Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 1 (2024)

Hate Speech Classification Using Rfdt, Bilstm, and Bilstm

P.Dhineshkumar[1], Dr.A.Nithya[2]

[1]Research Scholar, Department of Computer Science, Park's College, Chinnakkarai, Tirupur 641 605, Tamilnadu.

[2]Professor, Department of Computer Science, Park's College, Chinnakkarai, Tirupur 641 605, Tamilnadu.

Abstract

In this article, we will consider how to classify hate speech on social media. To avoid conflicts between citizens, it is necessary to automatically detect the spread of hate speech and offensive words on social media. Additionally, hate speech has targets, categories, and dimensions that need to be recognized in order for government agencies to prioritize which hate speech needs immediate action. This article describes the classification of hate speech and offensive words in texts on the social media: Twitter (X platform), English, and a mixture of both, down to types, categories, and levels. Classification of hate speech multi-label texts is studied using his RFDT and BiLSTM with pre-trained BERT models.

Keywords : *Social Media, Hate Speech Classification, RFDT Classification.*

Introduction

Hate speech is a direct or indirect statement directed at an individual or group that spreads hatred and is based on something specific to that individual or group. Factors commonly used as grounds for hatred include ethnicity, religion, disability, gender, and sexual orientation. Spreading hate speech is a very dangerous act and can lead to negative consequences such as discrimination, social conflict, and even genocide against people. 'Hate speech' is an emotive concept, and there is no generally accepted definition in international human rights law. Many people would argue that they can recognise "hate speech" when they see it, but that standard is often elusive or contradictory. Therefore, Brown's [1] idea is that the concept of "hate speech" may be a complex concept consisting of two basic concepts: hatred and language. According to the author, it can be divided into two main components.

_ Hated: Intense and intense feelings of shame, hostility, and disgust towards a person or group targeted because of certain actual or perceived protected characteristics are irrational feelings. "Hate" is not just prejudice; it must be discriminatory. Hatred is different from manifested behaviour because it indicates an emotional state or opinion.

_ Speech: Any utterance that conveys an opinion or idea—conveying a subjective opinion or idea to an external audience. It can take many forms, such as written, nonverbal, visual, or artistic, and can be distributed through any medium, such as the Internet, print media, radio, or television. Beyond these two basic elements, simply put, "hate speech" is any emotional expression of hatred towards people.

Literature review

Ricardo Martins et al. [1

In this task, a vocabulary baseline is established by applying a classification method using a dataset annotated for this purpose. Characteristically, our system uses natural language processing (NLP) techniques to add emotional information to the original dataset, making it available for machine learning classification. Using

emotional information in text improves the accuracy of hate speech detection. This claim is based on the success rate, which increased from 41% in the original study to 80.64% in the test. This almost 100% improvement can be said to be the result of our proposal.

Muhammad Okky Ibrohim et.al[2] To avoid conflicts between citizens, it is necessary to automatically detect the spread of hate speech and offensive words on social media. This study uses a machine learning approach using Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest to detect attacks, including target, category, and scope detection of hate speech on Twitter in Indonesia. Multi-label text classification for textual language and hate speech detection Decision Tree (RFDT) classifiers and Binary Relevance (BR), Label Powersets (LP), and Classifier Chains (CC) as data transformation methods. The FGD results led us to the conclusion that addressing the problem of hate speech on social media is more than just finding out whether a text or document is hate speech. There are several other tasks that authorities need to complete to address hate speech issues, including: B. Identifying buzzers, thread starters, and fake accounts that spread hate speech.

Muhammad Okky Ibrohim et.al.[3] Hate speech and abusive language on social media must be detected because they can trigger conflict among citizens. Not only on social media, but HSAL also often triggers conflict in real life. This paper discusses a summary of Indonesian HSAL detection research conducted by utilising the Kitchenham systematic literature review method. Based on our summary, we found that most Indonesian HSAL research still uses the classic machine-learning approach with classic text representation features that were experimented with on the Twitter text dataset. As far as we know, there is still no HSAL detection benchmark website, even for English as the major language. Therefore, building an HSAL detection benchmark website for HSAL research in Indonesian, English, or other languages is innovative and very helpful to the community..

