**INDIVIDUAL REPORT**

Sai Abhishree

G26078396

**Introduction-**

In the intricate landscape of financial markets, sentiment analysis emerges as a pivotal tool, offering insights into market trends and investor attitudes. This project delves into the realm of sentiment analysis within the financial sector, specifically focusing on the Financial PhraseBank dataset. Our objective is to discern the underlying sentiment - positive, negative, or neutral - expressed in financial news and reports, utilizing a blend of traditional machine learning algorithms and cutting-edge transformer models.

The Financial PhraseBank dataset, known for its collection of financial phrases and corresponding sentiment labels, presents a unique opportunity to explore and compare various approaches to sentiment analysis:

1. Naive Bayes: As a foundational model in text classification, Naive Bayes offers simplicity and efficiency. Despite its assumption of feature independence, it stands as a robust baseline model, especially useful for its speed and effectiveness in handling large volumes of text.

2. Logistic Regression: Another cornerstone of traditional machine learning, Logistic Regression, will be employed for its ability to provide probabilistic outputs. Its strength lies in offering interpretable results, making it a valuable tool for understanding which features most influence the sentiment classification.

3. DistilBERT: Representing the cutting-edge in NLP, DistilBERT – a distilled version of the larger BERT model – offers a balance between performance and resource efficiency. By leveraging pre-trained knowledge, DistilBERT can grasp complex language nuances, a crucial aspect in financial contexts where the sentiment can be subtly embedded in the text.

4. RoBERTa: As a robustly optimized version of BERT, RoBERTa fine-tunes the training approach and model architecture to achieve superior performance in text classification tasks. Its application to financial sentiment analysis aims to harness deep contextual representations of language, capturing the intricacies and domain-specific jargon inherent in financial texts.

Throughout this project, we aim to not only implement and compare these diverse methodologies but also to explore the nuances of financial sentiment analysis. The challenges here are unique - financial language is often laden with jargon and can be subtler in expressing sentiment compared to general language usage.

By analyzing the performance, strengths, and limitations of each model, we hope to shed light on the effectiveness of both traditional and state-of-the-art NLP techniques in financial sentiment analysis. This project stands to contribute valuable insights to both the fields of finance and natural language processing, showcasing the practical applications and potential of sentiment analysis in understanding and predicting market dynamics.

**Description of individual work.**

I worked on the classical NLP techniques such as the naïve bayes and logistic regression.

Some of the steps I have taken

1. Preprocessing: Text data is preprocessed to convert it into a suitable format. The text is first tokenized, removal of stopwords (common words with little semantic value), and lemmatization. The goal is to distill the text down to its most informative elements.
2. Vectorization: The preprocessed text is then converted into numerical features through vectorization. I used the count vectorization, which turn text into vectors representing word counts.
3. Model Training and Prediction: The Naive Bayes model is trained on these vectorized features.

Creating the Model: Instantiate a Naïve Bayes model. You can use scikit-learn's naivebayes class.

Training the Model: Fit the model on the training data (vectorized text and corresponding labels).

The same goes for logistic regression.

1. Prediction and evaluation- The accuracy of the model is then tested, and I got 0.84 and 0.94 for the naïve bayes model and logistic regression model respectively.
2. I also worked on sing streamlit to create the user interface for the distilBERT .

**RESULTS-**

A screenshot of a computer

Description automatically generated

A chart with different colors

Description automatically generated with medium confidence

FOR NAÏVE BAYES

A screenshot of a computer

Description automatically generated

A chart with different colored squares

Description automatically generated

FOR LOGISTIC REGRESSION

**Summary/Major Takeaways-**

* Different algorithms have varying levels of effectiveness for specific tasks. Naive Bayes and Logistic Regression, while simpler, can be surprisingly effective for text classification.
* The significance of data preprocessing in NLP is paramount. Effective tokenization, cleaning, and vectorization of text data directly influence the performance of machine learning models.
* Understood that real world data can be unbalanced sometimes and we have to use the right techniques to fix that, such as the Stratified K-Fold Cross-Validation.
* Deploying models via a Streamlit app demonstrated the importance of scalability and user accessibility.
* The field of NLP is rapidly evolving, especially with advancements in transformer models. Continuous learning and adaptation are crucial for staying relevant.

**Conclusion-**

This project not only provided practical insights into the application of various machine learning models for sentiment analysis but also emphasized broader lessons in NLP.

The accuracies of the models is 0.84 and 0.94 for the naïve bayes model and logistic regression model respectively.