximately 12 billion hours and a person actively spends an average of 2 hours and 20 minutes daily in social media. Shared posts may contain information that gives emotional scars, misguides people, or deprives harmony among social media fanatics [2]. Women centered dissemination of offensive and discriminatory material through social media platforms has increased rapidly and has emerged as a significant concern. This affects the well being of women and the freedom of expression [3]. All around the world many women have reported and suffered abuse, discrimination and other sexist experiences in real life. The contribution of social networks is found to be more, considering the transmission of sexism and other disrespectful and hateful behaviours. Detection, alert generation and computing the frequency of sexist behaviours and discourses in social media platforms is considered an important and challenging task [4]. Discriminatory information on women, which is unethical, is common in such posts. It is challenging to locate sexist content like dominance, misogyny, and inequality which can come out in diverse forms [5]. The main platforms for social complaint, activism, etc. are considered to be the Social Networks where movements like #MeTwoo, #8M or #Time'sUp have spread rapidly [6].

EXIST 2024 aims to capture sexism in a broad sense, from explicit misogyny to other subtle expressions that involve implicit sexist behaviors. The shared task EXIST 2024 was a part of CLEF 2024, based on English and Spanish comments. The shared task intended to spot different categories of sexist content [7][8]. The task contained five subtasks namely Task 1 to Task 5 in which we as a team participated in two subtasks namely Task 1 and Task 2. **Task 1 - Sexism Identification** The first task is a binary

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[*] Gersome Shimi.

[†] These authors contributed equally.

gshimi2022@gmail.com (G. Shimi); jerinmahibha@gmail.com (J. Mahibha); d\protect1_theni@ssn.edu.in (D. Thenmozhi)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

classification, the system has to decide whether or not a given tweet contains or describes sexist expressions or behaviors.

Task 2 - Source Intention

This task aims to categorize the sexist messages according to the intention of the author in one of the following categories:

(i) Direct sexist message

(ii) Reported sexist messages and

(iii) Judgemental sexist message

The second task is a multiclass classification problem, where the system needs to identify the intention behind the tweet. The possible intentions are directly addressing sexism, reporting sexism conditions about women, and judging/condemning sexism.

Various models including a MPL classifier with Language Agnostic Sentence Embeddings, XGBoost, and ensemble CNN were used for implementing the subtasks namely Task 1 and Task 2. The results of all these were submitted for ranking. Considering the two tasks the training and evaluation of the proposed models were carried out using the corresponding dataset provided by the EXIST 2024 task organizers. This model was then tested with the testing dataset provided for the shared task, based on which the task was evaluated.

This paper is organized as follows: Section 2 explains the related work, Section 3 describes the dataset, the methodology used is described in Section 4, the results and discussions are provided in Section 5 and Section 6 provides the Conclusion.

2. Related Works

A machine learning model based on a bidirectional LSTM architecture was used for the classification of sexist and non sexist tweets by [9]. The model had effectively captured contextual information and achieved an F1-Score of 0.6355. As part of IberLEF 2022 Language agnostic model and multilingual BERT classification model were used to identify sexist and non-sexist text from English and Spanish text. It had been found that the Language agnostic model performed better with an F1 score of 0.753 [2]. [10] had applied transfer learning from a pre-trained multilingual DeBERTa (mDeBERTa) model and easits zero classification. The Concept of majority voting was used to combine the methods by which mDeBERTa achieved an accuracy of 76.09% and 66.26% for Task 1 and Task2 respectively. Different tranformer models like BERT, DistilBERT, and RoBERTa had been used for implementing the three tasks shared by SEMEVAL 2023. The BERT model, had shown a macro F1-score of 0.8073, 0.5876 and 0.3729 for Task A, Task B and Task C respectively [6].

Seoul metropolitan ciKNN, Naïve Bayes, SVM and GBDTvil complaint dataset in Korean language had been classified using Random forest and XGBoost, the result had proven that XGBoost Classifier outperformed Random forest classifier [11]. For crime prediction, after applying TF-IDF (Term frequency-inverse document frequency) the machine language models XGBoost, KNN (K-Nearest Neighbor), Naïve Bayes, SVM(Support Vector Machine), and GBDT(Gradient Boost Decision Tree) were implemented and found XGBoost Outperformed other Machine Learning algorithms with 0.923, 0.916 and 0.919 for Precision, Recall, and F1-score respectively [12].

