Unleashing the Power of Text: Developing an NLP Model for Targeted Ad Campaigns

Presented by Masood Dastan

Objectives

Cornerstone Financial Services (CFS), an emerging and versatile financial institution providing a comprehensive range of services, including banking, loans, financial investments, aims to optimize targeted advertising campaigns by delivering personalized ads based on customers online posts.

The main objective is to identify suitable advertisements for individuals, *focusing on distinguishing customers with an interest in investing and stock market from those who may benefit from information about other services such as credit and loan services*.

This project is conducted by **New Edge Solutions** in collaboration with CFS in order to develop an innovative text analysis model using advanced natural language processing techniques.

Data

Two weeks ago, we were approached to implement a model and evaluate its effectiveness as a preliminary step for the project's continuation.

Considering the time limitations and specific objectives of Cornerstone Financial Services (CFS), our team swiftly organized the data acquisition process.

To expedite data collection, we chose to utilize finance-related subreddits on Reddit. Among these subreddits, **Investing** and **Personal Finance** emerged as ideal choices for our task.

Data

Investing: Lose money with Friends!

Investing primarily focuses on discussions related to various investment opportunities, stock markets, companies, and financial trends.

Photo by Daniel Lloyd Blunk-Fernández on Unsplash

Data

Personal Finance:

Personal finance centers around discussions regarding personal financial management, budgeting, debt management, and general financial advice for individuals.

Although both subreddits are finance-oriented, their objectives and areas of emphasis differ.

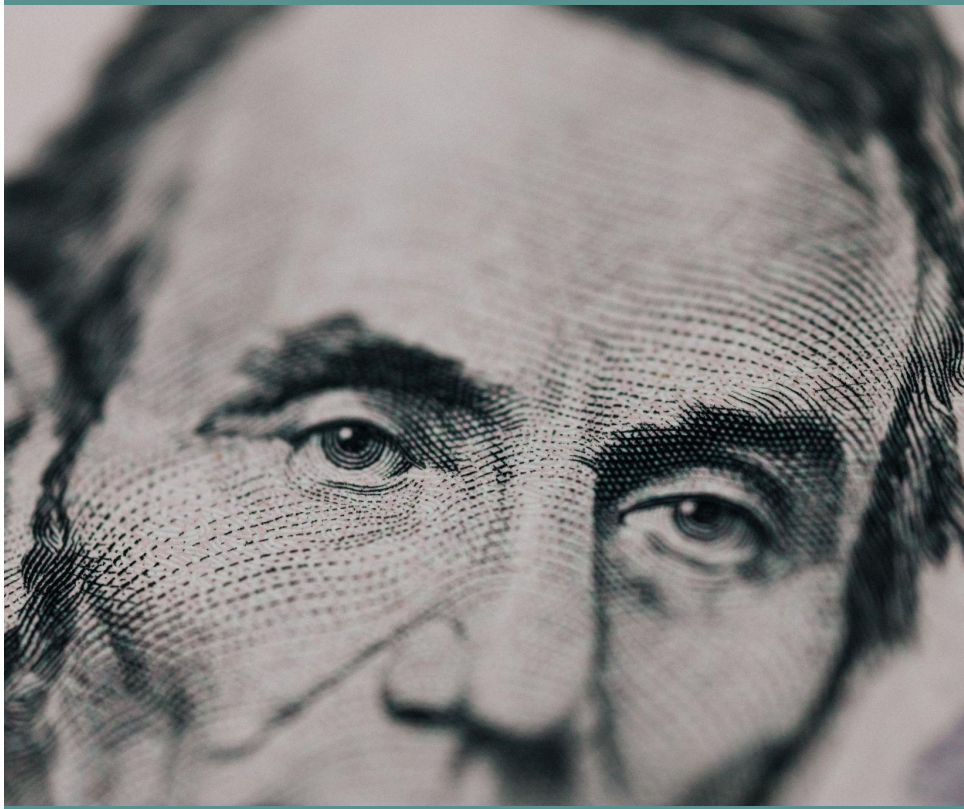
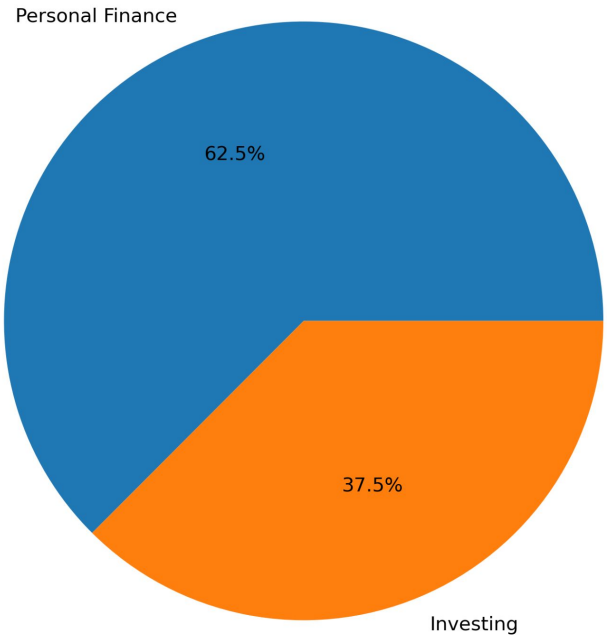


Photo by Karolina Grabowska Pexels

EDA

Around 62.5% of the data comes from Personal Finance and the remaining 37.5 percent belong to Investing.

