pin_cod_ at SemEval-2020 Task 12: Injecting Lexicons into Bidirectional Long Short-Term Memory Networks to Detect Turkish Offensive Tweets

Pınar Arslan
Independent Researcher
0707.pinar@gmail.com

Abstract

This paper describes a system (pin_cod_) built for SemEval 2020 Task 12: OffensEval: Multilingual Offensive Language Identification in Social Media (Zampieri et al., 2020). I present the system based on the architecture of bidirectional long short-term memory networks (BiLSTM) concatenated with lexicon-based features and a social-network specific feature and then followed by two fully connected dense layers for detecting Turkish offensive tweets. The pin_cod_'s system achieved a macro F1-score of 0.7496 for Sub-task A - Offensive language identification in Turkish.

1 Introduction

With the appearance of influential social media platforms, offensive language is becoming more prevalent and visible. Harbouring behind a physically invisible author, bullies—either an intimate partner or an absolute stranger to victims— spread abusive, offensive and hateful messages against a particular person or a group of people through many internet platforms. The scope of the victims is broad and especially teenagers, women and immigrants are among targets for bullies. Online bullying leads to mental health issues including anxiety disorder, depression, it reduces self confidence, and causes lower academic achievement.

Computational approaches to identify online bullying, hate speech, offensive and abusive language have gained acceleration and attention as a number of workshops have been organized for this purpose (Bosco et al., 2018; Fersini et al., 2018; Basile et al., 2019; Zampieri et al., 2019) towards generic users, women and/or immigrants in English, Italian, Spanish and German. SemEval 2020 Task 12 : OffensEval 2020 is probably the first workshop for identifying offensive language for Turkish language with Sub-task A - Offensive language identification (Çöltekin, 2020). I participated in this sub-task the main goal of which is to segregate offensive posts from not offensive ones (i.e., OFF for offensive, NOT for not offensive class). Offensive posts comprise insults, threats, and untargeted profanity. Non-offensive posts do not contain any offense or profanity. The pin_cod_'s system for this binary classification sub-task was based on BiLSTM networks incorporated with various lexicon-based features and the presence of user mentions (e.g. @username) and then followed by two fully connected dense layers. The scripts for preprocessing and the BiLSTM model can be found here[1]. The current study shows the features obtained from several lexicons and a social-network specific feature increased the performance of the BiLSTM model with a satisfactory F1-score of 0.7496.

This paper is organized as follows: related work in section 2, system description in section 3, results in section 4, error analysis and discussion in section 5, and conclusions in section 6.

2 Related Work

A spate of studies investigated hate speech in recent decades (Kwok and Wang, 2013; Burnap and Williams, 2015; Silva et al., 2016; Corazza et al., 2018a). The phenomenon of hate speech has been investigated

[1] https://github.com/0707pinar/Offensive-language-identification

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 2117–2122
Barcelona, Spain (Online), December 12, 2020.

under specific aspects, for instance, cyberbullying (Xu et al., 2012; Dinakar et al., 2012; Zhong et al., 2016; Arslan et al., 2019), offensive language (Razavi et al., 2010; Corazza et al., 2018b; Zampieri et al., 2019), abusive language (Mubarak et al., 2017), insults and profanity (Sood et al., 2012). A typical Natural Language Processing methodology that deals with hate speech detection involves classifying large volumes of text using supervised machine learning approaches (Chen et al., 2012; Van Hee et al., 2015; Waseem and Hovy, 2016).

3 System Description

I followed a supervised learning approach to identify offensive language in Turkish. I implemented Bidirectional Long Short-Term Memory Networks using word embeddings on the Turkish Twitter dataset provided by the organizer in SemEval-2020 Task 12 OffensEval: Sub-task A - Offensive language identification (Çöltekin, 2020). To predict the labels for the given test set consisting of 3528 unlabeled tweets, I only used the provided training set containing 31756 labeled tweets which was split into two parts: 80% for training set and 20% for validation set.

