Trigger Detection in Social Media Text

Asha Hegde[1], Fazlourrahman Balouchzahi[2], Kavya G[3] and Hosahalli Lakshmaiah Shashirekha[4]

Abstract

Trigger detection in social media content refers to the process of identifying and flagging content that may potentially trigger negative emotions or psychological responses in individuals. The study contributes to the Trigger Detection shared task at PAN@CLEF 2023, which aims to automatically assign violence trigger warnings to narratives. Using pretrained Global Vectors for Word Representation (GloVe) embeddings to train the Long Short Term Memory (LSTM) model, the proposed models achieved macro F1 scores of 0.57 and 0.048 on the Validation and Test sets respectively.

Keywords

Trigger detection, Social media text, Pretrained word vectors, Long Short Term Memory

1. Introduction

Social media has become an integral part of our daily lives, allowing us to connect to others, share information with others, and to engage in online communities. Moreover, as a rich source of user-generated content, social media presents a vast amount of textual data that can be leveraged through text processing techniques to gain insights, understand users' behaviors, and enhance user experiences [1]. However, the content such as folk tales, fairy tales, children's and youth fiction and so on, which often contain violence and cruelty [2], shared on social media platforms can sometimes be disturbing or triggering for some individuals. Such situation could be avoided by generating warnings which alarm the individuals with prior knowledge of sensitive topics that could evoke emotional or psychological distress, allowing them to make an informed decision about whether to engage with such content or not. Such warnings are called as trigger warnings and these warnings which appeared for discussion way back in the early 2000's in online communities (Tumblr and Live-Journal) stressed the need for such warnings [3].

Trigger warnings are brief notices that are used to alert readers about potentially distressing or triggering content that may be present in the textual content such as articles, books, or online posts. By highlighting potentially distressing content, trigger warnings allow readers, especially

CLEF 2023 – Conference and Labs of the Evaluation Forum, 18-21 September 2023, Thessaloniki - Greece

hegdekasha@gmail.com (A. Hegde); fbalouchzahi2021@cic.ipn.mx (F. Balouchzahi); kavyamujk@gmail.com (K. G); hlsrekha@mangaloreuniversity.ac.in (H. L. Shashirekha)

https://sites.google.com/view/asha-hegde/home (A. Hegde); https://sites.google.com/view/fazlfrs/home (F. Balouchzahi); https://mangaloreuniversity.ac.in/shashirekha/ (H. L. Shashirekha)

CEUR Workshop Proceedings (CEUR-WS.org)

those who have been through trauma, mental health challenges, or other sensitivities, to make informed decisions about what they consume to protect their well-being and emotional safety [4]. The purpose of trigger warnings is to create more inclusive and considerate environment, promoting individual autonomy and reducing the risk of distress or discomfort for individuals who may encounter triggering content unexpectedly. However, it is important to note that the use of trigger warnings is a subject of ongoing debate and opinions on their effectiveness and implementation may vary [2, 5].

The analysis and detection of trigger warnings in Natural Language Processing (NLP) is a relatively underexplored area despite its importance. To address this gap, "Trigger Detection" shared task[1] in PAN@CLEF 2023 invites researchers to develop models to automatically assign violence trigger warnings to the given comments/posts. The task recognized the potential impact of violent content on individuals and sought to explore the feasibility of NLP techniques in identifying and flagging such triggers. This initiative highlights the growing recognition for automated methods to assist in the identification and labeling of potentially disturbing and triggering content. The outcomes of this task have the potential to inform future research and development in NLP applications related to trigger warnings and content moderation. To address the challenges of the trigger detection, in this paper, we describe Long Short Term Memory (LSTM) model utilizing Global Vectors for Word Representation (GloVe) pretrained embeddings to represent the text data.

The rest of the paper is arranged as follows: a review of related work is included in Section 2, and the methodology is discussed in Section 3. Experiments and results are described in Section 4 followed by concluding the paper with future work in Section 5.