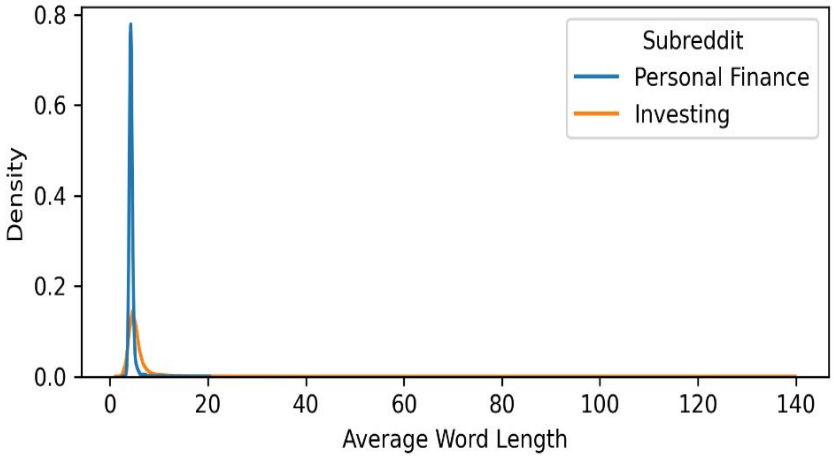


EDA

The average word length in personal finance posts is generally concentrated and tends to be shorter.

In contrast, investing-related posts exhibit a wider range of word lengths, often featuring unusually high word counts.

This disparity can be attributed to the fact that users of Investing are more inclined to include external source links, inflating the overall word length

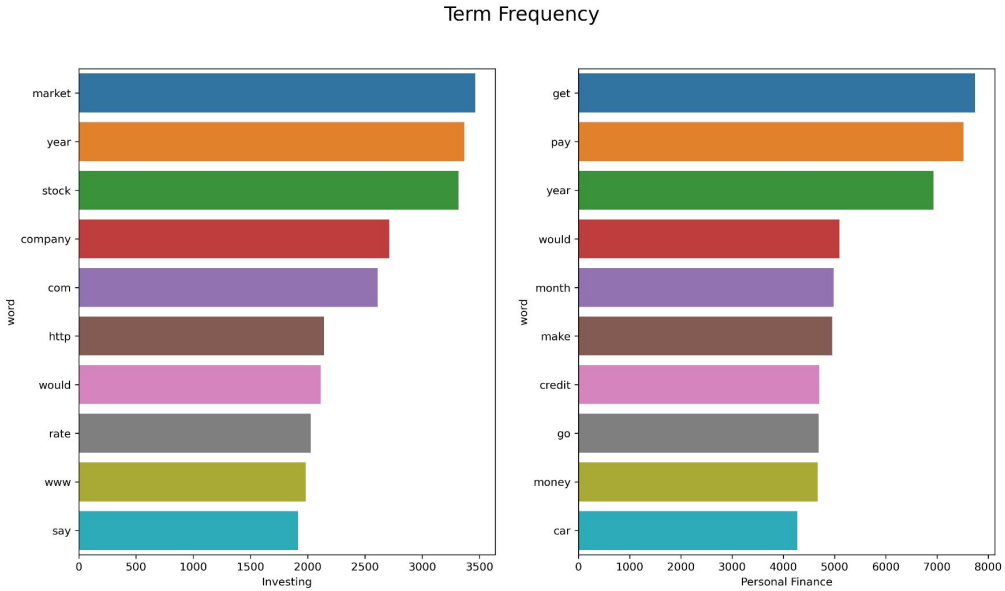


EDA - Term Frequency

There is a stark contrast in vocabulary between the two categories.

- The word *market* appeared approximately 3500 times in the Investing, whereas it did not rank among the top 10 frequently used words in the Personal Finance.

Also, the words used in the Investing tend to be more *specialized and focused on investment-related topics*. On the other hand, the vocabulary in the Personal Finance is *more closely associated with everyday financial matters*.

