SSN_NLP_MLRG at SemEval-2020 Task 12: Offensive Language Identification in English, Danish, Greek using BERT and Machine Learning Approach

Kalaivani A
Department of CSE
SSN College of Engineering, INDIA
kalaiwind@gmail.com

Thenmozhi D
Department of CSE
SSN College of Engineering, INDIA
theni_d@ssn.edu.in

Abstract

Offensive language identification is to detect the hurtful tweets, derogatory comments, swear words on social media. As an emerging growth of social media communication, offensive language detection has received more attention in the last years. We focus to perform the task on English, Danish and Greek languages. We have investigated which can be effect more on pre-trained models BERT (Bidirectional Encoder Representation from Transformer) and Machine Learning Approaches. Our investigation shows the performance between the three languages and to identify the best performance is evaluated by the classification algorithms. In the shared task SemEval-2020, our team SSN_NLP_MLRG submitted for three languages that are Subtasks A, B, C in English, Subtask A in Danish and Subtask A in Greek. Our team SSN_NLP_MLRG obtained the F1 Scores as 0.90, 0.61, 0.52 for the Subtasks A, B, and C in English, 0.56 for the Subtask A in Danish and 0.67 for the Subtask A in Greek respectively.

1 Introduction

Offensive language detection is the process of identifying and detecting the user generated offensive content or comments that are insult, hurt, profanity, and racism and targeted the individual or group in the massive social media (Marcos Zampieri et al., 2019). As an immense growth of social media, user generated content increasingly occurs on the social media platforms and much more attention to deal with the offensive content. Detecting the offensive content is helpful for the field of sentimental analysis (kalaivani A and Thenmozhi D, 2019), abusive identification (Kenneth Steimel et al., 2019), aggregation and cyber bullying. Several research works have been performed to identify the offensive language in social media. Research workshop has reported in NLP communities that are GermEval 2018 Shared Task on the Identification of Offensive Language (Michael Wiegand et al., 2018). OffensEval@SemEval2019 (Zampieri et al., 2019b) shared task focuses on identification and categorization of offensive content in social media. OffensEval@SemEval2020 (Zampieri et al., 2020) shared task focuses on multilingual identification and categorization of offensive content for five different languages in social media. Recent research area has reported to identifying the offensive content (Zeses Pitenis et al., 2020; Gudbjartur Ingi Sigurbergsson and Leon Derczynski, 2019) and abusive comment (Kenneth Steimel et al., 2019) in different languages. Zeses Pitenis (2020) used machine learning and deep learning techniques for the identification of offensive content in Greek annotated Tweet dataset. (Gudbjartur Ingi Sigurbergsson and Leon Derczynski, 2019) used Logistic regression and more different attention in BiLSTM models to detect and categorize the offensive language in the English and Danish annotated Facebook, reddit dataset. (Kenneth Steimel et al., 2019) used machine learning approaches and sampling method to detect the abusive language in the English and German dataset.

Our team SSN_NLP_MLRG participated in the shared task OffensEval@SemEval2020 task 12 for the three languages that are English, Greek and Danish languages. It focuses on the multilingual offensive language detection, categorization of offensive language and target identification.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 2161–2170
Barcelona, Spain (Online), December 12, 2020.

We have employed traditional machine learning approaches and BERT model to identify and categorize the offensive language. The BERT model that are implemented for the three subtasks A, B, C of English, subtask A of Danish and subtask A of Greek. The machine learning approaches are implemented for the Subtask B, C of the English language. We have investigated which classifiers show the similarity between the three languages and also find the best performance model for all the subtasks in the three languages.

The challenges in the shared task are as follows; a) Choosing the limit to set the given text is offensive or not offensive from the average confidence and standard deviation confidence positive values for the three subtasks A, B, C in English language. b) Handling of huge data in subtask A in the English language. c) Imbalanced data in the subtask A in the Danish and Greek language. d) Removing the noisy data and handling the ungrammatical sentences in all the three languages. e) Determine the types of features for the classifier. f) Small datasets which is hard to build a complex models for training the data in the subtask A in the Danish and Greek language. We have used the Tf-idf Vector for the feature extraction for the classifiers namely Linear Regression, Gaussian Naive Bayes, Support Vector Machine and Random Forest. We have investigated the results of the three languages and compared the results of the machine learning approaches with the performance of the pre-trained Bert models. The rest of the paper is structured as follows: We discuss background in section 2 and the system overview in section 3. Section 4 presents our experimental setup and feature sets. Section 5 presents the results, and draws conclusions and discusses future work in section 6.

