SSN_NLP at SemEval-2019 Task 6: Offensive Language Identification in Social Media using Traditional and Deep Machine Learning Approaches

D. Thenmozhi, B. Senthil Kumar, Chandrabose Aravindan, S. Srinethe
Department of CSE, SSN College of Engineering, India
{theni_d,senthil,aravindanc}@ssn.edu.in
srinethe16108@cse.ssn.edu.in

Abstract

Offensive language identification (OLI) in user generated text is automatic detection of any profanity, insult, obscenity, racism or vulgarity that degrades an individual or a group. It is helpful for hate speech detection, flame detection and cyber bullying. Due to immense growth of accessibility to social media, OLI helps to avoid abuse and hurts. In this paper, we present deep and traditional machine learning approaches for OLI. In deep learning approach, we have used bi-directional LSTM with different attention mechanisms to build the models and in traditional machine learning, TF-IDF weighting schemes with classifiers namely Multinomial Naive Bayes and Support Vector Machines with Stochastic Gradient Descent optimizer are used for model building. The approaches are evaluated on the OffensEval@SemEval2019 dataset and our team SSN_NLP submitted runs for three tasks of OffensEval shared task. The best runs of SSN_NLP obtained the F1 scores as 0.53, 0.48, 0.3 and the accuracies as 0.63, 0.84 and 0.42 for the tasks A, B and C respectively. Our approaches improved the base line F1 scores by 12%, 26% and 14% for Task A, B and C respectively.

1 Introduction

Offensive language identification (OLI) is a process of detecting offensive language classes (Razavi et al., 2010) such as slurs, homophobia, profanity, extremism, insult, disguise, obscenity, racism or vulgarity that hurts or degrades an individual or a group from user-generated text like social media postings. OLI is useful for several applications such as hate speech detection, flame detection, aggression detection and cyber bullying. Recently, several research work have been reported to identify the offensive languages using social media content. Several workshops such as TA-COS[1], TRAC[2] (Kumar et al., 2018a), Abusive Language Online[3] and GermEval (Wiegand et al., 2018) have been organized recently in this research area. In this line, OffensEval@SemEval2019 (Zampieri et al., 2019b) shared task focuses on identification and categorization of offensive language in social media. It focuses on three subtasks namely offensive language detection, categorization of offensive language and offensive language target identification. Sub_Task_A aims to detect text as offensive (OFF) or not offensive (NOT). Sub_Task_B aims to categorize the offensive type as targeted text (TIN) or untargeted text (UNT). Sub_Task_C focuses on identification of target as individual (IND), group (GRP) or others (OTH). Our team SSN_NLP participated in all the three subtasks.

2 Related Work

Several research work have been reported since 2010 in this research field of hate speech detection (Kwok and Wang, 2013; Burnap and Williams, 2015; Djuric et al., 2015; Davidson et al., 2017; Malmasi and Zampieri, 2018; Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; ElSherief et al., 2018; Gambäck and Sikdar, 2017; Zhang et al., 2018; Mathur et al., 2018). Schmidt and Wiegand (2017) & Fortuna and Nunes (2018) reviewed the approaches used for hate speech detection. Kwok and Wang (2013) used bag of words and bi-gram features with machine learning approach to classify the tweets as “racist” or “nonracist”. Burnap and Williams (2015) developed a supervised algorithm for hateful and antagonistic content in Twitter using voted ensemble meta-

[1] http://ta-cos.org/

[2] https://sites.google.com/view/trac1/home

[3] https://sites.google.com/site/abusivelanguageworkshop2017/

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 739–744
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

classifier. Djuric et al. (2015) learnt distributed low-dimensional representations of social media comments using neural language models for hate speech detection. Davidson et al. (2017) used n-gram (bigram, unigram, and trigram) features with TF-IDF score along with crowd-sourced hate speech lexicon and employed several classifiers including logistic regression with L1 regularization to separate hate speech from other offensive languages. Malmasi and Zampieri (2018) used n-grams, skip-grams and clustering-based word representations as features with ensemble classifier for hate speech detection. ElSherief et al. (2018) performed linguistic and psycholinguistic analysis to detect the hate speech is either "directed" towards a target, or "generalized" towards a group. Gambäck and Sikdar (2017) used deep learning using CNN models to detect the hate speech as "racism", "sexism", "both" and "non-hate-speech". They used character 4-grams, word vectors based on word2vec, randomly generated word vectors, and word vectors combined with character n-grams as features in their approach. Zhang et al. (2018) used convolution-GRU based deep neural network for detecting hate speech.

Many research work have been carried out in aggression detection (Aroyehun and Gelbukh, 2018; Madisetty and Desarkar, 2018; Raiyani et al., 2018; Kumar et al., 2018b). Aroyehun and Gelbukh (2018) & Raiyani et al. (2018) used LSTM and CNN respectively to detect aggression in text. Kumar et al. (2018b) presented the findings of the shared task on aggression identification which aims to detect different scales of aggression namely "Overtly Aggressive", "Covertly Aggressive", and "Non-aggressive". Madisetty and Desarkar (2018) used CNN, LSTM and Bi-LSTM to detect the above scales of aggression. Waseem et al. (2017) & Park and Fung (2017) presented the methodologies on abusive language identification using deep neural networks.