3.1 Data Preprocessing

I applied the following preprocessing steps on the provided datasets: (1) removing hashtag symbols, (2) removing mention tag symbol, (3) removing punctuations, (4) tagging numbers, (5) lowercasing letters and (6) word tokenization. After preprocessing, '@example This is an Example, #example 1.' would be shown as ['example', 'this', 'is', 'an', 'example', 'example', 'number']. The preprocessing script was written in Python 3.5.9 on macOS Catalina (version 10.15.3).

3.2 Feature Description

The following text-driven features were used in the final model:

- **Word embeddings**: Turkish fastText embeddings (Joulin et al., 2018)[2] were employed in the BiLSTM model.

- **Sentiment features**: NRC Emotion Lexicon also called EmoLex (Mohammad and Turney, 2013) was used for counting the number of negative words per tweet. The EmoLex provides word-level emotion (i.e., anger, fear, anticipation, trust, surprise, sadness, joy, disgust) and sentiment (i.e., negative, positive) tags for Turkish emotion and sentiment bearing words. Upon a detailed investigation to unveil the contribution of the tags on the pin_cod_'s system, only the negative tag was decided to be used as a feature.

- **Hate speech-related features**: I used the HurtLex lexicon (Bassignana et al., 2018), which consists of negative stereotypes, hate words, slurs beyond stereotypes and other words and insults, for Turkish. Upon applying a detailed experimentation to reveal the sub-categories of HurtLex increasing the performance of the pin_cod_'s system, the following HurtLex sub-categories were used in the final system: DDP: cognitive disabilities and diversity, DMC: moral and behavioral defects, OR: plants, ASM: male genitalia, ASF: female genitalia, OM: words related to homosexuality, CDS: derogatory words. The number of each of these selected categories per tweet was used as a feature.

- **Offensive/Profane/Slang word lists**: I compiled various wordlists[3]. Revising each word in these lists, if necessary, I made some additions or removals (e.g. internet slangs with positive polarity 'panpa', 'kanks' which both mean 'mate', 'lads', 'dudes' and are used to refer a friend were removed.) Then, I checked the presence of the offensive, profane and slang words per tweet. If present, the number of these words was counted per tweet and used as a feature.

- **Social-network specific feature**: The presence of mention tags per tweet was used as a feature. If a tweet contains a mention tag, it is represented with 1, otherwise, shown as 0.

[2] https://github.com/facebookresearch/fastText

[3] https://en.wikibooks.org/wiki/Turkish/Slang#Offensive;https://github.com/ooguz/turkce-kufur-karaliste/blob/master/README.md;http://tanersezer.com/?p=239

3.3 Bidirectional Long Short-Term Networks

The core of my offensive language classification model was a bidirectional recurrent neural network, specifically bidirectional long short-term memory networks (Hochreiter and Schmidhuber, 1997) which is one of the most popular architectures used in natural language processing tasks such as text classification. BiLSTM was chosen since it provides further context to the network by training one LSTM for the input sequence and the second one for the reversed copy of the input sequence.

FastText embeddings were employed for the BiLSTM model which was implemented based on Keras 2.3.1 using TensorFlow 2.1.0 backend. The script was written in Python 3.5.9 on macOS Catalina (version 10.15.3). Dimensional vectors were set to 300. Input vector length was set to 82. Upon applying some grid search on the number of neurons, dropout regulations, number of epochs, batch sizes, optimization algorithms, activation functions and patience in the early stopping, the number of neurons in the BiLSTM layer was set to 200, both dropout and recurrent dropout were set to 0.2. The recurrent layer was concatenated with the dense text-driven features stated in the subsection 3.2. The concatenated layer was then fed to two fully connected dense layers with 100 neurons and rectified linear units activation function each. Final layer had 2 units representing the number of classes to be predicted and 'sigmoid' activation function which returned the probabilities of each class between the range of 0 and 1. I compiled the offensive language classification model by using 'binary crossentropy' loss, 'adam' optimizer and 'accuracy' metrics. Some callbacks, a set of functions to be applied at given stages of the training procedure, were applied. Validation accuracy was monitored and the number of epochs with no improvement after which training would be stopped was set to 5. The latest best model based on the monitored validation accuracy was saved. Then, the labels (NOT or OFF) were predicted for the test set.

