LaSTUS/TALN at SemEval-2019 Task 6: Identification and Categorization of Offensive Language in Social Media with Attention-based Bi-LSTM model

Lutfiye Seda Mut Altin, Àlex Bravo, Horacio Saggion
Large Scale Text Understanding Systems Lab / TALN Research Group
Department of Information and Communication Technologies (DTIC)
Universitat Pompeu Fabra
Tanger 122, Barcelona (08018), Spain
lutfiyeseda.mut01@estudiant.upf.edu, {alex.bravo, horacio.saggion}@upf.edu

Abstract

This paper describes a bidirectional Long-Short Term Memory network for identifying offensive language in Twitter. Our system has been developed in the context of the SemEval 2019 Task 6 which comprises three different sub-tasks, namely A: Offensive Language Detection, B: Categorization of Offensive Language, C: Offensive Language Target Identification. We used a pre-trained Word Embeddings in tweet data, including information about emojis and hashtags. Our approach achieves good performance in the three sub-tasks.

1 Introduction

As the amount of user generated content in social media is increasing at an exponential pace, detecting offensive language and harmful content automatically in an efficient way is a very important issue for the society. Recent work has shown that offensive language in various forms such as hate speech, cyberbullying, profanity and harassment has negative effects especially in adolescents (Hamm et al., 2015).

The shared task, Categorizing Offensive Language in Social Media (SemEval 2019 - Task 6), focuses on improving identification of offensive language by considering type and target of the offense into account (Zampieri et al., 2019b). The task is composed of three sub-tasks. Sub-task A aims to identify if a given tweet is **offensive** or **not** (annotated as **OFF** or **NOT**). Sub-task B aims to categorize the offense type in offensive tweets into two categories: **targeted** (**TIN**) or **untargeted** (**UNT**) meaning that if a tweet contains an insult or threat to an individual, a group or something else or if a tweet contains non-targeted offense such as general profanity or non-acceptable language. Lastly, Sub-task C aims to identify the **target type** of targeted offensive posts. The target type is supposed to be classified as individual, group or other for the rest (annotated as **IND**, **GRP** or **OTH**). We submitted three different runs for each sub-task.

The training dataset released by the shared task organizers, consists of 14,100 English tweets with one annotation layer per task with a hierarchical annotation scheme where each annotation level is related to an independent sub-task. The methods used to collect this dataset is described in (Zampieri et al. (2019a)). Examples from the dataset with annotations at the end are given below:

> *"@USER That shit weird! Lol OFF (offensive) - -"*
>
> *"@USER @USER You are an embarrassing citizen!! OFF TIN -"*
>
> *"@USER @USER Liberals ruin everything! OFF TIN GRP"*

This paper describes a bidirectional Long Short Term Memory network (biLSTM) model with an Attention layer to identify offensive language in Twitter. The rest of the paper is organized as follows: In section 2, we introduce an overview of the work related to identification of offensive language. In Section 3 we describe our model structure and differences between the different runs submitted for each sub-task. In Section 4 we provide the results and discuss the performance of the system. Finally, in Section 5 we conclude giving an outline for the future work.

2 Related Work

Identification of offensive language in user-generated content can essentially be considered as a classification task. Previous research on the

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 672–677
Minneapolis, Minnesota, USA, June 6–7, 2019.

issue has been carried out with approaches from different perspectives such as abusive language (Waseem et al., 2017) (Chu et al., 2017), hate speech (Davidson et al., 2017) (Schmidt and Wiegand, 2017) (Fortuna and Nunes, 2018) and cyberbullying (Hee et al., 2018).
It has been referred by (Kumar et al., 2018) that for identification of aggression in a more general manner, classifiers such as SVM and logistic regression can equalize the results of neural networks-based systems if the right features are selected. On the other hand, (Zhang et al., 2018) pointed out that a deep neural network model combining convolutional neural network and long short term memory network, performed better than state of the art, including SVM. Furthermore, indicated that automatically selected features performed better than manual features.

Recent research also includes the work of (ElSherief et al., 2018) focusing on the target of the hate speech found that in terms of word characteristics, such as frequency of specific words, differences can be observed between hate to individuals or to groups.

