SINAI at SemEval-2021 Task 5: Combining Embeddings in a BiLSTM-CRF model for Toxic Spans Detection

Flor Miriam Plaza-del-Arco, Pilar López-Úbeda
L. Alfonso Ureña-López, M. Teresa Martín-Valdivia

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{fmplaza, plubeda, laurena, maite}@ujaen.es

Abstract

This paper describes the participation of SINAI team at Task 5: Toxic Spans Detection which consists of identifying spans that make a text toxic. Although several resources and systems have been developed so far in the context of offensive language, both annotation and tasks have mainly focused on classifying whether a text is offensive or not. However, detecting toxic spans is crucial to identify why a text is toxic and can assist human moderators to locate this type of content on social media. In order to accomplish the task, we follow a deep learning-based approach using a Bidirectional variant of a Long Short Term Memory network along with a stacked Conditional Random Field decoding layer (BiLSTM-CRF). Specifically, we test the performance of the combination of different pre-trained word embeddings for recognizing toxic entities in text. The results show that the combination of word embeddings helps in detecting offensive content. Our team ranks 29th out of 91 participants.

1 Introduction

The advance of online communication has increased the use of offensive or toxic language in several websites, including social networks such as Instagram, Twitter, or YouTube. Consequently, this type of prejudiced communication could lead to negative psychological effects among Internet users, causing anxiety, harassment, and even suicide in extreme cases (Hinduja and Patchin, 2010).

Moderation is essential to promote healthy online communication. Therefore, governments, online communities, and social media platforms are continuously taking appropriate actions to implement laws and policies combating toxic language on the Web. In order to help to track this type of comments and due to the amount of data generated every day on the Web, automatic systems based on Natural Language Processing (NLP) techniques are required. In particular, offensive language detection and analysis has become an important area of research in NLP, resulting in several studies that are contributing to combating this website phenomenon (Plaza-del Arco et al., 2019; Zampieri et al., 2019a; Ranasinghe et al., 2019; Plaza del Arco et al., 2020).

In this paper[1], we present our proposal system as part of our participation in SemEval-2021 Task 5: Toxic Spans Detection (Pavlopoulos et al., 2021), which aims to identify entities that refer to a toxic language in the text. To accomplish the task, our team focused on detecting specific types of toxic entities in the text using a methodology based on the BiLSTM-CRF model showing that the combination of different pre-trained language embeddings succeeds in detecting toxic entities.

The rest of the paper is structured as follows. In Section 2 some previous related studies are introduced. In Section 3 we explain the data used in our methods and we describe the architecture of our proposed system to the Toxic Spans Detection task. In Section 4 we discuss the analysis and evaluation results for the experiments we performed. Finally, we conclude in Section 5 with remarks and future work.

2 Related work

Heretofore, several shared tasks have been organized in the NLP field to detect offensiveness on the Web for different languages. For instance, the well-known offensive language task OffensEval has held two editions in the International Workshop on Semantic Evaluation (SemEval) (Zampieri et al., 2019b,a) introducing as the main novelty in the second edition a multilingual dataset comprising 5

[1]NOTE: This paper contains examples of potentially explicit or offensive content which may be offensive to some readers. They do not represent the views of the authors.

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 984–989
Bangkok, Thailand (online), August 5–6, 2021.

languages. The GermEval shared task focused on the identification of offensive language in German tweets and comprised two tasks, a coarse-grained binary classification task and a fine-grained multi-class classification task (Wiegand and Siegel, 2018). For Spanish, as far as we know, the first task on offensive language appeared at the 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval) (Carmona et al., 2018) whose goal was to detect aggressiveness Mexican Spanish Tweets.

