UIT-ISE-NLP at SemEval-2021 Task 5: Toxic Spans Detection with BiLSTM-CRF and ToxicBERT Comment Classification

Son T. Luu
University of Information Technology
Vietnam National University
Ho Chi Minh City, Vietnam
sonlt@uit.edu.vn

Ngan Nguyen
University of Information Technology
Vietnam National University
Ho Chi Minh City, Vietnam
ngannlt@uit.edu.vn

Abstract

We present our works on SemEval-2021 Task 5 about Toxic Spans Detection. This task aims to build a model for identifying toxic words in whole posts. We use the BiLSTM-CRF model combining with ToxicBERT Classification to train the detection model for identifying toxic words in posts. Our model achieves 62.23% by F1-score on the Toxic Spans Detection task.

1 Introduction

Detecting toxic posts on social network sites is a crucial task for social media moderators in order to keep a clean and friendly space for online discussion. To identify whether a comment or post is toxic or not, social network administrators often read the whole comment or post. However, with a large number of lengthy posts, the administrators need assistance to locate toxic words in each post to decide whether a post is toxic or non-toxic instead of reading the whole post. The SemEval-2021 Task 5 (Pavlopoulos et al., 2021) provides a valuable dataset called Toxic Spans Detection dataset in order to train the model for detecting toxic words in lengthy posts.

Based on the dataset from the shared task, we implement the machine learning model for detecting toxic words posts. Our model includes: the BiLSTM-CRF model (Lample et al., 2016) for detecting the toxic spans in the post, and the ToxicBERT (Hanu and Unitary team, 2020) for classifying whether the post is toxic or not. Before training the model, we pre-process texts in posts and encode them by the GloVe word embedding (Pennington et al., 2014). Our model achieves 62.23% on the test set provided by the task organizers.

2 Related works

Many corpora are constructed for toxic speech detection problems. They consist of flat label and hierarchical label datasets. The flat label datasets only classify one label for each comment in the dataset (e.g., hate, offensive, clean), while hierarchical datasets can classify multiple aspects of the comment (e.g., hate about racism, hate about sexual oriented, hate about religion, and hate about disability). For flat label, we present several datasets including the two datasets which are provided by Waseem and Hovy (2016) and Davidson et al. (2017) in English, the dataset of Albadi et al. (2018) in Arabic, and the dataset of by Alfina et al. (2017) in Indonesian. For the hierarchical label, we introduce the dataset constructed by Zampieri et al. (2019) in English, the dataset provided by Fortuna et al. (2019) in Portuguese, and the CONAN dataset by Chung et al. (2019), which is the multilingual corpus (constructed in Italian, English, and French).

In addition, many of shared tasks about hate speech and abusive languages are organized, such as the SemEval-2019 Task 5 (Multilingual) (Basile et al., 2019), the SemEval-2019 Task 6 (English) (Zampieri et al., 2019), the PolEval 2019 Shared task 6 (Polish) (Ptaszynski et al., 2019), the GermEval 2018 (Germany) (Wiegand et al., 2018), EVALITA 2019 (Italian) (Bosco et al., 2018), Toxic Comment Classification Challenge[1], and VLSP 2019 Shared task (Vietnamese) (Vu et al., 2019).

Besides, SOTA approaches like deep learning (Badjatiya et al., 2017) and transformers models (Isaksen and Gambäck, 2020) are applied in the hate speech detection and toxic posts classification. However, these models only classify based on the whole posts or documents. For the Toxic Spans Detection task, we adapt the mechanism from Sequence tagging (Wang et al., 2020) and Name entities Recognition (Lin et al., 2017) for detecting toxic words from posts.

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 846–851
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

3 Dataset

The dataset is provided from the SemEval-2021 Task 5: Toxic Spans Detection (Pavlopoulos et al., 2021). It includes the training and the test sets. Both of them consist of two parts: the content of posts and the spans denoting the toxic words in the posts. Spans represent toxic words in the posts as a set of character indexes. Table 1 illustrates several examples from the training set.

