Identifying and Detecting Toxic Spans

Abhishek Sushil
CSAI, IIIT Delhi
abhishek21441@iiitd.ac.in

Ayush Srivastava
CSAI, IIIT Delhi
ayush21457@iiitd.ac.in

Kartik Gupta
CSE, IIIT Delhi
kartik21056@iiitd.ac.in

Manas Narang
CSAM, IIIT Delhi
manas21473@iiitd.ac.in

April 25, 2024

Contents

1 Introduction

With the expansion of social media users around the globe, these sites have become a platform for discussions surrounding several sensitive subjects such as politics, race relations etc. Many such conversations often derail into the territory of hate-speech. Curbing hate-speech whilst not censoring legitimate arguments has thus emerged as a new problem.
The sheer volume of data in general makes it difficult for human moderators to scan for hate-speech directly. Thus our project aims to ease this burden by not only classifying text as hate-speech but also identifying the relevant sections (or spans).
Our work first applies a filter approach to detect the text that actually qualify as hate-speech, i.e. a classification task. Then it regresses to a supervised sequence labelling task as we identify the specific section(/s) of hate-speech text that is "toxic".
Our project is motivated by the need to address hate speech on social media platforms while preserving the integrity of legitimate discussions.

2 Related Work

2.1 Detection of Propaganda Techniques in News Articles: Span Identification Subtask

One of the foremost tasks of Sem-Eval 2020, this challenge saw Team Hitachi achieve the top performance in this subtask. They used a BIO encoding, which is typical for related segmentation and labeling tasks.
For the main BIO tag prediction objective, they used an additional CRF layer, which helps improve the consistency of the output. A number of architectures were trained independently —using BERT, GPT-2, XLNet, XLM, RoBERTa, or XLM-RoBERTa—, and the resulting models were combined in ensembles.

2.2 Automated hate speech detection and span extraction in underground hacking and extremist forums

This paper by Linda Zhou, Andrew Caines, Ildiko Pete and Alice Hutchings describes a hate speech dataset composed of posts extracted from HackForums, an online hacking forum, and Stormfront and Incels.co, two extremist forums. They combined Their dataset with a Twitter hate speech dataset to train a multi-platform classifier.
Their evaluation showed that a classifier trained on multiple sources of data does not always improve the performance compared to a mono-platform classifier.
Finally, this was one of the first work on extracting hate speech spans from longer texts. The paper fine-tuned BERT (Bidirectional Encoder Representations from Transformers) and adopted two approaches – span prediction and

sequence labelling. Both approaches successfully extracted hateful spans and achieve an F1-score of at least 69

2.3 Toxic Spans Detection

This task was proposed as Task-5 in Sem-Eval 2021 and saw a large number of submissions pursuing a variety of approaches.

Lexicon-based approaches were very popular among the teams that participated in this task. The lexicon was handcrafted by domain experts (Smedt et al., 2020) and was simply employed as a list of toxic words for lookup operations (Palomino et al., 2021). Another approach compiled the lexicon using the set of tokens labeled as toxic in the given span-annotated training set and it was used as a lookup table (Burtenshaw and Kestemont, 2021), possibly also storing the frequency of each lexicon token in the training set (Zhu et al., 2021). The former two approaches'combination was also attempted (Ranasinghe et al., 2021). The least supervised lexicons were built with statistical analysis on the occurrences of tokens in a training set solely annotated at the comment level (toxic/nontoxic post) (Rusert, 2021) Despite the unsupervised nature of this approach and its lack of considering context, the lexicon-based approaches performed well, with F1 scores of up to 64.98% attained by Zhu et al[HITSZ-HLT].

Fine-tuned language models (LMs) formed the most popular category among the shared task submissions . Both the winner and the runner-up of the shared task were based on ensembles of fine-tuned LMs . Both submissions used LMs fine-tuned for sequence labeling with the BIO (Beginning, Inside, Outside) scheme, but Zhu et al. [24] also used an LM fine-tuned for span boundary detection. Others participants, such as Chhablani et al. , used models designed for extractive question answering.

