Identification of Offensive Language in Social Media

Lutfiye Seda Mut Altin
LaSTUS-TALN Research Group, DTIC
Universitat Pompeu Fabra
C/Tànger 122-140, 08018 Barcelona, Spain
(lutfiyeseda.mut01@estudiant.upf.edu)

Abstract: Recent work shows that offensive language in social media is a serious problem that affects especially vulnerable groups. Therefore, systems designed to detect offensive language automatically have been the focus of attention of several works. Various Machine Learning approaches have been utilised for the classification of offensive text data. Within the scope of this research we aim to develop a neural network system that will effectively classify offensive text considering different aspects of it. In addition, multilingual and multi-task learning experiments are planned.

Keywords: Offensive language, Social media, Neural network, Bi-LSTM

Resumen: El uso de lenguaje ofensivo en las redes sociales es un problema que afecta especialmente a las personas vulnerables. Es por esta razón que el desarrollo de sistemas automáticos para la detección de lenguaje ofensivo es una tarea de considerable importancia social. En esta investigación nos proponemos desarrollar sistemas basados en técnicas recientes de aprendizaje de maquina tales como las redes neuronales para la clasificación de lenguaje ofensivo. Así mismo nos proponemos realizar experimentos con datos multilingües (español e inglés) y la aplicación de técnicas multitarea que estén relacionadas con este problema.

Palabras clave: Identificación de lenguaje ofensivo, Redes sociales, Redes neuronales, Bi-LSTM

1 *Motivation and Background*

Social media has become one of the most important environments for communication among people. As user-generated content on social media increases significantly, so does the harmful content such as offensive language. Aggressiveness in social media is a problem that especially affects vulnerable groups (Hamm et al., 2015), (Kowalski and Limber, 2013). Within this context, the need for automatic detection of offensive content gains a lot of attraction.

Traditional methods to detect offensive language include use of blacklisted keywords and phrases based on profane words, regular expressions, guidelines and human moderators to manually review and detect unwanted content. However, these methods are not sufficient, particularly considering the users that tend to use more obfuscated and implicit expressions.

Automatic identification of offensive language is essentially considered as a classification task. Previous research on the topic include approaches from different perspectives, utilizing different data sets and focusing on various contents such as abusive language (Waseem et al., 2017) (Chu, Jue, and Wang, 2016), hate speech (Davidson et al., 2017) (Schmidt and Wiegand, 2017) (Fortuna and Nunes, 2018) and cyberbullying (Van Hee et al., 2018).

Where machine learning approaches are of concern, (Davidson et al., 2017) indicated using certain terms and lexicons are useful. (Zhang, Robinson, and Tepper, 2018) compared different approaches and 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 classifiers

Lloret, E.; Saquete, E.; Martínez-Barco, P.; Sepúlveda-Torres, R. (eds.) *Proceedings of the Doctoral Symposium of the XXXV International Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)*, p. 50–55
Bilbao, Spain, September 25th 2019.

such as SVM.

There are several previous shared tasks similar to offensive language detection. The shared task on Aggression Identification called 'TRAC' provided participants a dataset containing annotated Facebook posts and comments in English and Hindi (Kumar et al., 2018). Aiming to classify the text among three classes including nonaggressive, covertly aggressive, and overtly aggressive. The best-performing systems in this task used deep learning approaches based on convolutional neural networks (CNN), recurrent neural networks and LSTM (Majumder, Mandl, and others, 2018). The Spanish language has also been considered. For example, in the recent shared task, MEX-A3T 2018, regarding aggression detection in Mexican Spanish; among the methodologies proposed by participants, there were content based (bag of words, word n-grams, dictionary words, slang words etc.) and stylistic-based features (frequencies, punctuations, POS etc.) as well as approaches based on neural networks (CNN, LSTM and others); baselines were outperformed by the most participants (Álvarez-Carmona et al., 2018). Furthermore, other shared tasks focusing on aggression in other languages include Italian, German (Bosco et al., 2018),(Wiegand, Siegel, and Ruppenhofer, 2018). One of the most recent shared task on the topic is "Categorizing Offensive Language in Social Media" (SemEval 2019 - Task 6) (Zampieri et al., 2019b). Referring to the problem in a hierarchichal scheme including the target type of the offense. To classify offensive text, about 70 % of the participants used deep learning approaches. Among the top-10 teams, seven used BERT (Devlin et al., 2018).

2 Methodology and Proposed Experiments

After an extensive literature review, collection of additional previous datasets related to the topic and preliminary experiments; we started to experiments through shared tasks as described below.

