Final Project Proposal

Hate Speech Detection on Obscured Text Using Language Models

20160502 : 이재희 / 20160791: 권용빈

20180679: 최우진 / 20190074 : 김동근

Project Topic

Application Project

Motivation

In the age of the Internet, where millions of photo, video, and text is uploaded every second, content moderation is now more important than ever. One important aspect of moderation is hate speech. This includes simply censoring out curse words, to hateful posts against certain demographics. Hate speech detection is one of the major research topics in NLP (Natural Language Processing). In this project, we want to tackle a specific case of detecting hate speech, which is, intentionally obscured words or sentences. In order to avoid the relatively simple moderation systems today, many people intentionally obscure the words they write to avoid redaction. For example, one could add random punctuation symbols in between or make grammatical errors. We want to apply machine learning techniques to perform a similar process as spelling-error correction to identify the meaning of the obscured text, and classify whether they are hate speech or not. Furthermore, we want to explore what kind of services we can provide based on the model we've built.

Method

We intend to use transformer language models such as BERT, OpenAI-GPT, ELMO, to pre-process the text data into contextualized vector representation. Then, we will fine-tune this model using CNN, or other models. The goal is to classify the text into positive/negative labels, indicating whether it is hate speech or not. In the hidden layers, our model should predict the true meaning of the obscured text based on context.

Experiments

We intend to gather data such as tweets, posts and comments from forums, and other various social network posts. Because the data we require is specific (obscured or grammatically-incorrect text), we may have to gather data from crawling, or make our own.

After gathering data, we will split them into train, test, and validation sets, and train the model accordingly. The performance of the model will be measured on our datasets. We are also considering developing a simple web-interface for testing out our model, and also as a proof-of-concept for real world application of our project. For example, when a user inputs some text, it will tell you if it is hate speech or not.

Reference

[1] https://aclanthology.org/2020.acl-main.82.pdf <spelling-error correction using BERT>

[2] https://aclanthology.org/2020.semeval-1.271.pdf <BERT-CNN for offensive speech detection>