(Gambäck and Sikdar, 2017) investigated classification of different sub-categories of hate speech with different Convolutional Neural Network models founding that word2vec embeddings performed best. Davidson et al. worked on distinguishing hate speech and offensive language by training a multi-class classifier showing that using lexicons is useful in agreement with the previous research (Davidson et al., 2017).

Both for English and other languages similar shared tasks have been organized. At GermEval that aims to identify offensive language in German tweets; popular features were lexicons of offensive words, word embeddings and character n-grams. Between deep learning approaches and traditional supervised classifiers there was not a clear superior system in terms of the scores (Wiegand et al., 2018). EVALITA 2018 Hate Speech Detection Shared Task focused on Italian text on Facebook and Twitter. Best performed system in this shared task utilized polarity and subjectivity lexicon with word embeddings (Caselli et al., 2018).

3 Methodology and Data

This paper describes a neural network based on the model proposed by Zhou et al. (2016) for relation extraction. The model consist of a bidirectional Long Short-Term Memory Networks (biLSTM) model with an Attention layer on top. The model capture the most important semantic information in a tweet, including emojis and hashtags, to face the three sub-tasks. In Figure 1 a simplified schema of our model can be seen. In the following we explain how the model works.

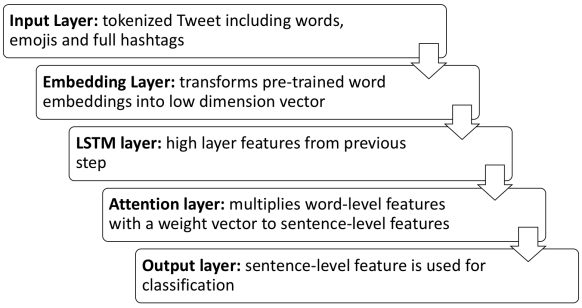

Figure 1: Simplified schema of the model

First, the tweets were tokenized removing punctuation marks and keeping emojis and full hashtags because can contribute to define the meaning of a tweet.

Second, the embedding layer transforms each element in the tokenized tweet (such as words, emojis and hashtags) into a low-dimension vector. The embedding layer, composed of the vocabulary of the task, was randomly initialized from a uniform distribution (between -0.8 and 0.8 values and with 300 dimensions). Recent studies have reported that pre-trained word embeddings are far more satisfactory than the randomly initialized embeddings (Erhan et al., 2010; Kim, 2014). For that reason, the initialized embedding layer was updated with the word vectors included in a pre-trained model based on all the tokens, emojis and hashtags from 20M English tweets (Barbieri et al., 2016), which were updated during the training.

Then, a biLSTM layer gets high-level features from previous embeddings. The LSTM were introduced by Hochreiter and Schmidhuber (1997) and were explicitly designed to avoid the long-term dependency problem. LSTM systems keep relevant information of inputs by incorporating a loop enabling data to flow from one step to the following. LSTM gets a word embedding sequentially, left to right, at each time step, produces a hidden step and keeps its hidden state through

time. Whereas, biLSTM, does the same process as standard LSTM, but processes the text in a left-to-right as well as right-to-left order in parallel. Therefore, gives two hidden state as output at each step and is able to capture backwards and long-range dependencies.

A critical and apparent disadvantage of seq2seq models (such as LSTM) is that they compress all information into a fixed-length vector, causing the incapability of remembering long tweets. Attention mechanism aims to overcome the limitation of fixed-length vector keeping relevant information from long tweet sequences. In addition, attention techniques have been recently demonstrated success in multiple areas of the Natural Language Processing such as question answering, machine translations, speech recognition and relation extraction (Bahdanau et al., 2014; Hermann et al., 2015; Chorowski et al., 2015; Zhou et al., 2016). For that reason, we added an attention layer, which produces a weight vector and merge word-level features from each time step into a tweet-level feature vector, by multiplying the weight vector. Finally, the tweet-level feature vector produced by the previous layers is used for classification task by a fully-connected layer.

Furthermore, we applied dropout regularization in order to alleviate overfitting. Dropout operation sets randomly to zero a proportion of the hidden units during forward propagation, creating more generalizable representations of data. As in Zhou et al. (2016), we employ dropout on the embedding layer, biLSTM layer and before the output layer. The dropout rate was set to 0.5 in all cases.