2. Related Work

The rapid growth of social media platforms has resulted in an increase in the volume of violent, offensive, and hate speech content being shared. This has created a pressing need for tools and models that can efficiently detect and address such unwanted content. Early detection of such content is crucial for maintaining the health and integrity of social media platforms, ensuring user safety, and fostering positive online communities. The development and deployment of robust detection systems can help mitigate the negative impact of harmful content, allowing platforms to take appropriate actions and create a healthier social media environment. Few of the relevant works which have discussed by the researchers to identify unhealthy content and trigger warnings in social media text are described below:

Sahoo et al. [6] discussed multilingual event and argument trigger detection and classification problem under the sequence labeling framework. They collected and annotated disaster related news data crawled from the online news portal in different low-resource Indian languages (Hindi, Bengali, and multilingual) for their experiments. The authors trained LSTM models considering word embeddings from FastText pretrained word vectors as features and their models obtained F1 scores of 0.42, 0.55, and 0.61 for Hindi, Bengali and multilingual texts respectively. Wiegmann et al. [7] created the Webis Trigger Warning Corpus 2022 with 41 million fan-fiction works data categorised into 36 different warning tags and the authors considered this problem as

[1]https://pan.webis.de/clef23/pan23-web/trigger-detection.html#related-work

multi-label classification task. The authors proposed four classification models (Support Vector Machine (SVM), Extreme Gradient Boost (XGBoost), Robustly Optimized Bidirectional Encoder Representations from Transformers (Roberta), and Long Forms (LF)) to benchmark their dataset and achieved a maximum F1 score of 0.47 for SVM model. Hegde and Shashirekha [8] describe the learning models submitted to "Sentiment Analysis and Homophobia Detection of YouTube Comments in Code-Mixed Dravidian Languages" shared task at Forum for Information Retrieval Evaluation (FIRE) 2022. Using preprocessing of converting emojis to text and removal of digits and stopwords, these models make use of Dynamic Meta Embedding (DME) to train LSTM model to perform Sentiment Analysis (SA) and detect Homopobic/Transphobic content in code-mixed Dravidian languages viz. Kannada, Malayalam, and Tamil. These models obtained 6^{th}, 4^{th}, and 9^{th} ranks for Tamil, Malayalam, and Kannada respectively in Task A and 1^{st}, 4^{th}, 1^{st}, and 5^{th} ranks for Tamil, English, Tamil-English, and Malayalam texts respectively in Task B.

Code-Mixing Offensive Language Identification (COOLI)-Ensemble and COOLI-Keras models are developed by Balouchzahi et al. [9] to identify offensive language in Dravidian languages at "Offensive Languages in Dravidian Languages 2021" shared task. Using term frequency of character sequences and words they trained ensembled Machine Learning (ML) classifiers (Logistic Regession (LR), eXtreme Gradient Boosting (XGB), and Multi Layer Perceptron) and Keras sequential model. COOLI-Ensemble model outperformed the other model securing 1^{st}, 6^{th}, and 4^{th} ranks for code-mixed Malayalam-English, Tamil-English, and Kannada texts respectively. Balouchzahi and Shashirekha [10] proposed three distinct models: i) ensemble of ML classifiers (RFC, LR, and Support Vector Classification (SVC)), ii) Transfer Learning (TL) classifier using Universal Language Model Fine-tuning (ULMFiT) model, and iii) ensemble of ML-TL models, to the HASOC 2020 shared task for identifying hate speech and offensive content in English, German and Hindi languages. The TL and ML-TL models exhibited macro F1 scores of 0.2517 and 0.4979 for English securing 5^{th} and 21^{st} ranks in Subtask A and Subtask B respectively. The ensembled ML classifier exhibited macro F1 score of 0.5044 for German securing 11^{th} rank in Subtask A and ML-TL model for Hindi language exhibited macro F1 score of 0.5182 securing 8^{th} rank in Subtask B.