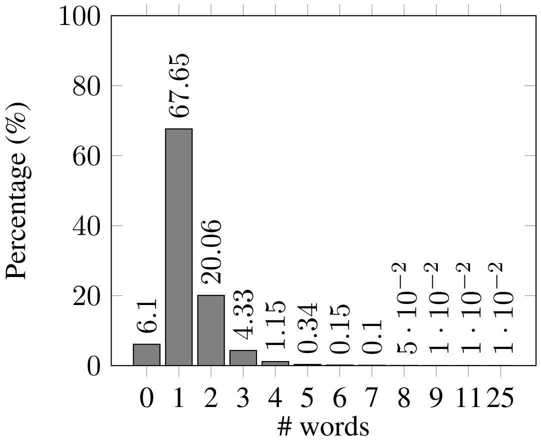

Figure 1: Number of toxic words in spans for each post in the training set.

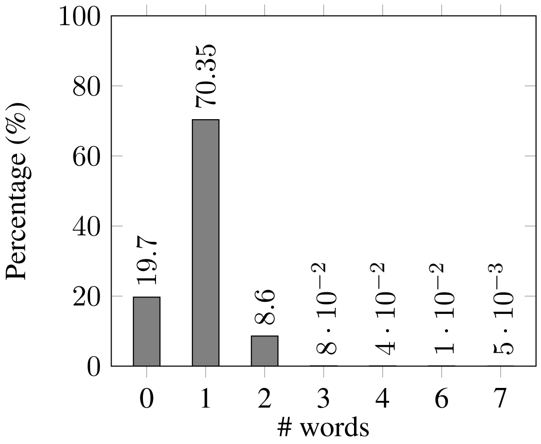

Figure 2: Number of toxic words in spans for each post in the test set.

According to Table 1, a post contains multiple spans of toxic words. For each span, it contains a single word, a phrase, or a sentence. As described in Figure 1, most of the spans in the training set are single words, which account for 67.65%, while only 20.06% of spans contains two words, and 6.1% of spans is empty. Posts whose spans contain more than two words in the dataset are few. Especially in the training set, there is a post in which spans contain 25 words.

Besides, Figure 2 illustrates the number of toxic words in spans per post for the test set. Spans containing single words account for the highest percentage (70.35%) in the test set, and are higher than in the training set, while the multiple-word spans are few. Also, the empty spans in the test set are higher than the training set, and the longest post in the test set contains only seven words.

4 System description

4.1 Data preparation

With the given dataset from the SemEval-2021 Task 5 about Toxic Spans Detection (Pavlopoulos et al., 2021), we firstly transform spans into a set of words. Then, we pre-process the posts as follows: (1) Segmenting the posts by the TweetTokenizer from nltk[2], and (2) Changing texts to lower case.

4.2 Feature extraction

We use the glove.twitter.27b.25d word embedding[3] to construct the dictionary and encode the text of posts. Posts are encoded by the dictionary of the word embedding. The $< UNK >$ tokens are added if a word in posts is not found in the dictionary. To make sure all vectors are the same length, we add the $< PAD >$ token. Then, we set the maximum length of vectors equal to 128. Spans are transformed into a one-hot vector corresponding to each word in posts where toxic words are denoted as 1 and others are denoted as 0. Table 2 illustrates an example of encoding data in our system.

4.3 Training models

Detection model: BiLSTM-CRF is a deep neural model used for Named-entity recognition task (Lample et al., 2016). We implement this model for the task of detecting toxic words in documents. The model includes three main layers: (1) The word representation layer uses embedding matrix from the GloVe word embedding, (2) The BiLSTM layer for sequence labeling, and (3) The Conditional Random Field (CRF) layer to control the probability of output labels. The output is a binary vector, in which each value determines whether

[2] https://www.nltk.org/api/nltk.tokenize.html

[3] https://nlp.stanford.edu/projects/glove/

No	Posts	Spans
1	What a **knucklehead**. How can anyone not know this would be offensive??	[7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]
2	I only use the word haole when **stupidity** and **arrogance** is involved and not all the time. Excluding the POTUS of course.	[31, 32, 33, 34, 35, 36, 37, 38, 39, 45, 46, 47, 48, 49, 50, 51, 52, 53]
3	**Such garbage logic by republicans** which will backfire and rush america into the great depression II	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32]
4	what a **hypocrite** of bs,, tell us loser how you live without gasoline, plastic, medical needs and medications, all from OIL,, but you cant of course so you **ignorant** fools in your **hypocrisy** spew this bs	[7, 8, 9, 10, 11, 12, 13, 14, 15, 155, 156, 157, 158, 159, 160, 161, 162, 178, 179, 180, 181, 182, 183, 184, 185, 186]
5	Exposing hypocrites like Trump and Pence is therapeutic for you? Good job!	[]

Table 1: Sample posts from the training set. The toxic span are highlighted as bold.

