LING 573 Project Report: D1

Levon Haroutunian*, **Yi-Chien Lin***, **Bridget Tyree***, **David Yi***
Department of Linguistics
University of Washington
{levonh, yichlin, btyree, davidyi6}@uw.edu

Abstract

1 Introduction

2 Task Description

This project focuses on detecting offensive speech in social media. The detection of offensive or abusive language has been a rich area of research in NLP, which is overviewed in Waseem et al. (2017). Automated abuse detection can be a helpful tool for content moderation, though it should be noted that such a tool is only one component of a system that can adequately address and prevent toxic behavior (Jurgens et al., 2019).

The tasks in this project are based on OffensEval 2020[1] (Zampieri et al., 2020), a shared task organized at SemEval 2020. Description regarding the primary task and adaptation task are presented respectively in the following subsections.

2.1 Primary Task

For our primary task, we tackle the subtask A of OffensEval 2020 – identifying offensive language in English (i.e. determining whether a given statement is offensive). More specifically, the affect type of this task is interpersonal stance and the target is sentiment. The identification of offensive language is obtained by implementing binary text classification on tweets.

We use the same training and test sets as the ones used in the subtask A of OffensEval 2020. OffensEval 2020 provides two datasets – Offensive Language Identification Dataset (OLID)[2] (Zampieri et al., 2019) and Semi-Supervised Offensive Language Identification Dataset (SOLID)[3] (Rosenthal et al., 2020). The former is the dataset from OffensEval-2019, which was annotated tweets using crowdsourcing based on a three-layer annotation scheme. This annotation scheme enables developers to use the same dataset for different subtasks. The latter was a newly created dataset specifically created for OffensEval 2020. SOLID was annotated using the same annotation scheme as OLID but was labelled in a semi-supervised manner.

Our training data for the primary task contains 9,089,140 English tweets, in which 1,448,861 are annotated as offensive (OFF) and 7,640,279 are labeled as not offensive (NOT). The provided test set contains 1,090 OFF tweets and 2,807 NOT tweets. For evaluation, we are planning to compute the Macro F1 score, which is the official evaluation metric of OffensEval 2020.

2.2 Adaptation Task

We selected subtask B of OffensEval 2020 as our adaptation task. Subtask B focuses on categorizing offensive language (Zampieri et al., 2020). That is, whether an offensive English tweet is untargeted or contains insults or threats that target a group or an individual. The dimension of this adaptation task highly overlaps with our primary task. The only difference is that the target of this adaptation task is entity, rather than sentiment.

OLID and SOLID datasets are also used in this adaptation task. The training data contains 188,974 English tweets, in which 149,550 are targeted (TIN) and 39,424 are not (UNT). The test set includes 850 TIN and 1,072 UNT tweets (1,922 tweets in total). The evaluation method is same as the one used for our primary task – calculating the F1 scores of the classification results.

*Equal contribution, sorted in alphabetical order.

[1] https://sites.google.com/site/offensevalsharedtask/results-and-paper-submission

[2] https://sites.google.com/site/offensevalsharedtask/olid

[3] https://sites.google.com/site/offensevalsharedtask/solid

3 System Overview

4 Approach

5 Results

6 Discussion

7 Conclusion

References

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*.

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 Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

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*.

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). *arXiv preprint arXiv:2006.07235*.