Group 825: Stock Market Sentiment Analysis

First Name	Last Name	Email address
Akshatha	Ganji	aganji1@hawk.iit.edu
Joel	Miranda	jmiranda11@hawk.iit.edu
Sree Ram Charan	Pammi	spammi@hawk.iit.edu
Varshitha	Chikkegowdanadoddi Somashekar	vchikkegowdanadoddis@hawk.iit.edu

Table of Contents

1. Introduction

On social media, information about public feelings has become abundant. Social media is transforming like a perfect platform to share public emotions about any topic and has a significant impact on overall public opinion. Twitter, a social media platform, has received a lot of attention from researchers in recent times. Twitter is a micro-blogging application that allows users to follow and comment on other users' thoughts or share their opinions in real time.

More than a million users post over 140 million tweets every day. This situation makes Twitter like a corpus with valuable data for researchers. Each tweet is 140 characters long and speaks public opinion on a topic concisely. The information exploited from tweets are very useful for making predictions. Sentiment analysis of twitter data and sentiment classification is the task of judging opinion in a piece of text as positive, negative, or neutral. In this project a method for predicting stock prices is developed using Twitter tweets about various companies. Sentiment analysis of the collected tweets is used as a prediction model for finding and analyzing correlation between contents of news articles and stock prices and then making predictions for future prices will be developed by using machine learning.
Sentiment analysis of "Tweets about the Top NASDAQ Companies from 2015 to 2020" dataset from Kaggle and examining their effects on stock market changes. Even headlines from different news sites can also be included. Awareness and click generation are important roles for business news headlines as well. The effect of news and tweets on the sentiment of the stock market can be understood. Also, comparison of a few Natural Language Processing (NLP) algorithms will be considered.

2. Data

Tweets about the Top NASDAQ Companies from 2000 to 2016
https://www.kaggle.com/code/tommyupton/twitter-stock-market-sentiment-analysis/input

In the analysis of stock market sentiment sourced from Yahoo Finance https://finance.yahoo.com/quote/%5EIXIC?p=^IXIC&.tsrc=fin-srch, labels were assigned to indicate market conditions, with '0' representing instances when the market was deemed low and '1' indicating periods of high market activity, as determined based on specific criteria for each date.

This dataset, as a part of the paper published in the 2020 IEEE International Conference on Big Data under the 6th Special Session on Intelligent Data Mining track, is created to determine possible speculators and influencers in a stock market. Although we used both tweet data and companies' market data in our project, we thought that it is a better choice to split our datasets into two parts while sharing in Kaggle. This dataset is helpful for those interested in tweets that are written about Amazon, Apple, Google, Microsoft, and Tesla by using their appropriate share tickers.

This dataset contains over 3 million unique tweets with their information such as tweet id, author of the tweet, postdate, the text body of the tweet, and the number of comments, likes, and retweets of tweets matched with the related company.Tweets are collected from Twitter by a parsing script that is based on Selenium.

3. Problems and Solutions

3.1 Problems

Stock exchange is a subject that is highly affected by economic, social, and political factors.
There are several factors e.g. external factors or internal factors which can affect and move the stock market. Stock prices rise and fall every second due to variations in supply and demand. Various Data mining techniques are frequently involved to solve this problem. But it has been observed it takes longer time, not accurate and complex routines involved in understanding the sentiment of stock market. Predicting the sentiment of the market can give lots of input in predicting stock prices and do investment accordingly.

3.2 Solutions

Techniques using machine learning will give a more accurate, precise and simple way to solve such issues related to stock and market sentiment and prices. Positive news and tweets in social media about a company would encourage people to invest in the stocks of that company and as a result the stock price of that company would increase. A prediction model for finding and analyzing correlation between contents of tweets and stock prices and then making predictions for future prices can be developed by using machine learning.

This project concentrates on using Python packages such as NLTK (Natural Language Toolkit) , pandas, numpy, sklearn, matplotlib and tensorflow.