4 Results

For OffensEval 2020: Sub-task A - Offensive language identification in Turkish, the architecture was built based on the bidirectional long short-term memory networks concatenated with a social-network specific feature, sentiment feature, lexicon-based features and word embeddings. Then, it was followed by two fully connected layers so that the test set consisting of Turkish tweets were predicted as either offensive (OFF) or not offensive (NOT). I did not use any external datasets apart from the training set provided by the organizer. The results of pin_cod_'s participation in the Sub-task A of OffensEval task on the test set are presented in Table 1. The confusion matrix for the test set classification is displayed in Table 2.

class	**precision**	**recall**	**f1-score**
NOT	0.88	0.95	0.91
OFF	0.71	0.50	0.59
macro avg	0.80	0.72	0.75
weighted avg	0.85	0.86	0.85

Table 1: pin_cod_'s system test results per class in sub-task A of OffensEval task for Turkish.

Predicted label			
		NOT	OFF
True label	NOT	2670	142
	OFF	361	355

Table 2: Confusion matrix for sub-task A of OffensEval task for Turkish.

The official evaluation metric in OffensEval 2020 was macro f1-score. The pin_cod_'s system ranked 23rd among 46 participants shown in Table 3.

5 Error Analysis and Discussion

The pin_cod_'s system detecting Turkish offensive tweets achieved a satisfactory result. Yet some misclassifications were obtained as presented in Table 2. The system yielded more false negatives than

Team name	**Ranking**	**Macro F1**
Galileo	1	0.82576383
lukez	15	0.772856039
pin_cod_	23	0.749619862
prhlt-upv	35	0.712691885
SpurthiAH	46	0.310887838

Table 3: pin_cod_'s system results relevant to other selected systems in sub-task A of OffensEval task for Turkish.

false positives, which could stem from the fact that the majority class of both training and test sets was not offensive (i.e. NOT). The misclassifications emanated from false negatives, which means that true label is offensive but the system predicts as not offensive, were mostly related to the following specific phenomena: (i) sarcasm (e.g. 'İmparator büyük takımlar dedi zaten *lig 4 uncusunden bahsetmedi ki*' - English translation: 'The emperor said big teams, *he did not mention the fourth of the league anyway*'), (ii) limited coverage of the lexicons (e.g. 'Gidin bi kurban kesin, hamama gidin.. bişey yapın. bu ne *cenabetliktir* arkadaş.' - English translation: 'Go sacrifice an animal, go to a bath.. do something. what an *impurity* friend.'), (iii) misspellings and whitespaces (e.g. '*Dıyer* kanal *lar* akp *nı* yalan *makınesı* [intended: Diğer kanallar AKP'nin yalan makinesi]' - English translation: '*Other* channel*s* AKP'*s polygraph*'), (iv) polysemy (e.g. 'Burada atıp tutacağına o kötü koşullarda 3 kuruşa sen çalışsana *yiyorsa*' - English translation: 'If you *dare*, work for 3 pennies under the bad conditions rather than swagger here'), (v) multiword expressions (e.g. 'Hepsi *ceplerini* çok güzel *dolduruyor* vatandaşta birbirine sayıyor işte' - English translation: 'They all *feather* fabulously *their own nests* folks are cursing each other'), (vi) metaphorical expressions (e.g. 'bazıları da var görüp cevap vermiyor, *bizim kangal mı daha insan yoksa onlar mı?*' - English translation: 'some of them see but do not respond, *is our kangal dog or are they more human?*'). The misclassified tweets with false positives were usually containing negative words and in generic or imperative form but they did not carry an offensive meaning (e.g. '#HayvanaSiddetSuctur *tecavüzcü* müebbet, *işkenceci* caydırıcı sürede hapis cezası almalı. Bu dünyanın sahibi ne sensin ne benim. Herkes haddini bilsin artık.' - English translation: '#ViolenceAgainstAnimalsIsCrime *the rapist* should be sentenced to life in prison, *the torturer* should receive a prison term in the deterrent period. Neither you nor I are the owner of this world. Everyone should know their limits.').