2.1 Participation to 'Categorizing Offensive Language in Social Media (SemEval 2019-Task 6)'

A bi-LSTM neural network model that has been developed (Altin, Serrano, and Saggion, 2019) within the context of the participation to shared task which is called 'Categorizing Offensive Language in Social Media (SemEval 2019 - Task 6)', focusing on identification of offensive language by considering type and target of the offense into account (Zampieri et al., 2019b).

This model consists of a bidirectional Long Short-Term Memory Networks (biLSTM) model with an Attention layer on top. The model captures the most important semantic information in a tweet, including emojis and hashtags. A simplified schema of our model can be seen in the following figure.

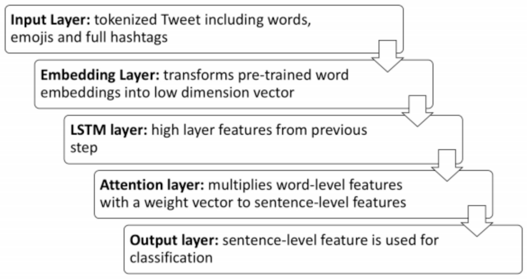
Input Layer: tokenized Tweet including words, emojis and full hashtags
Embedding Layer: transforms pre-trained word embeddings into low dimension vector
LSTM layer: high layer features from previous step
Attention layer: multiplies word-level features with a weight vector to sentence-level features
Output layer: sentence-level feature is used for classification

Figura 1: 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 longterm 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 longrange 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.

2.2 Experimenting with Multi-task Learning: Initial Experiments on Aggressiveness detection

In this work, we presented a bi-LSTM model with two dense layers at the end. We have developed a system in the context of the shared task: MEX-A3T: Authorship and aggressiveness analysis in twitter. Specifically, the Aggressiveness Identification track, which focuses on the detection of aggressive comments in tweets from Mexican users and the other related IberLEF 2019 shared tasks.

We have used data from different tasks in order to train more examples in the model. As we believe that the tasks of humor and sentiment analysis could help in detecting aggressive language, we have selected three additional task to train with MEX-A3T at the same time. The other tasks were IroSva, that aims investigating the recognition of irony in Twitter messages in three different Spanish variants (from Spain, Mexico, and Cuba); HAHA which we used the classification task related to identify if a Spanish tweet is a joke or not and TASS 2019 that focuses on the evaluation of polarity classification systems of tweets written in Spanish. We used the data related to this task, tweets written in the Spanish language spoken in Spain, Peru, Costa Rica, Uruguay and Mexico, which were annotated with 4 different levels of opinion intensity (Positive, Negative, Neutral and Nothing).

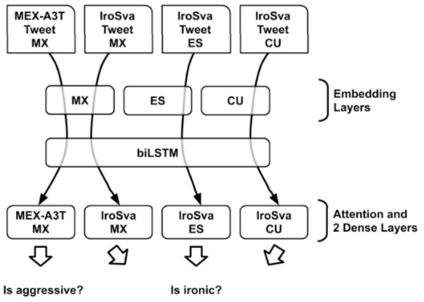

Figura 2: Simplified schema of the multi- task model

In this scenario, we defined an Embedding layer for each Spanish variant in IroSva task. Classification tasks with the same Spanish variant used the same Embedding layer during the training process. Furthermore, all task shared the biLSTM layer during training. For the moment this approach was not very successful; however this may be due to lack of data to train the different models.

3 *Current work*

Despite the progress in this shared task, there are potential issues for the future work. Future experiments were planned mainly in 2 groups:

First, improvement areas will be investigated for the efficiency of the classification model developed for SemEval 2019 - Task 6 shared task, with the same dataset that is called Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019a).

Initial experiments have been done taking only the words into account. Using additional features such as WordNet synsets, Part of Speech (POS) tags, frequencies, offensive word dictionaries and so on, is expected to improve the precision of the results.

Furthermore, changes in the methodology such as applying 'Bidirectional Encoder Representations from Transformers' (Devlin et. al,2018) is also another option.

Secondly, in a later phase of the study, it is planned to obtain a new dataset using Twitter's streaming API and crowdannotation and using the new dataset for the experiments including the metadata such as user-session time, whether it is a reply or a retweet.

For this purpose, first of all, a set of specific hashtags will be decided with a high potential of being associated with offensive tweets.

After pulling the data and deciding the annotation scheme, the data will be presented for crowd annotation.

After compilation of a corpus, model training will be carried out with the system given the most promising results for the OLID dataset.

