DA-LD-Hildesheim at SemEval-2019 Task 6: Tracking Offensive Content with Deep Learning using Shallow Representation

Sandip Modha
DA-IICT
Gandhinagar India
sjmodha
@gmail.com

Prasenjit Majumder
DA-IICT
Gandhinagar India
p_majumder
@daiict.ac.in

Thomas Mandl
Uni. of Hildesheim
Hildesheim Germany
mandl
@uni-hildesheim.de

Daksh Patel
LDRP-ITR
Gandhinagar India
dakshpatel68
@gmail.com

Abstract

This paper presents the participation of team DA-LD-Hildesheim of Information Retrieval and Language Processing lab at DA-IICT, India in Semeval-19 OffenEval track. The aim of this shared task is to identify offensive content at fined-grained level granularity. The task is divided into three sub-tasks. The system is required to check whether social media posts contain any offensive or profane content or not, targeted or untargeted towards any entity and classifying targeted posts into the individual, group or other categories. Social media posts suffer from data sparsity problem, Therefore, the distributed word representation technique is chosen over the Bag-of-Words for the text representation. Since limited labeled data was available for the training, pre-trained word vectors are used and fine-tuned on this classification task. Various deep learning models based on LSTM, Bidirectional LSTM, CNN, and Stacked CNN are used for the classification. It has been observed that labeled data was highly affected with class imbalance and our technique to handle the class-balance was not effective, in fact performance was degraded in some of the runs. Macro F1 score is used as a primary evaluation metric for the performance. Our System achieves Macro F1 score = 0.7833 in sub-task A, 0.6456 in the sub-task B and 0.5533 in the sub-task C.

1 Introduction

NLP researchers are developing innovative systems based on the input of the text data. The power of predictions has moved from simple sentiment classification task to much more advanced labeling of the content. The task related to hate, aggression, abusive or offensive speech currently attracts research more to algorithms making decisions which can also be ambiguous for humans. Due to the availability of standard datasets, such data collections are created based on social media data and are offered at forum like TRAC [1] (Kumar et al., 2018), GermEval [2], and SemEval OffenEval 2019 (Zampieri et al., 2019a)[3].

The exponential rise in social media user-base backed by the cutting edge mobile data technologies leads to the inorganic growth in the posts related to hate speech or offensive speech. Researchers working in the area of domain-specific sentiment analysis move to the problem of domain specific or open domain hate or offensive speech detection. They are reshaping the hate speech problem into the new notion like abusive, aggressive, or offensive speech. Such categorization of social media posts, help law-enforcement agencies with the surveillance of the social media.

The shared task in SemEval-OffenEval 2019 was introduced as a 3-level classification task (Zampieri et al., 2019b). In the first level, sub-task A, systems are required to classify tweets into two class, namely: Offensive (OFF) and Non-offensive (NOT). In the second level, sub-task B, offensive tweets are further required to be categorized into two labels, namely :targeted (TIN)-post which contain threat/insult to the targeted entity and untargeted (UNT), respectively. In the sub-task C, target of insults and threats are further classified to Individual (IND), Group (GRP), or Other (OTH) classes. Table 1 presents the statistic about the dataset. One can observe that classes in dataset, particularly for the sub-task B and sub-task C, is highly imbalanced.

Our approach for this shared task is based on distributed word representation and deep learning. fastText pre-trained word embedding (Mikolov et al., 2018) is used to initialize embedding layer or first layer of the model and fine tuned for classification task. The rest of the model is still needed

[1] https://sites.google.com/view/trac1/home
[2] http://https://projects.fzai.h-da.de/iggsa/
[3] https://competitions.codalab.org/competitions/20011

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 577–581
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

Details	# Tweets Train Dataset	# Tweets in Test Dataset
Total Posts in Sub-task A	13240	860
Offensive posts	4440	240
Non-offensive posts	8800	620
sub-task-B : Targeted (TIN) posts	3876	213
Non-Targeted (UNT) posts	524	27
Sub-task C: Individual	2407	100
Group	1074	78
Other	395	35

Table 1: Dataset statistics

to be trained from scratch.(Howard and Ruder, 2018) termed this techniques as shallow representation against the hierarchical representation.

The rest of this paper is organized as follows. In section 2 we briefly discuss the related work in this area. In section 3, we present our method and model. In section 4, we present results and give the brief analysis. In section 5, we will give our final conclusion along with future works.

2 Related Work

Hate Speech Detection research attracts researchers from diverse background like computational Linguistic, computer science, and social science. The actual term hate speech was coined by (Warner and Hirschberg, 2012). Various Authors used different notion like offensive language (Razavi et al., 2010), cyberbullying (Xu et al., 2012), aggression (Kumar et al., 2018). (Davidson et al., 2017) studied tweet classification of hate speech and offensive language and defined hate speech as following: language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group. Authors observed that offensive language is often miss-classified as hate speech. They have trained a multi-class classifier on N-gram features weighted by its TF-IDF weights and PoS tags. In addition to these, features like sentiment score of each tweet, number of hashtags, and URLS, mentions are considered. Authors concluded that Logistic Regression and Linear SVM are better than NB, Decision Tree, and Random Forests. (Schmidt and Wiegand, 2017) perform comprehensive survey on hate speech. They have identified features like surface features, sentiment, word generalization, lexical, linguistics etc. can be used by classifier.