Some inconsistencies and incorrect annotations were also noticed in the gold labels. The tweet 'Yeeni yıla bak *anasını satıyım*6 Günündeeyiz 5 günü tatildiGeeldee seevmee böylee yılıafeerin yeeni yılafeerin' meaning 'Look at the new year *what the heck* we are on the 6th day it was 5 days of vacation how can I not love such a year well done new year well done' had an offensive label. In fact, it carries a positive meaning although it consists of a slang word. Some words such as 'manyak' meaning 'maniac' existed in both offensive and non-offensive tweets. Semantically similar contexts containing this word were labeled differently in the gold standard, which might have caused the system to misclassify some tweets. For instance, 'Bana sevgili degilher şeyi beraber yapabilecegim *manyak* bi insan lazım' meaning 'I do not need a sweetheart I need a *maniac* person with whom I can do everything together' and '*Manyak* işte Allah bilir kafaya neyi takmıştır.' meaning '*Maniac* huh God knows what was in his mind.' were both labeled as offensive, while 'İnsan sevdiği için tescilli *manyak* olabiliyorsa seviyordur bizi sınamayın' meaning 'If a person can become a registered *maniac* for his sweetheart, he loves her do not try us' was labeled as not offensive.

A conclusion might be made that an algorithmic model of detecting offensive language cannot be solely limited to detecting bad words or slang words. Certain insulting terms might sound differently to a friend or an absolute stranger. Other factors such as the use of emojis along with such insulting terms might make a message offensive or not offensive. On the contrary, highly offensive messages might not necessarily include any toxic and hurtful words while they contain some metaphors or metonymies. If the predicting model is restricted to the textual analysis of content, it will likely boost the chances of yielding false negatives and false positives.

6 Conclusions

In this paper, I presented the pin_cod_'s system I have developed as part of my participation in SemEval-2020 Task 12: OffensEval 2020: Multilingual Offensive Language Identification in Social Media. Specifically, I have participated in Sub-task A - Offensive language identification for Turkish language. To bring a solution for this task, I adopted bidirectional long short-term memory neural networks incorporating lexical features from a polarity lexicon, hate speech-related lexicon and a compiled offensive/profane/slang wordlist as well as a social-network specific feature. Then, this concatenated layer was fed to two fully connnected dense layers. Although a satisfactory result is obtained for this task, I plan to improve the system's performance by adopting knowledge base and more social-network specific features in the future.

References

Pinar Arslan, Michele Corazza, Elena Cabrio, and Serena Villata. 2019. Overwhelmed by negative emotions? maybe you are being cyber-bullied! In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 1061–1063.

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.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

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.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Çağrı Çöltekin. 2020. A Corpus of Turkish Offensive Language on Social Media. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*. ELRA.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE.

Michele Corazza, Stefano Menini, Pinar Arslan, Rachele Sprugnoli, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018a. Comparing different supervised approaches to hate speech detection.

Michele Corazza, Stefano Menini, Pinar Arslan, Rachele Sprugnoli, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018b. Inriafbk at germeval 2018: Identifying offensive tweets using recurrent neural networks.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):1–30.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@ SEPLN*, pages 214–228.

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

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on arabic social media. In *Proceedings of the first workshop on abusive language online*, pages 52–56.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Tenth International AAAI Conference on Web and Social Media*.

Sara Owsley Sood, Judd Antin, and Elizabeth Churchill. 2012. Using crowdsourcing to improve profanity detection. In *2012 AAAI Spring Symposium Series*.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 672–680.

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.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval)*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.

Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. 2016. Content-driven detection of cyberbullying on the instagram social network. In *IJCAI*, pages 3952–3958.