TAC at SemEval-2020 Task 12: Ensembling Approach for Multilingual Offensive Language Identification in Social Media

Talha Anwar **Mirza Omer Beg**
Department of Computer Science
National University of Computer and Emerging Sciences
Islamabad, Pakistan.
{i181509,omer.beg}@nu.edu.pk

Abstract

Usage of offensive language on social media is getting more common these days, and there is a need of a mechanism to detect it and control it. This paper deals with offensive language detection in five different languages; English, Arabic, Danish, Greek and Turkish. We presented an almost similar ensemble pipeline comprised of machine learning and deep learning models for all five languages. Three machine learning and four deep learning models were used in the ensemble. In the OffensEval-2020 competition our model achieved F1-score of 0.85, 0.74, 0.68, 0.81, and 0.9 for Arabic, Turkish, Danish, Greek and English language tasks respectively.

1 Introduction

With the growing and easy access to social media without check and balance, ethical and moral values are somehow put aside. One of the main issues is abusive behaviour, cyberbullying and offensive language. Recently a lot of work (Kumar et al., 2018) (Zampieri et al., 2019),(i Orts, 2019) was performed to tackle this problem. Most of the research work was done for languages like English and Arabic (Al-Hassan and Al-Dossari, 2019), and less work for languages like Danish, Turkish and Greek. The SemEval 2020 task 12 (Zampieri et al., 2020) consisted of identification of offensive words in different languages such as Arabic, Turkish, English, Greek and Danish. This paper proposed an ensemble approach for offensive language identification in social media text. We have participated in all languages tasks. For Arabic, Turkish, Danish and Greek language, we used FastText embeddings. For English subtask A, we used FastText embeddings combined with GloVe embeddings. The datasets were skewed, for which we used the focal loss as it gives more weightage to less occurred class in the dataset.

This paper is organized as follows. Section 2 presents related work in the area of offensive language detection. Section 3 deals with more detail of provided dataset and the methodology used for the tasks. Results are discussed in section 4. Section 5 concludes the paper and presents some ideas for future work.

2 Related Work

AA Altowayan and Ashraf (2017) trained sentiment specific embedding using FastText both for the bag of words (CBOW) and skip-gram models to study syntactic and semantic information of Arabic language. They used four different datasets. Embeddings were trained on one dataset and tested on the other three datasets. F1 score of 82.47%, 82.22% and 70.62% was achieved on the rest of three datasets respectively. Similarly, using skip-gram model, F1 score of 82.31%,86.25% and 72.37% was achieved. Mohaouchane et al. (2019) achieved F1 score of 83.46% by training convolution neural network on the dataset comprised of 15,000 YouTube comments in Arabic labeled as offensive or not. They used Tree-Structured Parzen Estimator (TPE) algorithm; a Bayesian hyper-parameter optimization technique for tuning the hyper-parameter of neural network. Ghallab et al. (2020) did an extensive literature review

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 2177–2182
Barcelona, Spain (Online), December 12, 2020.

of Arabic sentiment analysis comprised of papers published during 2013-18. Analysis and detection of religious hate speech in tweets (Albadi et al., 2018) was one of the first work in detection of religious hate speech. The authors annotated the twitter data and applied lexicon-based, n-grams and deep learning models to classify it. Using gated recurrent neural network they achieved an F1 score of 77%. The Arabic language sub-task organizer (Mubarak et al., 2020) used support vector machine and obtained an F1-score of 79.7% while classifying offensive tweets. They built the largest offensive language corpus for Arabic language and categorized the tweets as offensive or non offensive. They further split the offensive tweets as vulgar or hate speech. They used lexicon based features, pre-trained embeddings, embedding train on their own data and contextualized multilingual BERT-base. The best result was achieved using pre-trained embeddings with SVM classifier.

Using artificial neural network (Bozyiğit et al., 2019) on a dataset comprised of 3000 Turkish tweets, F1 score of 91% was obtained. A study (Özel et al., 2017) comprised of twitter tweets and Instagram post showed that using multinomial Naive Bayes cyberbullying can be classified with an accuracy of 84 % using TF-IDF as features. Turkish language sub-task organizer (Çöltekin, 2020) collected the tweets and classified them as offensive vs non offensive tweets with an F1-score of 77.3%. They categorized the offensive tweets into further categories as targeted or non targeted. If the category is targeted, it is further categorized as group targeted, individual targeted or other. They used both character grams and words grams as features and linear SVM as classifier.

