NAYEL at SemEval-2020 Task 12: TF/IDF-Based Approach for Automatic Offensive Language Detection in Arabic Tweets

Hamada A. Nayel
Department of Computer Science
Faculty of Computers and Artificial Intelligence
Benha University
hamada.ali@fci.bu.edu.eg

Abstract

In this paper, we present the system submitted to "SemEval-2020 Task 12". The proposed system aims at automatically identify the Offensive Language in Arabic Tweets. A machine learning based approach has been used to design our system. We implemented a linear classifier with Stochastic Gradient Descent (SGD) as optimization algorithm. Our model reported 84.20%, 81.82% f1-score on development set and test set respectively. The best performed system and the system in the last rank reported 90.17% and 44.51% f1-score on test set respectively.

1 Introduction

The tremendous usage of social media platforms makes it important to apply different Natural Language Processing (NLP) tasks on these platforms. Different tasks, such as cyberbullying identification, hate speech detection, sarcasm detection and offensive language detection attracted NLP researchers to concentrate on automation of these tasks (Kwok and Wang, 2013). One of these tasks which gained a research interests is automatic offensive language detection. Offensive language is widespread in social media. Computational offensive language detection is a solution to identify such hostility and has shown promising performance (Nayel and L, 2019).

Arabic is a significant language having an immense number of speakers as it is the official language of 22 countries (Guellil et al., 2019). It is recognized as the 4th most used language of the Internet (Boudad et al., 2018). The research in NLP for Arabic is constantly increasing (Nayel et al., 2019). Automatic offensive language detection becomes an important NLP task due to the overwhelming usage of social media. Automatic offensive language identification in Arabic is a challenge due to the complexity of Arabic language (Nayel, 2019).

In this paper we describe the model that has been submitted to the offensive language detection shared task "OffensEval 2020" (Zampieri et al., 2020). Given a tweet, then the task in brief is to determine whether it contains an offensive language or not. The first version, "OffensEval 2019", was held at SemEval 2019 (Zampieri et al., 2019b). A dataset containing English tweets and annotated using a hierarchical three-level annotation model has been used in "OffensEval 2019" (Zampieri et al., 2019a). In "OffensEval 2020", in addition to English, four more languages have been added to the dataset namely, Arabic, Danish, Greek and Turkish. We participated in "OffensEval 2020" for Arabic. A machine learning based approach has been used to develop our submission. Term Frequency/ Inverse Document Frequency (TF/IDF) vector space model has been used to represent the given tweets.

2 Related Work

Recently, offensive language detection has gained significant attention and a lot of contributions have been recorded in this area (Waseem et al., 2017; Davidson et al., 2017; Kumar et al., 2018; Mubarak et al., 2017; Malmasi and Zampieri, 2018; Mandl et al., 2019). Zampieri et al presented a dataset with annotation of type and target of offensive language (Zampieri et al., 2019a). They implemented SVM, Convolutional

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 2086–2089
Barcelona, Spain (Online), December 12, 2020.

Neural Network (CNN) and Bidirectional Long Short-Term-Memory (BiLSTM) for offensive language detection. Nayel and Shashireka used classical machine learning algorithms to detect hate speech for multi-lingual tweets (Nayel and L, 2019).

3 Task Description

Given a tweet, the objective of the task is to determine if the tweet contains offensive language or not. Suppose $\mathcal{C} = \{\texttt{NOT}, \texttt{OFF}\}$, a set of two classes where **NOT** is the class of non-Offensive tweets and **OFF** is the class of offensive tweets. We have formulated the task as a binary classification problem that assigns one of the two predefined classes of $\mathcal{C}$ to a new unlabelled tweet.

4 Methodology

Our approach depends on TF/IDF vector space model, convert the tweet into a vector and then apply the linear classifier on the vector space. Linear classifier is a simple classifier that uses a set of linear discriminant functions to distinguish between different classes (Theodoridis and Koutroumbas, 2009).

4.1 General Framework

The general framework of the proposed model consists of the following stages:

4.1.1 Preprocessing

Preprocessing was the first stage in our pipeline. In this stage the following steps have been applied to tweets:

1. **Abbreviation Removal**
   '@USER', 'URL' and '<LF>' were commonly used in tweets. These are English abbreviations and refers to private information about users.

2. **Punctuation and Digit Elimination**
   Punctuation marks such as {'+', '_', '#', '$'.. } and digits {'0','1','2',..,'9', '٠', '١', '٢',... '٩'} have been removed. These are increasing the dimension of feature space with redundant features.

3. **Elongation Elimination**
   Majority of Arabic tweets are free of following the standard rules of Arabic language. A common manner of users is to repeat a specific letter in a word. Elongation elimination encompasses removing this redundancy to reduce the feature space. In our experiments, the letter is assumed to be redundant if it is repeated more than two times. For example the words "مبروووووووك" [pronounced "*mabrook*" and the meaning is congratulation] and "عاااااجل" [pronounced *"aaagel"* and the meaning is "urgent") containing redundant letter and will be reduced to "مبروك" and "عااجل" respectively.