BLSTM-C, a hybrid model of BLSTM and Convolutional Neural Network performed well with the Chinese language dataset for text classification. The BLSTM-C had been coded with two layers of LSTM and one layer of CNN to obtain the accuracy of 0.962 [13].

Few research works were carried out on sexism identification and related text classification tasks had been explored. It is found that continuous research is being carried out in related fields like identifying insulting comments, hate speech, toxic comments, and intent classification which can be used as a base for identifying comments representing sexism from social media text. It could also be observed that the tweet and its contents have inconsistent structure, data preprocessing will helps to improve the accuracy of the training model.

Table 1
Sample Data

Sample Data	Language
@ultimonomada_ Si comicsgate se parece en algo a gamergate pues muy bien por el acoso. Y si se está haciendo un sabotaje porque hay personajes que no os gustan entonces gracias por darme la razón. Sois unos lloricas ofendidos.	Spanish
$@Geek$_{p}ride@kathrynstimpson@medicalpokeEverydaySexismWouldn't$ work for women who get assaulted at home or work. Also would give the government the ability to track anyone for any reason.	English

Table 2
Task 1 Dataset Distribution

	Source	Size	English	Spanish	sexist	Non sexist
Training	Twitter	6920	3260	3660	3553	3367
Evaluation	Twitter	1038	489	549	559	479
Testing	Twitter	12456	5868	6588	–	–

Table 3
Task 2 Dataset Distribution

	Source	Size	English	Spanish	Direct	Reported	Judgemental	No
Training	Twitter	6920	3260	3660	3141	1298	1035	1446
Evaluation	Twitter	1039	489	549	452	215	168	203
Testing	Twitter	12456	5868	6588	–	–	–	–

3. Dataset

The dataset used to implement Task 1 and Task 2 of EXIST 2024 was the training, evaluation and the test dataset, that were provided by the organizers of the shared task. All the datasets for the shared task Exist 2024 were given in the JSON format from which the important features required for implementing Task 1 and Task 2 were selected. This includes features like id_EXIST, tweet, annotators, and labels_task1 for Task 1 and id_EXIST, tweet, annotators and labels_task2 for Task 2. Other features like gender_annotators, age_annotators, ethnicities study_levels_annotators, countries were identified as unimportant features and were eliminated. Table 1 shows sample instances from the dataset considering both the languages English and Spanish. Twitter is the source of all the instances in the dataset.

The data distribution in the training, evaluation and testing dataset is represented by Table 2 and 3. The training dataset for Task 1 and Task 2 had 6920 instances of which 3260 tweets were in English and 3660 instances were in Spanish. Considering Task 1 there were 3553 instances under the Sexist category and 3367 instances under Non Sexist category. Considering the Task 2, the number of instances was 3141, 1298 and 1035 under the categories Direct, Reported, and Judgemental respectively. The test dataset had 12456 instances of which 5868 were in English and 6588 were in Spanish.

4. Methodology

The proposed system uses XGBoost Classifier, LSTM-CNN Classifier, MLP classifier with Language Agnostic embeddings for Task 1 which is a binary classification problem to detect Sexist and Non Sexist comments. The Task 2 was implemented using XGBoost Classifier and MLP classifier with Language Agnostic embeddings, a multi class classification problem with three class labels namely Direct sexist message, Reported sexist message and Judgemental sexist message. The proposed architecture of the system is shown in Figure 1.

