IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HATE SPAN DETECTION IN SOCIAL MEDIA CONTENT

[1]Alisha Kamalapurkar , [2] Rashmi Rachh, [3], Sanjana Kavatagi
[1]PG Student, [2]Associate Professor, [3]Research Scholar
[1,2,3] Department of Computer Science and Engineering,
[1,2,3] Visvesvaraya Technological University, Belagavi, Karnataka, India,

Abstract: In the era of digital communication, social media platforms have become vital, facilitating global connectivity. However, they also confront severe challenges, such as the proliferation of hate speech and offensive content. Hate speech poses substantial risks by targeting individuals or groups based on attributes like religion, race, or gender. Addressing this issue necessitates a combination of technological, community-driven, and policy-oriented solutions. Due to the lengthy posts, it would be beneficial to identify the specific span of text containing hateful content to assist site moderators with removing hate speech. This paper investigates the application of machine learning and deep learning techniques, specifically Bidirectional Long Short-Term Memory (BiLSTM) networks, for automating the identification of hate speech and offensive content. Using BiLSTM in conjunction with GloVe embedding's, we achieved a considerable F1-score of hate span detection, marking a significant improvement over previous methods. Our automated system represents a critical step toward a safer and more inclusive online environment. By seamlessly integrating machine learning and deep learning, it enhances our ability to detect and combat hate speech effectively. This research underscores the importance of technological advancements in addressing complex social challenges within the digital realm. While significant progress has been made, ongoing efforts in technological innovation, community engagement, and policy development are necessary to create a more respectful and harmonious online space.

Index Terms **- hate speech, hate span, embedding's, Bi-LSTM.**

I. INTRODUCTION

Digital platforms used by people, groups, and organizations to share and exchange material are referred to as social media. These platforms make it possible to interact with one another, communicate, and create content utilizing text, photos, videos, and links. Modern communication is increasingly dominated by social networking, which has an impact on how individuals interact, share experiences, express ideas, and participate globally. Applications of social media include communication, networking, research, data analysis, relationship maintenance, and social awareness promotion [1].

Social media does, however, come with drawbacks, such as cyberbullying, radicalization online, and content regulation. Insensitive language that attacks people or groups based on characteristics like race, religion, or gender is a serious problem [3]. It may take many different forms across platforms, ranging from vulgar language to overt threats of violence Technology, community involvement, regulations, and awareness campaigns must all be used to combat hate speech [4]. Promoting polite dialogue, educating users about its effects, enabling bystanders to report it, and enhancing content moderation systems using AI are all worthwhile endeavours. Given the abundance of material, it is essential to automate the identification of hate speech, with deep learning playing a key role.

Sequential data, including text and time series, is a strong suit for recurrent neural networks (RNNs), including LSTM. In order to better capture long-term relationships, LSTM solves the vanishing gradient problem. By processing input both forward and backward, bidirectional LSTM enhances context understanding even further.

Textual information may be captured using a variety of text representation approaches, including n-grams, TF-IDF, GloVe embedding's, and Word2Vec. The decision is based on the task's parameters and the data's properties. By giving words numerical representations, these strategies improve the quality of the input for machine learning models.

As an example, take into account an automatic text categorization system that makes use of a Bidirectional LSTM model with pre-trained GloVe features. The procedure involves loading of the data, pre-processing, feature extraction from the TF-IDF, label preparation, model construction, and integration of GloVe embedding's, training, assessment, and analysis. Combining machine learning and deep learning approaches to strengthen natural language understanding and classification tasks, the objective is to improve the accuracy of categorizing pertinent text spans in a given context.

II. LITERATURE REVIEW

Various methods for automated identification and classification are used to identify hate speech. It has been customary to use binary categorization, such as for classifying racist and sexist tweets [6]. According to several research, there should be many labels used to categories hate speech, offensive language, and non-hateful communications [5]. Logistic regression and support vector machines are examples of common machine learning classifiers [7]. Both methods use word or character n-grams, while TF-IDF prioritizes useful words [8].

Researchers have employed a variety of methods, including as deep neural networks like CNNs, RNNs (LSTM, GRU), and transformers like BERT [6] .RNNs acquire sequence information, CNNs gather local text properties, and transformers employ attention methods to grasp word relationships [8]. Combining these networks has led to improved results [9]. As input features, word embedding's like Word2Vec and Fast Text are employed [12].

With most training and testing done on lone datasets like Twitter, generalization is still difficult [3]. We have investigated multi-platform classifiers trained on various sources [15]. Investigative work is being done on multi-platform classifiers trained on diverse platforms including Twitter, dark net hacking, and extreme forums [16].