The best performing team (HITSZ-HLT) formulated the problem as a combination of token labeling and span extraction (Zhu et al., 2021).
For their token labeling approach, the team used two systems based on BERT (Devlin et al., 2019). Both systems had a Conditional Random Field(CRF) layer (Sutton and McCallum, 2006) on top, but one of the two also had an LSTM layer (Hochreiter and Schmidhuber, 1997) between BERT and the CRF layer. In both approaches, word-level BIO tags were used, i.e., words were labelled as B (beginning word of a toxic span), I (inside word of a toxic span), or O (outside of any toxic span).
For their span extraction approach, the team also used BERT to produce probabilities indicating how likely it is for each token to be the beginning or end of a toxic span. A heuristic search algorithm, originally developed for target extraction in sentiment analysis by Hu et al. (2019), selects the best combinations of candidate begin and end tokens, aiming to output the most likely set of toxic spans per post.

3 Methodology

Our first task lay in resolving the issue of false positives in the previous approaches mentioned in the literature review.
This occurred as the models would often report completely non toxic comments. To resolve this issue, we began by creating a BERT classifier model that would classify a post as toxic or non-toxic. This allowed for a starting point, to filter out sentences that would otherwise end up as false positive spans. We remain fairly confident in our classifier as it achieved an high accuracy when applied to SemEval 2020 Task 11 : Toxic Comment Classification, thus ensuring we don't significantly increase the false negatives in the process.

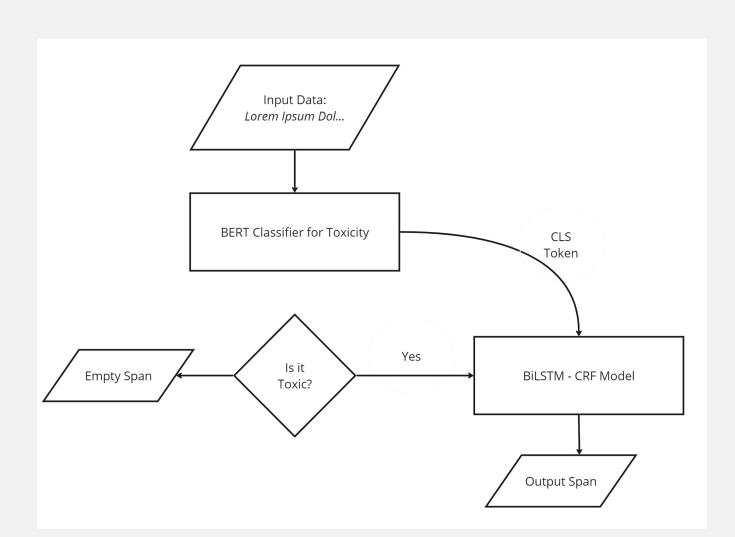

Figure 1: Procedure followed during the project

Once the number of false positive cases were reduced we were left with the task of actually detecting the spans.
We then extracted the CLS tokens from the BERT classifier and combined them with word2vec embeddings. This hybrid embedding scheme was then utilized as input for a Bilstm-Crf Model, which integrates Bidirectional Long Short-Term Memory (Bilstm) with Conditional Random Fields (CRF) for sequence labeling tasks.
In this model architecture, the Bilstm component facilitates bidirectional processing of the input text, enabling comprehensive capture of both forward and backward contextual information. This deep understanding of the text's syntactic structure and semantic context enhances the classification process.
On top of the Bilstm layer, the CRF layer is employed to model the dependencies between adjacent labels, enabling inference of the most probable sequence of la-

bels for the input text. By combining Bilstm and CRF components, the model effectively identifies and labels the spans of toxicity within the text, thereby enhancing moderation capabilities for online content.

