Title: Analyzing the Impact of Reddit Post Sentiment on Stock Market Trends

Steven Clark

Project Goal:

My goal is to investigate the influence that sentiment expressed in Reddit posts has on stock market trends. This project is important because it will help me understand the growing impact of online discussions on financial markets, as evidenced by recent events surrounding meme stocks and trading communities such as r/wallstreetbets. Gaining insights into the relationship between online sentiment and stock market trends can help investors, regulators, and financial institutions make better-informed decisions.

Data Sources:

The primary data for this project consists of Reddit posts and comments from the following relevant subreddits: r/wallstreetbets, r/investing, and r/stocks. Reddit's API provides access to post and comment data, including timestamps, content, and user engagement metrics. Additionally, I will source stock market data from a dataset found on Kaggle.

Basic Aspects of the Data:

The Reddit data I will work with consists of posts and comments from the aforementioned subreddits, including their titles, text content, author information, timestamp, upvotes, downvotes, and comment counts. The stock market data will include historical stock prices and trading volumes.

Data Selection:

To address my project goal, I will use the following aspects of the data:

Reddit posts and comments: Text content of posts and comments for sentiment, while using metadata such as timestamps and upvotes to measure user engagement and post popularity.

Stock market data: Company ticker name, daily opening/closing price, daily high/low price, and trading volumes associated with their respective dates.

> STOCK DATA **(***From Kaggle)*
>
> ***"Context:*** *High-quality financial data is expensive to acquire and is therefore rarely shared for free. Here I provide the full historical daily price and volume data for all US-based stocks and ETFs trading on the NYSE, NASDAQ, and NYSE MKT. It's one of the best datasets of its kind you can obtain.*
> ***Content:*** *The data (last updated 11/10/2017) is presented in CSV format as follows: Date, Open, High, Low, Close, Volume, OpenInt. Note that prices have been adjusted for dividends and splits."*

Data Processing Steps:

I will undergo the following data processing steps to extract results:

1. **Combine multiple datasets:** I will merge Reddit post data with stock market data based on timestamps to associate posts and comments with relevant stock market events.

2. **Clean data by recoding variables:** I will recode variables in the datasets, such as converting timestamps to a common format and normalizing stock prices for easier comparison.

3. **Use the Reddit API:** I will utilize the PRAW Reddit API to fetch post and comment data from relevant subreddits and gather stock market data using financial data APIs.

4. **Read in data from JSON format:** I will process and parse JSON data obtained from the APIs to extract relevant information.

5. **Reformat dates:** I will ensure that date formats in both Reddit and stock market data are consistent, enabling accurate merging of datasets based on time.

6. **Use regular expressions:** I will apply regular expressions (REGEX) to extract relevant information from post and comment text, such as stock tickers, and filter out unrelated content.

7. **Sentiment Analysis:** I will follow these steps to analyze the sentiment of Reddit posts and comments related to stocks and trading:

    Text preprocessing: To prepare the text data for sentiment analysis, I will perform several preprocessing tasks, including:

    a. Lowercasing: Convert all text to lowercase to ensure consistency and facilitate further processing.

    b. Tokenization: Break the text into individual words (tokens) to enable the analysis of each word separately.

    c. Stopword removal: Remove common words (e.g., "the", "and", "in") that do not contribute to the overall sentiment.

    d. Lemmatization: Reduce words to their base form (lemma) to combine similar words and reduce dimensionality.

    Feature extraction: To represent the text data in a format suitable for machine learning models, I will extract relevant features, such as:

    a. Term Frequency-Inverse Document Frequency (TF-IDF): Calculate the TF-IDF scores for each token in the dataset to weigh the importance of words based on their frequency within a document and across the entire corpus.

    b. Word embeddings: Generate word embeddings using distilBERT, a lightweight version of the pre-trained model BERT to represent words as dense vectors capturing semantic information.

    Sentiment classification: To determine the sentiment of each Reddit post and comment, I will utilize the pre-trained sentiment analysis model VADER (Valence Aware Dictionary and sEntiment Reasoner. Vader is suitable for the following reasons:

    a. Specifically designed for social media text, including slang, acronyms, and emoticons.

    b. Considers intensity of sentiment by using a valence-based approach, which can provide more nuanced sentiment scores.

c. Does not require training, as it is based on a pre-built lexicon and rule-based sentiment analysis.

Aggregating sentiment scores: Once the sentiment of each post and comment is determined, I will aggregate the sentiment scores at the following levels:

a. Per post: Calculate the average sentiment score for each post by averaging the scores of all associated comments.

b. Per stock ticker: Group sentiment scores by stock ticker to determine the overall sentiment for each stock.

c. Per time period: Aggregate sentiment scores over specific time periods (e.g., daily, weekly, or monthly) to analyze trends and correlations with stock market data.

Time series analysis: I will conduct multivariate time series analysis on stock market data and sentiment scores to identify potential causal relationships, correlations, or dependencies.

Data visualization: I will generate visualizations to illustrate the relationship between Reddit post sentiment and stock market trends, such as line charts, scatter plots, and correlation heatmaps. Tools I plan to use include Tableau, Gephi, python (matplotlib), and/or Excel.

Through this project, I aim to provide valuable insights into the relationship between online discussions and stock market trends, which can aid decision-making processes for various stakeholders in the financial industry.