Shankar Biradar et.al[4] This review aims to investigate the performance of transformer models, such as IndicBERT and Multilingual Bidirectional Encoder Representation (mBERT), and transfer learning from pre-trained language models, such as ULMFiT and Bidirectional Encoder Representation (BERT), to reduce hateful content in Hinglish. I found. Furthermore, this study proposes a transformer-based interpreter and feature extraction model for deep neural networks (TIF-DNN). Experimental results reveal that our proposed model outperforms existing state-of-the-art methods for identifying hate speech in Hinglish with 73% accuracy..

Muhammad Okky Ibrohim et.al[5] In this review, we describe two methods for translating (with and without speech recognition) multilingual text classification and an untranslated method for multilingual hate speech recognition (including Hindi, English, and Indonesian) using machine learning approaches. Adapt and compare. Using word N-grammes and character N-grammes (character N-grammes) as feature extraction, several classification algorithms (classifiers) use.

Methods and materials

This section describes the methodology used in this study. This research method consists of six phases: data collection, data pre-processing, feature extraction, classification, and evaluation.

Data collecting

The data used in this study is hate speech survey data that is publicly accessible on the Kaggle website. Three datasets were acquired in raw form, and pre-processing steps were applied.

Data pre-processing

After the data is acquired, pre-processing occurs before the data is extracted. The pre-processing step helps filter noisy and irrelevant content. All duplicate tweets are filtered first, as they do not provide any information to the model. The pre-processing used in this study includes text cleaning, text normalisation, lowercase text conversion, stemming, and stop word removal. Text sanitization is done by removing usernames, URLs, RTs (retweets), the "@" character, and special characters that can affect classification performance. Additionally, the text normalisation phase converts words with insufficient or ambiguous vocabulary into words with rich vocabulary. Finally, the pre-processed tweets are converted to lowercase to avoid ambiguity.

Feature Extraction

After pre-processing the dataset, we extract multiple features from the dataset and add them to a feature vector. Tweets are raw text data. [6] Must be extracted to be trained in machine learning algorithms. The feature extractors used are frequency-term word-n-gram and acter-n-gram. A word N-gram maps words by their frequency. Meanwhile, character n-gram mapping maps characters with their frequency. [6] N-gram is a statistical modelling language for processing a text consisting of a sequence of items, where n indicates the length of the series (if n = 1 is called a Uni-gram, n = 2 is called a m, and n = 3 n = 3 is a). The tweet representation process begins with modelling the N-gram feature, including Unigram, Bigram, and Trigram.. [11]

Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical approach that weights words according to their importance in a corpus of documents. In the feature extraction phase, the keyword-based tweets are weighted with his TF-IDF score. After calculating the W weight for each document, we sort the W weights to determine the similarity between documents and keywords. [9] In this case, the second vectorization method uses TF-IDF. The weight W in the TF-IDF estimate can be found in equation 1, with t_{fij} as the quantity of words investigated in an article and Id f_j as the inverse article incidence.

$$W_{ij} = t_{ij} \times Id\ f_j \qquad (1)$$

Global Vectors (GloVe)

GloVe (Global Vectors for Word Representation) This is another popular word embedding model. [8] GloVe is a model that utilises count data while capturing general, meaningful linear substructures with a log-bilinear prediction-based method. When compared to Word2vec (with the same corpus, vocabulary, window size, and training time), GloVe consistently outperforms Word2vec. A study comparing GloVe with other word embedding models such as Continuous Bag-of-Words, Skip-Gramme, and Hellinger PCA found that not only did GloVe produce better and faster results, but GloVe also outperformed other word embedding models. Produced the best results regardless of speed compared to GloVe is superior to other models and is the best model..

Building Similarity Corpus

A corpus is created based on each word contained in the tweet data. A similarity corpus is created from each of these words using GloVe. Tweet data and a combination of both are used to find similarities between words..

Classification

This research use RFDT, BiLSTM, to classified abusive and hate speech.

Random Forest Decision Tree (RFDT)

Random Forest

Random forests are a machine learning technique used to classify large amounts of data. Random forests combine many trees in the training data to achieve a high level of accuracy. Random forest formation uses the Gini index value to determine the split and is used as a node in the following formula:

$$Gini\ (S) = \ 1 - \sum_{i=1}^{k} pi^2 \qquad (2)$$

K is the k trees, and pi is the probability that S belongs to class i. After calculating the Gini value, the next step is to calculate the Gini gain value using the following formula:

$$GiniGain\ (S) = \ Gini\ (S) - \ Gini\ (A,S) = Gini\ (S)\ \sum_{i=1}^{n} \frac{|S_i|}{|S|} Gini\ (S_i) \qquad (3)$$

where S_i is the partition of S caused by attribute A.