Wolska et al. [2] created a labelled corpus for trigger warning assignment by extracting narrative works hosted on Archive of Our Own (AO3) from a well known fan-fiction site. The authors focus on assigning a most frequent type of trigger warning i.e., violence and carried out a binary classification task at document-level. Further, they categorized both the corpora into four categories: small easy, small difficult, large easy, and large difficult based on the size and the complexity of the corpora. In their study, the authors utilized SVM model trained on Term Frequency-Inverse Document Frequency (TF-IDF) features of word uni-grams and bi-grams, as well as Bidirectional Encoder Representations from Transformers (BERT) models, to detect trigger warnings. They found that the SVM model outperformed the BERT model, achieving F1 scores of 0.798, 0.676, 0.780, and 0.686 for small easy, small difficult, large easy, and large difficult corpora respectively, indicating its superior performance in trigger warning detection across different dataset sizes and difficulty levels.

Trigger detection in social media text is rarely explored. The evolving nature of the user generated text on social media platforms, trigger content, and the need for nuanced understanding of context present ongoing challenges. Continued research and innovation in this area are essential to enhance the effectiveness and accuracy of detection systems, enabling social media

platforms to stay ahead of evolving threats and ensure the well-being of their users.

3. Methodology

The methodology employed in this study is visualized in Figure 1 and the steps involved in the methodology are briefly described below:

3.1. Preprocessing

To prepare the textual data for analysis, various preprocessing functions (stripping HTML tags, removing accented characters, expanding contractions, lemmatizing, stemming, removing special characters and digits, converting text to lowercase, and removing stopwords) are used.

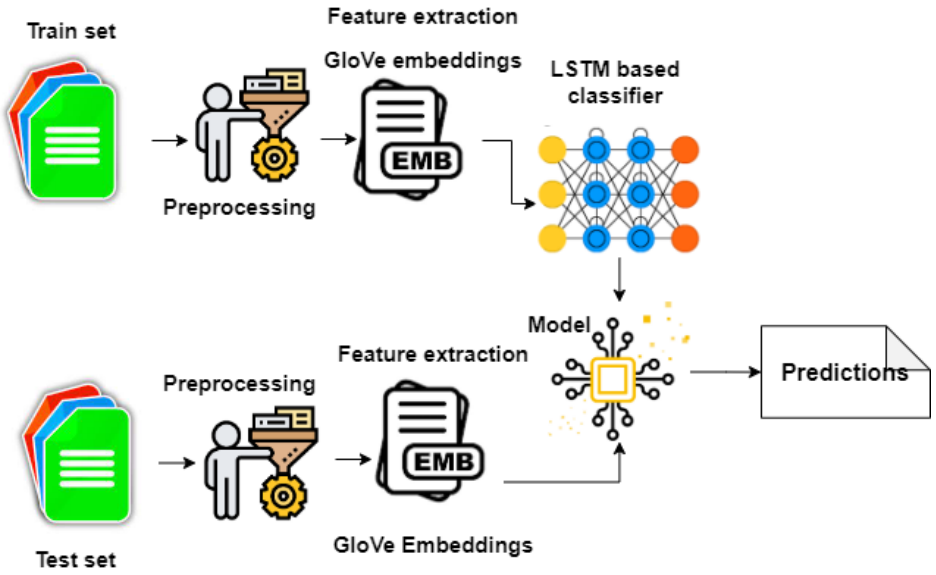

Figure 1: Framework of the proposed model

3.2. Text Representation

GloVe[2] embeddings are a type of word representation model used in NLP [11]. These are dense vector representations that capture semantic and syntactic information about words based on their distributional properties in a given corpus. GloVe embeddings are available with 25, 50, 100, 200, and 300 dimensions and this work utilizes GloVe embeddings with dimension 100. While GloVe embeddings are effective in mapping a large number of words present in the training set, there is still a possibility of encountering Out-Of-Vocabulary (OOV) words that are not included in the mapping. The proposed model is trained using the GloVe features and OOV words are handled with '0' vectors of size 100.