Sigurbergsson and Derczynski (2020), the task organizers for Danish language, got an F1 score of 70% while classifying offensive vs non offensive text using logistic regression. Authors have categorized their data in a similar hierarchical way as proposed by (Zampieri et al., 2019). They collected the data from twitter, Facebook and reddit. Their data is skewed as 88% of the posts were labelled as not offensive and rest as offensive. In low resource language like Danish, because of over fitting in term of word distribution on training data, accuracy drop when moving from one domain such as film/movie reviews to another domain such as company reviews. So an offline approach to solve this problem was proposed (Elming et al., 2014). The authors trained a base model, created k copies of training data, then corrupted these copies based on weights of base model. Finally they train on corrupted training data. This tended to increase performance on unseen data and help in inter domain classification.

Using combination of lexicon and word2vec trained on Greek Wikipedia (Giatsoglou et al., 2017), support vector machine classifier tended to achieve an F1 score of 77.85%. The dataset collected from Greek e-shopping website consisted of product reviews. Using TF-IDF uni-gram features and linear support vector machine (Pitenis et al., 2020) macro F1 score of 80% was achieved. They tried different deep learning models and LSTM and GRU with attention mechanism resulted in F1 score of 89% .

English offensive language identification task is similar with the task held last year (Zampieri et al., 2019), where more than 100 team submitted their papers. Team NULI (Liu et al., 2019) achieved highest F1 score using BERT uncased model.

Hassan and Al-Dossari (2019) did a literature review on detection of hate speech on social media. They covered multilingual corpus in their survey.

3 Methodology and Data

3.1 Dataset

The SemEval shared task Multilingual Offensive Language Identification in Social Media (Zampieri et al., 2020) consisted of 5 languages; English, Arabic, Turkish, Greek and Danish. Dataset comprised of offensive and non offensive tweets were provided for each language. All the dataset were imbalanced to some extent. Danish dataset is the smallest dataset and the most imbalanced dataset. For English task, two datasets were used. OLID dataset (Zampieri et al., 2019) and SOLID dataset (Rosenthal et al., 2020). Unlike other datasets, SOLID dataset was labelled in a semi-supervised way by using machine

The online implementation of our work is available at `https://github.com/talhaanwarch/OffenseEval2020`.

Language	OFF	NOT	Total
Arabic (Mubarak et al., 2020)	1589	6411	8000
Turkish (Çöltekin, 2020)	6131	25625	31756
Danish (Sigurbergsson and Derczynski, 2020)	348	2576	2960
Greek (Pitenis et al., 2020)	2486	6257	8743
English (Rosenthal et al., 2020)	1446768	7628650	9,075,418
English (Zampieri et al., 2019)	4640	9460	14100

Table 1: Training Dataset Size

learning algorithms.For English language task, we participated in sub task 1 only *i.e.* offensive language identification. Table 1 shows the training data distribution for each language task.

3.2 Preprocessing

In order to make our pipeline independent of language, almost similar preprocessing for all languages dataset was performed. @USER, URL and digits were removed from all tweets. In Arabic RT and <LF>words were also removed.

3.3 Methodology

Both machine learning and deep learning techniques were used to classify offensive tweets. For machine learning logistic regression was used. Count vectorizer, TF-IDF word level and character level features were fed to logistic regression model. Ten thousands most frequent words are selected from the corpus and length of each tweet is set to 200. If a tweet has less than 200 words, zero padding was used otherwise the tweet was truncated to 200 words. Grid-search technique was used to find the optimal hyper parameter of logistic regression classifier. Only parameter C of logistic regression was tuned with a lower bound 10 and upper bound as 25. Grid-search then selected the best C value of logistic regression from these lower and upper bounds.