Research on identifying offensive languages has been focused on non-English languages like German (Wiegand et al., 2018), Hindi (Kumar et al., 2018b), Hinglish: Hindi-English (Mathur et al., 2018), Slovene (Fišer et al., 2017) and Chinese (Su et al., 2017). Wiegand et al. (2018) presented an overview of GermEval shared task on the identification of offensive language that focused on classification of German tweets from Twitter. Kumar et al. (2018b) focused on the shared task to identify aggression on Hindi text. Mathur et al. (2018) applied transfer learning to detect three classes namely "nonoffensive", "abusive" and "hate-speech" from Hindi-English code switched language. Fišer et al. (2017) presented a framework to annotate offensive labels in Slovene. Su et al. (2017) rephrased profanity in Chinese text after detecting them from social media text.

3 Data and Methodology

In our approach, we have used OLID dataset (Zampieri et al., 2019a) given by OffensEval@SemEval2019 shared task. The dataset is given in $.tsv$ file format with columns namely, ID, INSTANCE, SUBA, SUBB, SUBC where ID represents the identification number for the tweet, INSTANCE represents the tweets, SUBA consists of the labels namely Offensive (OFF) and Not Offensive (NOT), SUBB consists of the labels namely Targeted Insult and Threats (TIN) and Untargeted (UNT) and SUBC consists of the labels namely Individual (IND), Group (GRP) and Other (OTH). The dataset has 13240 tweets. All the instances are considered for Sub_Task_A. However, we have filtered and considered the data that are labelled with "TIN/UNT" and "IND/GRP/OTH" for Sub_Task_B and Sub_Task_C respectively by ignoring the instances labelled with "NULL". Thus, we have obtained 4400 and 3876 instances for Sub_Task_B and Sub_Task_C respectively. We have preprocessed the data by removing the URLs and the text "@USER" from the tweets. Tweet tokenizer [4] is used to obtain the vocabulary and features for the training data.

We have employed both traditional machine learning and deep learning approaches to identify the offensive language in social media. The models that are implemented for the three sub-tasks are given in Table 1.

In deep learning (DL) approach, the tweets are vectorized using word embeddings and are fed into encoding and decoding processes. Bidirectional LSTMs are used for encoding and decoding processes. We have used 2 layers of LSTM for this. The output is given to softmax layer by incorporating attention wrapper to obtain the OffensEval class labels. We have trained the deep learning models with a batch size 128 and dropout 0.2 for 300 epochs to build the model. We have em-

[4] https://www.nltk.org/

Tasks	Models	Description
Task A	Task_A_DL_NB	Deep learning with Normed Bahdanau attention
	Task_A_DL_SL	Deep learning with Scaled Luong attention
Tak B	Task_B_DL_NB	Deep learning with Normed Bahdanau attention
	Task_B_DL_SL	Deep learning with Scaled Luong attention
	Task_B_TL_MNB	Traditional Machine Learning with Multinomial Naive Bayes
Task C	Task_C_DL_NB	Deep learning with Normed Bahdanau attention
	Task_C_DL_SL	Deep learning with Scaled Luong attention
	Task_C_TL_SVM	Traditional Machine Learning with Support Vector Machine and Stochastic Gradient Descent optimizer

Table 1: Models for the Tasks

ployed two attention mechanisms namely Normed Bahdanau (NB) (Sutskever et al., 2014; Bahdanau et al., 2014) and Scaled Luong (SL) (Luong et al., 2015, 2017) in this approach. These two variations are implemented to predict the class labels for all the three sub tasks. These attention mechanisms help the model to capture the group of input words relevant to the target output label. For example, consider the instance in Task C: “we do not watch any nfl games this guy can shove it in his pie hole”. This instance clearly contains the offensive slang “pie hole” and about watching the “nfl games”. The attention mechanism captures these named entities or group of words and correctly map to the label “GRP”. Also, it is evident from the earlier experiments (Sutskever et al., 2014; Thenmozhi et al., 2018) that bi-directional LSTM with attention mechanism performs better for mapping input sequences to the output sequences.

In traditional learning (TL) approach, the features are extracted from the tokens with minimum count of two. The feature vectors are constructed using TF-IDF scores for the training instances. We have chosen the classifiers namely Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) with Stochastic Gradient Descent optimizer to build the models for Task B and Task C respectively. These classifiers have been chosen based on the cross validation accuracies. The class labels namely “TIN/UNT” and “IND/GRP/OTH” are predicted for Task B and Task C using the respective models.

4 Results

We have evaluated our models using the test data of OffensEval@SemEval2019 shared task for the three sub tasks. The performance was analyzed using the metrics namely precision, recall, macro-averaged F1 and accuracy. The results of our approaches are presented in Tables 2, 3 and 4 for Task A, Task B and Task C respectively. We have obtained the best results for Task_A_DL_SL, Task_B_DL_NB, Task_C_TL_SVM models for Task A, Task B and Task C respectively.