4. KDD

4.1. Data Processing

These are the steps for data preprocessing which are performed.

- Utilized a custom function (**remove_urls**) to eliminate URLs from the 'body' column of the tweet dataset.
- Converted the 'timestamp' column to a datetime format ('new_date') for better readability.
- Sorted the dataset based on key columns such as 'tweet_id,' 'writer,' 'post_date,' 'body,' 'comment_num,' 'retweet_num,' 'like_num,' and 'new_date.'
- Applied grouping by date and selected the top 25 tweets per date using the 'head' function.
- Saved the resulting dataset, focusing on the top 25 tweets per date, to a new CSV file named 'preprocess.csv.'
- Aggregated tweets by date and transposed the data, transforming it from a vertical to a horizontal format.
- Saved the transposed dataset, capturing aggregated tweets per date, to 'preprocess.csv' for subsequent analysis. (as the data set was 104k rows it was done in chunks and finally, merged all the preprocessed files to a single file named ‘Data.csv’.

4.2. Data Mining Methods and Processes

1. **Imbalance Issue identification & Oversampling:**

   Examined the distribution of labels in the dataset and identified an evident imbalance between classes. Recognized the potential impact of imbalance on machine learning model performance. Utilized the RandomOverSampler from the imblearn library. Oversampled the minority class to achieve a more balanced distribution.

2. **Dataset Splitting:**

   The initial step involves dividing the dataset into two subsets: the training set and the test set. This is typically done based on a temporal criterion, where data before a certain date is assigned to the training set, and data after that date is used for testing.

3. **Text Preprocessing:**

   Punctuation Removal: Special characters and punctuation in the text data are removed to ensure a cleaner dataset.

Punctuation Removal: Special characters and punctuation in the text data are removed to ensure a cleaner dataset.

4. **Text Normalization:**

   Convert to Lowercase: All text data, specifically headlines in this case, is converted to lowercase. This normalization step ensures consistency and uniformity in the text.

5. **Text Transformation:**

   Concatenation: The processed text data is then concatenated to form cohesive headlines, facilitating further analysis.

6. **Feature Extraction - Bag of Words:**

   The Bag of Words model is implemented using the CountVectorizer from scikit-learn.

   N-gram Range: The model is configured to consider bigrams (2,2), capturing pairs of adjacent words.

   Transformation: The model is applied to the preprocessed headlines, resulting in a matrix representation of the text data.

7. **The algorithms that we used in our analysis are:**

   Fine-tuned algorithmic parameters across various Data Mining (DM) and Machine Learning (ML) models, executing multiple iterations to identify the optimal configuration for enhanced performance.

Multinomial Naive Bayes Classifier:

Utilizes the Bayes theorem with an assumption of independence among features to classify based on the probability of a feature's occurrence given a class.

Logistic Regression:

An algorithm based on regression that estimates the likelihood that an instance belongs to a specific class is used for binary classification.

K-Nearest Neighbor Classifier:

Classifies instances according to the most common class among its k-nearest neighbors in the feature space, as decided by the majority vote of its neighbors.

Support Vector Machine Classifier:

The algorithm searches for the best boundary between classes by building a hyperplane or collection of hyperplanes in a high-dimensional space.

Random forest Classifier:

Using an ensemble learning technique, multiple decision trees are built during training, and the output is either the class mode for classification tasks or the average prediction for regression tasks.

8.Visualization:

Incorporated word clouds as a visualization technique to gain insights into the sentiment expressed in the news headlines related to stock movements.

Text Preprocessing:

- Tokenizing each news headline by words.
- Removing common English stop words to focus on more meaningful terms.
- Stemming the words to capture their root form.

Building the Corpus:

- Constructing a corpus of processed news headlines.

Word Clouds for Different Sentiments:

- Segregating headlines based on the stock's upward or downward movement.
- Generating separate word clouds for instance indicating a fall in stock prices and those indicating a rise.