Four deep learning (DL) models were used to identify the offensive tweets. Two of them were variants of recurrent neural network and the other two were convolution neural networks. FastText embeddings of dimension 300 were used for word representation in all languages task except English where GloVe embedding of 300 dimension combined with FastText was used. The first model we used is a bidirectional LSTM followed by Bidirectional GRU. Embedding input followed by spatial dropout layer was fed to bidirectional LSTM (biLSTM) layer. The biLSTM layer was followed by bidirectional GRU layer. Number of units in both layer was 50. Average pooling layer, max pooling layer and biGRU layers were concatenated to capture maximum information. The model is made in such way that output from biLSTM was fed to biGRU, and pooling layers as input. In the second model attention mechanism was added. Output from biLSTM was fed to biGRU. The attention layer, global max pooling and global average pooling received input from BiGRU layer. Then the attention layer, global max pooling and global average pooling layers were concatenated. Third model used was temporal convolution network (Bai et al., 2018) proved to better than LSTM and GRU in many cases specially in seq2seq task. We used TCN to classify offensive language. Embeddings were passed to TCN layer of 128 length followed by a layer of 64 length with dilation of 1,2 and 4. The TCN layers were followed by global average pooling and global max pooling which were concatenated. After concatenation layer, there were two dense layer of 64 neurons with ReLU activation functions before final output layer. Convolution neural network proposed by (Kim, 2014) was the fourth deep learning model used.

For English task only, BERT with a batch size of 16 and epoch 3 was trained on 0.2 M balanced tweets from SOLID dataset. Fixed learning rate of 2e-6 was used. Binary Cross entropy was used as loss function. Maximum sequence length is set to 160. Five fold cross validation was not used for BERT model.

3.4 Model training and Ensembling

Five fold cross validation was used to avoid over fitting. Number of epochs was set to 10, optimizer to adam and batch size to 64. In all deep learning model, instead of using binary cross entropy, focal loss proposed by (Lin et al., 2017) was used as a loss function. Focal loss is mostly used in computer vision application for imbalanced dataset. Focal loss down weights the well classified examples and give more importance to those example which are difficult to classify. So net learning of classifier tends toward hard examples. Learning rate plays an important role during training of deep learning classifier. Optimum learning rate result in better and fast convergence. Small learning rate, result in slow learning and large learning rate result in divergence and the model does not reach to global minima. So, cyclic learning rate propose by (Smith, 2017) was used. In this approach an upper and lower bound learning rate is defined and learning rate oscillate between these bound.

Training data was cross validated for each model. Labels of test data was predicted from each model and weight averaged according to average validation F1 score. Bert model was used in English task only with other models. We did not train any of our models on English SOLID dataset because of hardware limits. Instead we trained our models on OLID dataset, and ensembled with BERT which was trained on 0.2M tweets of SOLID dataset. So English task comprised of an ensemble machine learning and deep learning models trained on OLID dataset along with BERT trained on 0.2 million tweets of SOLID dataset. BERT was not used in any other language except English. Figure 1 shows our ensemble pipeline.

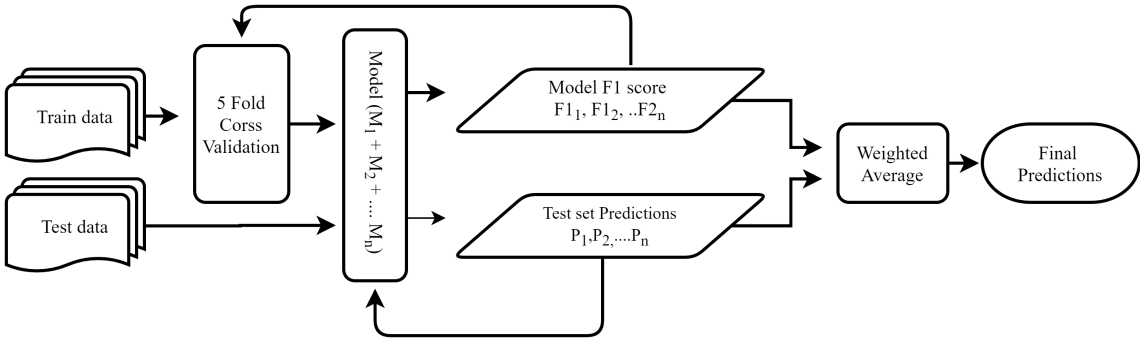

Figure 1: Ensemble pipeline flow chart

4 Results

F1 macro was used as evaluation criteria for all of the task, as it is a good metric for imbalanced dataset classification.

