**News Headlines Analysis Project (Stock Market)**

**Data Management Plan**

1. **Project Description:**

* This report/research tries to find the relation between the sentiment of news headlines and tweets and stock prices by determining all factors that may affect the stock prices in the short and long term.
* Is it possible to build a model that accurately predicts stock prices depending solely on the sentiment of financial news headlines/tweets? Or should we include the different types of news headlines that cover all factors that affect stock prices? (Nature and weather news, political news, cultural news, ... etc.).

1. **Project Goal:**

* To create a platform/tool to help people, who are interested in the stock market, make their investment decision based on the insights they get from our platform (Fin-LAFIO).
* Fin-LAFIO provides insights extracted from the following factors influencing the stocks prices:
* News Headlines.
* Major analysts` tweets and analysis.
* Gossip- to be added later on.

1. **Problem Statement and Value Proposition:**

Sometimes it is overwhelming for investors to decide to buy, sell or keep a stock since there are many aspects to consider when deciding. The stock price is determined by supply and demand, which are affected positively or negatively by many factors including; news headlines, company value/performance, key people analysis, and gossip in the market. **Fin- LAFIO's role** is to gather all these elements in one place and make the process easier and more insightful for users (One-Stop-Shop).

1. **Model Data Inputs:**

* Date
* Time / Periods (of news/tweets)
* News Headlines
* Tweets
* Sentiment
* Prices (change in the price depending on the sentiment)
* Stock (company, geographical location, industry)

1. **Model Features:**

* Correlation between news headlines and stock prices.
* Accurate (Depends on the model & the application)
* The model needs to be used for a different financial market, not the USA market only, the main issue is to have data that can predict the changes.
* Analyze the behavior (sentiment) of all users (not only the behavior of the American users).
* Short and long-term prediction.

1. **Questions:**

* What is the impact of the frequency of financial news headlines?
* What is the effect of Quarterly financial reports on stock prices?
* Is there a relationship between timely news and stock prices?
* How many sources do we have for financial news?
* Does news (other than financial) affect the stock prices?
* Can stock prices be predicted in the short and long-term depending on the news headlines?
* Which comes first (news, sentiment then change in the price) Or (change in the price in the major/most important stocks then news spread)?
* What is the relationship between the investor type (investment strategy used) and stock prices?
* What are micro and macro reasons/causes of a change in the stock prices
* Do news headlines have the same effect/impact on all users in the different geographical areas? (The way users deal with the news; does it differ from one location to another?)
* Are there cultural differences in dealing with the news?
* Can stock prices be predicted depending on sentiments? (Can a mathematical model be used to predict stock prices?
* What are the different sources of news?
* What is the impact of the source of the news on the stock prices (if it is a governmental statement, or a company and industry news, or gossip) (Effect of news types (such as noise news) on the stock price)?
* Does the news have the same effect on all users/stocks? (Bad news for some, good news for others).
* Which industries are more (sensitive to/stable against) news? (What is the relation between the news and stock prices in a certain industry/sector?)
* Is there a relation between the way (medium) and speed of news (how the news is transmitted) (using tweets, conferences, web...etc.) and stock prices?
* Do seasonal trends affect stock prices?
* What is the effect of market makers (setters) on changing stock prices? (Impact of the market makers as the source of the news).

1. **Workflow:**

* We chose one stock to study as a start- Apple Company.
* Got news headlines for Apple company for the same period.
* Used a dataset of the price & volume of the Nasdaq index to compare with the changes in the Apple stock.
* Data was collected and studied to determine the relationship (if exists) between the stock price and news headlines.

1. **Future Steps:**

* Study stocks of the main competitors of Apple Co.
* Add key people analysis and tweets to determine their impact on stock prices (Scrap data from Twitter, blogs, and news websites).
* Expand to include new sectors (other than technology) to study the results of news headlines on stock prices (bad news for one sector, are good news for other sectors).
* Use Natural Language Processing (NLP) to analyze text data scraped from Twitter and news websites.
* Analyze the news and opinion (headlines) using sentiment analysis (describe the headlines by positive or negative by using certain words- bag of words).

1. **Data Management Plan**
   1. **Data Collection and Description:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Short Description** | **Data Type** | **Derived from** | **Time Covered** | **Records** | **Columns** | **Missing Values (Null)** |
| News Headlines | Includes news headlines and opinions from different sources for many companies. Only news and opinions related to Apples were used | CSV-  Unstructured | Kaggle | 2012-2020 | 221,562 | 9 | none |
| (Apple) Stock | Stock price & traded volume | CSV- Structured | Yahoo | 2012- 2020 | 2,537 | 13 | none |
| Nasdaq Index | Index price & traded volume | CSV- Structured | Yahoo | 2012- 2020 | 2,556 | 7 | none |

* 1. **Other sources of data currently working on**
* Reddit
* Twitters
* New Websites (CNBC finance, Bloomberg, Reuters news, wall street journal, NYSE stock market, S&P 500)
  1. **Work Done & Challenges with Data:**
* Data was clean and didn't need much work.  Data were derived from Kaggle and Yahoo, no scraping technologies were used, but we will be using them to get data from Twitter and news websites.
* Data of Apple Stock and Nasdaq Index are Structured, no cleaning was done. When we start dealing with text data (Unstructured Data) there will be a lot of cleaning steps that need to be done, in addition to NLP.
  1. **Challenges:**
* Finding the relationship between the stock price and the news (to find the link between them).
* The split factor caused a difference in the stock price (before and after the split factor).
* Describing the headlines as positive or negative was done based on the change in the stock Close price (Stock Diff column), so the impact of the news/opinion headlines on the stock price wasn`t determined for now. This is to be done.
  1. **Data Distribution:**
* All data derived from the open-source could be shared and distributed according to the best ethical practices as don’t need any pre-approval from the publisher.
* All outsourced articles, blogs, sites will be referenced to the owner as a part of our ethical commitments.
* Any other newly generated data will be shared as part of the subscription plan ( type & size ) and services.
  1. **Data Availability:**
* All free accounts will have limited access to all data.
* Paid users will have access to data based on several options of subscription plans.
  1. **Data Preservation:**
* Data will be preserved and stored using cloud platforms and servers with automatic backup.

1. **Fin-LAFIO Project Team:**

* Lina Shbaitah - Domain Expert
* Ayman Khalaf - IT Specialist
* Firas Abu Nimreh - IT Specialist
* Issa Baitouni - Domain Expert
* Omar Ja`bari - Domain Expert

1. **Contact Details:**

* **E-mail:** finlafio@gmail.com