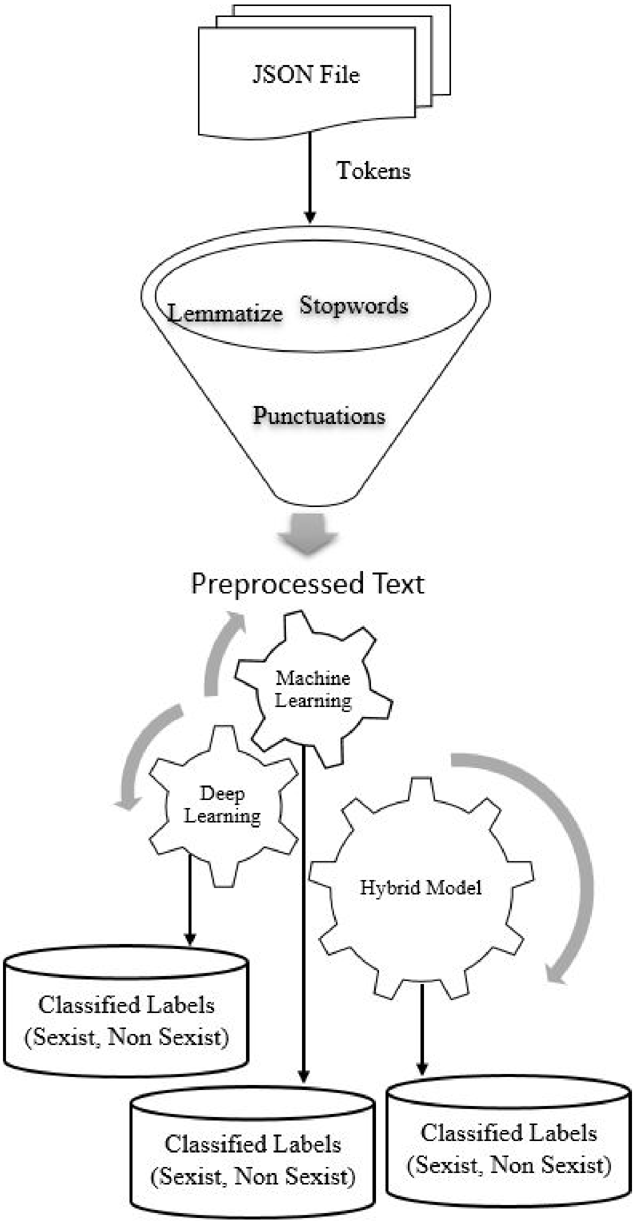
JSON File
Tokens
Lemmatize
Stopwords
Punctuations
Preprocessed Text
Machine Learning
Deep Learning
Hybrid Model
Classified Labels (Sexist, Non Sexist)
Classified Labels (Sexist, Non Sexist)
Classified Labels (Sexist, Non Sexist)

Figure 1: System Architecture

4.1. Preprocessing

The dataset instances in JSON format, was read and cleaned by preprocessing techniques. Preprocessing is the technique of removing unimportant information from texts, which are not used during the classification process. It is performed by removing stop words, symbols, and special characters in addition to that root words are extracted using stemmer and lemmatization algorithms before the dataset is fed to the model.

The class label associated with each of the tweets was not provided directly. Instead the labels are provided by six different annotators as Hard Labels and Soft Labels. We chose Hard Label for our implementation. As a part of preprocessing, the approach of majority voting was applied to the provided information to decide the class label associated with the tweet. This was done for both Task 1 and Task 2.

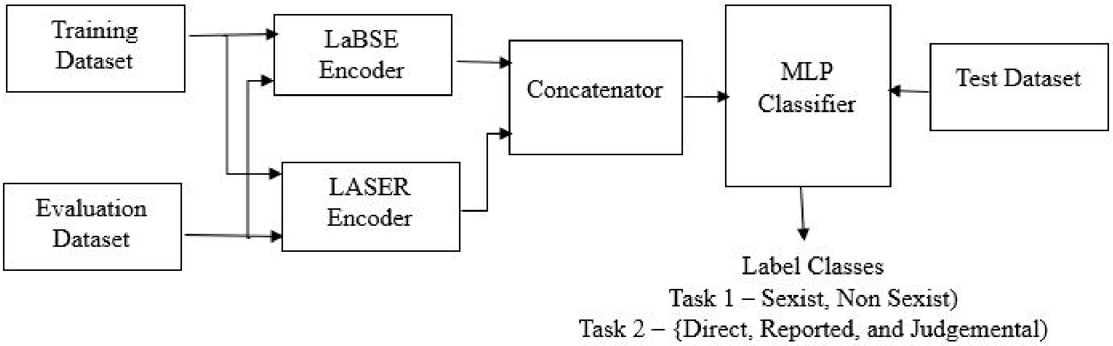

Figure 2: Proposed System

4.2. MLP classifier with Language Agnostic Embeddings

The proposed system used a MLP classifier for which custom generated embedding was provided as input. Language agnostic sentence transformer was used to generate text embeddings. As the Language agnostic sentence transformer is multilingual in nature and support both English and Spanish languages, the same model was used to generate the embeddings for all the given tweets. Similarly Laser encoder pipeline was used to generate LASER embeddings for all the tweets. Both these embeddings were concatenated to generate a final set of embeddings using which the MLP classifier was trained. The hyper parameters associated with the MLP classifier are: random state was set as 42, the maximum iteration was set as 300, relu activation function was used, the parameter alpha was set as 0.05, learning rate as adaptive and solver as adam. The working of this model is represented in Figure 2.

The proposed model when evaluated using the evaluation dataset, it provided an accuracy and Macro F1 Score 0.77.