Technique	Arabic	Turkish	Danish	Greek	English*
Count Vectorizer features	0.80	0.71	0.75	0.76	0.67
TF-IDF Word level features	0.79	0.69	0.71	0.75	0.68
TF-IDF Char level features	**0.85**	**0.73**	**0.76**	**0.80**	0.69
Bi LSTM followed by Bi GRU	0.84	0.72	0.72	0.78	**0.72**
Bi LSTM followed by Bi GRU with attention	0.84	0.71	0.73	0.77	**0.72**
Temporal Convolution Network	0.84	0.72	0.62	0.79	0.71
Convolution Neural Network	0.80	0.71	0.58	0.74	**0.72**

Table 2: 5 Fold cross-validation F1 score. *Training was performed on OLID dataset

Table 2 shows that validation F1 score achieved. Logistic regression trained on TF-IDF character level feature resulted in 0.85 0.73,0.76 and 0.8 F1 score in Arabic, Turkish, Danish and Greek language. Character level feature based logistic regression tend to the best feature model than rest of others. No validation is performed for BERT English model, because of resource limitation. Table 3 shows F1 score

Language	Arabic	Turkish	Danish	Greek	English
F1 Score	0.85191	0.74772	0.682	0.814	0.90925
Rank	15/53	24/46	29/39	17/37	32/85

Table 3: Test set result. F1 score obtained on test set for multilingual languages with correspondent ranking on leader-board

on test data. F1 score of 0.85,0.74,0.68,0.81 and 0.9 was obtained for Arabic, Turkish, Danish, Greek and English language task, respectively. We believe using BERT on 0.2M tweets of SOLID dataset, helped to obtain high score as compared to validation score. For Danish, it seemed that temporal convolution network and convolution neural network lower the overall score on test dataset. Rank shows the position we achieved in numerator and total teams in denominator. Team TAC got a position of 15 out of 53 teams in Arabic task, 32 out of 85 in English Task, 24 out of 45 in Turkish, 17 out of 37 teams in Greek. The lowest rank TAC team achieved is in Danish which is 29 out of 39 teams.

5 Conclusion and Feature Work

In this paper, we classified offensive language in multiple languages text such as English, Arabic, Danish, Greek and Turkish. We proposed a similar ensemble pipe line for all language tasks. We use count vectorizer, word level TF-IDF, character level TF-IDF as feature for logistic regression. Two variants of recurrent neural network and convolution neural network each are applied to classify offensive language. We found that character level TF-IDF work better in almost all cased, except English, where training was performed on OLID dataset. In deep learning methodology, instead of using traditional binary cross entropy as loss functions, we tried focal loss. Instead of fixed learning rate, cyclic learning rate was used. As future work, we will try multilingual BERT and language specific BERT model such as AraBERT (Antoun et al., 2020). For English language task, we will try to overcome hardware limitations and train our ensemble pipeline on larger dataset.

Acknowledgements

I would like to thank the task organizer for their quick responses via email and my subject instructor Dr. Omer Baig for teaching me the course with a research perspective.

References

Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: a survey on multilingual corpus. In *6th International Conference on Computer Science and Information Technology*.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76. IEEE.

A Aziz Altowayan and Ashraf Elnagar. 2017. Improving arabic sentiment analysis with sentiment-specific embeddings. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4314–4320. IEEE.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Alican Bozyiğit, Semih Utku, and Efendi Nasiboğlu. 2019. Cyberbullying detection by using artificial neural network models. In *2019 4th International Conference on Computer Science and Engineering (UBMK)*, pages 520–524. IEEE.

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.

Jakob Elming, Barbara Plank, and Dirk Hovy. 2014. Robust cross-domain sentiment analysis for low-resource languages. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–7.

Abdullatif Ghallab, Abdulqader Mohsen, and Yousef Ali. 2020. Arabic sentiment analysis: A systematic literature review. *Applied Computational Intelligence and Soft Computing*, 2020.

Maria Giatsoglou, Manolis G Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, and Konstantinos Ch Chatzisavvas. 2017. Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214–224.

Òscar Garibo i Orts. 2019. Multilingual detection of hate speech against immigrants and women in twitter at semeval-2019 task 5: Frequency analysis interpolation for hate in speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 460–463.

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 Cyberbullying (TRAC-2018)*, pages 1–11.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Ping Liu, Wen Li, and Liang Zou. 2019. Nuli at semeval-2019 task 6: transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91.

Hanane Mohaouchane, Asmaa Mourhir, and Nikola S Nikolov. 2019. Detecting offensive language on arabic social media using deep learning. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 466–471. IEEE.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Selma Ayşe Özel, Esra Saraç, Seyran Akdemir, and Hülya Aksu. 2017. Detection of cyberbullying on social media messages in turkish. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 366–370. IEEE.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive Language Identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive Language and Hate Speech Detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.

Leslie N Smith. 2017. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

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*.