GL at SemEval-2019 Task 5: Identifying hateful tweets with a deep learning approach.

Gretel Liz De la Peña Sarracén
Universitat Politècnica de València, Spain

Abstract

This paper describes the system we developed for SemEval 2019 on Multilingual detection of hate speech against immigrants and women in Twitter (HatEval - Task 5). We use an approach based on an Attention-based Long Short-Term Memory Recurrent Neural Network. In particular, we build a Bidirectional LSTM to extract information from the word embeddings over the sentence, then apply attention over the hidden states and finally feed this vector to another LSTM model to get a representation from de data. Then, the output obtained with this model is used to get the prediction of each of the sub-tasks with models based on neural networks and linguistic characteristics.

1 Introduction

Nowadays, the number of content generated by users on social networks is growing rapidly. In this context, the problem of detecting and limiting the dissemination of the Hate Speech is becoming a matter of great importance. Therefore, many efforts are dedicated to studying and treating this phenomenon. A large number of workshops on this topic have been developed in recent years, which reflects the interest of many researchers.

Some examples are the Workshop on Trolling, Aggression and Cyberbullying (Kumar et al., 2018), that included a shared task on aggression identification; the tracks on Automatic Misogyny Identification (AMI) (Fersini et al., 2018a) and on Autohorship and Aggressiveness Analysis (MEX-A3T) (Álvarez-Carmona et al., 2018) proposed at IberEval 2018; the Automatic Misogyny Identificationtask at EVALITA 2018 (Fersini et al., 2018b), the Workshop on Abusive Language (Waseem et al., 2017) and the GermEval Shared Task on the Identification of Offensive Language (Wiegand et al., 2018).

The proposed works have used different features and models. Among them, models based on deep learning, such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) have been widely used.

This paper presents a strategy based on RNN, which is an extension of previous models proposed for the tasks MEX-A3T and EVALITA 2018 (Cuza et al., 2018; la Peña Sarracén et al., 2018). Those models use an Attention-based LSTM inspired by the work (Yang et al., 2016). In this work the authors use a hierarchical attention network for document classification and their experiments show that the architecture outperforms previous methods. This last model has two levels of attention mechanisms applied at the word and sentence level, enabling it to attend differentially to more and less important content when constructing the document representation.

The aim of the proposed extension in this work is not only the hate language identification, but also the study of other features of hateful messages. In this case, the model based on the Attention mechanism and the LSTM is used as a representation of the input. This representation is then used to detect the hate language with a fully connected network. In addition, it is combined with some linguistic characteristics to analyze the other features of hateful messages. The objective has been defined for the HatEval[1] shared task on SemEval 2019.

The HatEval (Basile et al., 2019) task consists in Hate Speech identification in messages from Twitter in Spanish and English. The main objective focuses on detecting the hate expressed against women and immigrants in particular. The task is divided into two related subtasks: a binary classification as a first sub-task about Hate Speech de-

[1] https://competitions.codalab.org/competitions/19935

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 416–419
Minneapolis, Minnesota, USA, June 6–7, 2019.

tection, and another one where other features of hateful contents is investigated:

- TASK A - Hate Speech Detection against Immigrants and Women: Predicting whether a tweet with a given target (women or immigrants) is hateful or not (HS).
- TASK B - Aggressive behavior and Target Classification: Classifying hateful tweets as aggressive or not (AG), and second identifying the target harassed as individual or generic (TR).

The paper is organized as follows. Section 2 describes the proposed methodology. Experimental results are then discussed in Section3. Finally, we present our conclusions with a summary of our findings in Section 4.

2 Methodology

In this work, a simple preprocessing is performed in which the text is cleaned. First, emoticons, hashtags, urls, and other strings that do not represent alphabetic sequences are eliminated. Then, the texts are represented as vectors with a word embedding model. We used pre-trained word vectors of Glove (Pennington et al., 2014), trained on 2 billion words from Twitter for English. On the other hand, for Spanish we used the word vectors of fasttext (Bojanowski et al., 2017).

In general, we propose a model that consists of a Bidirectional LSTM neural network (Bi-LSTM) at the word level in the input. At each time step the Bi-LSTM gets as input a word embedding. Afterward, an attention layer is applied over each hidden state. The attention weights are learned using the concatenation of the current hidden state of the Bi-LSTM and the past hidden state of another LSTM. In this way, a representation (R) of the input is obtained from the last LSTM to get the prediction of each of the subtasks, as shown in Figure 1.

2.1 Sub-task A

For the sub-task A, which consists in the identification of hate in tweets (HS), R is used as input of a Fully Connected Neural Network (FCNN) of two dense layers with the relu activation function. The class (hatefull or not) is obtained in an output layer with two units, relative to the number of classes, with the softmax activation function.