Two objectives from SemEval-2020 and SemEval-2021 concentrate on recognizing offensive language in the context of span identification. Large pertained transformer models, such as BERT and RoBERTa, were used in the solutions [7] [19] [18]. Sequence labelling has been done using conditional random fields (CRFs) [20]. Success has been gained using complex neural networks that combine language characteristics and previously trained models [21]. The complexity of semantic comprehension makes SE activities tough. The HateXplain dataset [23] examples of hate speech and their reasons. Bi-RNN and BERT were used to create attention vectors for justifications [23]. The effectiveness of several span identification methods was assessed, with F1 values indicating difficulty [18] [22].The identification of a new data corpus for hate speech is suggested.

Given the difficulty of the problem, hostile span removal is examined using the described methodologies' span prediction and sequence labelling algorithms [18]. Identification of hate speech is considered difficult but may be easier than that of propaganda [22]. It is more complex than just toxicity detection to identify hostile spans since it demands deep semantic comprehension.

III. RESEARCH METHODOLOGY

Dataset Description

The dataset is provided from the SemEval-2021 Task 5: Toxic Spans Detection dataset, which contains both training and test sets for the task of identifying toxic spans (words or phrases) within text posts: Dataset Overview: The dataset is used for the SemEval-2021 Task 5, specifically for the task of Toxic Spans Detection. It includes two main parts for each post: the content of the post and the spans denoting the toxic words or phrases in the post. Which presumably provides a detailed breakdown of the dataset. Train Dataset, there should be columns indicating the content of posts and the corresponding toxic spans, which are represented as character indices within the text. Training Dataset: The training dataset consists of 7940 rows or examples. Within these 7940 rows, there are 4438 unique toxic spans in the text. The maximum frequency of a specific span appearing in the text is 485. This means that a particular toxic word or phrase occurred 485 times across different posts in the training set. In Test Dataset, which is used for evaluating the performance of toxic spans detection models. It consists of two columns: "Spans" and "Text. "Test Dataset: The test dataset consists of 2001 rows or examples. In this dataset, there are 1034 unique toxic spans in the text. The maximum frequency of a specific span appearing in the text is 394. This indicates that a particular toxic word or phrase occurred 394 times across different posts in the test set. To provide further analysis or insights into the dataset, it would be helpful to know more about the specific format of the data, the task's goals, and any additional statistics or patterns that were observed during the analysis. This dataset appears to be focused on the problem of detecting and categorizing toxic spans within text, which is important for various natural language processing tasks, including content moderation and sentiment analysis. The goal of the SemEval-2021 Task 5 is to develop models and algorithms that can accurately detect and categorize toxic spans within text, which is a crucial task for addressing issues related to online harassment and hate speech. Researchers and practitioners use such datasets to train and evaluate models designed to automate the detection of toxic content in various online platforms [23].

Data pre-processing

In the context of the SemEval-2021 Task 5 dataset on Toxic Spans Detection, several crucial pre-processing steps are employed to prepare the text data for analysis and modelling. Initially, the toxic spans, represented as character indices within the text, are transformed into actual words or phrases. This conversion facilitates easier pinpointing of toxic content within the text. To handle the informal language and symbols often found in social media posts, the Tweet Tokenizer from the Natural Language Toolkit (nltk) is applied to segment the text content. This tokenizer is specifically designed to cater to the unique characteristics of social media language. Lowercasing of all text content is performed to standardize the text and ensure that words with different capitalizations are treated uniformly. This standardization reduces the dimensionality of the vocabulary, which is beneficial for subsequent analysis and modelling.

Text cleaning steps involve the removal of URLs using regular expressions, as these often do not contribute to the text's meaning and can be considered noise. Additionally, punctuation marks such as periods, commas, and exclamation points that are not essential for categorization are eliminated, simplifying the text further. Although not explicitly mentioned, it is common practice to remove stopwords, which are common words like "and," "the," and "is" that typically do not provide meaningful information for categorization tasks. Stopword removal helps reduce noise in the data. Tokenization is employed to break down the text into individual words or tokens, a crucial step for subsequent text vectorization techniques like TF-IDF. Lastly, stemming and lemmatization are applied to reduce words to their base or root form, standardizing the language and further reducing vocabulary dimensionality. These pre-processing steps are essential in natural language processing (NLP) tasks to clean and optimize text data for subsequent modelling. The specific steps chosen should align with the characteristics and objectives of the NLP task being undertaken, ensuring that the data is well-suited for analysis and modelling purposes.