[2]https://github.com/stanfordnlp/GloVe

Table 1
Dataset statistics

Dataset	# of comments
Train set	3,07,102
Development set	17,104
Test set	17,040

3.3. Model construction

LSTM is a type of Recurrent Neural Network (RNN) architecture that has achieved remarkable success in various NLP tasks. Its ability to effectively capture and model sequential dependencies makes it well-suited for applications such as language modeling, sentiment analysis [8], machine translation [12], and named entity recognition [13], among others. LSTM's capacity to handle long-term dependencies and its ability to retain and forget information over extended sequences have made it a popular choice in the NLP community.

The proposed LSTM model that aims to detect trigger content in social media text utilizes the Tokenizer[3] class from the Keras library to tokenize the text data. In order to ensure uniform input dimensions to train the classification model as per the requirement of the model, the tokenized sequences are padded or truncated to a fixed length of 5,000. The proposed LSTM model is initialized by defining an input layer that expects the maximum length of the input sequences. Further, an embedding layer using the Embedding function that maps the input sequences to dense vector representations is created. An LSTM layer containing 128 hidden units is constructed and is applied to the embedding_layer input. Eventually, the final dense layer is created with 32 units (as the task contains 32 labels) and a sigmoid activation function to make the target prediction. It takes the output of the LSTM layer as input and performs non-linear transformations on it.

4. Experiments and Results

The organizers of "Trigger Detection" task provided Train, Development, and Test sets and the statistics of the dataset is shown in Table 1. This dataset exhibits a long-tailed frequency distribution indicating that there are few labels that occur frequently, while the majority of labels are increasingly rare. This can pose challenges for training the learning models, as they struggle to accurately learn from the infrequent labels due to their limited representation in the dataset. Glimpse of the label distribution in the dataset is shown in Figure 2.

The proposed model is trained for 10 epochs with a batch size of 128 to minimize the binary cross entropy loss function and is evaluated on the Test set provided by the shared task organizers. Performances of the models submitted by all the participants of the shared task are ranked based on macro F1 score. The proposed model is placed 7^{th} in the task with macro F1 scores of 0.57 and 0.048 on Validation and Test sets respectively.

[3]https://www.tensorflow.org/api_docs/python/tf/keras/ preprocessing/text/Tokenizer