Decision Tree

Decision trees are supervised machine learning algorithms suitable for solving classification and regression problems. [10] The best split increases the purity of the amount obtained from the split. Gini is defined if L is a data set with j different class labels,

$$Gini\ (L) = \ 1 - \sum_{i=1}^{j} pi^2 \quad (4)$$

Where pi is the comparative occurrence of class i in L. If the dataset is split on characteristic A into two subsets with sizes and, respectively, Gini is calculated as

$$Gini\ _B(L)\ \frac{N_1}{N}\ Gini\ (L_1) + \frac{N_2}{N}\ Gini\ (L_2) \quad (5)$$

Reduction in impurity is calculated as

$$\Delta\ Gini\ (B) = Gini\ (L) - \ Gini_B\ (L) \quad (6)$$

BiLSTM

BiLSTM is a type of recurrent neural network and an improvement over the LSTM network. It contains a memory block to process the sequential information for temporal behaviour modelling. An LSTM cell consists of input gate I_t, forget gateF_t, output gate O_t and a memory cell state C_t. The input gate at timestamp t, I_t, controls the flow of information in a cell and update its state to a new value using Eq. 7 whereas forget gate decides the amount of information to be erased at time t using Eq. 8. The Eq. 9 computes the candidate cell value, $\tilde{C}_t$. Similarly, the current cell state valueC_t, the output O_t from the output gate, and the final output ht of the LSTM cell at time t are calculated using Equation 10 and Equation 11, respectively. [7] In these equations, Ft represents the input to BiLSTM at time t obtained from high-level attention, and w, b, σ, tanh represent the weight vector, bias vector, sigma function, and hyperbolic tangent function represents. Or, in addition, $\oplus$ performs element-wise multiplication.

$$I_t = \sigma(W_i.\left[h_{t-1,}F_t\right] + b_i) \quad (7)$$

$$F_t = \sigma(W_f.\left[h_{t-1,}F_t\right] + b_f) \quad (8)$$

$$\tilde{C}_t = tanh(W_c.\left[h_{t-1,}F_t\right] + b_c) \quad (9)$$

$$C_t = \ F_t \oplus \mathrm{C_{t-1}} + i_t \ \oplus \tilde{C}_t \quad (10)$$

$$O_t = \ \sigma\left(W_o\ .\left[h_{t-1,}F_t\right] + b_o\right) \quad (11)$$

$$h_t = \ O_t \oplus \tanh(C_t) \quad (12)$$

Here, we use BiLSTM instead of LSTM to capture context information in both directions. BiLSTM has a LSTM pair; the forward LSTM performs a left-to-right sequence to obtain the future context, and the backward LSTM performs right-to-left sequential information to obtain the historical context. The proposed model passes this encoded information to the attention layer and assigns variable weights.

Experimental results

Classification is a technique for developing models based on labelled datasets. This means that each record in the training dataset is assigned a class label. This model is later used to predict class labels for new or unseen data. The predictive accuracy of a classification model is its ability to accurately predict class labels for unseen data. The confusion matrix shows how often behaviour is correctly recognised and classified as a class. In a confusion matrix, results that are correctly classified into the positive class are called true positives (TP), and results that are correctly classified into the negative class are called true negatives (TN). On the other hand, the positive class is classified as false negative (FN), and the negative class is classified as positive false positive (FP). Common metrics for measuring the accuracy of a classification model are confusion matrix, precision, recall, and precision. First, a confusion matrix is created, from which all other metrics can be easily calculated.

Dataset Details

This dataset uses Twitter data (X platform) and is used to study hate speech detection. It is important to note that, due to the nature of the study, this dataset includes some texts that may be considered hate speech and some that are not. The used hate speech dataset contains 18,396 tweets with a total of 12 parameters. Run the dataset obtained from the Kaggle public website with Twitter (X Platform) to enrich the data with 5,227 new tweets.

Accuracy

Accuracy is a metric used to evaluate classification models. Informally, accuracy is the percentage of predictions that the model makes that are correct. Formally, the definition of precision is: Accuracy = number of correct predictions Total number of predictions.

Accuracy = (TP+TN) / (TP+FN +FP+TN) (13)

Recall

Recall is a metric that evaluates a model's ability to predict true positive results for each available category.

