Abusive Chat Detection

Samuel Sears
Summer 2021
https://github.com/ssears219

Which Domain?

This project will deal with the customer service chat domain. As a former customer service representative on both the voice and messaging channels, I am familiar with the different types of abuse encountered. Upset customers can frequently turn to downright offensive or abusive. When it happens, reps are typically responsible for either de-escalating the situation or forwarding the conversation to a manager.

1. https://www.whoson.com/customer-service/when-chatters-attack-dealing-with-abusive-customers/ - When chatters attack: dealing with abusive customer
2. https://www.intercom.com/blog/how-to-cut-the-cord-on-inappropriate-customer-conversations/ - Cutting the cord on inappropriate customer conversations
3. https://www.researchgate.net/publication/326177182_CHAT_APPLICATION_WITH_ABUSIVE_CLASSIFICATION_MODELS_AND_THEIR_COMPARATIVE_STUDY - Chat application with abusive classification models and their comparative study
4. https://www.dongnguyen.nl/publications/vidgen-alw2019.pdf - Challenges and frontiers in abusive content detection
5. https://arxiv.org/pdf/1801.04433.pdf - Detecting offensive language in Tweets using deep learning
6. http://www.eiti.uottawa.ca/~diana/publications/Flame_Final.pdf - Offensive language detection using multi-level classification
7. https://link.springer.com/chapter/10.1007/978-3-319-73706-5_15 - Automatic classification of abusive language and personal attacks in various forms of online communication
8. https://www.frontiersin.org/articles/10.3389/fdata.2019.00008/full - Abusive language detection
9. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data - Toxic comment classification
10. https://arxiv.org/abs/1408.3934 - Detecting messaging abuse in short text messages using linguistic and behavioral patterns

Which Data?

Dataset: https://www.kaggle.com/arashnic/7-nlp-tasks-with-tweets

This Kaggle dataset contains labeled tweets for 7 different NLP tasks; emoji prediction, emotion prediction, hate speech identification, irony detection, offensive chat classification, sentiment scoring, and stance. For this project, I will be using the hate speech and offensive chat datasets.

Each dataset is split into training, testing, and validation sets. Each set is also split between text and labels. For the two datasets I am using, the labels are binary; either they are offensive or hateful, or they are not. The offensive speech training set has over 11,000 tweets and the hate set has over 9,000 tweets.

Research Questions? Benefits? Why analyze these data?

In the chat space, there is a need for automatic detection and handling of abusive chats. Even when a customer is talking to a chat bot, this detection could be utilized. When a chat bot cannot handle a user inquiry, it will typically escalate that chat to a live representative. If the customer is using abusive language, escalating the chat to a representative would want to be avoided. Allowing abuse of customer service representatives effects the work atmosphere poorly and hurts moral. Being able to detect abusive messages prior to escalating the chat could save the organization from putting the customer service representative in a vulnerable position.

What Method?

I will be training a model using the Twitter data to detect offensive language and hate speech. Along with that, I will be adding in another layer to the pipeline to simply determine if the message contains profanity. It will be assumed that offensive messages, hate speech, and profanity are unacceptable customer behaviors that are desired not to be escalated to a chat representative.

Since several of the paper's I have read have found success in using neural networks with LSTM based classifiers, that is the type of model I will be exploring. I will experiment with training hate speech and offensive speech models separately and combined, where the prediction would be whether the message was either hateful or offensive. In the beginning stages of the project, I will also conduct and analysis of variance on unigrams, bigrams, and trigrams to examine what words and phrases seem to be indicators of this sort of abusive language.

Potential Issues?

The main issue I foresee is being able to test and prove the model in it's use within a chat bot setting. The dataset is composed of tweets, not customer service transcripts. Through several hours of searching, I was not able to find a customer service dataset with abusive language labeled. One solution I am thinking through would be to use unlabeled customer service messages, which I can find, and adding in a few examples of abusive language manually. Regardless, since tweets are short texts messages, they do take on a form similar to chat messages. For this solution to be effective, I will need to ensure a very low false positive rate. Blocking customers incorrectly would be unacceptable. Letting a couple abusive customers through would still be an improvement on having no abusive chat detection.

Concluding Remarks

In this project, I will analyze and create an abusive chat detection service which could be utilized along side a chat bot. The service will detect when a customer's chat is abusive and will give the organization an opportunity to prevent an escalation to a vulnerable chat representative. The service will include model(s) trained to detect hate speech and abusive language along with a profanity check. Models will be trained using a large Twitter dataset which contains labels for offensive and hate tweets. Effectiveness of the model will be evaluated using both the Twitter data and customer service data.