2 Back ground

2.1 Task setup

We have used OLID dataset (Zampieri et al., 2020) in the shared task Semeval 2020 task 12. The shared task focuses on multilingual identification and categorization of offensive language in social media. It focuses on five different languages that are English, Greek, Danish, Turkish and Arabic. Our team SSN_NLP_MLRG have participated in the English, Danish and Greek languages. In the English language, we have focuses on three subtasks as follows

Subtask A: Offensive Language Identification

- (NOT) Not Offensive: A post containing no offensive language.
- (OFF) Offensive: A post with offensive language or a targeted offense.

Subtask B: Automatic Categorization of Offense Types

- (TIN) Targeted Threats: A post containing a threat to an individual, a group, or others.
- (UNT) Untargeted: A post containing untargeted offense.

Subtask C: Offense Target Identification

- (IND) Individual: The target of the offensive tweet is an individual.
- (GRP) Group: The target of the offensive tweet is a group of people like religious.
- (OTH) Other: The target of the offensive post does not belong to IND and GRP.

In the Danish language (Gudbjartur Ingi Sigurbergsson and Leon Derczynski, 2020), we have focused the Sub task A: Offensive language identification (NOT) Not Offensive and (OFF) Offensive. In the Greek language, we have focused the Sub task A: Offensive language identification (NOT) Not Offensive and (OFF) Offensive. Table 1, 2, 3 presents the annotated tweets for the English, Danish, Greek languages from the OLID dataset.

Tweets	**Subtask A**	**Subtask B**	**Subtask C**
He ain't gone learn til you stab his ass	**OFF**	**TIN**	**IND**
@USER □ I know and they scam the shit out of these people	**OFF**	**TIN**	**GRP**
I feel like shit and it shows	**OFF**	**TIN**	**OTH**
he s just chilling with bro	**NOT**	NULL	NULL
that s stupid	**OFF**	**UNT**	NULL

Table 1: English Annotated tweets

Tweets	Sub task A
Hvil i fredog æret være hans minde Ja mojn du hol æ kæft	**NOT** **OFF**

Table 2: Danish Annotated tweets

Tweets	Sub task A
Μωρή Λίβερπουλ αυτό φέτος δεν χάνεται με τίποτα AVLLIV εκ φορες μαλακας η μια θα σε πειραξει	**NOT** **OFF**

Table 3: Greek Annotated tweets

2.2 Related work

Offensive Language: (Hamdy Mubarak et al, 2020) reported the lexical features, static and deep contextualized embedding for the Support Vector Machine classifiers to detect Arabic offensive language and determine the topics, dialects, genders which are associated with the offensive tweets. (Ltekin, 2020) presents the corpus of offensive language in the turkish language for the shared task semeval 2020. (Zampieri et al, 2019; Thenmozhi D et al, 2019) reported the shared task SemEval 2019 task 6 for the offensive language identification and catergorization in social media. (Hui-Po Su et al, 2017)presented rephrasing rules to revise the input sentence by using not offensive words on real-word social websites. (Ritesh Kumar et al,2018) reported the Workshop on Trolling, Aggression and Cyberbullying (TRAC) and presents Aggression Identification in Social Media.

Cyberbullying: (Karthik Dinakar et al, 2011) used the YouTube comments and evaluated by the binary classifiers and analysed the various features techniques for the detection of textual cyberbullying. (Jun-Ming Xu et al, 2012) presented the evidence of cyberbullying in social media and used the role labelling and machine learning approaches. (Maral Dadvar et al, 2013) used the user context and analysed the content based, user based features for the detection of cyberbullying. (Sandip Modha et al, 2019) used deep learning approach LSTM to detect the hate speech and offensive content in the English, Hindi and German.