4.3. XGBoost Classifier

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting to solve many data science problems in a fast and accurate way. The system uses XGBoost Classifier which gets the output from the TF-IDF(Term frequency-inverse document frequency) model. The preprocessed text is fed to the TF-IDF model to find the term frequency and document inverse frequency. TF-IDF algorithm [13] works on the frequency of the occurrence of the word in the document. The importance of a word is determined by the number of times a word appears in a document and is inversely proportional to the number of times it appears in the entire document set. Term Frequency is calculated by the formula:

$$TF_{i,j} \equiv \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$n_{i,j}$ denotes occurrence of w_i in document x_j
$\sum_k n_{k,j}$ denotes sum of all entries in document x_j

$$IDF_j \equiv log_2 \frac{|D|}{|\{j : w_i \epsilon x_j\}| + 1}$$

TF-IDF of the word w_i is calculated by the formula
$TF - IDF_{wi} = TF_{i,j} * IDF_j$

The XGBoost Classifier model is tuned by the hyperparameters learning_rate, max_depth, n_estimators, use_label_encoder, eval_metric with the values 0.7,10,80, False, rmse respectively.

Table 4
Performance metrics for Task 1 - Evaluation Dataset

Model	Accuracy	Recall	Precision
MLP Classifier with Language Agnostic Embeddings	0.75	0.75	0.75
XGBoost Classifier	0.72	0.72	0.72
ensemble CNN Classifier	0.56	0.56	0.56

The evaluation dataset of Task 1 when evaluated using XGBoost, achieved an accuracy of 0.72 and Macro F1 Score of 0.71. The evaluation dataset of Task 2 is evaluated using XGBoost and achieved an accuracy of 0.48 and Macro F1 Score 0.38.

4.4. Ensemble CNN Classifier

CNN model is one of the baseline models in Natural Language Processing and can be used to classify sentences and text. It processes the data sequences and enables them to evaluate the perspective of a given sentence and classify it based on the predefined labels [14]. The ensemble CNN model is used to classify the EXIST 2024, shared Task 1. After performing sequence padding the data is fed to the LSTM and CNN model, tuning the hyperparameters optimizer, loss with values Adam, binary_crossentropy respectively. LSTM model is coded by activating one LSTM layer, one Embedding layer and two dense layers. CNN model is coded by activating one Embedding layer, Conv1D layer and GlobalMaxPooling1D layer, and two dense layers with activation function relu and sigmoid respectively. This ensemble model is trained with epochs=10 and batch_size=32. The evaluation dataset of Task 1 when evaluated using ensemble CNN, achieved an accuracy and Macro F1 Score of 0.56.

The performance metrics associated with the evaluation of the different models using the evaluation dataset are represented in Table 4 and Table 5.

5. Results and Discussions

The metrics considered for the evaluation of Task 1 were ICM-Hard, ICM-Hard Norm and F1_Yes. The metrics considered to evaluation Task 2 are ICM-Hard, ICM-Hard Norm and Macro F1. The values of these performance metrics for the different models submitted are shown in Table 6 and Table 7.

On testing the model with the test dataset the MLP classifier with language agnostic embedding provided an ICM-Hard value of 0.3220, ICM-Hard Norm value of 0.6623 and F1_YES value of 0.7044 for Task 1. The same model applied for task2, it achieved a value of -2.0626 for ICM-Hard, 0.2115 for ICM-Hard Norm and 0.1200 for Macro F1. It could be found that the MLP classifier with Language Agnostic Embeddings outperformed the other models.

When the XGBoost Classifier model was tested using the test dataset, the model provided an ICM-Hard value of 0.2905, ICM-Hard Norm value of 0.6460 and F1_YES value of 0.6946 for Task 1. Considering Task 2, the same model achieved an ICM-Hard value of -0.8873, ICM-Hard Norm value of 0.2115 and Macro F1 value of 0.3148. The XGBoost Classifier outperformed other models for Task2.

When Task 1 is implemented using ensemble CNN model, it achieves an ICM-Hard value of -0.3410, ICM-Hard Norm value of 0.3286 and F1_YES value of 0.4922.

The MLP classifier with language agnostic embedding resulted in an F1_Yes score of 0.7044 based on which Task 1 was evaluated and we were ranked 48 on the leader board. Task 2 resulted in a macro-F1 score of 0.32 using the XGBoost Classifier, by which we were ranked 37 on the leader board.