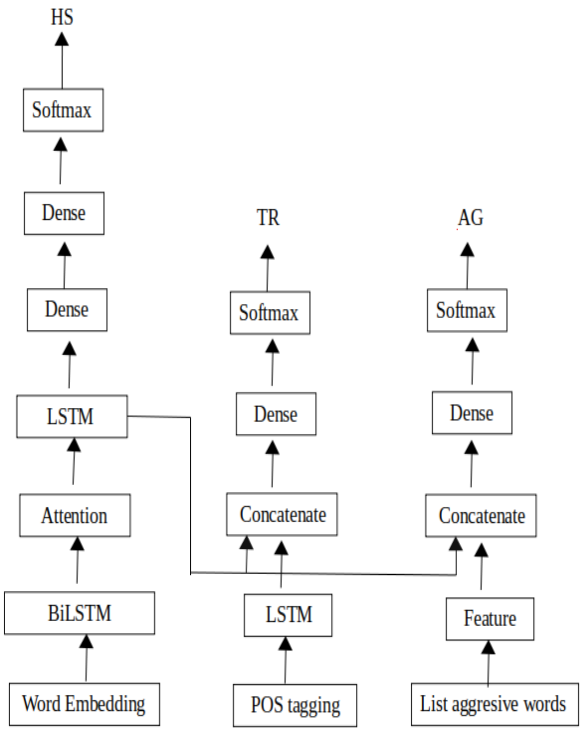

Figure 1: General architecture

2.2 Sub-task B

In the sub-task B the aim is to analyze some features of hateful messages, as discussed above. On the one hand, identifying against whom is directed the hate language (TR) and on the other hand, detecting whether a message with hate language is also aggressive.

For the task of detecting the target of hate (TR), the information of the part-of-speech tagging process of the tweets is used. The sequence of labels is analyzed with a LSTM RNN, obtaining a vector. Then, this vector is concatenated with R (from subtask A) and it is used as input to a dense layer with the relu activation function. Finally, the output layer has two neurons with the softmax activation function. In this way, the prediction corresponding to the offensive target in the tweets is obtained.

In the case of the task of classifying a hateful tweet in aggressive or not, a linguistic resource that consists of a dictionary of aggressive words is used. In this way, a linguistic characteristic is added to R according to the number of words in the tweet that appear in the dictionary. In addition, a one hot vector corresponding to the POS tags present in the tweet is added. With this new vector, the prediction is obtained in a similar way to

the previous task, using another dense layer with the relu activation function.

3 Results

For the evaluation of the results of the task, different strategies and metrics are applied. For the sub-task A, systems are evaluated using standard evaluation metrics, including accuracy, precision, recall and F1-score. The submissions are ranked by F1-score. For the sub-task B, systems are evaluated with two criteria: partial match and exact match (EMR). In partial match, each dimension to be predicted (HS, TR and AG) is evaluated independently of the others using standard evaluation metrics including accuracy, precision, recall and F1-score, and then combined. In exact match, all the dimensions to be predicted are jointly considered. The submissions are ranked by the EMR measure.

Table 1 shows the results obtained for each of the languages in the sub-task A. In addition, the results of the system positioned in the first place of the ranking are shown for each of the metrics. In the same way, the results obtained for the subtask B are shown in Table 2.

Language	Task A			
	Acc	P	R	F1-score
English	0.453	0.545	0.516	0.39
Best-English	0.653	0.69	0.679	0.651
Spanish	0.723	0.717	0.722	0.718
Best-Spanish	0.731	0.734	0.741	0.73

Table 1: Results for the sub-task A

Language	Task B	
	F1-score	EMR
English	0.532	0.268
Best-English	0.467	0.57
Spanish	0.74	0.618
Best-Spanish	0.755	0.705

Table 2: Results for the sub-task B

As can be seen, the results for English are very far from the best results obtained in the competition, reaching a low position in the ranking of the participating systems. On the other hand, the results for Spanish are better, reaching the eighth position in the ranking for task A and the eleventh for task B. However, these results are not good enough, which reveals that possibly the complexity of the model used is a problem. Therefore, a better approach might be to simplify the model in order to reduce the number of parameters to train. On the other hand, we think that it is very important for improving the results, to find discriminatory linguistic features, able of capturing the nature of the texts with hate; and to make a fined tunned of paramaters for each language.

4 Conclusion

In this paper we presented our solution for HatEval task in SemEval 2019. We used an approximation that combines a BiLSTM RNN with an attention mechanism to obtain a text representation vector. This vector is used as input in each of the models designed for each of the subtasks in which HatEval is divided. The results obtained were not very good, far away from the results obtained by the best system in the competition. As a conclusion of the analysis, it is believed that a better approximation would be to simplify the proposed model to reduce the number of parameters to train. Also, searching for discriminating linguistic features, which manage to capture the nature of texts with hate, should help to obtain better results.

References

Miguel Álvarez-Carmona, Estefanıa Guzmán-Falcón, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villasenor-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 *Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain*, volume 6.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, 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 (SemEval-2019)*. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Carlos Enrique Muniz Cuza, Gretel Liz De la Pena Sarracén, and Paolo Rosso. 2018. Attention mechanism

for aggressive detection. In *CEUR Workshop Proceedings*, volume 2150, pages 114–118.

Elisabetta Fersini, Maria Anzovino, and Paolo Rosso. 2018a. Overview of the task on automatic misogyny identification at ibereval. 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). CEUR Workshop Proceedings. CEUR-WS. org, Seville, Spain*.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018b. Overview of the evalita 2018 task on automatic misogyny identification (ami). *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA18), Turin, Italy. CEUR. org*.

Ritesh Kumar, Atul Kr Ojha, Marcos Zampieri, and Shervin Malmasi. 2018. Proceedings of the first workshop on trolling, aggression and cyberbullying (trac-2018). In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*.

Gretel Liz De la Peña Sarracén, Reynaldo Gil Pons, Carlos Enrique Muñiz-Cuza, and Paolo Rosso. 2018. Hate speech detection using attention-based LSTM. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018.*

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault. 2017. Proceedings of the first workshop on abusive language online. In *Proceedings of the First Workshop on Abusive Language Online*.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *14th Conference on Natural Language Processing KONVENS 2018*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.