Hate speech: (Irene Kwok and Yuzhou Wang, 2013) used supervised machine learning approach and detect the hate speech which the twitter content has to learn the binary classifier. (Pete Burnap and Matthew L. Williams, 2015) used probabilistic, rule-based, statistical modelling and spatial-based classifiers with a voted ensemble meta-classifier for the detecting the hate speech. (Nemanja Djuric et al, 2015) addressed the issues of high-dimensionality and sparsity and overcome by used the neural language models to identify hate speech content. (Stéphan Tulkens et al, 2016) used dictionary-based approach to detection racism in Dutch language in the social media comments and evaluated by multiple Support vector machines and Word2vec for embedding's. (Ross et al, 2016) used binary classifiers to detect the hateful messages and rate the degree of hateful messages to measuring the reliability. (Schmidt et al, 2017) explored the automation of hate speech detection using the types of utterances in natural language processing. (Schofield et al, 2017) reported comparison between the supervised and unsupervised learning techniques with different feature types for the task. Naive Bayes classifier with Tf-idf features used to detection of hate speech. (Thomas Davidson et al, 2017) used machine learning approaches to detection of the multi-class crowd-sourced hate speech lexicon and to collect tweets containing hate speech keywords. (Shervin Malmasi and Marcos Zampieri, 2017) used supervised classification algorithms with the character n-grams, word n-grams and word skip-grams for detection of profanity and hate speech. (Ziqi Zhang et al, 2018) used Convolution-GRU Based Deep Neural Network to detect the hate content in social media. (Shervin Malmasi and Marcos Zampieri, 2018) used ensemble classifiers with the features that are n-grams, skip-grams and clustering-based word

representations to discriminating the hate comments in social media. (Valerio Basile et al, 2019) reported the shared task of Semeval-2019 task 5: to detecting the presence of hate speech against woman in twitter.

Abusive language: (Chikashi Nobata et al, 2016) performed the machine learning approach and deep learning approach for detected the abusive language in online content. (Hamdy Mubarak et al, 2017) used the unigrams, bigrams features for identification of abusive Arabic language in social media. (Björn Gambäck and Utpal Kumar Sikdar, 2017) used n grams, 4 grams, word2Vec embedding's, conventional neural network for detection and cauterization of abusive language tweets. (Zeerak Waseem et al, 2017) performed a typology model that captures central similarities and differences between subtasks to detect the cyberbullying and abusive comments. (Antigoni-Maria Founta et al, 2018) used incremental and iterative methodology to characterize the abusive behaviour in twitter. (Mai ElSherief et al, 2018) performed Target-based Linguistic Analysis that is directed hate to individual or group, generalized to religious hate speech in Social Media.

3 System Overview

We have employed both the traditional machine learning approaches and BERT pre-trained models to identify and categorize the offensive language in the English, Danish and Greek languages. The training dataset for the English subtasks A, B with columns namely, ID, TEXT, AVERAGE, STD. ID refers the identification number for the tweet, TEXT refers the tweets, and AVERAGE refers average of the confidences to belong to the positive class for that subtask. The positive class is OFF for subtask A, and UNT for subtask B. STD refers confidences standard deviation from AVG_CONF for a particular instance. The English subtask C with columns namely ID, TEXT, AVG_ IND, AVG_ GRP, AVG_ OTH, STD_IND, STD_GRP, STD_OTH. AVG_IND refers average of the confidences of individual target likewise Group target and other target. In the Danish and Greek languages, the training data with the columns namely ID, TWEET, SUBTASK_A. Here SUBTASK_A refers the labels OFF and NOT. First, we have started with the machine learning approaches of the subtask B and subtask C in English language namely logistic regression, Multinomial Naive Bayes (NB), Random forest and Linear Support vector machine (SVC) respectively.

We have used the Doc2vec, Tf-Idf Vectorise for the feature extraction and trained the models using linear regression. Doc2Vec is not giving good performance in the ungrammatical sentences and Tf-Idf performs well when compared with Doc2Vec feature selection. Tf-Idf vectorise of the English subtask B is (188974, 55585) and the English subtask C is (188973, 55585). We have cross validate the data with the train test split size is 0.25 and maximum number of iterations is 100.

Models	**Subtask B**	**Subtask C**
Logistic Regression	**0.80**	**0.86**
Random Forest	0.77	0.84
Linear SVC	0.79	0.85
Multinomial NB	0.72	0.81

Table 4: Accuracies of the English Subtasks B and C.

Second, we have trained the models using Multinomial Naive Bayes, Random Forest and Linear Support vector machines. We have cross validated the each model separately. Accuracies of the models are shown in the Table 4. In both the sub task B and C, Logistic Regression Performs well when compared with the others models. Based on the performance, we have chosen the logistic regression for building the model and cross validate the trained model. The classification metrics for the subtask B using Logistic regression are shown as in Table 5.

