Final Year Project Proposal 2023

**"Stock Sentiment Analysis with LLm Models”**

**Group Members**

Arsalan Bakhtiar : 20jzbcs0117

Afzaal Ahmad : 20jzbcs0119

**Supervisor**

Dr. Laeeq Ahmad

Table of Contents

[1. Introduction 4](#_Toc151365028)

[2. Motivation 4](#_Toc151365029)

[3. Methodology 5](#_Toc151365030)

[3.1 Data Collection 5](#_Toc151365031)

[3.2 Data Preprocessing 5](#_Toc151365032)

[3.3 Sentiment Analysis: 5](#_Toc151365033)

[3.4 Evaluation Metrics: 5](#_Toc151365034)

[3.5 Real-world Testing: 5](#_Toc151365035)

[4. Dataset’s/Other Resources 6](#_Toc151365036)

[4.1 Data Sources 6](#_Toc151365037)

[4.2 Data Preprocessing 6](#_Toc151365038)

[4.3 Feature Engineering for Stock Sentiment Analysis Using LLm Model 6](#_Toc151365039)

[4.4 External Libraries and Tools Used in This Project 7](#_Toc151365040)

[5. Evaluation Plan 7](#_Toc151365041)

[5.1 Data Splitting 7](#_Toc151365042)

[5.2 Performance Metrics 7](#_Toc151365043)

[5.3 Cross Validation 7](#_Toc151365044)

[5.5 Ensemble Learning Evaluation 7](#_Toc151365045)

[5.6 Real-world Testing 8](#_Toc151365046)

[5.7 Error Analysis 8](#_Toc151365047)

[6. Tools 8](#_Toc151365048)

[6.0 Python 8](#_Toc151365049)

[6.1 TensorFlow and PyTorch: 8](#_Toc151365050)

[6.2 Natural Language Toolkit (NLTK) 8](#_Toc151365051)

[6.3 Scikitlearn 8](#_Toc151365052)

[6.4 Pandas 8](#_Toc151365053)

[6.5 Matplotlib and Seaborn 8](#_Toc151365054)

[6.6 Jupyter Notebooks 9](#_Toc151365055)

[6.7 Spyder Integrated Development Environment (IDE) 9](#_Toc151365056)

[7. Conclusion 9](#_Toc151365057)

[8. References 10](#_Toc151365058)

# 1. Introduction

In today's online world, people share a lot of thoughts and feelings. We want to understand how they feel about stocks, but we're doing it in a new way. Instead of using complicated computer programs, we're using something called a Large Language Model (LLM).

Think of the LLM as having a smart friend who knows a lot about words. This friend helps us figure out what people are saying about the stock market. We're not using fancy machine learning this time—just the simple and powerful LLM.

Our project is all about gathering information from different places like the stock market, Twitter, financial news websites, and social media. It's like putting together puzzle pieces from various sources to get the full picture.

In our project, we'll explain everything, share what we find, and talk about how our simple and smart ideas can be useful in real life. We're excited about this and hope it leads to more interesting discoveries in the future.

# 2. Motivation

We're aware of the complexities of the stock market, often resembling a challenging puzzle. Determining whether stocks will rise, or fall can feel like interpreting cryptic messages. Recognizing the difficulty people face in making sense of this, we're introducing a specialized tool. This tool aims to decode the stock market using information gleaned from news, social media, and more.

Picture it as having an intelligent companion who simplifies the stock market for you. This friend utilizes technology to comprehend the meaning behind all the words, providing insights into whether stocks are likely to go up or down. Our goal is to leverage this technology, specifically the capabilities of Large Language Models (LLMs). We aim to assist you and others in making informed decisions in the financial realm.

Believing that understanding words and data can demystify the stock market, we aspire to empower everyone to make smarter choices. Our tool acts as a knowledgeable friend, analyzing vast amounts of information and offering prompt observations. It succinctly communicates, "Based on the data I've examined, it appears that these stocks might decline soon." This way, you can make quicker and more informed decisions without the need for extensive research.

