VerbaVisor@Multimodal Hate Speech Event Detection 2023: Hate Speech Detection using Transformer Model

Sarika Esackimuthu, Prabavathy Balasundaram

Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Chennai - 603110, Tamil Nadu, India
{sarika2010128, prabavathyb}@ssn.edu.in

Abstract

Thapa et al. (2023) task focuses on identifying hate speech or not from text-embedded images and also identify the targets of hate speech.Hate speech detection has emerged as a critical research area in recent years due to the rise of online social platforms and the proliferation of harmful content targeting individuals or specific groups.This task highlights the importance of detecting hate speech in text-embedded images.By leveraging deep learning models,this research aims to uncover the connection between hate speech and the entities it targets.

1 Introduction

Hate speech detection plays a crucial role in fostering a safer and more inclusive digital landscape. In today's interconnected world, where social media and online platforms dominate communication, the spread of hate speech can have far-reaching and detrimental consequences. Detecting and addressing hate speech helps protect vulnerable communities from harm, prevents the escalation of conflicts, and promotes constructive dialogue.Moreover, the integration of multimodal techniques, combining both textual and visual information, has further enhanced hate speech detection systems.

In recent years, the detection of hate speech has witnessed significant advancements, driven by the rapid progress in natural language processing (NLP) and computer vision technologies. Machine learning algorithms, particularly deep learning models, have revolutionized the field, allowing for more accurate and efficient hate speech detection. By analyzing text-embedded images and their associated textual content, algorithms can uncover hidden patterns and better identify hateful content targeted at specific entities or communities.

This research paper introduces an investigation into the detection of hate speech and identifying hate speech targets by employing NLP transformers, specifically the ALBERT base model.

2 Related works

Farooqi et al. (2021) research paper proposes an innovative method for hate speech detection in Hindi-English code-mixed conversations on Twitter. Their neural network approach leverages transformer's cross-lingual embeddings, fine-tuned for low-resource hate speech classification in transliterated Hindi text. The best-performing system, a hard voting ensemble of Indic-BERT, XLM-RoBERTa, and Multilingual BERT, achieved an impressive macro F1 score of 0.725. This highlights the method's effectiveness in accurately identifying hate speech, considering context and addressing challenges posed by code-mixing on social media platforms. The findings offer valuable insights for hate speech detection in multilingual settings.

In Jafri et al. (2023) the authors introduce a new dataset called IEHate, comprising 11,457 manually annotated Hindi tweets related to the Indian Assembly Election Campaign. They perform a comprehensive analysis of hate speech prevalence and its various forms in political discourse. The dataset is benchmarked using machine learning, deep learning, and transformer-based algorithms. Among the models, RoBERTa (multilingual) and BERT (HAM) achieved the highest F1-scores of 0.725 and 0.706, respectively. Transformer-based models outperformed machine learning and deep learning models.

Tiţa and Zubiaga (2021) research paper focuses on hate speech detection in a cross-lingual setting, emphasizing the importance of addressing this issue on global online platforms. The study utilizes fine-tuned altered multi-lingual Transformer models (mBERT, XLM-RoBERTa) with cross-lingual training between English and French and within

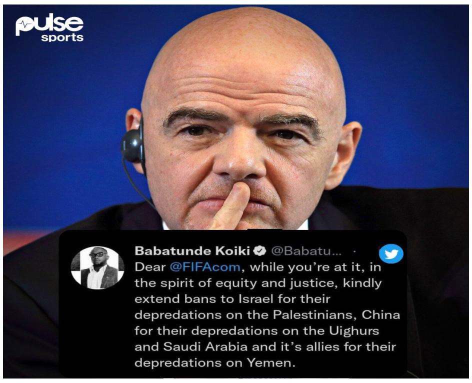
Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, pages 79–83, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-089-2_011

each language independently. The results indicate that multi-lingual BERT outperforms XLM-RoBERTa in two out of three language pairs, showing significantly higher macro average scores for both English-only and French-only data. However, the fine-tuned altered XLM-RoBERTa performs poorly in the monolingual setting, with scores less than 0.5. The findings highlight the importance of selecting appropriate models and training strategies for effective hate speech detection in cross-lingual contexts.

3 Task and Dataset Description

3.1 Hate Speech Detection

The objective of this task is to discern the presence of hate speech in text-embedded images. These images constitute the dataset Bhandari et al. (2023) utilized for this subtask and are accompanied by annotations that denote the extent of hate speech prevalence.An example of text-embedded image used in dataset is shown in Figure 1. The features of the dataset is given in the table 1.



Babatunde Koiki @Babatu... ·
Dear @FIFAcom, while you're at it, in the spirit of equity and justice, kindly extend bans to Israel for their depredations on the Palestinians, China for their depredations on the Uighurs and Saudi Arabia and it's allies for their depredations on Yemen.