Labels	Precision	Recall	F1-Score	Support	Macro F1
UNT TIN	0.73 0.84	0.66 0.88	0.69 0.86	15724 31520	**0.77**
Avg / Total	0.80	0.81	0.80	47244	**0.77**

Table 5: Cross validation scores for the English subtask B

Labels	Precision	Recall	F1-Score	Support	Macro F1
IND GRP OTH	0.76 0.88 0.62	0.51 0.97 0.15	0.61 0.92 0.24	6127 38201 2916	**0.59**
Avg / Total	0.84	0.86	0.84	47244	**0.59**

Table 6: Cross validation scores for the English subtask C

In the subtask A for the English language, it have huge amount of data when compared with the others datasets. We have used the BERT base uncased pre-trained model (Devlin et al, 2018) for building the model. The model has 12-layer, 768-hidden, 12-heads, 110M parameters. The training data is in .tsv files. We have split the training data into train and dev tsv files. We have added the separate special characters in both the train and dev data tsv files.

We download the Bert vocabulary package for evaluation purpose. We pass the train, dev file for build the training model. In that we have used the cola processer to build and predict the model with the batch size is 32 and the trained epochs are 4. BERT model performs good compare with the machine learning approaches. We have decided to use BERT model for all three subtasks A, B, C in English, Subtask A in Danish and Subtask A in the Greek. We have to investigate the similarity between the three languages using BERT model.

4 Experimental setup

The test data is given in the tsv file format by the organizer of the SemEval 2020 shared task. The format of the test data with the columns namely ID, TWEET for the three languages. We have pre-processed the data by removing the URLs and the text "@USER" from the tweets. Word tokenizer is used for the Subtask A in English, Tweet tokenizer is used to obtain the vocabulary and features for the training data. We have used the NTLK libraries and Gensim libraries for pre-processing.

Data Split	English			Danish	Greek
Subtask	A	B	C	A	A
Train Dev Test	7260334 1815084 3887	151179 37795 1442	151178 37795 850	2368 592 330	`6994 1544 1749
Total	**9079305**	**190416**	**189823**	**3290**	**10287**

Table 7: BERT models data split up for the three languages.

In the English language subtask A, we used the BERT model to predict the label as OFF, NOT. OFF refers the offensive Content which insults someone or groups and NOT refers the Non-offensive comments which will not affect the persons. We have split the data into the train, dev and test data for development to build and predict the model as shown in the Table 7. In the English subtask B and C, we

have used both the Logistic Regression and BERT models to predict the labels as TIN or UNT. TIN refers the targeted offensive comments which hurt the individual or group. UNT refers the untargeted offensive comments. The performance of the English subtask B is good in BERT model than the Logistic Regression.

Task	**Precision**	**Recall**	**F1 Macro**	**Baseline F1**
English subtask A	0.889	0.944	**0.909**	0.419
English subtask B	0.689	0.625	**0.618**	0.374
English subtask C	0.360	0.380	0.317	0.270
Danish subtask A	0.631	0.555	**0.567**	0.514
Greek subtask A	0.745	0.745	**0.747**	0.426

Table 8: BERT models for the three languages

The English subtask C gives better performance in Logistic Regression than the BERT models. In the Danish and Greek Languages, we have used the BERT models to predict the labels as OFF and NOT.

Task	**Precision**	**Recall**	**F1 Macro**	**Baseline F1**
English subtask B	0.321	0.366	0.300	0.374
English subtask C	0.622	0.511	**0.526**	0.270

Table 9: Logistic regression models for the English Subtask B and C

5 Results

We have used the evaluation metrics as precision, recall and F1 macro. The performances of the three languages of BERT model as shown in the Table 8. The performances of the English subtask B and subtask C of Logistic Regression as shown in the Table 9. Based on the evaluation metrics, the English subtask A and B performs well in BERT model. The English subtask C performs well in the Logistic Regression. The Greek language outperforms the Danish subtask A, English subtask B and C.

5.1 Analysis

Overall the English subtask A gives good performance than the other languages; because training data is very large when compare with the others languages and other subtask in English. The performance of the English subtask A, B are good than Subtask C, because the binary classification works well in BERT model than Multi label classifications. Multi label classification is good in Logistic Regression in the machine learning approach than the other machine learning approach what we used. The F1 macro results of the English subtasks A, B and C have improved by 116.8%, 65.2% and 94.5% respectively when compared with the baseline F1 results as shown in Table 8 and 9. The F1 macro result of the Danish subtask A has improved by 11.17% when compared with the baseline F1 result as shown in Table 8. The F1 macro result of the Greek subtask A has improved by 77.8% when compared with the baseline F1 result as shown in Table 8 respectively.