# 3. Methodology

Our methodology integrates Large Language Model (LLM) techniques to conduct stock sentiment analysis, steering away from traditional Natural Language Processing (NLP) and machine learning models. While NLP and machine learning have their merits, the LLM model provides a more streamlined and effective approach to our analysis. Here's a breakdown of our comprehensive methodology:

## 3.1 Data Collection

We utilize LLM techniques for data collection, focusing on gathering information from various sources, including financial news articles, social media posts, and historical stock data. This ensures a diverse and comprehensive data set for our analysis.

## 3.2 Data Preprocessing

Employing LLM-based preprocessing methods, we ensure the cleanliness and consistency of the data. This involves tasks such as text cleaning, stop word removal, and stemming or lemmatization to prepare the textual data for analysis.

## 3.3 Sentiment Analysis:

The core of our methodology lies in sentiment analysis using the LLM model. By leveraging the inherent language understanding capabilities of the model, we aim to provide nuanced insights into stock market sentiment without the need for traditional machine learning algorithms.

## 3.4 Evaluation Metrics:

Our evaluation metrics focus on the performance of the LLM model in sentiment analysis. Metrics such as accuracy, precision, recall, and F1 score will be used to gauge the effectiveness of the model in providing accurate sentiment predictions.

## 3.5 Real-world Testing:

To ensure the practical applicability of our LLM-based sentiment analysis, we conduct real-world testing scenarios. This involves applying the model to actual market data and assessing its performance in predicting sentiment in real-time.

By incorporating the LLM model into every step of our methodology, we aim to streamline and enhance the process of stock sentiment analysis, providing more accurate and nuanced insights into market trends.

# 4. Dataset’s/Other Resources

## 4.1 Data Sources

We compile a diverse range of data sources to create a comprehensive dataset for our analysis:

Financial News Articles: We gather textual data from reputable financial news websites, capturing insights into market trends and sentiment from financial experts.

Social Media Posts: Social media platforms, particularly Twitter, provide a rich source of real-time data, reflecting public sentiment and opinions related to stocks.

Stock Market Data: Historical stock data, including price movements and trading volumes, is essential for correlating market performance with sentiment.

Moreover, we understand the importance of diverse data sources. Instead of relying on a single medium like twitter 〚 1 〛, we draw insights from multiple channels. This approach ensures that our analysis is not limited to one perspective, increasing the richness and robustness of our predictions.

## 4.2 Data Preprocessing

In the landscape of language models, effective data preprocessing is pivotal for the success of LLm models. These Large Language Models, renowned for their prowess in natural language understanding and generation, require meticulous preparation. Cleaning the dataset, tokenization, and removing stop words are foundational steps. The adept handling of missing values and exploration of advanced techniques contribute to the model's optimal performance. In essence, mastering the intricacies of data preprocessing is the key to unlocking the full potential of LLm models.

## 4.3 Feature Engineering for Stock Sentiment Analysis Using LLm Model

In the context of stock sentiment analysis, efficient feature engineering is pivotal when employing LLm models. This process involves crafting meaningful input variables, such as sentiment scores from news articles and social media trends, to enhance the model's ability to discern nuanced market sentiment. These carefully curated features empower LLm models to provide more accurate predictions, bridging the gap between textual information and stock market dynamics with precision.

## 4.4 External Libraries and Tools Used in This Project

In this project focused on LLm model development for stock sentiment analysis, several external libraries and tools play a crucial role in enhancing efficiency and functionality. Notable libraries include TensorFlow and PyTorch for building and training the LLm model, NLTK for natural language processing tasks, and Scikit-learn for feature engineering and data preprocessing. Additionally, tools like Jupyter Notebooks facilitate interactive and collaborative development, while Pandas and NumPy streamline data manipulation. The integration of these external resources ensures a robust and effective implementation of the LLm model for accurate stock sentiment analysis.

## 5. Evaluation Plan

Our evaluation plan is a critical component of our stock sentiment analysis project, outlining how we intend to assess the performance and effectiveness of our methodologies and models. Here's a comprehensive overview:

## 5.1 Data Splitting

We will divide our dataset into training and testing sets to ensure a fair evaluation. A typical split might involve using 70% of the data for training and reserving the remaining 30% for testing.

## 5.2 Performance Metrics

To gauge the performance of our sentiment analysis models, we will employ a range of metrics, including:

Accuracy: The ratio of correctly predicted sentiments to the total predictions.