Table 5
Performance metrics for Task 2 - Evaluation Dataset

Model	Accuracy	Recall	Percision
MLP Classifier with Language Agnostic Embeddings	0.65	0.41	0.47
XGBoost Classifier	0.48	0.40	0.38

Table 6
Performance metrics for Task 1 - Test Dataset

Model	Language	Rank	ICM-Hard	ICM-Hard Norm	F1_YES
MLP Classifier with Language Agnostic Embeddings	All	48	0.3230	0.6623	0.7044
	English	51	0.3097	0.6581	0.6731
	Spanish	46	0.3199	0.6600	0.7279
XGBoost Classifier	All	50	0.2905	0.6460	0.6946
	English	53	0.2624	0.6339	0.6568
	Spanish	50	0.2983	0.6492	0.7225
ensemble CNN	All	62	-0.3410	0.3286	0.4922
	English	66	-0.3667	0.3129	0.4617
	Spanish	62	-0.3286	0.3357	0.5179

Table 7
Performance metrics for Task 2 - Test Dataset

Model	Language	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
MLP Classifier with Language Agnostic Embeddings	All	43	-2.0626	0.0	0.1200
	English	45	-2.2094	0.0	0.0990
	Spanish	46	0.3199	0.6600	0.7279
XGBoost Classifier	All	37	-0.8873	0.2115	0.3148
	English	40	-1.0372	0.1411	0.2694
	Spanish	42	-1.9744	0.0	0.1375

6. Conclusion

Sexism detection has become a current research area as it is interlinked with different applications like sentiment analysis, opinion mining, offensive and hate speech detection. Having this in mind CLEF 2024 had come up with the task of sexism detection, EXIST 2024. As per the requirement of shared task by EXIST 2024, the proposed system implemented the MLP classifier with Language Agnostic Embeddings, XGBoost Classifier, and ensemble CNN classification model for Task 1 and MLP classifier with Language Agnostic Embeddings and XGBoost Classifier for Task 2. It was found that MLP classifier with Language Agnostic Embeddings performed well for Task 1 compared to the other models with an F1 score of 0.70. In Task2 XGBoost Classifier model performed well with an F1 score of 0.32. Usage of hybrid approaches where different deep learning models are combined can also facilitate efficient detection of sexism from the text. Often it could be observed that sexism is not in the text, but could be detected from the intonation or facial expression, which has made multimodel sexism detection also a promising research area.

References

[1] R. Briandana, C. M. Doktoralina, S. A. Hassan, W. N. W. Hasan, Da'wah communication and social media: The interpretation of millennials in southeast asia, International Journal of Economics and Business Administration 8 (2020) 216–226.

[2] G. Shimi, J. Mahibha, D. Thenmozhi, Sexism identification in social media using deep learning models (2022).

[3] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023–learning with disagreement for sexism identification and characterization, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 316–342.

[4] A. Sheth, V. L. Shalin, U. Kursuncu, Defining and detecting toxicity on social media: context and knowledge are key, Neurocomputing 490 (2022) 312–318.

[5] D. Felmlee, P. Inara Rodis, A. Zhang, Sexist slurs: Reinforcing feminine stereotypes online, Sex roles 83 (2020) 16–28.

[6] C. J. Mahibha, C. Swaathi, R. Jeevitha, R. P. Martina, D. Thenmozhi, Brainstormers_msec at semeval-2023 task 10: Detection of sexism related comments in social media using deep learning, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 1114–1120.

[7] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[8] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview), in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[9] A. Chaudhary, R. Kumar, Sexism identification in social networks, Working Notes of CLEF (2023).

[10] H. T. Ta, A. B. S. Rahman, L. Najjar, A. F. Gelbukh, Transfer learning from multilingual deberta for sexism identification., in: IberLEF@ SEPLN, 2022.

[11] J.-E. Ha, H.-C. Shin, Z.-K. Lee, Korean text classification using randomforest and xgboost focusing on seoul metropolitan civil complaint data, The Journal of Bigdata 2 (2017) 95–104.

[12] Z. Qi, The text classification of theft crime based on tf-idf and xgboost model, in: 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 2020, pp. 1241–1246. doi:10.1109/ICAICA50127.2020.9182555.

[13] Y. Li, X. Wang, P. Xu, Chinese text classification model based on deep learning, Future Internet 10 (2018) 113.

[14] R. Sujatha, K. Nimala, Classification of conversational sentences using an ensemble pre-trained language model with the fine-tuned parameter., Computers, Materials & Continua 78 (2024).