Capstone Report

Necky Tran

2023-04-09

Introduction

Sentiment analysis is a natural language processing approach used to identify the emotional tone behind a body of text. It is utilized by organizations to categorize opinions about a product, service or idea. Generally, these tools analyze text data from online sources such as product reviews, blog posts and forum comments to determine whether customers liked or disliked a product.

For my Brainstation Capstone Project, I created sentiment models to identify the emotional tone of Reddit comments. Video game companies such as Valve and Riot constantly release new balance/patch notes for their online competitive video games which are often discussed in their communities' subreddits. Sentiment models can be used here to identify the community's overall reaction to these gameplay changes in order to build a better player experience which would in turn generate more revenue.

Dataset - GoEmotions

The NLP research community has made several open source datasets for the purposes of emotion/sentiment classification. However, they were relatively small and only focused on 6 main emotions. A Google Research team created a large-scale dataset (GoEmotions) that covered a more extensive set of emotions with the hopes that it could be used for a broader scope of future potential applications.

GoEmotions is a dataset of 58k unique Reddit comments extracted from popular English-language subreddits and labeled with 27 emotional categories. The dataset was created from Reddit comments from 2005 to 2019. Curation measures were applied to remove internet bias of extensive offensive language and to limit each comment length to 3-30 tokens. Each comment was scored for emotions by multiple reviewers who were native English speakers from India.

Processing

- Raw dataset 200k, multiple reviewers per comment
- aggregated emotional scorings to score each comment for positive/negative/amb
- end dataset, ggplot of sentiment

Insights

- trained and optimized all classification models
- logreg performed the best based on texting accuracy
- show table of
- also R shiny app

Summary

- model does not perform well for sarcasm, this is a generally problem for a lot of online conversational text
- does not perform as well as a pre-trained model, probably trained on a lot more data
- next steps, probably just use a pre-trained model