4 Dataset

4.1 Description

The founders of Civil Comments, in collaboration with researchers from Google Jigsaw, undertook an effort to open source the collection of more than two million comments that had been collected. After filtering the comments to remove personally identifiable information, a revised version of the annotation system of Wulczyn et al. (2017) was used on the Appen crowd rating platform to label the comments using a number of attributes including 'toxicity', 'obscene', 'threat'. The dataset comprises around 10K comments extracted from the Civil Comments Dataset and annotated using crowd-raters. The trial dataset consists of 690 texts, whereas the training dataset consists of 7939 texts. The test set on which our system was finally evaluated consisted of 2000 text samples.

Post	**Offensive Spans**
Stupid hatcheries have completely fucked everything	[0, 1, 2, 3, 4, 5, 34, 35, 36, 37, 38, 39]
Victimitis: You are such an asshole.	[28, 29, 30, 31, 32, 33, 34]
So is his mother. They are silver spoon parasites.	[]
You're just silly.	[12, 13, 14, 15, 16]

Figure 2: Four Comments along with their annotations. Offensive words are highlighted in red

A similar collaboration between Civil Comments and Jigsaw was featured in the 2020 SemEval Task 11 : Toxic Comment Classification, leaving aside the span detection part.
The data set consisted of comments, followed by a multi-class categorization into the classes mentioned above, followed by the detected span(/s) if any. The collected text data is preprocessed to remove noise, such as special characters, emojis, and HTML tags. Text normalization techniques are applied to standardize the text, including lowercasing, tokenization, and lemmatization.
We then used the given span(/s) to annotate the text with BIO (Beginning, Inside, Outside) labels to indicate the boundaries of hateful speech spans. We then store in a dictionary format with each id corresponding to another dictionary that consists of the comment as it is and the corresponding BIO encoded sentence.

4.2 Files

Link for span detection dataset.

Link for classification dataset.

5 Experimental Setup and Results

Our experimental setup focused on addressing the issue of false positives identified during the literature review. The previous approaches often failed to report empty spans for entirely non-toxic comments, leading to false positives. To tackle this, we devised a BERT classifier model for hate-speech, excluding span detection. This initial step allowed us to filter out sentences prone to false positive spans.

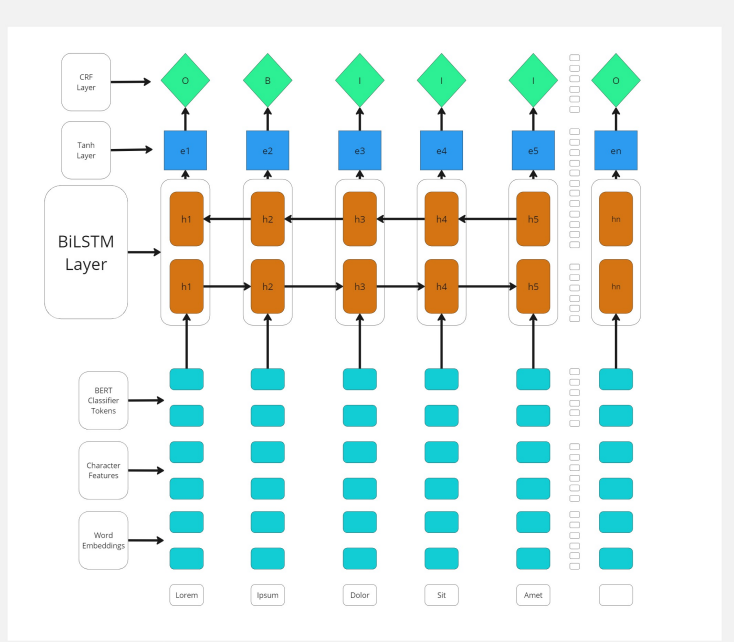

Figure 3: Architecture of BiLSTM CRF Model Used

We validated the effectiveness of our classifier by applying it to the SemEval 2020 Task 11: Toxic Comment Classification dataset, where it demonstrated high accuracy, instilling confidence in its ability to minimize false negatives without a significant increase in false positives.
Following the reduction in false positive cases, our attention shifted to the task of span detection. Leveraging the outputs from the classifier, we proceeded to apply a BiLSTM-CRF model. This model architecture integrates BiLSTM, renowned for its ability to encode contextualized word embeddings, with Conditional Random Fields (CRF), ideal for sequence labeling tasks. By combining these components, we aimed to capture both the semantic meaning and syntactic structure of the text while modeling dependencies between adjacent labels

