Adrar University
Faculty of Science and Technology
Department of Mathematics and Computer Science

Initiation to Research (Course)

2nd Year Master (S3)
2020/2021

Arabic Hate Speech Classification using Pre-trained Multi-task Learning Models

Aldjanbi wassen [1]

Instructor: Dr. Abdelghani DAHOU [2]

February 15, 2021

[1]Email: wassen.eldjanabi@gmail.com
[2]Email: dahou.abdghani@univ-adrar.edu.dz

Contents

LIST OF FIGURES

List of Tables

1 Abstract

social media platforms has provided opportunities for people to connect but has also opened the door for misuse with the spread of hate speech, according to the Cambridge dictionary hate speech is defined as "public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation". Detecting such content is crucial for understanding and predicting conflict,in this research proposal we tackled this problem by adapting an existing multi-task learning for deep learning model (MTL-DNN) to perform Arabic hate speech classification, we used the AraBert model with the addition of MTL to get better performance.

2 Introduction

Hate speech was found to negatively impact the psychological well-being of individuals and to deteriorate inter-group relations on the societal level (Tynes et al., 2008)(1). As such, detection and prevention mechanisms should be setup to deal with such content. Machine learning algorithms can be employed to automatically detect these behaviors by relying on recent techniques in natural language processing that have shown a significant performance.

A small number of works targeted this problem but by the detecting both hate and offensive speech in Arabic simultaneously,for exemple Haddad et al.(2019) targeted the problem of hate and offensive speech detection for the Tunisian dialect using Support Vector Machine (SVM) and Naive Bayes classifier.In the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4) Marc Djandji et al.(2020) developed a multi-task learning using AraBert model.Ibrahim Abu-Farha et al(2020) in the same shared task proposed a multitask learning architecture using CNN-BiLSTM model.

The models that we will experiment is on fine-tuning the Arabic Bidirectional Encoder Representation from Transformer (AraBERT) model (AUBMind-Lab, 2020),the aim of this study is to train our model on multiple corpora.

3 Related Works

.

- Multi-Task Deep Neural Networks for Natural Language Understanding: In this paper (Xiaodong Liu et al(2020))(2)combined multi-task learning and language model pre-training and proposed a new Multi-Task Deep Neural Network (MT-DNN) we took this model as a based model and adapted it to perform on Arabic hate speech classification
- Hate Speech Detection: very few works in the literature target the problem of Arabic hate speech detection (Albadi et al. (2018))(3)introduced the first dataset containing 6.6K Arabic hate-speech tweets targeting religious groups,they compared a lexicon-based classifier, SVM classifier trained with character n-gram features.
- Hate and Offensive Speech Detection:(Haddad et al. (2019))(4) created a dataset of 6K tweets containing hate and offen- sive speech in the Tunisian dialect. For binary

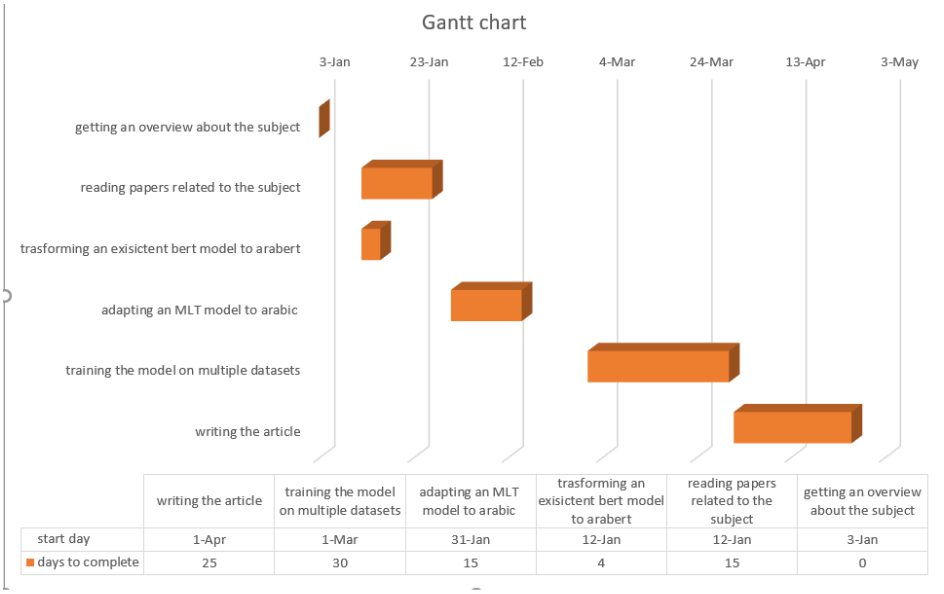
(offensive, non-offensive) and multi-class (offensive, hate, or normal) classification of hate and offensive speech, the authors ex- tracted several n-gram features from each tweet and applied Term Frequency (TF) weighing to select the most effective features. The extracted features were then used to develop an SVM and Naive Bayesian (NB) classifier.(Marc Djandji et al(2020))(5) proposed a model based on AraBERT with Multitask Learning, which solves the data imbalance problem by leveraging information from multiple tasks simultaneously.(Ibrahim Abu-Farha(2020))(6) developed a multitask learning architecture, based on CNN-BiLSTM, that was trained to detect hate-speech and offensive language and predict sentiment

4 METHODOLOGY/RESEARCH METHODS

We based our approaches on the recently released AraBERT model with the multitask learning,our model consists of two components: a part that gets trained by all the tasks data in order to extract a general feature representation for all the tasks and a task-specific part that gets trained only by the task-specific

for the test dataset we used the SemEval 2020 Task 12 Arabic offensive language dataset (OffensEval2020, Subtask A), contain 10K tweets that were annotated for offensiveness with labels (OFF or NOT OFF) and hate speech with labels (HS or NOT HS). and few data preprocessing was performed where user mentions were replaced with @USER, URLs were replaced with URL, and empty lines were replaced with <LF>

due to the lack of the hate speech dataset we trained our model on multiple datesets.



	writing the article	training the model on multiple datasets	adapting an MLT model to arabic	trasforming an exisictent bert model to arabert	reading papers related to the subject	getting an overview about the subject
start day	1-Apr	1-Mar	31-Jan	12-Jan	12-Jan	3-Jan
days to complete	25	30	15	4	15	0

Figure 5.1: Gantt chart

5 Project Timeline

This work has Started from january 3rd, 2020 it started by having an overview of the subjet and reading papers related to the current researches and techniques. For now we are trying to adabt the MTL model to arabic classification, then we will try to get to our objective with is training our model on multiple corporas .

REFERENCES

1 TYNES, B. M. et al. Online racial discrimination and psychological adjustment among adolescents. *Journal of adolescent health*, Elsevier, v. 43, n. 6, p. 565–569, 2008.

2 LIU, X. et al. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*, 2019.

3 ALBADI, N.; KURDI, M.; MISHRA, S. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In: IEEE. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. [S.l.], 2018. p. 69–76.

4 HADDAD, H.; MULKI, H.; OUESLATI, A. T-hsab: A tunisian hate speech and abusive dataset. In: SPRINGER. *International Conference on Arabic Language Processing*. [S.l.], 2019. p. 251–263.

5 DJANDJI, M. et al. Multi-task learning using arabert for offensive language detection. In: *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*. [S.l.: s.n.], 2020. p. 97–101.

6 FARHA, I. A.; MAGDY, W. Multitask learning for arabic offensive language and hate-speech detection. In: *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*. [S.l.: s.n.], 2020. p. 86–90.