4.1.2 Feature Extraction

The second stage in our pipeline was feature extraction. TF/IDF with range of n-grams has been used to represent all the tweets in the training set. TF/IDF has been calculated as given in (Nayel and Shashirekha, 2017). We used range of 3-grams model, i.e. unigram, bigram and trigram terms. For example the sentence "الدورى يا زمالك" [pronounced "*eldawry ya zamalek*" and the meaning is League oh Zamalek[1]] has following set of features {"الدورى", "يا", "زمالك", "الدورى يا", "يا زمالك", "الدورى يا زمالك"}.

4.1.3 Training Classifier

In this phase, we used the features that have been extracted in previous phase to train the classifier. We tried a set of different classifiers, namely, linear classifier, Support Vector Machines (SVM), Multilayer Perceptron (MLP), as well as ensemble approach. According to the task's rules only one run can be submitted. The output of the best performed classifier on the development set has been submitted.

[1]Zamalek is one the most famous sports club in Arab world and Africa

4.2 Dataset

The dataset that was used to build the model has been distributed by organizers contains a set of tweets and divided into training, dev and test set (Mubarak et al., 2020). A statistics about the training and development sets is given in table 1, and the test set contains 2000 unlabeled tweets.

	OFF	**NOT**	**Total**
Training set	1410	5590	7000
Development set	179	821	1000
Total	1589	6411	8000

Table 1: Statistics of training and development sets

5 Experiments and Results

In the proposed models, the Stochastic Gradient Descent (SGD) optimization algorithm has been used for optimizing the parameters of linear classifier. The loss function used in linear classifier was "Hinge" loss function (Rosasco et al., 2004). Linear kernel has been used for SVM classifier. In MLP classifier the logistic function has been used as activation function using 20 neurons in the hidden layer. We used hard voting approach for ensembles the output of all classifiers. The performance of the proposed classifiers on development, and test set is represented as f1-score and given in Table2.

	Development set	**Test set**
Linear Classifier	0.8421	0.8182
SVM	0.8115	0.8043
MLP $(n = 60)$	0.8033	0.7831
Voting	0.8265	0.8129

Table 2: F1-score of implemented classifiers on development set and test set

The local context representation of tweets, TF/IDF, affected the performance of our model negatively. In addition, the usage of classical classification algorithms limits the performance of the proposed models. Deep learning models show improvement in different NLP tasks, where deep models depend on the word embeddings (a semi-supervised approach for global word representation).

6 Conclusion

In this working notes, a model which performs satisfactorily in the given task has been presented. The model is based on a simple framework, where TF/IDF was used as as weighting scores and classical machine learning algorithms as classifiers. The improvement of our work can be done using deep learning architecture with better word representation. Another hitch of the model is that it does not use any external data other than the provided dataset which may affects results based on the small size of the data. Investment of the related domain knowledge may improve the performance of the model.

References

Naaima Boudad, Rdouan Faizi, Rachid [Oulad Haj Thami], and Raddouane Chiheb. 2018. Sentiment analysis in arabic: A review of the literature. *Ain Shams Engineering Journal*, 9(4):2479 – 2490.

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

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2019. Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*.

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.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *AAAI Conference on Artificial Intelligence*.

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

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. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, pages 14–17.

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

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*.

Hamada A. Nayel and Shashirekha H. L. 2019. DEEP at HASOC2019: A machine learning framework for hate speech and offensive language detection. In Parth Mehta, Paolo Rosso, Prasenjit Majumder, and Mandar Mitra, editors, *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 336–343. CEUR-WS.org.

Hamada A. Nayel and H. L. Shashirekha. 2017. Mangalore-University@INLI-FIRE-2017: Indian Native Language Identification using Support Vector Machines and Ensemble Approach. In Prasenjit Majumder, Mandar Mitra, Parth Mehta, and Jainisha Sankhavara, editors, *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, December 8-10, 2017.*, volume 2036 of *CEUR Workshop Proceedings*, pages 106–109. CEUR-WS.org.

Hamada A. Nayel, Walaa Medhat, and Metwally Rashad. 2019. BENHA@IDAT: Improving Irony Detection in Arabic Tweets using Ensemble Approach. In Parth Mehta, Paolo Rosso, Prasenjit Majumder, and Mandar Mitra, editors, *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 401–408. CEUR-WS.org, December.

Hamada A. Nayel. 2019. NAYEL@APDA: Machine Learning Approach for Author Profiling and Deception Detection in Arabic Texts. In Parth Mehta, Paolo Rosso, Prasenjit Majumder, and Mandar Mitra, editors, *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 92–99. CEUR-WS.org, December.

Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. 2004. Are loss functions all the same? *Neural Comput.*, 16(5):1063–1076, May.

Sergios Theodoridis and Konstantinos Koutroumbas. 2009. Chapter 3 - Linear Classifiers. In *Pattern Recognition (Fourth Edition)*, pages 91 – 150. Academic Press, Boston, fourth edition edition.

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*.

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

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, 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*.