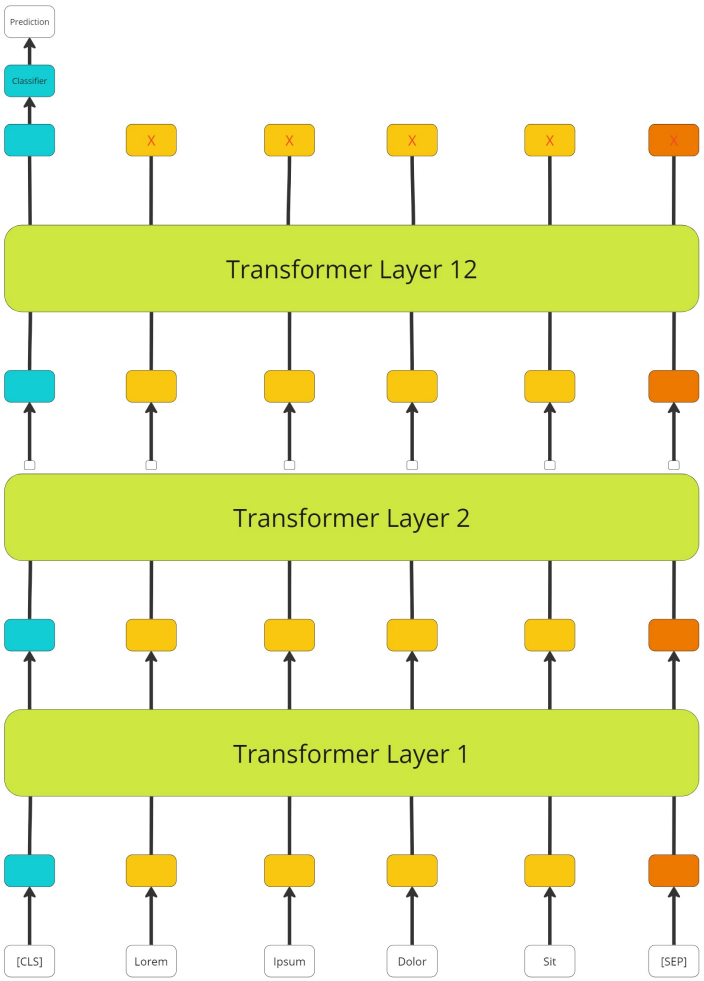

Figure 4: Architecture of BERT Classifier used

to infer the most probable sequence of labels for the input text.

6 Observations and Analysis

The addition of the initial BERT Classifier model helped greatly reduce the number of false positives we would have encountered (**1210 out of 7539 potentially toxic comments**).

We can see that the toxicBERT classifier performs well on each category of toxic speech.

We also noted that this greatly helped in reducing the time and hence, increasing the training efficiency of the BiLSTM-CRF model.

	precision	recall	f1-score
severe_toxic	0.57	0.50	0.53
obscene	0.82	0.86	0.84
threat	0.43	0.75	0.55
insult	0.66	0.81	0.73
identity_hate	0.57	0.57	0.57
micro avg	0.71	0.80	0.75
macro avg	0.61	0.70	0.64
weighted avg	0.71	0.80	0.75
samples avg	0.05	0.05	0.05

Figure 5: Results of the BERT Classifier Model

We also observe steady training and validation of out BiLSTM CRF Model used for Span detection.

7 Conclusion and Future Work

In conclusion, our project aimed to tackle the pervasive issue of hate speech detection by employing said two-step approach. Through this combined methodology, we not only effectively identified toxic content but also delineated its specific occurrences, contributing to the ongoing efforts in mitigating online hate speech and fostering safer digital environments.