We used an additional annotated dataset for the sub-task A. This additional dataset was released with Shared Task on Aggression Identification organized as part of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC - 1) (Kumar et al., 2018). This dataset is composed of 15,000 aggression-annotated Facebook Posts that were annotated as Overtly Aggressive, Covertly Aggressive, and Non-aggressive texts. For this sub-task A, posts with aggressive annotations were considered as offensive (OFF) and Non-aggressive annotation as not (NOT).

For every sub-task, three different runs were submitted following the same scheme of the neural network. Specifically, for sub-task A, we submitted 2 runs taking into account the additional dataset (using Adam in the Run1A and RMSProp optimizer in the Run3A). The third run (Run2A) was obtained using only the dataset provided by the organizers and using Adam as optimizer.

For sub-tasks B and C, we did not use additional data for training. Instead, we weighted the classes in the training giving major relevance to unbalanced classes. For the rest of the runs, some parameters were changed in order to obtain different results. Specifically, the Run1B and Run1C the Adam optimizer was used with 50 units in the LSTM. The RMSProp optimizer was used in the Run2B and RUN2C with the previous number of LSTM units. Finally, in the Run3B and Run3C was also applied the RMSProp optimizer but with 100 units in the LSTM. Additionally, to improve the model performance and reducing the overfitting for sub-task C, which contains the smallest number of instances for training, the LSTM layer included a weight regularization (L1 and L2).

4 Results

F1 scores and accuracies of our three different submissions for sub-task A are shown in Table 1. For this sub-task we have achieved the highest score with the system we did not train with an additional dataset. F1 scores and accuracies of all submissions for the subsequent tasks are seen in Table 2 and 3 respectively.

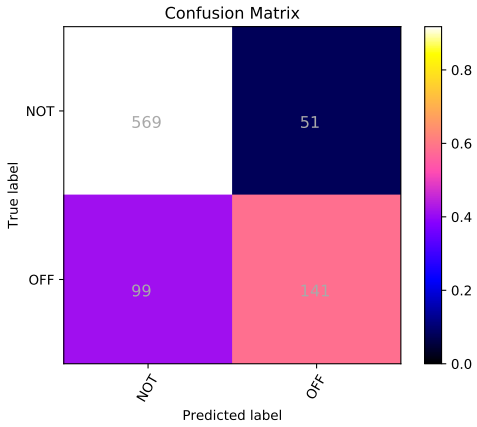

Figure 2: Confusion matrix of our best performed model for Sub-task A (biLSTM with specific configuration - Run2A)

The confusion matrix of our best performed model for the first task (see Figure 2) illustrates that between the two classes, NOT (not offensive) class achieves the best result where the majority of the data being correctly classified.

For sub-task B, classification of TIN (targeted

System	F1 (macro)	Accuracy
All NOT baseline	0.4189	0.7209
All OFF baseline	0.2182	0.2790
LaSTUS/TALN - Run1A	0.7406	0.7860
LaSTUS/TALN - Run2A	**0.7682**	**0.8256**
LaSTUS/TALN - Run3A	0.7411	0.7872

Table 1: Results of different submissions for **Sub-task A**.

System	F1 (macro)	Accuracy
All TIN baseline	0.4702	0.8875
All UNT baseline	0.1011	0.1125
LaSTUS/TALN - Run1B	0.6425	0.8458
LaSTUS/TALN - Run2B	0.6150	0.8292
LaSTUS/TALN - Run3B	**0.6618**	**0.8542**

Table 2: Results of different submissions for **Sub-task B**.

insult and threads) and UNT (untargeted) content from a sub-set of offensive tweets, the confusion matrix demonstrates that our best performed model has the highest precision for class TIN as shown in Figure 3.