Visual Representation:

- Utilizing the Word Cloud library to create visually appealing representations.
- Displaying the word clouds to highlight the most frequent and impactful words associated with each sentiment.

5. Evaluations and Results

5.1. Evaluation Methods

To assess the efficacy of the final model, we employed three principal metrics: accuracy, AUC curve and F1 score. These metrics were chosen due to their widespread acceptance and reliability in evaluating the performance of classification models.

The dataset spans from the year 2000 to 2016. For the purpose of training, we utilized data ranging from 2000 to 2015. The test set comprised data subsequent to December 31, 2014, ensuring a robust evaluation of the model’s predictive capabilities on unseen data. By thoroughly evaluating each classifier manually, the hyperparameters were chosen with the goal of maximizing the model's performance on the training set while conserving test set generalization.

The hyperparameter is structured with a concentration on key terminology and a clear description of the procedures used for processing data and model evaluation, ensuring formal reporting and clarity.

5.2. Results and Findings

The Logistic Regression achieved highest accuracy, AUC score and F-1 score of 86% among others. KNN consistently shows lower accuracy, AUC, and F1-score compared to other classifiers.

6. Conclusions and Future Work

6.1. Conclusions

With the use of several machine learning models, this research attempted to do sentiment analysis or categorization on textual input. To anticipate attitudes or classify text, the project required preprocessing data, training models, and evaluation. By counting the number of true positives, true negatives, false positives, and false negatives, the confusion matrix can be examined to get insight into the prediction performance of the model.

Additionally, word clouds were employed to do sentiment analysis. These clouds provide visual insights into the words associated with increasing and dropping stock labels, which can help decipher sentiment patterns linked to stock movements. A basis for comprehending and evaluating sentiment patterns and classifier performance in a stock-related context is provided by the evaluation metrics and visuals produced by this analysis.

6.2. Limitations:

External Factors: The model's performance may be influenced by evolving trends, thereby necessitating periodic retraining and updates to ensure its relevance over time.

Data Limitations: Potential inaccuracies in the dataset could arise if the data source lacks transparency regarding data collection methods.

Data Annotation Costs and Time Constraints: It frequently takes a lot of time and resources to annotate or label textual material. With limited resources, the amount or quality of labeled data available for training may be impacted by restrictions on the scope or depth of data annotation.

Non-stationary Nature of Financial Data: Financial data often exhibits non-stationary behavior, meaning that statistical properties such as mean, and variance can change over time. ML models, especially those that assume stationarity, may struggle to adapt to these changing conditions. **Data Quality and Feature Selection:** The quality of predictions heavily depends on the quality of input data. Noisy or unreliable data can lead to inaccurate predictions. Additionally, selecting relevant features is crucial, and if important variables are omitted or the wrong features are chosen, the model's performance may suffer.

Black Swan Events: ML models are typically trained on historical data, and they might not be well-equipped to handle unforeseen events or "black swan" events—unexpected and extreme occurrences that can significantly impact financial markets but were not present in the training data.

6.3. Potential Improvements or Future Work

1. Feature Engineering:

Experiment with different financial indicators, technical indicators, and alternative data sources. Consider creating custom features that capture specific market dynamics. Use domain knowledge to identify relevant features that may contribute to better predictions.

2.Validation and Testing:

Establish a monitoring system to detect model degradation or shifts in performance. Schedule regular model reviews and updates to incorporate new information and improve predictions.

3.Temporal Aspects and Non-stationarity:

Implement time-series analysis techniques to capture temporal dependencies in the data. Consider models designed specifically for handling non-stationary time series data. Regularly update models to adapt to changing market conditions.

4.Explainability and Interpretability:

Choose models that offer interpretability, such as decision trees or rule-based models. Use techniques like SHAP (SHapley Additive Explanations) to understand feature contributions. Ensure that stakeholders can comprehend and trust the model's predictions.