Precision: The percentage of true positive predictions out of all positive predictions.

Recall: The percentage of true positive predictions out of all actual positive cases.

F1Score: A harmonic mean of precision and recall, providing a balanced measure of model performance.

## 5.3 Cross Validation

We will use cross validation techniques, such as k-fold cross validation, to ensure that our model's performance is consistent across different subsets of the data. This reduces the risk of overfitting.

## 5.5 Ensemble Learning Evaluation

For the ensemble learning approach, we will assess the effectiveness of combining individual models to enhance sentiment analysis accuracy. We will consider metrics specific to ensemble methods, such as ensemble accuracy and diversity among individual models.

## 5.6 Real-world Testing

To evaluate the practical application of our sentiment analysis, we will conduct real-world testing. This involves using the models and methodologies in real market scenarios to make predictions and assess their reliability in an operational environment.

## 5.7 Error Analysis

In-depth error analysis will be carried out to understand the types of misclassifications made by the models. This analysis will help identify areas for improvement and refinement.

# 6. Tools

## 6.0 Python

Python is our primary programming language, known for its versatility in data analysis and machine learning. We leverage Python for coding, data manipulation, and model development.

## 6.1 TensorFlow and PyTorch:

These deep learning frameworks are instrumental in building and training the LLm model, providing a foundation for intricate language understanding.

## 6.2 Natural Language Toolkit (NLTK)

NLTK is an invaluable library for natural language processing. It provides a wide array of tools and resources for text analysis, including stop word removal, stemming, lemmatization, and sentiment analysis.

## 6.3 Scikitlearn

Scikitlearn is a comprehensive machine learning library in Python. It offers various tools for classification, regression, clustering, and model evaluation, making it an essential part of our model development process.

## 6.4 Pandas

Pandas is a robust library for data manipulation and analysis. It simplifies the handling of large datasets, enabling efficient data preprocessing and feature engineering.

## 6.5 Matplotlib and Seaborn

Matplotlib and Seaborn are libraries for data visualization. These tools are crucial for creating visual representations of our data, facilitating insights and understanding.

## 6.6 Jupyter Notebooks

Jupyter Notebooks provide an interactive and collaborative coding environment. They enhance our research process, making it efficient and transparent.

## 6.7 Spyder Integrated Development Environment (IDE)

Spyder IDE is another powerful tool for coding, data analysis, and model development. It offers features that enhance productivity and code management.

These tools collectively support our research from data collection and preprocessing to machine learning model development and performance evaluation. Their integration ensures the efficiency and effectiveness of our stock sentiment analysis project.

# 7. Conclusion

In our efforts to understand stock sentiment, we've used the power of technology and data. Instead of relying on just one model, we've teamed up multiple models to improve our accuracy. This teamwork has made our predictions more reliable for investors and traders.

As we finish this part of our journey, we know that finance is always changing. We're staying committed to innovation, ready to make our methods even better. Our goal is to help people make smarter investment decisions by combining technology and different models. The future holds the promise of even more precise insights, making the financial world more accessible to everyone.

# 8. References

1. Johnson, L., "Twitter Data Mining for Sentiment Analysis," Proceedings of the International Conference on Natural Language Processing, 2019.

2. Brownlee, J. (2021). "How to Develop a Deep Learning Bag-of-Words Model for Sentiment Analysis (Text Classification)." Machine Learning Mastery. [[Link]](https://machinelearningmastery.com/deep-learning-bag-of-words-model-sentiment-analysis/)

3. Rouse, M. (2021). "Natural Language Processing (NLP)." SearchEnterpriseAI. [[Link]](https://searchenterpriseai.techtarget.com/definition/Natural-Language-Processing-NLP)

4. Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 12, 2825–2830. [[Link]](https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf)

5. Chollet, F., et al. (2015). "Keras." GitHub Repository. [[Link]](https://github.com/keras-team/keras)

6. PyTorch Documentation. (2021). [[Link]](https://pytorch.org/docs/stable/index.html)

7. Bird, S., Klein, E., & Loper, E. (2009). "Natural Language Processing with Python." O'Reilly Media. [[Link]](https://www.nltk.org/book/)