Carpe Data

Dan Le

Data Science Capstone

Individual Project Summary

Jun 12, 2023

Abstract

This project was sponsored by Carpe Data, an insurance technology company. The goal of this project was to train a machine learning model to classify businesses with the labels “entertainment” and “traffic” based on factors such as review content and business name. The main achievements from this project was training a KNN model that achieved a max average precision of 96% for classifying traffic and a max average precision of 72% for classifying entertainment using the Universal Sentence Encoder embeddings.   My role in this project involved taking notes for both student and sponsor meetings every week, researching embedding methods for our models, and implementing a decision tree model to classify the businesses.

Overview

Carpe Data provided a dataset of 487,945 different reviews with 17 individual variables. Initially, we thought that our initial question was “Which machine learning model would give us the highest performance metrics?”. However, we eventually realized that it was not the models that were the main problem. It was the inputs, or the embedding methods, used for model training.

Data. Of the 17 variables in the dataset, only 5 were used: the labels for entertainment, labels for traffic, business id, business name, and customer review content. Other variables such as the state, zip code, city, and ratings were not considered for the model training. One of the challenges we came across with the dataset was the imbalance of classes. A business could be traffic only, entertainment only, both entertainment and traffic, or neither. The majority of businesses were labeled as traffic, but there were not many records for entertainment. Predicting entertainment labels was more challenging than predicting traffic labels.

Methodology.

1. Before starting any model training, we explored the variables in the dataset. We looked at the ratings, zip code, and hours of each business.
2. The business name and customer review content were combined together for each business. The combined text was then tokenized into individual words and stemmed. Stop words were also removed from the text.
3. Once the text was preprocessed, it was fed into three different embedding methods: TF-IDF, Sentence Bidirectional Encoder Representations (Sentence-BERT), and Universal Sentence Encoder (USE).
4. For TF-IDF, three dimension reduction techniques were researched: SVD, NMF, and the TF-IDF importance threshold. Ultimately, the TF-IDF importance threshold was chosen because it was the least expensive in terms of computational time. No dimension reduction method was needed for Sentence-BERT or USE.
5. Six different models were trained with the embeddings: KNN, Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and Light GBM.
6. The optimal model was selected using the Precision-Recall curves and average precision metric.

Findings.

1. Our first main finding was the imbalance of classes in the dataset. About 83% of all entries in the dataset were labeled as traffic, but only about 15% were labeled as entertainment. Our initial performance metrics such as accuracy and auc-roc curve were not a good fit for this imbalance because they were biased towards the majority class. The Precision-Recall curve and average precision metric were chosen to account for this class imbalance.
2. Our second main finding was that all the models performed similarly for all embedding methods. There was no significant difference between any of the models. It was hard to find which model performed the best.
3. Our third main finding was that the deep learning embedding methods Sentence-BERT and USE both performed better for predicting entertainment labels. Sentence-BERT and

USE both averaged 70% average precision for entertainment, while TF-IDF averaged 45% average precision. The reason for this increase in performance was because of the behaviors of TF-IDF. TF-IDF does not account for the context a word is used in. It only tracks how many times a word occurs in a document and the whole corpus. For example, consider the sentences “I want to play” and “I am going to see a play”. The word “play” has two different meanings in these sentences, but TF-IDF would give the same embedding for both. Sentence-BERT and USE both account for the context differences by looking at the surrounding words.

Outcomes

1. **Poster Presentation:** https://docs.google.com/presentation/d/e/2PACX-1vQNsqQYxvJL81U2cRlvMmB8zCrPzOqsoq7q7kG4lB26Kz23_4_XHOARzZdKJY4P3fXh2JM2dvGuzSxs/pub?start=false&loop=false&delayms=3000

2. **Github Repository:**

   https://github.com/dantle1/CarpeDataCapstoneProject

3. **Weekly Meeting Slides:** https://drive.google.com/drive/u/1/folders/1-BMM-GdJkozzxkdDddu0KQhqrF4_9B9-

4. **Meeting Notes:** https://drive.google.com/drive/u/1/folders/1B4YRMWoLmfsoVh-MRCidxajrbthvAIUs

Personal Contributions

1. Notetaker

   During student and sponsor meetings, I took notes. I wrote down comments and advice from our sponsors and mentor and any questions my team had for them. I also presented the notes to guide the project.

2. Decision Tree Model

I was responsible for training the decision tree model. I researched the hyperparameter tuning for decision trees, and cross validated my model.

3. Sentence-BERT research

I was responsible for researching BERT embeddings. I learned about how BERT embeddings worked and