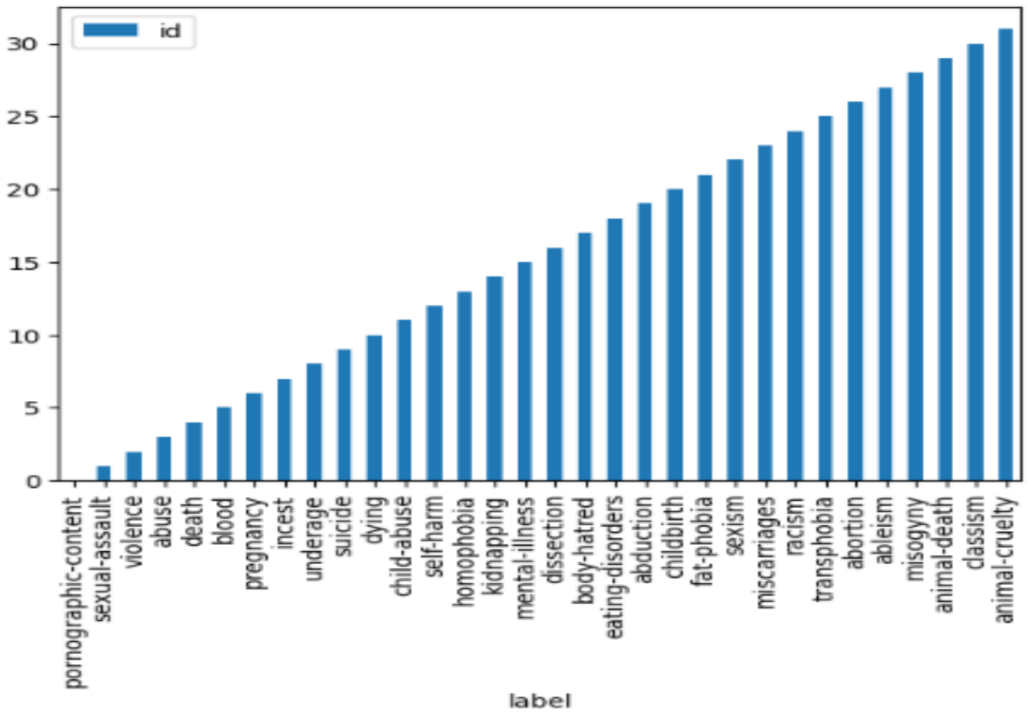

Figure 2: Classwise distribution of the labels in the Train set

5. Conclusion

This paper describes the model submitted to "Trigger Detection" shared task in PAN@CLEF 2023 to detect trigger content in social media text. Utilizing LSTM model trained with GloVe embeddings, the proposed model achieved macro F1 scores of 0.57 and 0.048 on the Validation and Test sets respectively, indicating its effectiveness in detecting and classifying triggering content. Various feature extraction techniques will be explored further to enhance the performance of ML models to identify the triggering content.

References

[1] S. Butt, S. Sharma, R. Sharma, G. Sidorov, A. Gelbukh, What Goes on Inside Rumour and Non-rumour Tweets and their Reactions: A Psycholinguistic Analyses, in: Computers in Human Behavior, 2022, p. 107345.

[2] M. Wolska, C. Schröder, O. Borchardt, B. Stein, M. Potthast, Trigger Warnings: Bootstrapping a Violence Detector for FanFiction, in: arXiv preprint arXiv:2209.04409, 2022.

[3] E. J. Knox, Trigger Warnings: History, Theory, Context, 2017.

[4] A. Hegde, S. Coelho, A. E. Dashti, H. Shashirekha, MUCS@ Text-LT-EDI@ ACL 2022: Detecting Sign of Depression from Social Media Text using Supervised Learning Approach,

in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, 2022, pp. 312–316.

[5] A. Hegde, M. D. Anusha, H. L. Shashirekha, Ensemble based machine learning models for hate speech and offensive content identification, in: Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org, 2021.

[6] S. K. Sahoo, S. Saha, A. Ekbal, P. Bhattacharyya, A Multi-task Model for Multilingual Trigger Detection and Classification, in: Proceedings of the 16th International Conference on Natural Language Processing, 2019, pp. 160–169.

[7] M. Wiegmann, M. Wolska, C. Schröder, O. Borchardt, B. Stein, M. Potthast, Trigger Warning Assignment as a Multi-Label Document Classification Problem, in: Genre, 2022, pp. 90–5.

[8] A. Hegde, H. L. Shashirekha, Leveraging Dynamic Meta Embedding for Sentiment Analysis and Detection of Homophobic/Transphobic Content in Code-mixed Dravidian Languages, in: Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org, 2022.

[9] F. Balouchzahi, B. Aparna, H. L. Shashirekha, MUCS@ DravidianLangTech-EACL2021: COOLI-code-mixing Offensive Language Identification, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 323–329.

[10] F. Balouchzahi, H. L. Shashirekha, LAs for HASOC-Learning Approaches for Hate Speech and Offensive Content Identification, in: FIRE (Working Notes), 2020, pp. 145–151.

[11] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, in: arXiv preprint arXiv:1301.3781, 2013.

[12] A. Hegde, I. Gashaw, H. L. Shashirekha, MUCS@mixwmt-Machine Translation for Dravidian Languages using Stacked Long Short Term Memory, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, 2021, pp. 340–345.

[13] H. A. Nayel, H. L. Shashirekha, H. Shindo, Y. Matsumoto, Improving Multi-word Entity Recognition for Biomedical Texts, in: arXiv preprint arXiv:1908.05691, 2019.