(Malmasi and Zampieri, 2018) tried to address the problem of discriminating profanity from hate speech in the social media posts. N-grams, skip-gram and clustering based word representation features are considered for the 3-class classification. The author uses SVM and advance ensemble based classifier for this task and achieved 80% accuracy. (Gambäck and Sikdar, 2017) performed 4-class classification on Twitter messages using CNN with word embedding generated through Word2vec and character n-grams. Authors claim that word embedding generated through Word2vec outperformed random vector and n-gram characters. (Zhang et al., 2018) proposed a new method based on CNN and LSTM with drop out and pooling for hate speech detection. Authors concluded that their method achieved improvement on F1 score of most of the hate speech datasets.

3 Methodology and Data

Since the social media data suffers from the data sparsity problem, classifier based on the BoW features might not be appropriate as compared to distributed word representation. Our previous work (Majumder et al., 2018) also supported this intuition. Empirical evidence (Majumder et al., 2018) suggest that pre-trained vector trained on huge corpus provides better word embedding than embedding generated from a limited training corpus. Some authors (Howard and Ruder, 2018) termed it as a shallow-transfer learning approach. In this method, first layer or embedding layer of deep neural network is initialized with pre-trained vectors and the rest of the network is trained from scratch. Since fastText generates word embedding for a word which is unseen during the training by using the subword or n-gram of the word, it is the better choice than Word2vec and Glove. As discussed in the previous section, there is substantial class imbalance particularly in sub-task B and C. To address this issue, class weights are incorporated into the cost function of the classifier which gives higher weight to minority class and lower weights to the majority class. Four deep learning based models: Bidirectional LSTM, Single LSTM, CNN and stacked CNN are designed for the classification.

Pre-processing : Track organizers have partially pre-processed tweets in the dataset. User mention, URL are replaced with standard tags. We did not perform any sort of further pre-processing or stemming on the texts.

Word embedding : fastText pre-trained word vectors with dimension 300 are used to initialize the embedding layer. This model is trained on 600B tokens of commonly crawled corpus.

3.1 Model Architecture and Hyperparameters

In this sub-section, we briefly describe our models used for the classification. The first model is based on Bidirectional LSTM model includes the embedding layer with 300 dimensions, Bidirectional LSTM layer with 50 memory units followed by one-dimensional global max pooling and dense layer with softmax/sigmoid activations. Hyperparameters are as follows: Sequence length is fixed at 30. Number of features is equal to the half of total vocabulary size in each task. Models are trained for 10 epoch. Adam optimization algorithm is used to update network weights.

The second model is based on LSTM. The model includes embedding layer with 300 dimensions, LSTM layer with 64 memory units, followed by two dense layers with softmax/sigmoid activations. A dropout layer is added to the hidden layer to counter the overfitting. Hyperparameters of the model is the same as the first model.

Rest of the two models are based on Convolution Neural Network, includes embedding layer with embed size of 300, followed by a one-dimensional convolution layer with 100 filters of height 2 and stride 1 to target biagrams. In addition to this, Global Max Pooling layer is added. Pooling layer fetches the maximum value from the filters which are feeded to the dense layer. There are 256 nodes in the hidden layer without any dropout. The last model is same as previous CNN model except three one-dimensional convolution layer are stacked together. Different one-dimensional filters with height 2,3,4 to target bigrams, trigrams, and four-grams features. After convolution layers and max pool layer, model concatenate max pooled results from each of one-dimensional convolution layers, then build one output layer on top of them. (Majumder et al., 2018). Hyperparameters of the model is the same as the first model.

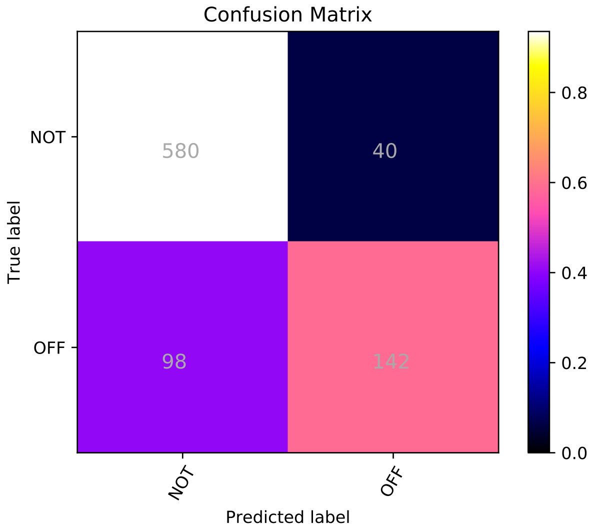

Figure 1: Sub-task A, Confusion Matrix:BLSTM classifier

4 Results

In this section, we report the results obtained by the model discussed in the previous section. Table 2, 3, 4 display results of sub-task A, B, and C, respectively. We have randomly splitted the dataset into 80% training and 20% validation. By and large, results on test dataset are better than cross-validation. F1-macro score is the primary metric for the evaluation. Results are comparable with the top team and substantially outperforms all the random baselines. Figure 1, 2, and 3, show the confusion matrices for all the sub-tasks.