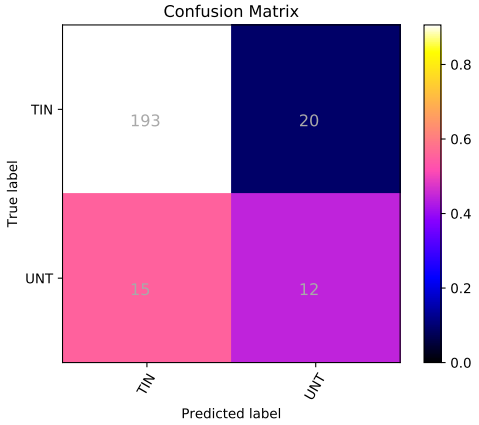

Figure 3: Confusion matrix of our best performed model for Sub-task B (biLSTM with specific configuration - Run3B)

The confusion matrix of our best performed model for sub-task C can be seen in Figure 4. It includes three classes as GRP (group), IND (individual) and OTH (other) for a sub-set of tweets containing targeted offense. The system achieves the highest precision for IND. The color range also shows the level of precision from darker to lighter.

Overall, we have achieved competitive results and rankings for each sub-task. Out of all our submissions, best performed ones for each sub-task and their comparison with the winner system of the shared task are given in Table 4.

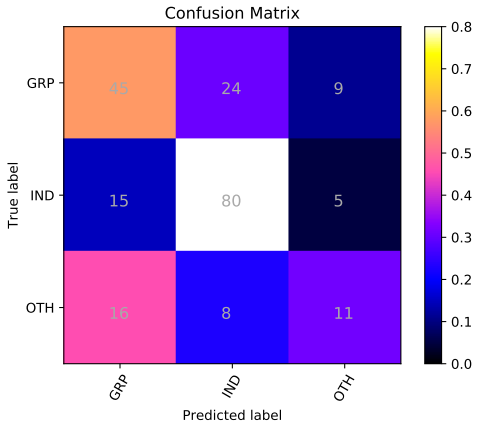

Figure 4: Confusion matrix of our best performed model for Sub-task C (biLSTM with specific configuration - Run3C)

5 Conclusion

In this paper, participation of LaSTUS/TALN to OffensEval: Identifying and Categorizing Offensive Language in Social Media (SemEval 2019 - Task 6) has been presented. We described and evaluated our system which is a bidirectional LSTM (biLSTM) model with an Attention layer on top, to classify if a tweet contains offensive language and the type and target of the offense for the offensive content.

For the future work, more detailed analyses on integration of linguistic annotations into neural network can be considered. In addition, a larger amount of data and also meta-data such as whether a tweet is a response to another tweet can represent contextual information and used to improve

System	F1 (macro)	Accuracy
All GRP baseline	0.1787	0.3662
All IND baseline	0.2130	0.4695
All OTH baseline	0.0941	0.1643
LaSTUS/TALN - Run1C	0.5631	0.6432
LaSTUS/TALN - Run2C	**0.5686**	**0.6385**
LaSTUS/TALN - Run3C	0.5480	0.6150

Table 3: Results of different submissions for **Sub-task C**.

	sub-task A	sub-task B	sub-task C
Best performer - F1(macro)	0.829	0.755	0.660
LaSTUS/TALN - F1(macro)	0.768	0.662	0.569
LaSTUS/TALN Ranking/Submissions	30 / 104	21 / 79	16 / 66

Table 4: Overall results and best rankings

performance.

Acknowledgements

Our work is partly supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502). We thanks two reviewers for their constructive comments.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 531–535. ACM.

Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. Evalita 2018: Overview on the 6th evaluation campaign of natural language processing and speech tools for italian. In *EVALITA@CLiC-it*.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585.

T. Y. Chu, Kylie Jue, and Max L. Wang. 2017. Comment abuse classification with deep learning.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. *arXiv preprint arXiv:1804.04257*.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90.

Michele P. Hamm, Amanda S. Newton, Annabritt Chisholm, Jocelyn Shulhan, Andrea Milne, Purnima Sundar, Heather Ennis, Shannon D. Scott, and Lisa Hartling. 2015. Prevalence and effect of cyberbullying on children and young people: A scoping review of social media studies. *JAMA pediatrics*, 169 8:770–7.

Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. In *PloS one*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC)*, Santa Fe, USA.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection Using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics*, pages 1–10, Valencia, Spain.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Langauge Online*.

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. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

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 (SemEval)*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *Lecture Notes in Computer Science*. Springer Verlag.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 207–212.