We have observed that the BERT model performed well in the three languages of the English subtask A, B, Danish subtask A and Greek subtask A for Binary Classification and Logistic Regression performed well in the English subtask C for multi label classification. Officially we submitted the BERT models Scores. The confusion matrixes for the three languages are shown in the Figures 1, 2, 3, 4 and 5. we have obtained the best scores for BERT models in the English subtask A ,B ,the Danish subtask A, the Greek subtask A and Logistic regression in the English subtask C respectively.

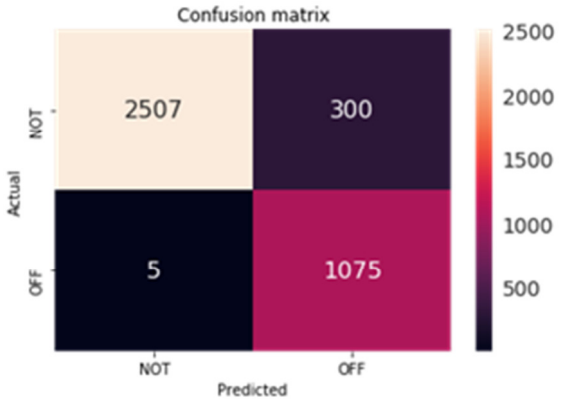

Figure 1: English Subtask A – BERT

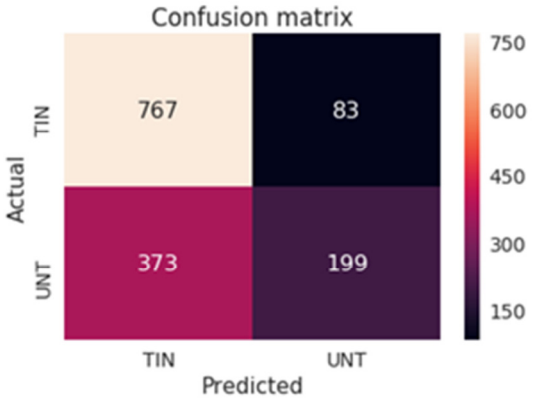

Figure 2: English Subtask B - BERT

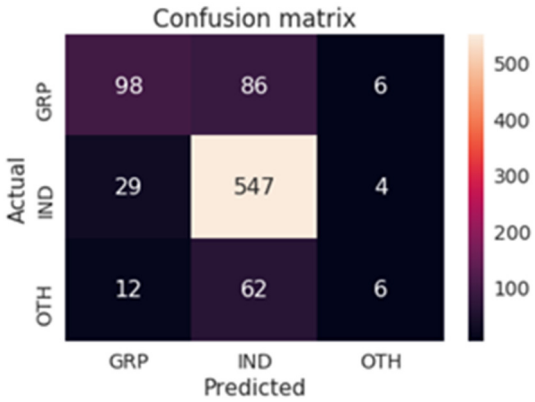

Figure 3: English Subtask C- Logistic Regression

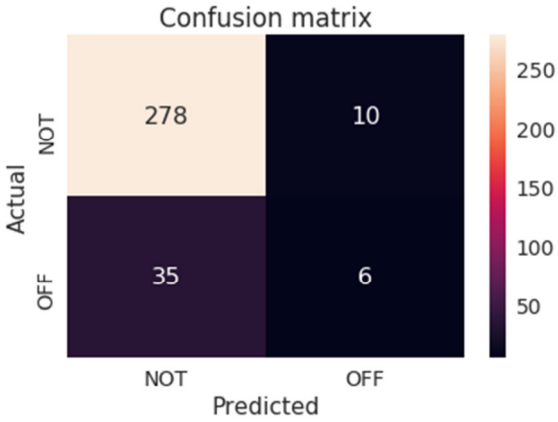

Figure 4: Danish Subtask A - BERT

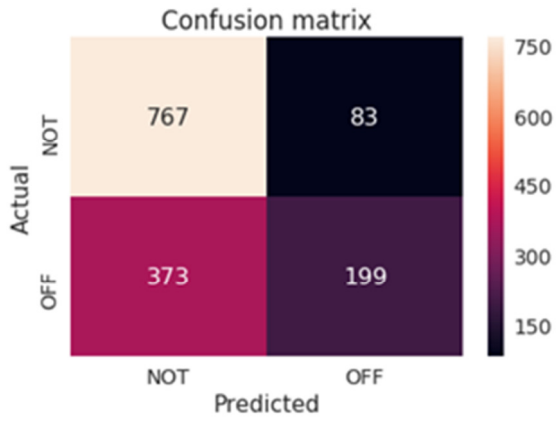

Figure 5: Greek Subtask A - BERT

6 Conclusion

We have implemented both the traditional machine learning and BERT pre-trained model to identify the offensive language for the English subtask A, Danish subtask A, Greek Subtask A and categorization of offensive languages for the English subtask B and C from social media. The approaches are evaluated on OffensEval@SemEval2020 dataset. The given tweets are pre-processed and vectorised using Tf-Idf in machine learning models. The classifiers namely Multinomial Naive Bayes, Support Vector Machine, Logistic regression and Random forest were employed to build the models for sub tasks B and C. We have employed BERT model to build the model for all the subtasks in the three languages. Logistic regression performs better results for the subtask C in English. BERT model give better results for subtask A in the English, Danish, Greek and subtask B in English respectively. Our models outperform the base line for all the three tasks. The performance may be improved further by evaluating different models with the adoptable features.

References

Kenneth Steimel, Daniel Dakota, Yue Chen, and Sandra Kübler. 2019. Investigating Multilingual Abusive Language Detection: A Cautionary Tale, *Proceedings of Recent Advances in Natural Language Processing*, pages 1151–1160,Varna, Bulgaria, Sep 2–4, 2019. https://doi.org/10.26615/978-954-452-056-4_132.

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)*.

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-HLT 2019*, Association for Computational Linguistics, pages 1415–1420, June 2 - June 7, 2019.

Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2019. Offensive Language and Hate Speech Detection for Danish arXiv:108.04531 [cs.CL], ACM. August 14, 2019.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, ltekin,. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020, *Proceedings of SemEval2020*.

Mubarak, Rashed Hamdy, Ammar, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic Offensive Language on Twitter: Analysis and Experiments, arXiv preprint arXiv:2004.02192.

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

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

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A Large-Scale Semi-Supervised Dataset for Offensive Language Identification, arxiv.

ltekin. 2020. A Corpus of Turkish Offensive Language on Social Media, *Proceedings of the 12th International Conference on Language Resources and Evaluation.*

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*), Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval).*