	Original	Transformed
Text	I only use the word haole when **stupidity** and **arrogance** is involved and not all the time. Excluding the POTUS of course.	['i', 'only', 'use', 'the', 'word', 'haole', ..] **Vector:** [12, 216, 718, 15, 894,..]
Spans	[31, 32, 33, 34, 35, 36, 37, 38, 39, 45, 46, 47, 48, 49, 50, 51, 52, 53]	['i', 'only', 'use', 'the', 'word', 'haole', 'when', **'stupidity'**, 'and', **'arrogance'**, 'is', ...] **Vector:** [0, 0, 0, 0, 0, 0, 0, **1**, 0, **1**, 0, 0 ...]

Table 2: Example of encoding data into vectors.

a word is toxic or non-toxic. The architecture of BiLSTM-CRF is described in Figure 3.

Classification model: The ToxicBERT model (Detoxify) is introduced by Hanu and Unitary team (2020) with the purpose to stop online abusive comments. It is a pre-trained model and is easy to use by using transformers library[4]. The model is trained on the Toxic Comments Classification Challenge datasets provided by Jigsaw.

Our system combines the detection and classification model together. The detection model (BiLSTM-CRF) returns the toxic spans from the post, while the classification model (ToxicBERT) classifies whether a post is toxic or non-toxic. If a post is non-toxic, the classification model returns an empty span. By contrast, it reserves the spans of the detection model. Then, predicted spans are decoded to character indexes for submission. Our system is illustrated in Figure 4

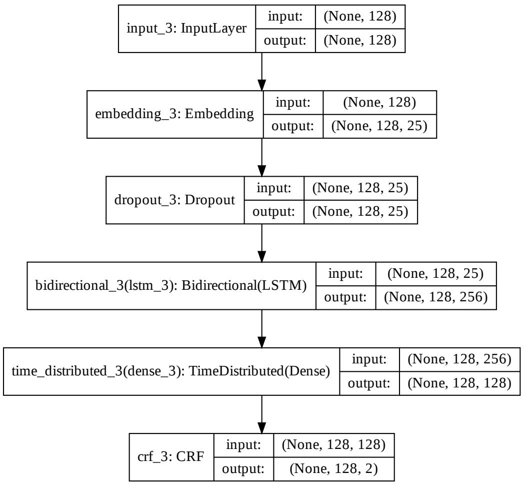

Figure 3: The BiLSTM-CRF model architecture.

[4] https://huggingface.co/unitary/toxic-bert

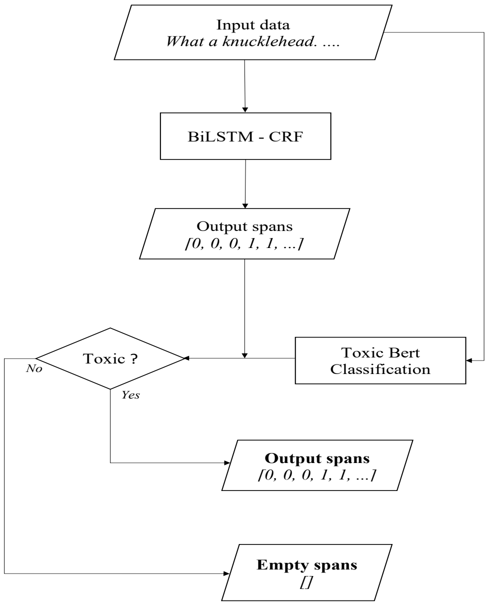

Figure 4: Our system architecture.

5 Experimental results

5.1 Evaluation metric

The variant version of F1-score is used to evaluate the results of the competition (Da San Martino et al., 2019). Let T is the total of post in the dataset, $T = [t_1, t_2, ..., t_n]$, n is the number of posts, A is spans given by the model, and G is ground truth spans.

