Erdos Data Science Bootcamp May 2024 Executive Summary

Team Members: Vinicius Ambrosi, Gilyoung Cheong, Dohoon Kim, & Hannah Lloyd,
Github: https://github.com/dhk628/erdos-companydiscourse

Overview:

In the age of digital communication, a wealth of information exists in the discourse surrounding companies and their products on social media platforms and online forums. This project utilizes natural language processing (NLP) and machine learning (ML) techniques to construct predictive models capable of assessing and rating comments provided by consumers. By employing these advanced analytical methods, we aim to enhance the correctness and effectiveness of sentiment analysis in understanding and forecasting consumer behavior. This approach is computationally efficient, while maintaining contextual integrity in the data and leveraging complex analytical techniques to gauge audience sentiment through online discourse.

Objective:

Our primary goal is to collect Google review data for our target company, Costco, and build ML models with the following key components:

1. **Vector Indexing:** Using Sentence Transformers and the GTE pre-trained model to convert textual data into numerical vectors.
2. **Sampling Techniques:** Employing various sampling techniques to address data imbalance.
3. **Evaluation of ML Models:** Assessing the performance of different sampling techniques and ML models using accuracy and cross-entropy.

Methodology:

Dataset Pre-Processing:

- Collected and pre-processed 788,766 comments from Google Reviews about Costco.
- Vectorized comments into 1024-dimensional vectors using GTE Sentence Transformers.

Dataset Sampling

- Addressed the imbalance in the data, which predominantly consisted of 4 and 5-star ratings, by using both the original sample and implementing random undersampling.

Machine Learning Models:

- Built models using logistic regression, k-nearest neighbor (kNN), support vector machine (SVM), XGBoost, and feedforward neural networks (FNN).
- Trained and validated the models using 80% of the data, with 11-fold cross-validation.

Model Evaluation:

- Evaluated model performance based on accuracy, cross-entropy, and normalized correlation.

Results and Advanced Methods:

Model Comparison (accuracy)

- Logistic Regression: 0.7410
- kNN: 0.7347
- XGBoost: 0.7402
- FNN: 0.7386

Model Comparison (cross-entropy)

- Logistic Regression: 0.6569
- kNN: 0.8564
- XGBoost: 0.6532
- FNN: 0.6515

Sampling Comparison

- Random undersampling led to a decrease in accuracy but produced more balanced confusion matrices, suggesting better performance in predicting less frequent ratings.

Conclusions and Future Directions:

Conclusions:

- This project highlights the potential of advanced NLP and ML techniques in improving sentiment analysis and provides a foundation for future research and application in different domains and datasets.
- Leveraged random undersampling to train models that preserve contextual integrity and computational efficiency.
- Discovered the importance of evaluating model performance beyond common metrics like accuracy, emphasizing the need for a more nuanced approach.

Future Directions:

- Test models on new datasets without ratings (e.g., Reddit) and for a new company. Preliminary results are promising on similarly structured datasets (i.e., reviews from the Costco website)
- Leverage alternative metrics for more rigorous model evaluation.
- Explore alternative text embedding techniques to enhance this approach further.