5 Conclusion

In this paper, we have presented our deep learning based approach for multi-level offensive text classification. The system reports reasonable performance. Macro f1 and accuracy score around 78.3% and 84% in sub-task A. In sub-task B, our system performs the worst in UNT class(offensive post without target). The reason behind this underperformance is few number of training examples for the UNT class. Similar case happened in the

	Test Dataset		Cross Validation	
System	**F1 (macro)**	**Accuracy**	**F1 (macro)**	**Accuracy**
All NOT baseline	0.4189	0.7209		
All OFF baseline	0.2182	0.2790		
Bidirectional LSTM	**0.7833**	0.8395	0.75	0.79
CNN	0.7800	0.8337	0.76	0.7915
LSTM-balanced	0.7500	0.8047	0.74	0.7708
Top-run: pliu19	0.829			

Table 2: Results for Sub-task A.

	Test Dataset		Cross Validation	
System	**F1 (macro)**	**Accuracy**	**F1 (macro)**	**Accuracy**
All TIN baseline	0.4702	0.8875		
All UNT baseline	0.1011	0.1125		
CNN	**0.6456**	0.9042	0.60	0.8875
LSTM-balanced	0.5471	0.825	0.60	0.867
CNN-balanced	0.6455	0.8917	0.55	0.8943
Top Team jhan014	0.755			

Table 3: Results for Sub-task B

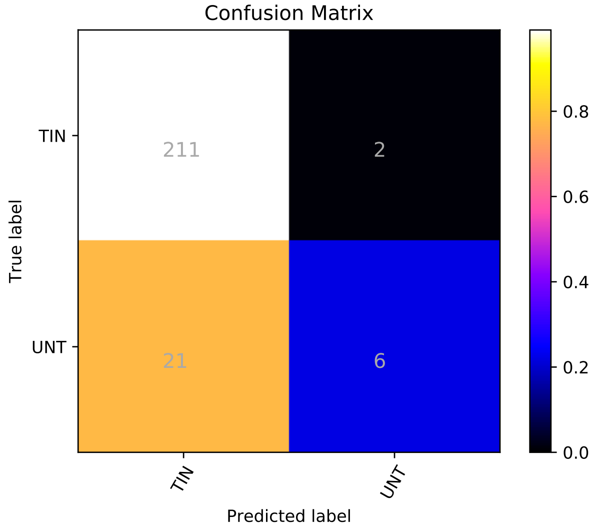

Figure 2: Sub-task B, Confusion matrix : CNN classifier

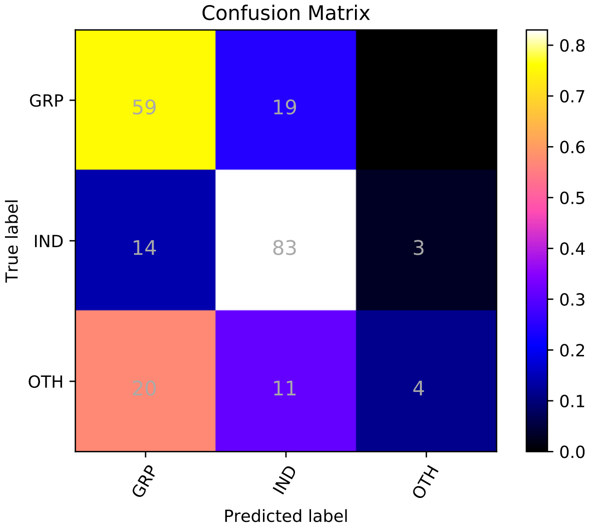

Figure 3: Sub-task C, Confusion Matrix:CNN Classifier

	Test Dataset		Cross Validation	
System	**F1 (macro)**	**Accuracy**	**F1 (macro)**	**Accuracy**
All GRP baseline	0.1787	0.3662		
All IND baseline	0.2130	0.4695		
All OTH baseline	0.0941	0.1643		
CNN-balanced	**0.5533**	0.6854	0.5231	0.695
BLSTM-balanced	0.4829	0.662	0.5223	0.6959
Stacked CNN	0.5198	0.662	0.5074	0.6920
Top Team vradivchev	0.66			

Table 4: Results for Sub-task C

sub-task C. We have set class weights in the cost function of the model. Unfortunately, it did not work. In the future, we will try to address imbalance class problem using external vocabulary augmentation. we would like to explore various transfer learning model like BERT, ELMO and ULMFit for this multi-level classification problem.

References

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

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.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

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.

Prasenjit Majumder, Thomas 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.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1–16.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

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.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics.

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. 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.