System	F1 (macro)	Accuracy
All NOT baseline	0.4189	0.7209
All OFF baseline	0.2182	0.2790
Task_A_DL_NB (527733)	0.5166	0.614
Task_A_DL_SL (527740)	**0.5341**	**0.6349**

Table 2: Results for Sub-task A.

System	F1 (macro)	Accuracy
All TIN baseline	0.4702	0.8875
All UNT baseline	0.1011	0.1125
Task_B_DL_NB (532649)	**0.4800**	**0.8375**
Task_B_DL_SL (532651)	0.4558	0.8375
Task_B_TL_MNB (532654)	0.4558	0.7792

Table 3: Results for Sub-task B.

System	F1 (macro)	Accuracy
All GRP baseline	0.1787	0.3662
All IND baseline	0.2130	0.4695
All OTH baseline	0.0941	0.1643
Task_C_DL_NB (536200)	0.2462	**0.4507**
Task_C_DL_SL (536201)	0.2663	0.4178
Task_C_TL_SVM (536203)	**0.3001**	0.4178

Table 4: Results for Sub-task C.

The attention mechanism Scaled Luong performs better when more data is available for training. Normed Bahdanau attention mechanism performs better even for a small dataset. However, deep learning gives poor results than traditional learning approach for Task C, because only 3876 instances were considered for model building. The deep learning model could not learn the features appropiately due to less domain knowledge imparted by the smaller data set. Thus, traditional learning performs better with the given data size when compared to deep learning for Task C. The confusion matrix for our best run in the three sub tasks are depicted in Tables 5, 6 and 7. These tables show that the true positive rate of "NOT", "TIN" and "IND" classes are good as the number of samples for those classes are more in training set. Our approaches show improvement over the base line systems for all the three tasks. We have obtained 12% and 14% improvement on F1 and accuracy respectively for Task A when compared with the base line. For Task B, we have obtained 26% and 34% improvement on F1 and accuracy respectively. Also, Task C results have been improved by 14% and 7% for F1 and accuracy when compared to base line results.

	OFF	NOT
OFF	73	147
NOT	167	473

Table 5: Confusion Matrix for Task_A_DL_SL.

	TIN	UNT
TIN	200	26
UNT	13	1

Table 6: Confusion Matrix for Task_B_DL_NB.

	GRP	IND	OTH
GRP	16	26	7
IND	62	71	27
OTH	0	3	2

Table 7: Confusion Matrix for Task_C_TL_SVM.

5 Conclusion

We have implemented both traditional machine learning and deep learning approaches for identifying offensive languages from social media. The approaches are evaluated on OffensEval@SemEval2019 dataset. The given instances are preprocessed and vectorized using word embeddings in deep learning models. We have employed 2 layered bi-directional LSTM with Scaled Luong and Normed Bahdanau attention mechanisms to build the model for all the three sub tasks. The instances are vectorized using TF-IDF score for traditional machine learning models with minimum count two. The classifiers namely Multinomial Naive Bayes and Support Vector Machine with Stochastic Gradient Descent optimizer were employed to build the models for sub tasks B and C. Deep learning with Scaled Luong attention, deep learning with Normed Bahdanau attention, traditional machine learning with SVM give better results for Task A, Task B and Task C respectively. Our models outperform the base line for all the three tasks. The performance may be improved further by incorporating external datasets (Kumar et al., 2018a; Davidson et al., 2017), lexicons and dictionaries.

References

Segun Taofeek Aroyehun and Alexander Gelbukh. 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech

Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 29–30. International World Wide Web Conferences Steering Committee.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. *arXiv preprint arXiv:1804.04257*.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal Framework, Dataset and Annotation Schema for Socially Unacceptable On-line Discourse Practices in Slovene. In *Proceedings of the Workshop Workshop on Abusive Language Online (ALW)*, Vancouver, Canada.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hatespeech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018a. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC)*, Santa Fe, USA.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018b. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting Tweets Against Blacks. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Minh-Thang Luong, Eugene Brevdo, and Rui Zhao. 2017. Neural machine translation (seq2seq) tutorial. *https://github.com/tensorflow/nmt*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Sreekanth Madisetty and Maunendra Sankar Desarkar. 2018. Aggression detection in social media using deep neural networks. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 120–127.

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

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

Kashyap Raiyani, Teresa Gonçalves, Paulo Quaresma, and Vítor Nogueira. 2018. Fully connected neural network with advance preprocessor to identify aggression over facebook and twitter. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. TRAC-2018-ACL.

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.

Huei-Po Su, Chen-Jie Huang, Hao-Tsung Chang, and Chuan-Jie Lin. 2017. Rephrasing Profanity in Chinese Text. In *Proceedings of the Workshop Workshop on Abusive Language Online (ALW)*, Vancouver, Canada.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

D Thenmozhi, B Senthil Kumar, and Chandrabose Aravindan. 2018. Ssn_nlp@ iecsil-fire-2018: Deep learning approach to named entity recognition and relation extraction for conversational systems in indian languages. *CEUR*, 2266:187–201.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Langauge Online*.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval*.

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.