As a result, most of the studies and resources in offensive language research have been developed specifically for binary and multi-class classification tasks (Ranasinghe et al., 2019; Plaza-del Arco et al., 2019, 2020). However, other tasks such as Named Entity Recognition (NER) play an important role in this research and are essential to identify the entities that make a text toxic. Highlighting these toxic spans can help human moderators to interpret and identify easily this type of content on the Web instead of relying on a system that generates a score of unexplained toxicity per post. NER aims to identify and classify named entities mentioned in unstructured text into predefined categories. The earliest systems developed for addressing this task did not use training data but worked based on handcrafted features, heuristics, and a set of rules (Nadeau and Sekine, 2007; Collins and Singer, 1999; López-Ubeda et al., 2018). However, the cost of manual feature tagging and the poor obtained results lead to deep learning-based techniques as the most suitable choice to tackle the task by discovering patterns and learning the features in an end-to-end manner (López-Úbeda et al., 2020). Existing state-of-the-art approaches for sequence labeling have proven that Recurrent Neural Networks (RNNs) are capable of learning useful representations automatically as they enable the modeling of long-distance dependencies between words in a sentence (Limsopatham and Collier, 2016; Wintaka et al., 2019). Inspired by these studies, we have developed a system based on BiLSTM-CRF model along with the combination of different types of word embeddings to address the toxic spans detection task in text.

3 Named Entity Recognition Methodology

To address the toxic detection task, we focus on recognizing and extracting specific types of toxic entities in the text. Specifically, we follow a methodology proposed by (Huang et al., 2015) implementing a BiLSTM-CRF model for the NER task.

3.1 Word Embeddings

As input layer of the BiLSTM-CRF neural network we have combined the following word embeddings:

- **Static Word Embedding**. We use GloVe embeddings which are static and word-level, i.e. each distinct word gets exactly one pre-computed embedding. This type of embeddings is context-independent (Pennington et al., 2014).

- **Contextual Word Embedding**. For our experiments, we tested two different contextual pre-trained word embeddings: BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) and XLM (Lample and Conneau, 2019). Unlike the previous ones, they are context-dependent which means they produce word representations that are dynamically informed by the words around them. They are based on the well-known Transformer (Vaswani et al., 2017), an attention mechanism that processes the entire text input simultaneously to learn contextual relations between words (or sub-words). Specifically, we used the *xlm-mlm-en-2048* model and the *bert-base-cased model* provided in HuggingFace (Wolf et al., 2019).

3.2 BiLSTM-CRF architecture

We use the combination of bidirectional LSTM and CRF to identify the toxic spans. The context of each word in the sentence is captured by the BiLSTM and then the predictions on the entities are simultaneously performed in the CRF layer (Sutton and McCallum, 2006). The architecture of BiLSTM-CRF model is illustrated in Figure 1. This architecture follows a sequence of layers as follows:

- Embedding layer. Each word of the sentence is mapped to a vector of concatenated embeddings. As mentioned above, in our experiments, we use XLM, BERT, and GLOVE embeddings.

- BiLSTM layer. A bidirectional LSTM recurrent network takes as input the embeddings. In sequence tagging tasks, for a specific time

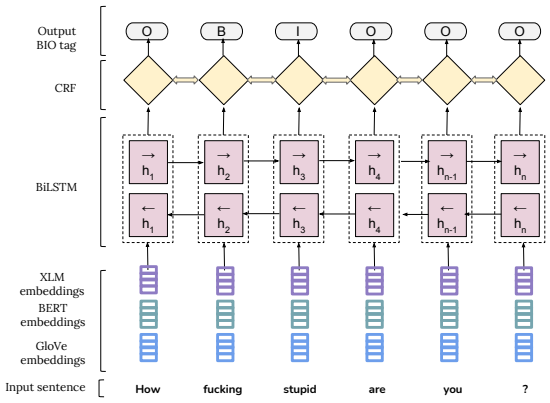

Figure 1: Proposed system architecture based on a BiLSTM-CRF neural network.

frame, this layer enables the hidden states to capture both historical and future context information and then to label a token.

- CRF layer. It allows to use efficiently historical and future tags to predict the current tag.

4 Data and Experimental Setup

4.1 Dataset preprocessing

We use the English dataset provided by the organizers in SemEval 2020 Task 5: Toxic Spans Detection. The dataset is split into three different subsets: train, trial, and test, consisting of 7,939, 690, and 2,000 instances, respectively.