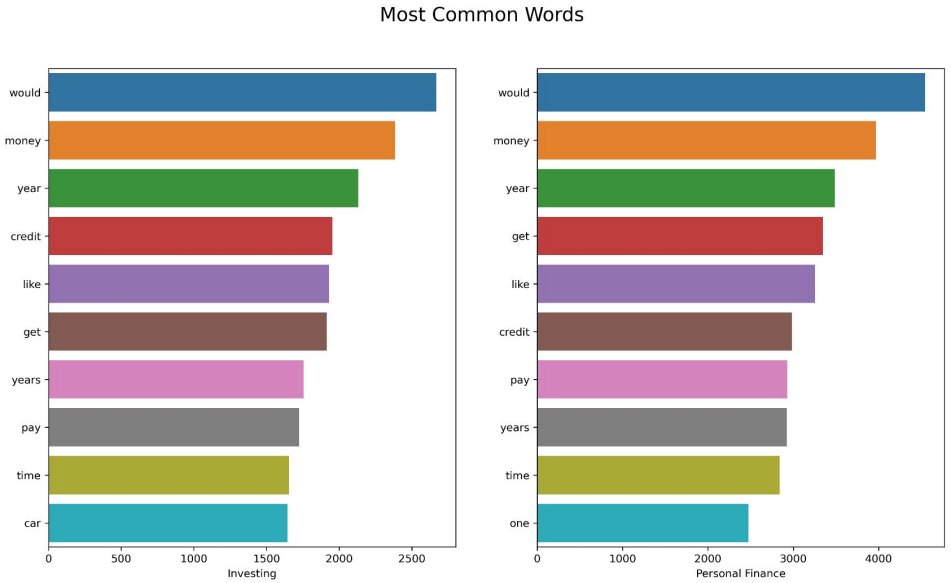


EDA - Most Common Words

In addition to examining the most frequently used words, another common approach is to analyze the repetition of words across the posts.

Interestingly, the common words in the Investing and Personal Finance are almost identical.

This observation holds implications for the development of an optimal text analysis model.



Data Preparation

Data was collected over a span of 6 days and resulting in a dataset containing approximately ***11,825*** unique posts.

Posts with missing content were identified and replaced with an empty string.

All mentions of the subreddit names (*investing and personal finance*) were removed from the dataset to prevent the models from receiving excessive information that could potentially aid their predictions.

The next step involved vectorizing the text inputs using count vectorization and TF-IDF methods. These vectorization techniques allowed us to transform the textual information into numerical features that were compatible with machine learning algorithms.

Modelling

Five classification models, namely *Naïve Bayes, Logistic Regression, Random Forest, Support Vector Machine, and Gradient Boosting*, were evaluated using a *cross-validated randomized search on the vectorized data*. This search aimed to identify the model with the best hyperparameters that yielded optimal performance.

The models' performance was measured based on accuracy and AUC scores.

After comparing the models, the ***Support Vector Classifier with TF-IDF vectorizer*** emerged as the top performer in both accuracy (91%) and AUC score (96%). Therefore, it was selected as the model of choice.

Model Evaluation

In the last step, the selected model was trained once again using the entire available training data.

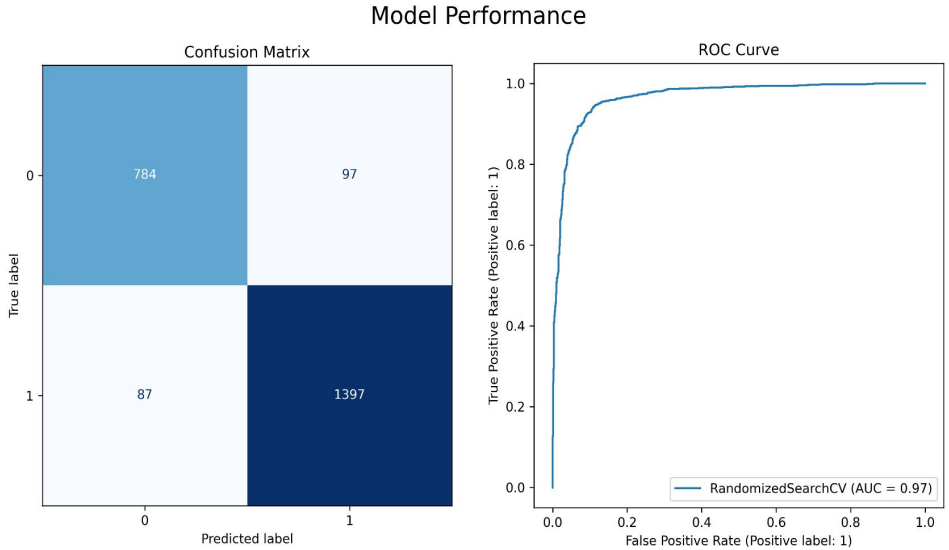
By utilizing the complete dataset, the model can benefit from a wider range of examples, enabling it to capture more intricate patterns and enhance its understanding of the underlying relationships.

We then tested the model against unseen data. This evaluation allows us to make informed decisions about the model's generalizability and its suitability for classifying texts in real life scenarios.

Model Evaluation

The model surpasses the benchmark by around **30 percentage** points. Achieves **97% AUC score**, *92% accuracy, a recall rate of 94%, and specificity rate of 89%.*

The model's performance is highly promising, and there is potential for even greater performance by incorporating a larger and more up-to-date dataset.



Limitations

Unfortunately, the project faced an unexpected challenge due to changes in Reddit's API access rules during the project's initiation. These changes limited our daily access to Reddit data, with constraints such as a maximum of 1000 posts per query and the inability to search across time. These restrictions significantly impacted our data collection process.

In a real-world application where timely text data is available, we anticipate that our model can be further fine-tuned to improve its performance. With access to up-to-date data, we can enhance the model's training and optimize its predictive capabilities, ultimately leading to better results and increased effectiveness in targeted advertising campaigns.

Questions