P1 – Project Statement

1. Project title

Sentiment Analysis of the Apple Company using Tweets

2. Names of the member(s) of the group

Jiaye Tang, Xi Rao, Kris Lin

3. Problem Statement

Stock price analysis is an important topic in the financial industry and the reasons that cause the changes in stock price can be various. According to the Efficient Market Hypothesis (EMH), it states that stock market prices are largely driven by new information and follow a random walk pattern [1]. Therefore, our project will explore how different sentiments on social media affect stock price change. To specify our research, the project will only investigate the stock price change of Apple Company, which is one of the biggest technology companies in the world. Analyzing the impact of sentiments on Apple Company's stock price will help companies gain insights from social media platforms and find out if public opinions will cause the stock price to change or be affected by it.

In our project, we will apply classic natural language processing methods and dictionaries to investigate the following questions, including the impact of different sentiments on the 5-year change of stock prices of Apple company, different sentiments distribution of the year/month at the highest/lowest stock price, as well as the stock price at the highest and lowest sentiment point. Moreover, each year Apple hosts a new generation iPhone event in September or October, we will especially analyze sentiments and the stock price change during those special periods. Furthermore, we will explore the correlation between stock prices and different sentiments using a linear regression model.

4. Possible Approaches

Our datasets, which are Tweet.csv, Company_Tweet.csv, and Company.csv, include tweets that are written about Amazon, Apple, Google, Microsoft, and Tesla by using their appropriate share tickers from 2015 to 2019. It contains over 3 million unique tweets with their information such as tweet id, author of the tweet, post-date, the text body of the tweets matched with the related company. We will select tweets for Apple to do the sentiment analysis. Additionally, another dataset, CompanyValues.csv, contains the daily stock price of Apple from 2015 to 2019.

a. Preprocessing

Before the sentiment analysis, we will first preprocess the tweets data. For example: remove tweets handles, remove special characters, remove stop words, and remove URLs and stemming.

Since there are three datasets: Company_Tweet.csv, Tweet.csv and company.csv. Both the company.csv and tweet.csv have the same feature tweet_id, then we could merge these two datasets via tweet_id. For the stock price dataset and tweets dataset, we will use ticket_symbol in

and day_date in company value dataset and the corresponding ticket_symbol in tweets dataset to join them for future analysis and correlation.

NLTK, an open-source library for natural language processing will be used to deal with the preprocessing. We find the body of the tweets are string type and contain tweet handles, URLs, punctuations etc. Then the natural language preprocessing steps are [4]:

Remove punctuations like #, @,

Remove URLs

Remove stop words

Lower casing

Tokenize

Stemming

b. Dictionary

There are two main ways to analyze sentiments, Dictionaries (or so-called lexicons) and Machine Learning. Dictionaries are ruled-based sentiment analysis tools that can be used to process unlabeled data, and those kinds of the pre-prepared sentiment lexicon should contain a word and corresponding sentiment score to it. It is a relatively simple and cost-efficient approach to implement sentiment analysis, since it does not require model training.

VADER (Valence Aware Dictionary and Sentiment Reasoner) Sentiment Analyzer will be applied to the dataset to classify words into three sentiments, including “Positive”, “Negative”, and “Neutral”. As we will analyze tweets, VADER is a suitable tool. VADER is specifically attuned to sentiments expressed in social media and works well on texts from other domains. VADER is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. VADER maps lexical features to emotion intensities known as sentiment scores. It is available in the Python NLTK toolkits.

c. Correlation

As for another dataset, stock market values of Top NASDAQ companies from 2010 to 2020, we will select the Apple Inc. stocks from 2015 to 2019 as our data. After classifying and analyzing the sentiment of the tweets, we will compare it with the variance of the company stock price and try to find out the correlation between sentiments and stock prices. Additionally, we will analyze the stock data with sentiment analysis through some visualization, such as correlation heatmap and scatter plots.

5. Project plan

Oct 17 - Oct. 28: P1 submission and Data preprocessing

Oct 31 - Nov 11: Tag words and correlation analysis

Nov 14 - Nov 28: Draft the report

Nov 31- Dec 2: Prepare the presentation ppt

Dec 2 - Dec 7: Refine and submit the report

Reference

[1] Anshul Mittal, Arpit Goel, Stock Prediction Using Twitter Sentiment Analysis, Standford University, CS229, 2011.

[2] Kaggle, Tweets about the Top Companies from 2015 to 2020, https://www.kaggle.com/datasets/omermetinn/tweets-about-the-top-companies-from-2015-to-2020?resource=download&select=Tweet.csv.

[3] Deepanshi. (2022, July 19). Text preprocessing NLP: Text preprocessing in NLP with python codes. Analytics Vidhya. Retrieved October 24, 2022, from https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/

[4] Amrita Shelar and Ching-yu Huang, Sentiment Analysis of Twitter Data, School of Computer Science, Kean University, Union, NJ 07083, USA, 2018.

[5] Shihab Elbagir and Jing Yang, Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment, International MultiConference of Engineers and Computer Scientists, 2019.

[6] C.J. Hutto and Eric Gilbert, VADER: A Parsimonious Rule-based Model for Sentiment Analysis of social media Text, Georgia Institute of Technology, Atlanta, GA 30032, 2014.