The F1-score over the dataset is defined as:

$$\frac{1}{|T|}\sum_{t}^{T} F_1^t = 2 * \frac{P^t(A,G) * R^t(A,G)}{P^t(A,G) + R^t(A,G)} \qquad (1)$$

In the Equation 1, P^t determines the precision, and R^t determines the recall of the post t. The precision and recall are calculated as Equation 2 and Equation 3, respectively. The S^t in both Equation 2 and Equation 3 is set of toxic characters of post t (span).

$$P^t(A,G) = \frac{|S_A^t \cap S_G^t|}{S_A^t} \qquad (2)$$

$$R^t(A,G) = \frac{|S_A^t \cap S_G^t|}{S_G^t} \qquad (3)$$

5.2 Main result from task

Model	Result (%)
BiLSTM-CRF	61.32
BiLSTM-CRF + ToxicBERT	**62.23**
Top 1 rank	**70.83**

Table 3: Experimental results obtained by our system in comparison with Top 1 rank result.

According to Table 3, when only BiLSTM-CRF is used, the result by F1 score is 61.32%. The result increases up to 62.23% when we applied ToxicBERT Classifier, and this is our final result of the shared task (ranked 63^{th} among 92 teams).

5.3 Error analysis

According to Table 4, the appearance of the ToxicBERT classifier can make a better prediction for the non-toxic posts (See example No. 1). This increases the performance of our system, however, not significantly, because the number of empty toxic span comments in the test set is not too much (as described in Section 3). Apart from empty spans, wrong predictions are fell into the case as example No.3. Although the ToxicBERT model predicts this example as non-toxic, the BiLSTM-CRF model predicts it as empty spans, and thus the result is empty spans according to the result of the BiLSTM-CRF model. For example No.2, the system returns wrong spans for the first span in the post. Finally, our system cannot predict well for spans that contain more than two words. It returns spans, but not enough, as shown in examples No.4 and No.5 from Table 4.

6 Conclusion

We use the BiLSTM-CRF and ToxicBERT models for detecting toxic words in the posts. Our model achieves 62.23% by F1-score from the competition. From the error analysis, we found that our model predicts well just for single-word spans and empty spans.

In further researches, we improve the performance of the detection model by applying the attention mechanism and using the character-level representation combining with word-level representation. Character-level models like CharBERT (Ma et al., 2020) is a potential approach to increase the performance of toxic spans detection tasks.

No.	Posts	Gold span	Prediction span: BiLSTM-CRF	Prediction span: BiLSTM-CRF+ ToxicBERT
1	Indeed, people the world over, all know that President Trump is a loser!	[]	[66, 67, 68, 69, 70]	[]
2	Idiot ”miner” in the photo isn’t even wearing a dust mask. MSHA needs to inspect this operation for more worker safety violations.	[0, 1, 2, 3, 4]	[-1, 0, 1, 2, 3]	[-1, 0, 1, 2, 3]
3	People who are anti-immigration are weak, lying, racist whiners.	[49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62]	[]	[]
4	Uh-No, keep voting for failed Liberal idiocy that guarantees results ala Detroit, Chicago, etc. You’ll wish your body had only some crap rather than gang-banger gunfire.	[38, 39, 40, 41, 42, 43]	[38, 39, 40, 41, 42, 43, 133, 134, 135, 136]	[38, 39, 40, 41, 42, 43, 133, 134, 135, 136]
5	What is he going to do about those toxic mercury florescent bulbs Bush and Gore pushed on the stupid American public?	[94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115]	[94, 95, 96, 97, 98, 99]	[94, 95, 96, 97, 98, 99]

Table 4: Several wrong predictions on the test set by our system.

References

N. Albadi, M. Kurdi, and S. Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76.

I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW ’17 Companion*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Cristina Bosco, Dell’Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meet-*

ing of the Association for Computational Linguistics, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Vebjørn Isaksen and Björn Gambäck. 2020. Using transfer-based language models to detect hateful and offensive language online. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, Online. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Bill Y. Lin, Frank Xu, Zhiyi Luo, and Kenny Zhu. 2017. Multi-channel BiLSTM-CRF model for emerging named entity recognition in social media. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 160–165, Copenhagen, Denmark. Association for Computational Linguistics.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. CharBERT: Character-aware pre-trained language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 39–50, Barcelona, Spain (Online). International Committee on Computational Linguistics.

John Pavlopoulos, Léo 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*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. 2019. Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter.

Xuan-Son Vu, Thanh Vu, Mai-Vu Tran, Thanh Le-Cong, and Huyen T M. Nguyen. 2019. HSD shared task in VLSP campaign 2019: Hate speech detection for social good. In *Proceedings of VLSP 2019*.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195, 1st virtual meeting. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota.