Toxic Speech Classification Project Proposal

Chengyuan Zhou
czhou690@usc.edu

Ruichao Ma
ruichaom@usc.edu

Tongkai Yang
tongkaiy@usc.edu

Shihan Xu
shihanxu@usc.edu

Jiawen Song
jiawenso@usc.edu

Abstract

This document is a proposal for group project of course CSCI544 Applied Natural Language Processing. The project tries to classify toxic speech by using NLP techniques including text cleaning and pre-trained language models. There has been many research on speech recognition and text classification. However, toxic speech such as insult, threat, discrimination and identity hate is not a well explored area in both speech and text field. Therefore, we decided to fine-tune a multi-language pre-trained model by using datasets provided Jigsaw that contains large amount of annotated toxic comments. Then combine the classifier with a speech recognition model to deliver a toxic speech classifier.

1 Introduction

We have noticed that many people will express their anger or madness via vocal language and they may use aggressive words or curse at somebody else. However, such "name-calling" behaviors might not be detected by text-based sentiment analysis or words-filters. More and more people start realizing this and use speech instead of text when they want to use the toxic/aggressive language.

Our idea is to implement a speech-based toxic content recognizer that can detect toxic words or aggressive words in the speech so that people can better behave themselves even when they're talking.

2 Related Work

It is a very wide-spread and meaningful topic in our real life. Prior scientists have conducted many research in this area. Two professors at Stanford University, Animesh Koratana and Kevin Hu, claimed that previous research on speech recognition mainly used "classical methods" that require manual feature engineering, including logistic regression, Bayesian model and random forest; it is proven to be successful but has its own shortcomings. Feature engineering processes take enormous efforts and are always inexhaustive, as people can easily wrap their message and find ways around being detected by the system. However, by using deep learning and NLP models, this process can be much easier since feature extraction is automated and can perhaps capture patterns and trends that humans cannot think of. In their experiment, the LSTM model is able to produce higher F1 and accuracy compared to a fairly strong logistic regression baseline.

Another scientist researcher, Sandjai Bhulai also proved the efficiency of NLP algorithms in toxic word detection. He deployed neural networks where the word embeddings were initialized with either random embeddings or GloVe embedding. LSTM performed best and embeddings learned from deep neural network models when combined with gradient boosted decision trees lead to best accuracy. However, there also exist some limitations for NLP in toxic words detection. Hate speech is a difficult phenomena to define, making it even harder for users to classify a text as hate. It could easily cause subjective biased problems. They have future work worthy to do. In the future, to enable successful execution of the research it was first necessary to understand what toxic sentences are. To accomplish this, an overview of this topic has been conducted. Here it can be concluded that a toxic sentence has several definitions, all coming from different platforms. Toxic sentence detection is a classification-related task, and that is why further literature was reviewed to understand the idea

behind Natural Language Processing and the application of various techniques. Previous work showed that deep learning models improve the state-of-art approaches within hate speech classification tasks. Therefore, NLP techniques deserve to be widely used.

3 Datasets

For this project, we mainly use a dataset from Jigsaw's API, with URL https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification/data, which focuses on toxicity models and others in a growing set of languages. Over the past year, the field has seen impressive multilingual capabilities from the latest model innovations, including few- and zero-shot learning. We're excited to learn whether these results "translate" to toxicity classification. Our data will be around 250000 rows of English data consisting of all toxic comments. Theses comments are from Wikipedia talk pages and Civil Comments dataset. Then, we add some features to classify them, including "toxic", "severe toxic", "obscene", "threat", "insult", "identity hate". Since our dataset structure is similar to our homework, we will use standardized preprocessing and cleaning steps as we used in homework1.

4 Methodology

Training model directly using speech can be tough and is hard to implement, as we need to find a way to encode the speech input (presumably it should be audio files in different format), however, encoding audio file to a format that can be used to train a model is much more harder than encoding texts. Alternatively, we can first convert the audio file to text and use a text-based model to classify contents. To sum up, because training a classifier that takes speech as input directly requires lots of training samples as well as encodings, which we are lacking, we will divide this project into following two major parts.

First, we will build a converter that converts speech to text. There are many available speech-to-text packages or APIs, what we can do is to simply make a call to the package or cloud services such as IBM Watson speech to text services and store the returned texts. We can use Lambda to connect speech input and text output. The speech to text module will handle the conversion for us. For example, the webpage here is the introduction of IBM Watson Speech to text module, it provides multiple platforms as well as languages. IBM Watson Speech to Text technology enables fast and accurate speech transcription in multiple languages for a variety of use cases, `https://www.ibm.com/cloud/watson-speech-to-text#:~:text=What%20is%20Watson%20Speech%20to,agent%20assistance%20and%20speech%20analytics`.

Second, we will build a text-based classifier that identifies toxic contents. This can be accomplished by training a classifier using labeled data in forms of texts, which is much more easier to obtain and train than speech. Such models may utilize different techniques relating to NLP, include but not limited to LSTM, SVM, Logistic Regression as multivariate classification (because we have predictions for different levels of toxic words and different toxic words). We can even train multiple models on the training dataset and pick the one with best performance.

By combining two parts, we will be able to build a pipeline that takes input as speech and return result based on the text.

5 Technical Challenge

Technical challenges can mainly be divided into following three categories: data cleaning and processing, NLP techniques and speech recognition of toxic words. First, data cleaning and preprocessing would be a difficult step because data consists of lots of punctuation and emoji. Especially for emoji, we want first to judge whether some emoji convey useful information or are meaningless at all. Then we want to remove meaningless emoji to help improve prediction accuracy. But how to judge whether emoji is useful or not is inherently difficult because sometimes emoji don't represent what the speaker really wants to express, or even it will cause misleading effects. Also, since our dataset is abundant, we cannot assure that one algorithm or judging criteria can fit all circumstances so we need to specialize them.

Second, some toxic comments are so long. Under such circumstances, contextual words and phrases and homonyms would be a more difficult problem. Since one comment may have lots of the same words but they may have different meanings according to the context of the sentence. More-

over, irony and sarcasm present problems for machine learning models because they generally use words and phrases that, strictly by definition, may be positive or negative, but actually connote the opposite. Since toxic comments are longer, ambiguity can be more troublesome. For instance, one word could be used as a verb, noun, or adjective. How to judge them may need us to update complex algorithms. Also, semantic ambiguity and syntactic ambiguity will cause us hard to classify toxic comments correctly. Sometimes even for humans, one sentence alone is difficult to interpret without context or surrounding text. Misspelled or misused words can create problems for text analysis. Auto-correct and grammar correction applications can handle common mistakes, but don't always understand the writer's intention. With spoken language, mispronunciations, different accents, stutters, etc., can be difficult for a machine to understand.

Synonyms also will cause confusion for us while predicting classification accuracy. some of these words may convey exactly the same meaning, while some may be levels of complexity (small, little, tiny, minute) and different people use synonyms to denote slightly different meanings within their personal vocabulary. So, for building NLP systems, it's important to include all of a word's possible meanings and all possible synonyms.

Thirdly, how to transfer speech content into text also causes troubles for us. To be specific, There are many words that sound similar or identical (like "to", "too","two") but mean very different things (they are homophones). We need to know which word (and meaning) the speaker intends. Also, sometimes people speak too fast (or too slow) — they don't stop or slow down at the end of a sentence before they start a new one. The sentences sound like a continuous long stream of words (it's hard to "hear" the sentence structure from sound alone) and it's unclear when one word ends and another begins. To sum up, machine learning requires a lot of data to function to its outer limits – billions of pieces of training data. The more data NLP models are trained on, the smarter they become. That said, data (and human language!) is only growing by the day, as are new machine learning techniques and custom algorithms. All of the problems above will require more research and new techniques in order to improve on them.

References

Koratana, A., Hu, K. 2018. *Toxic Speech Detection. In Neural Inf. Process. Syst. (p. 9).*

Biere, S., Bhulai, S., Analytics, M. B. 2018. *Hate speech detection using natural language processing techniques. Master Business AnalyticsDepartment of Mathematics Faculty of Science*.