Figure 1: Text-embedded image

Table 1: Features of the dataset

Field	Description
filename	name of the file with index value
text	text extracted from text-embedded images

3.2 Target Detection

The objective of this subtask is to discern the specific targets of hate speech within a given text-embedded image containing hateful content. The text-embedded images in this dataset are meticulously annotated to identify the targets of hate speech, categorized into ”community,” ”individual,” and ”organization” labels. To facilitate the detection process, the text within these images is extracted using sophisticated techniques, enabling the subsequent analysis for hate speech identification.The text-embedded images employed in this study were subjected to text extraction techniques to extract the textual content present within the images.Features of the dataset is given in table 1.

Label	Train
Hate	1,942
Not Hate	1,658
Total	3,600

Table 2: Data Distribution of Hate Speech Detection

Label	Train
Individual	823
Community	784
Organization	335
Total	1,942

Table 3: Data Distribution of Target Detection

4 Methodologies used

In this study, we employed the deep learning model transformers, specifically the ALBERT (A Lite BERT) Base v1 and Artificial Neural Network(ANN).

4.1 ALBERT Base v1

Albert base v1 (Lan et al., 2019) is a type of deep learning model, specifically an ”ALBERT” (A Lite BERT) model, designed for natural language processing tasks, such as text classification. In this case, it is being used to detect hate speech in texts. The ALBERT model uses a technique called transfer learning to understand the underlying patterns and structures in the text data. It is pre-trained on a large corpus of text data to learn the general features of language.ALBERT tokenizes the input text, breaking it down into smaller units called tokens. Each token represents a word or subword in the text.Each token is mapped to a high-dimensional vector representation called an embedding.ALBERT utilizes a self-attention mechanism to assess token relationships in the text,thus grasping dependencies and long-range associations between words within the context.The ALBERT model is further fine-tuned on a labeled dataset of texts. During the fine-tuning process, the model adjusts its parameters to make accurate predictions based on the specific characteristics of hate speech present in the training data.The architecture is shown in figure 2

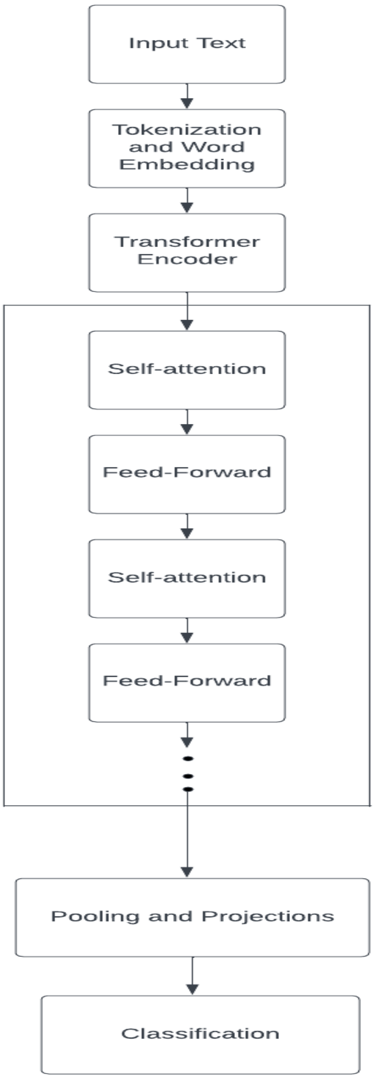


Figure 2: Architecture of Proposed System

4.2 Artificial Neural Network

Artificial Neural Networks (ANNs) have emerged as a pivotal element in the field of machine learning and artificial intelligence, owing to their ability to effectively model complex and non-linear relationships within data. Before feeding the data into the ANN, the texts are preprocessed. This includes tokenization, where each text is broken down into individual words or subwords. These words are then converted into numerical representations.The ANN is constructed using layers of interconnected artificial neurons. : The training process is where the ANN learns to detect hate speech. The training data, which consists of the numerical representations of texts and their corresponding labels, is used to adjust the internal parameters (weights and biases) of the neurons in the ANN.During training, the training data is fed into the ANN, and it performs a forward pass. This means the data flows through the layers of the network, and computations are performed to generate predictions. The predictions are then compared to the actual labels using a loss function, which measures the difference between the predicted and true labels.The architecture is shown in figure 3

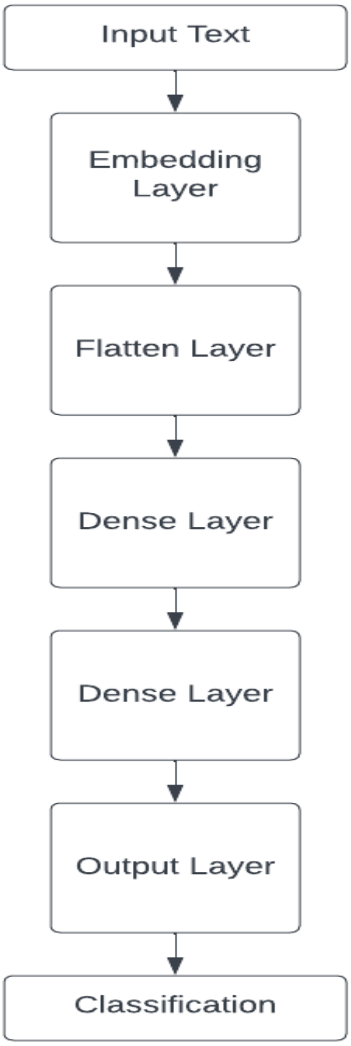


Figure 3: Architecture of Proposed System

5 Result Analysis of Hate Speech Detection Task

This section discusses about the implementation of Transformer Model and ANN with the analysis of the results using evaluation metrics.