Additional improvements for the system design and other potential features will be experimented considering the performance of the preliminary tests.

4 Specific Issues of Investigation

The main research questions that are intended to answer with this work are as the following:

•What algorithms are those that provide us with greater accuracy to identify offensive language in a text?

•What characteristics should be taken into account in the process of analysing text in terms of aggressiveness?

•What type of metadata would be useful to increase the accuracy while analysing the text?

•Finally, how would be the overall system for this classification task that will bring the highest accuracy?

5 Thesis Objectives

The main objectives of the research can be listed as follows:

I.Executing preliminary experiments to classify offensive messages in social media (particularly tweets and short messages) datasets.

There are several published datasets belonging previous researches that is annotated as Offensive or within the similar context such as cyberbullying, hate speech related, misogyny [1,2,3].According to the specific annotation scheme and the content, handcrafted features might have an important parameter for the performance. Experimenting on these previous datasets will help understanding the strengths and weaknesses of different design specifications and features and eventually help optimization of them.

II. Experiments to improve the performance of the current system with fine-tuned system design and feature engineering.

The neural network system for the initial experiments took only words into account. However, there is a potential to improve the results of this system with additional feature extraction. Furthermore, detailed analysis on integration of linguistic annotations into neural network and other models like convolution can be considered to improve the performance.

III. Creating a new dataset with crowd annotation. There are several crowd annotation platforms such as: Mechanical Turk[4] , crowdflower[5] , crowdtruth[6] . By uploading the data and deciding the rules of annotation these platforms help annotating the data by human annotators.

To crowd-annotate tweet data, first of all, the data will be pulled from Twitter API according to certain hashtags. Hashtags will be decided for certain contexts such as political debate hashtags or hashtags related to sportive rivalry. After that, annotation schema will be decided. Annotation schema of previous datasets are usually in hierarchical order and contains additional information such as target or for instance if it contains aggression whether it is cyberbullying or not.

IV. Experiments on the new dataset with various approaches on the system and features.

A new dataset can give the opportunity to

[1] https://www.kaggle.com/alternacx/hateoffensive-speechdetection

[2] https://www.amnesty.org/en/

[3] https://zenodo.org/record/1184178.XTBv2pMzaRt

[4] https://www.mturk.com/

[5] https://www.figure-eight.com/

[6] http://crowdtruth.org/

reproduce previous well-performed systems designed. Moreover, majority of the related datasets published do not include metadata. With the new dataset collected through Twitter API it will be possible to obtain metadata, as well. Therefore, user-related features such as the frequency of profanity in previous messages can be obtained and it would help understand the importance of metadata on the performance.

References

Altin, L. S. M., À. B. Serrano, and H. Saggion. 2019. Lastus/taln at semeval-2019 task 6: Identification and categorization of offensive language in social media with attention-based bi-lstm model. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 672–677.

Álvarez-Carmona, M. Á., E. Guzmán-Falcón, M. Montes-y Gómez, H. J. Escalante, L. Villasenor-Pineda, V. Reyes-Meza, and A. 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.

Bosco, C., D. Felice, F. Poletto, M. Sanguinetti, and T. 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.

Chu, T., K. Jue, and M. Wang. 2016. Comment abuse classification with deep learning. *Von https://web. stanford. edu/class/cs224n/reports/2762092. pdf abgerufen.*

Davidson, T., D. Warmsley, M. Macy, and I. Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media.*

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Fortuna, P. and S. Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Hamm, M. P., A. S. Newton, A. Chisholm, J. Shulhan, A. Milne, P. Sundar, H. Ennis, S. D. Scott, and L. Hartling. 2015. Prevalence and effect of cyberbullying on children and young people: A scoping review of social media studies. *JAMA pediatrics*, 169(8):770–777.

Kowalski, R. M. and S. P. Limber. 2013. Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of Adolescent Health*, 53(1):S13–S20.

Kumar, R., A. K. Ojha, S. Malmasi, and M. Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Majumder, P., T. Mandl, et al. 2018. Filtering aggression from the multilingual social media feed. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 199–207.

Schmidt, A. and M. 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*, pages 1–10.

Van Hee, C., G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste. 2018. Automatic detection of cyberbullying in social media text. *PloS one*, 13(10):e0203794.

Waseem, Z., T. Davidson, D. Warmsley, and I. Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899.*

Wiegand, M., M. Siegel, and J. Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Zampieri, M., S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019a. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666.*

Zampieri, M., S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

Zhang, Z., D. Robinson, and J. Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European Semantic Web Conference*, pages 745–760. Springer.