Recall = (TP) / (TP+FN) (14)

Precision

Accuracy refers to the percentage of true positives that the model cancels out. The accuracy estimate for the class is given by:

Precision = TP / (TP + FP) (15)

Method	Accuracy	Recall	Precision
RFDT	76.45	72.23	74.46
BiLSTM	69.23	68.26	67.17
RFDT+BiLSTM	73.68	70.49	71.53

Table 1: Confusion matrix on hate speech classification

Based on the results, the above table, which represented the RFDT, which provides the accuracy of the RFDT method, was 76.45, recall was 72.23, and precision was 74.46. BiLSTM method accuracy was 69.23, recall was 72.23, and precision was 74.46. The combination of RFDT and BiLSTM produced an accuracy of 73.68, and recall was 70.49 and 71.53.

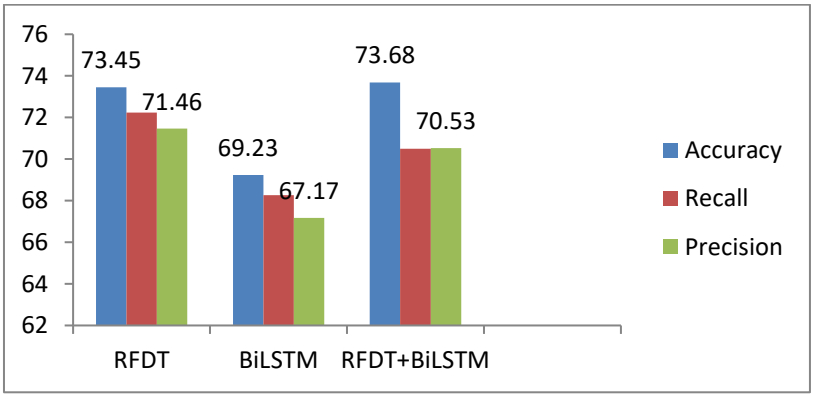

Figure 1: Confusion matrix on hate speech classification

Conclusion

This article describes the classification of hate speech and offensive words in text in the Twitter social media dataset. The models proposed to classify insults and hate speech are RFDT and BiLSTM. In addition, feature extraction, such as word n-grammes and character n-grammes, is used for classification. In this experiment, we can conclude that TF-IDF is the best choice as the term weighting scheme. Based on the experimental results and the above discussion, the researchers concluded that the hierarchical approach can improve the classification performance. To the best of our knowledge, this was one of the first experiments to detect hate speech on social media in a scenario that included examples without hate speech.

Future work

Future work on classification studies in the field of abusive language and hate speech will use pre-trained BERT models and multilingual BERT models.

References

[1] Ricardo Martins, Marco Gomes, "Hate speech classification in social media using emotional analysis", 2019.
[2] Shakir Khan, Mohd Fazil, "BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection", 2022.
[3] Muhammad Okky Ibrohim, "Hate speech and abusive language detection in Indonesian social media: Progress and challenges", 2023.
[4] Shankar Biradar, Sunil Saumya, "Fighting hate speech from bilingual hinglish speaker's perspective, a transformer- and translation-based approach", 2022.
[5] Muhammad Okky Ibrohim, Indra Budi, "Translated vs Non-Translated Method for Multilingual Hate Speech Identification in Twitter", Vol.9 (2019) No. 4.
[6] Faizal Adhitama Prabowo, Muhammad Okky Ibrohim, "Hierarchical Multi-label Classification to Identify Hate Speech and Abusive Language on Indonesian Twitter", 2019.
[7] . Arup Baruah, Ferdous Ahmed Barbhuiya, "ABARUAH at SemEval-2019 Task 5: Bi-directional LSTM for Hate Speech Detection", 2019.
[8] Hind Saleh, Areej Alhothali & Kawthar Moria, "Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model", 2023.
[9] Febiana Anistya, Erwin Budi Setiawan, "Hate Speech Detection on Twitter in Indonesia with Feature Expansion Using GloVe", Vol. 5 No. 6 (2021).
[10] Bahzad Taha Jijo, Adnan Mohsin Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning", Vol. 02, No. 01, pp. 20 – 28 (2021).
[11] Sandy Kurniawan, Indra Budi, "Indonesian Tweets Hate Speech Target Classification using Machine Learning", 2021.