Each instance in the dataset comprises two fields, the text and a list of toxic spans. A toxic span is defined as a sequence of characters in words that attribute to the text's toxicity. If the text does not contain toxic spans, the span list is empty. An example of two instances in the dataset is provided in Table 1. In the first example, the word "crap" is labeled as toxic in the text, which has character offsets from 15 to 18. The second example includes the toxic span "idiot" which has character offsets from 4 to 8.

Text	Spans
What a load of **crap**.	[15, 16, 17, 18]
You **idiot**. The media went to war against truth.	[4, 5, 6, 7, 8]

Table 1: Two instances in the dataset. Toxic words are highlighted in the text.

To perform our experiments, we preprocess the subsets of the dataset in the following way. First, we used the nltk.tokenize package[2] to tokenize the text. Then, we generated the following features for each text in the subset: the word, the position of the beginning and end of the word in the text, and the NER tag. In order to perform the NER tagging, we follow the BIO annotation scheme to label multi-token named entities (Ratinov and Roth, 2009), which represents that the label is the beginning of a span (B-Toxic), inside the span (I-Toxic), or belongs to no span (O). This scheme is the most popular in the NER task. Figure 2 shows an example of the features generated for the following example in the training set: "How fucking stupid are you?", spans: [4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17].

```
How       0   3   O
fucking   4   11  B-Toxic
stupid   12   18  I-Toxic
are      19   22  O
you      23   26  O
?        26   27  O
```

Figure 2: Example of training set instance with generated features using BIO annotation scheme.

4.2 Experiments

During the pre-evaluation period, we trained our models on the train set and evaluated our different approaches on the trial set. During the evaluation period, we trained our models on the train and trial sets and tested the model on the test set.

Flair's framework (Akbik et al., 2019) builds directly on Pytorch was used to design the BiLSTM-CRF network. We used the default hyperparameter setting in Flair with the following configuration: learning rate as 0.1, batch size as 32, dropout probability as 0.01, and maximum epoch as 300. All experiments (training and evaluation) were performed on a node equipped with two Intel Xeon Silver 4208 CPU at 2.10GHz, 192GB RAM, as main processors, and six GPUs NVIDIA GeForce RTX 2080Ti (with 11GB each).

Our team (SINAI) submitted 4 runs for the Toxic Spans Detection task and each run evaluates the word embeddings as an input to the BiLSTM-CRF network, as explained in Section 3.

Run 1. GloVe embeddings.

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

Run 2. BERT embeddings.

Run 3. XLM embeddings.

Run 4. BERT + XLM + GloVe embeddings.

5 Results

In this section, we present the results obtained by our proposed system. In order to evaluate them, we use the official competition metric F1-score.

The results of our participation in the Toxic Spans Detection task during the evaluation phase are shown in Table 2. In particular, we list the performance of the four runs submitted using the BiLSTM-CRF model along with the combination of different word embeddings. If we analyze the results of the first 3 runs (each embeddings independently), we notice that they sightly differ, the best result is achieved by the contextual embedding XLM. However, training the model on the combination of static and contextual embeddings (GloVe, BERT, and XLM) leads to enhanced performance with a 0.6727 F1-score. Therefore, our results show the success of the combination of embeddings we chose to solve the task of toxic spans detection in comments using the proposed model.

Run	Word embeddings	F1-score
1	GloVe	0.6618
2	BERT	0.6606
3	XLM	0.6635
4	**BERT + XLM + GloVe**	**0.6727**

Table 2: Systems test results achieved by SINAI in SemEval Task 5: Toxic Spans Detection.

Table 3 shows the official rank in the competition. As we can see, we are ranked 29th out of 91 participating teams obtaining an F1-score of 0.6727 with our system. The best result was obtained by the team HITSZ-HLT with an F1-score of 0.7083, which differs from our results achieved by 3.56%. In general, low results for the task are obtained which shows the Toxic Spans Detection as a challenge to be addressed by the NLP community and, therefore, further research is needed to advance on this specific task. We also observe that the number of participants in this task is high (91) which shows the importance and interest of the NLP community in contributing to addressing this challenge.