System architecture

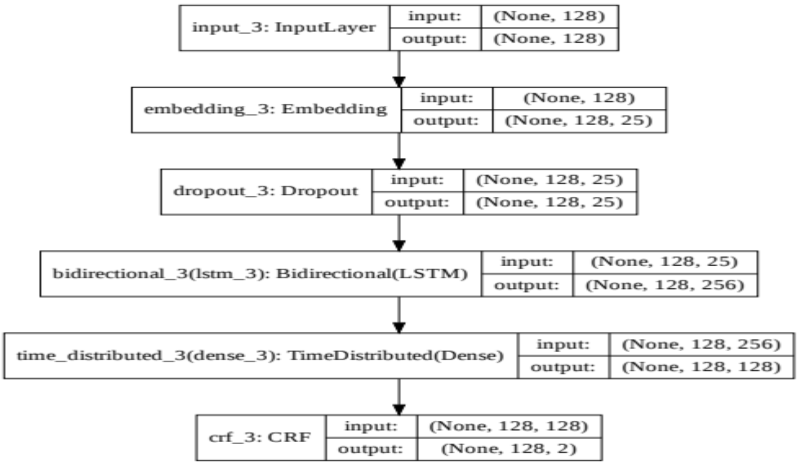


Fig: the bilstm-crf model architecture.

[22]The integration of the BiLSTM-CRF model and the Toxic BERT model represents a comprehensive approach to identifying and categorizing harmful content within documents or comments.

The BiLSTM-CRF model serves the primary purpose of detecting toxic words or phrases within the text. It operates in several layers: first, it employs pre-trained word embedding's from "GloVe" to represent words, allowing it to capture the semantic meaning of individual words. Then, the Bidirectional Long Short-Term Memory (BiLSTM) layer processes these word embedding's sequentially, capturing contextual information by considering the surrounding words. Finally, a Conditional Random Field (CRF) layer is applied to model label dependencies and determine the likelihood of output labels. This results in a binary vector that indicates whether each word is harmful or not.

On the other hand, the Toxic BERT model, based on the BERT architecture, is designed for classifying entire comments as either toxic or non-toxic. It has been pre-trained on the Jigsaw's Toxic Comments Classification Challenge dataset, making it adept at recognizing toxic or abusive language patterns.

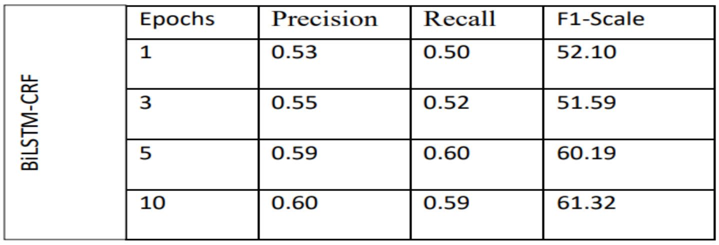
The combination of these models is where their complementary strengths shine. The BiLSTM-CRF model identifies hazardous spans or sequences of words within a given post, providing granular information about where the toxicity lies. Meanwhile, the Toxic BERT model assesses the overall toxicity of the entire post, allowing it to classify posts as hazardous or non-toxic. If a post is determined to be non-toxic, the classification model returns an empty span, effectively signaling that the entire post is safe. However, if a post is labeled as hazardous, the hazardous spans identified by the detection model are retained.

Ultimately, this combined architecture enables a more thorough and nuanced approach to handling harmful content in documents or comments. It leverages the strengths of both models to provide a comprehensive assessment, making it a valuable tool for content moderation and ensuring a safer online environment.

IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive

Table1: results obtained for bilstm-crf model



	Epochs	Precision	Recall	F1-Scale
BiLSTM-CRF	1	0.53	0.50	52.10
	3	0.55	0.52	51.59
	5	0.59	0.60	60.19
	10	0.60	0.59	61.32

The model's performance across different epochs:

1. Precision and Recall Balance: As the number of epoch's increases, the precision values tend to improve, indicating a higher proportion of correctly classified positive predictions (spans). Simultaneously, recall values fluctuate; initially, they are lower, indicating missed positive instances, but then rise, suggesting that more positive instances are being detected as the model refines its learning.
2. F1 Score Trend: The F1 score, which balances precision and recall, displays a somewhat inconsistent trend. It appears to peak around epoch 5, suggesting that at this point, the model strikes a better balance between accurate positive predictions and effective identification of all actual positive instances.
3. Epoch Impact: While both precision and recall may improve initially with more epochs, the ultimate goal is to achieve a balanced and high F1 score. The tradeoff between precision and recall is crucial, as reflected in the F1 score, which helps to assess overall model effectiveness.

REFERENCES

[1] Biradar, Shankar, Sunil Saumya, and Arun Chauhan. &quot;Fighting hate speech from bilingual hinglish speaker's perspective, a transformer-and translation-based approach.&quot; Social Network Analysis and Mining 12.1 (2022): 87.

[2] S. Kavatagi and R. Rachh, &quot;A Context Aware Embedding for the Detection of Hate Speech in Social Media Networks,&quot; 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Pune, India, 2021, pp. 1-4, doi:10.1109/SMARTGENCON51891.2021.9645877.

[3] Biradar, Shankar, et al. &quot;Pradvis vac: A socio-demographic dataset for determining the level of hatred severity in a low-resource Hinglish language.&quot; ACM Transactions on Asian and Low-Resource Language Information Processing (2022).

[4] Biradar, S., Saumya, S. &amp; chauhan, A. Fighting hate speech from bilingual hinglish speaker's perspective, a transformer- and translation-based approach.. Soc. Netw.Anal. Min. 12, 87 (2022). https://doi.org/10.1007/s13278-022-00920-w

[5] Shankar Biradar, Sunil Saumya, and Arun Chauhan. &quot;mBERT based model for identification of offensive content in south Indian languages.&quot; Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation (Online). CEUR. 2021.

[6] Kavatagi, S., Rachh, R., &amp; Mulimani, M. (2022). VTU_BGM at CheckThat! 2022: An Autoregressive Encoding Model for Detecting Check-worthy Claims.

[7] UN. (2020). https://www.un.org/en/genocideprevention/documents/UN

[8] de Gibert O., Perez N., García-Pablos A. and Cuadros M. (2018). Hate speech statistics collected from a forum for white supremacy. In Proceedings 2nd Workshop on Abusive Language Online. Association for Computational Linguistics, pp. 11–20

[9] Waseem Z. and Hovy D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In Proceedings of the NAACL Student Research Workshop. Association for Computational Linguistics, pp. 88–93.

[10] Davidson T., Warmsley D., Macy M. and Weber I. (2017). Detecting hate speech automatically and the issue with offensive language.The International AAAI Conference on Web and Social Media Proceedings, vol. 11, pp. 512–515.

[11] Kwok I. and Wang Y. (2013). Locate the hate: Detecting Tweets against blacks. In Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press, pp. 1621–1622.

[12] Devlin J., Chang M., Lee K. and Toutanova K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pp. 4171–4186.

[13] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser ŁL. and Polosukhin I. (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., pp. 6000–6010.

[14] Zhang Z., Robinson D. and Tepper J.A. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In Proceedings of The Semantic Web, pp. 745–760.

[15] Mozafari M., Farahbakhsh R. and Noël C. (2019). A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. In International Conference on Complex Networks and Their Applications, pp. 928– 940.

[16] Zhu Y., Kiros R., Zemel R., Salakhutdinov R., Urtasun R., Torralba A. and Fidler S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). IEEE Computer Society, pp. 19–27.

[17] Bengio Y., Ducharme R., Vincent P. and Janvin C. (2003). A neural probabilistic language model. Journal research on machine learning 3, 1137– 1155.

[18] Mikolov T., Chen K., Corrado G. and Dean J. (2013). Efficient Estimation of Word Representations in Vector Space. ICLR Workshop.

[19] Bojanowski P., Grave E., Joulin A. and Mikolov T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5, 135–146.

[20] Bruwaene D.V., Huang Q. and Inkpen D. (2020). A multi-platform dataset for detecting cyberbullying in social media. Language Resources and Evaluation 54, 851–874.

[21] Corazza M., Menini S., Cabrio E., Tonelli S. and Villata S. (2019). Crossplatform evaluation for Italian hate speech detection. In CLiC-it 2019 – 6th Annual Conference of the Italian Association for Computational Linguistics, vol. 2481.

[22] Da San Martino G., Barrón-Cedeño A., Wachsmuth H., Petrov R. and Nakov P. (2020). SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. published in The Fourteenth Workshop on Semantic Evaluation Proceedings. Comité international de linguistique computation, pp. 1377–1414.

[23] John Pavlopoulos, Leo Laugier, Jeffrey Sorensen, and ´ Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection (to appear). In Proceedings of the 15th International Workshop on Semantic Evaluation