5.1 Implementation

The implementation of the ALBERT Base v1 model for the Transformer-based classification task is achieved through the utilization of the simpletransformers library. The ClassificationModel class from the library is employed, specifying 'albert' as the model type and 'albert-base-v1' as the specific pre-trained ALBERT model variant. The model is configured to handle a binary classification and multilabel classification task. To optimize performance, several arguments are set, such as enabling input data reprocessing, disabling the use of cached evaluation features. Additionally, the model is trained for a specified number of epochs, with the option to increase this value for enhanced performance.

The ANN model is built using the Sequential API from Keras, which allows creating a sequential stack of layers.The first layer is an Embedding layer, which is used to convert the numerical tokens into dense vectors (embeddings). It maps each token to a 64-dimensional vector, which represents the meaning and context of the word in the text.The Flatten layer is used to convert the 2D tensor output from the Embedding layer into a 1D tensor, as ANN models require a 1D input.The next layer is a Dense layer with 32 units and a ReLU activation function, which introduces non-linearity and allows the model to learn complex patterns in the data.Finally, there is another Dense layer with 1 unit and a sigmoid activation function.The model is then trained on the training data for 10 epochs (iterations), with a batch size of 32.

5.2 Results

The dataset is partitioned into training and testing sets, and the evaluation results are presented in a table 4. The table contains the performance metrics and assessment outcomes for the model used in the study. The division of the dataset into training and testing sets enables to evaluate the effectiveness and generalization capabilities of their proposed hate speech detection models. The assessment table provides valuable insights into the model's performance.

Parameters	Score
Accuracy	0.7680
F1-score	0.7679
Recall	0.7681
Precision	0.7678

Table 4: Assessment of Models using Evaluation Metrics of ALBERT Base

Assessment using Artificial neural network(ANN) gave poor results as compared to ALBERT Base. The metrics are given in table 5

Parameters	Score
Accuracy	0.699
F1-score	0.733
Recall	0.729
Precision	0.738

Table 5: Assessment of Models using Evaluation Metrics of ANN

The evaluation result for the test dataset is given in table 6

Parameters	Score
Accuracy	0.7856
F1-score	0.7821
Recall	0.7806
Precision	0.7849

Table 6: Evaluation metrics of ALBERT Base for Hate Speech Detection Task

6 Result Analysis of Target Detection Task

6.1 Results

The training dataset is divided into training and testing sets to evaluate the proposed target of hate speech detection model effectively. The evaluation results, including performance metrics and assessment outcomes, are presented in a table 7 and 9.

The performance of the Artificial Neural Network (ANN) model was found to be inferior when compared to the ALBERT Base model. The evaluation metrics, presented in the table 8, clearly indicated that ALBERT Base outperformed the ANN in various aspects.

Parameters	Score
Accuracy	0.640
F1-score	0.6394
Recall	0.6401
Precision	0.6403

Table 7: Assessment of Models using Evaluation Metrics

Parameters	Score
Accuracy	0.560
F1-score	0.470
Recall	0.473
Precision	0.483

Table 8: Assessment of Models using Evaluation Metrics of ANN

7 Conclusion

We constructed an ALBERT base Model to perform hate speech detection. Preprocessing all the models with NLTK was considered essential in creating a robust model. However, accurately gauging the emotion of social media posts depends on individual perception, making it challenging for conventional models to achieve high accuracy. Another contributing factor to reduced accuracy is the imbalanced data distribution among the output class labels. To address these challenges, we plan to explore various transformer models and data augmentation techniques to enhance the performance of our hate speech detection system.

Parameters	Score
Accuracy	0.7149
F1-score	0.6805
Recall	0.6777
Precision	0.6841

Table 9: Evaluation metrics of ALBERT Base for Target Detection Task

References

Aashish Bhandari, Siddhant Bikram Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zaki Mustafa Farooqi, Sreyan Ghosh, and Rajiv Ratn Shah. 2021. Leveraging transformers for hate speech detection in conversational code-mixed tweets. *arXiv preprint arXiv:2112.09986*.

Farhan Ahmad Jafri, Mohammad Aman Siddiqui, Surendrabikram Thapa, Kritesh Rauniyar, Usman Naseem, and Imran Razzak. 2023. Uncovering political hate speech during indian election campaign: A new low-resource dataset and baselines. *arXiv preprint arXiv:2306.14764*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Surendrabikram Thapa, Farhan Ahmad Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka-Wei Lee, and Usman Naseem. 2023. Multimodal hate speech event detection - shared task 4, case 2023. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*.

Teodor Tiţa and Arkaitz Zubiaga. 2021. Cross-lingual hate speech detection using transformer models. *arXiv preprint arXiv:2111.00981*.