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, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages(1415—1420).

Dinakar, Karthik and Reichart, Roi and Lieberman, and Henry. 2011. Modeling the detection of Textual Cyberbullying, *The Social Mobile Web*, pages (11—17).

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media, *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.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context, *Advances in Information Retrieval*, pages (693—696), Springer.

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

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, pages (223—242), Wiley Online Library.

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

St'ephan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A Dictionary-based Approach to Racism Detection in Dutch Social Media, *Proceedings of the Workshop Text Analytics for Cybersecurity and Online Safety (TA-COS).*

Bj"orn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis, *Proceedings of the Workshop on Natural Language Processing for Computer-Mediated Communication (NLP4CMC).*

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content, *Proceedings of the 25th International Conference on World Wide Web*, pages(145—153), International World Wide Web Conferences Steering Committee.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection Using Natural Language Processing, *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics,* Pages (1—10).

Alexandra Schofield and Thomas Davidson. 2017. Identifying Hate Speech in Social Media, *XRDS: Crossroads, The ACM Magazine for Students*, volume-24, Pages (56—59), ACM.

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

Hamdy Mubarak, Darwish Kareem and Magdy Walid. 2017.Abusive Language Detection on Arabic Social Media, *Proceedings of the Workshop on Abusive Language Online (ALW).*

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

Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media, *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP).*

Bj"orn Gamb"ack and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-speech, *Proceedings of the First Workshop on Abusive Language Online.*

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

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network, Lecture Notes in Computer Science, Springer Verlag.

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.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profanity from Hate Speech, *Journal of Experimental and Theoretical Artificial Intelligence*, pages(1—16), volume 30, issue 2, Taylor and Francis.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis.2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior, arXiv preprint arXiv:1802.00393.

Kalaivani A and Thenmozhi D. 2019. Sentimental Analysis using Deep Learning Techniques, International journal of recent technology and engineering, ISSN: 2277-3878.

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

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

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

D Thenmozhi, Senthil Kumar B, Srinethe Sharavanan, and Aravindan Chandrabose. 2019. SSN NLP at SemEval-2019 Task 6: Offensive language identification in social media using machine learning and deep learning approaches. In Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval).

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, *Proceedings of the 13th International Workshop on Semantic Evaluation.*

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, *Proceedings of the 11th Forum for Information Retrieval Evaluation.*