In the future, we could try to improve upon our model by incorporating context better in the embeddings we train our models on. Contextualized embeddings like ELMo or Flair embeddings could help capture the meaning of a word based on its surrounding context, potentially aiding in identifying toxic language patterns. Our model's better understanding of context could help us potentially

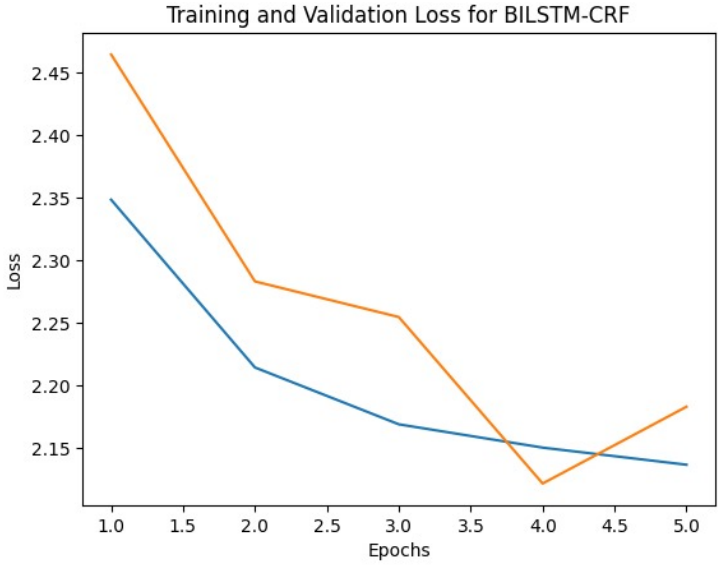

Figure 6: Training and Validation Loss

reduce both false positives and negatives as the model would be able to differentiate between a word being offensive in one context and not being offensive in another.

Moreover, the quality and quantity of the data we are training our model on could be improved further to build a larger and more refined dataset. A vast dataset covering a broader range of topics, forums, and demographic groups will help the model fit better. This will enhance the model's generalization capabilities and may also help mitigate any potential biases in the original model.

We could also potentially expand our task of toxic span detection- rather than just detecting the toxic span, we could further use another model to do hate speech implication generation. This would involve generating the underlying biases present in a toxic comment on a social media platform. It might help us better understand what sort of biases and stereotypes are generally perpetuated and combat them accordingly.

Lastly, we could cover more stakeholders and make a better impact if we incorporated multiple languages in our model. This way, we can ensure that toxic posts don't escape censorship and spread hate speech if they are in other languages than english. A multi-lingual approach would also mean catering to a higher number of users and protecting them from negative comments spreading hate against specific demographic groups or just cyber-bullying that might negatively impact someone's mental health.

Code

Link for code repository.

Model Checkpoints

Drive Link for Model Checkpoint.

References

- Abdessamad Benlahbib, Hamza Alami, and Ahmed Alami. 2021. LISAC FSDM USMBA at SemEval 2021 Task 5: Tackling toxic spans detection challenge with supervised spanBERT-based model and unsupervised LIME based model. In SemEval.
- Ben Burtenshaw and Mike Kestemont. 2021. UAntwerp at SemEval-2021 Task 5: Spans are spans, stacking a binary word level approach to toxic span detection. In SemEval
- Maggie Cech. 2021. macech at SemEval-2021 Task 5: Toxic spans detection. In SemEval.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL, pages 4171–4186.
- Marco Palomino, Dawid Grad, and James Bedwell. 2021. An ensemble approach to identify toxicity in text. In SemEval.
- Thakur Ashutosh Suman and Abhinav Jain. 2021. AStarTwice at SemEval-2021 Task 5: Toxic span detection using RoBERTa-CRF, domain specific pretraining and self-training. In SemEval.
- Charles Sutton and Andrew McCallum. 2006. An Introduction to Conditional Random Fields for relational learning. Introduction to statistical relational learning, 2:93–128.