User name (ranking)	F1-score
HITSZ-HLT (1)	0.7083
lmazxn (10)	0.6893
SINAI (29)	**0.6727**
UIT-ISE-NLP (63)	0.6223
ramya.akula01 (85)	0.1968
AmrHendy (91)	0.0675

Table 3: System Results per participating team in Task 5: Toxic Spans Detection.

6 Conclusions and Future Work

This paper presents the participation of the SINAI research group in Task 5: Toxic Spans Detection at SemEval 2021.

In this paper, we use a deep learning-based approach for NER to identify spans that make a text toxic, which focuses on the use of a BiLSTM-CRF neural network where different word embeddings are tested. The model is trained on the dataset provided by the organizers of the task (Pavlopoulos et al., 2021) and preprocessing techniques are carried out to tokenize and tagged the dataset by using the BIO scheme.

Our results show that the sophisticated BiLSTM-CRF architecture which has been successfully used for other tasks such as biomedical entity recognition or part-of-speech tagging, but also achieves remarkable results when addressing tasks related to the identification of offensive language in comments. Besides, we find that this architecture with our proposed combination of embeddings for word representation provides useful insights for the learning phase of the neural network achieving better results than training the network with a single type of word embedding.

For future work, we plan to study the performance of our proposed method using a variety of linguistic features, including emotions that are inextricably linked to offensive language.

Acknowledgments

This work has been partially supported by a grant from the Ministry of Science, Innovation, and Universities (Scholarship FPI-PRE2019-089310), Fondo Europeo de Desarrollo Regional (FEDER), and LIVING-LANG project (RTI2018-094653-B-C21) from the Spanish Government.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Flor Miriam Plaza-del Arco, M. Dolores Molina-González, Maite Martin, and L. Alfonso Ureña-López. 2019. SINAI at SemEval-2019 task 6: Incorporating lexicon knowledge into SVM learning to identify and categorize offensive language in social media. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 735–738, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Flor Miriam Plaza del Arco, M. Dolores Molina González, Alfonso Ureña-López, and Maite Martin. 2020. SINAI at SemEval-2020 task 12: Offensive language identification exploring transfer learning models. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1622–1627, Barcelona (online). International Committee for Computational Linguistics.

Flor Miriam Plaza-del Arco, M Dolores Molina-González, L Alfonso Ureña-López, and M Teresa Martín-Valdivia. 2020. Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

Miguel Ángel Álvarez Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Verónica Reyes-Meza, and Antonio Rico Sulayes. 2018. Overview of MEX-A3T at IberEval 2018: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 74–96. CEUR-WS.org.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sameer Hinduja and Justin W Patchin. 2010. Bullying, cyberbullying, and suicide. *Archives of suicide research*, 14(3):206–221.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Nut Limsopatham and Nigel Collier. 2016. Bidirectional LSTM for Named Entity Recognition in Twitter Messages.

Pilar López-Ubeda, Manuel C Dıaz-Galiano, Marıa Teresa Martın-Valdivia, and L Alfonso Urena-López. 2018. SINAI en TASS 2018 Task 3. Clasificando acciones y conceptos con UMLS en MedLine. *Proceedings of TASS*, 2172.

Pilar López-Úbeda, MC Dıaz-Galiano, MT Martın-Valdivia, and L Alfonso Ureña-López. 2020. Extracting Neoplasms Morphology Mentions in Spanish Clinical Cases through Word Embeddings. *Proceedings of IberLEF*.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

John Pavlopoulos, Léo Laugier, Jeffrey Sorensen, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation*.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi. 2019. BRUMS at HASOC 2019: Deep Learning Models for Multilingual Hate Speech and Offensive Language Identification. In *FIRE (Working Notes)*, pages 199–207.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Charles Sutton and Andrew McCallum. 2006. An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning*, 2:93–128.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Michael Wiegand and Melanie Siegel. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of KONVENS 2018*.

Deni Cahya Wintaka, Moch Arif Bijaksana, and Ibnu Asror. 2019. Named-Entity Recognition on Indonesian Tweets using Bidirectional LSTM-CRF. *Procedia Computer Science*, 157:221–228.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. 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)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.