Project Proposal

(Information Retrieval and Text Mining)

Corpus

The corpus I am planning to use is part of a dataset that consists of all the tweets Elon Musk posted from his official account “elonmusk”. The data is divided into 13 files, one per year(from 2010-2022). As of now, Elon Musk has posted a bit over 17 000 tweets and I assume the data has a similar amount of datapoints.

Link to the data: https://www.kaggle.com/datasets/vidyapb/elon-musk-tweets-2015-to-2020

Idea

Currently, Elon Musk has around 81 million followers and he is on 8th position of most followed people on Twitter. I have always been inspired by what he was able to achieve and I am interested to analyze what he says on social media. To complete the idea and to set a more interesting goal for the project I would like to explore how what he says influences people. To do that I would connect his post to the price of his stock or the stocks he mentions in his tweets. I would assume that most of his followers are interested in technology, and most people interested in technology are more likely to invest in technology stocks.

Main Research Question

How Elon Musk tweets influence the people’s opinion about specific entity?

Methods

The analyzes of the data will be performed by extracting the named entities in a tweet. Furthermore, the sentiment of the tweet will be extracted and mapped to the entities. Then if the entities are famous companies, most likely they have stocks. There are various libraries in Python that provide historical stock data. One example is “yfinance”, this library allows users to get data for any stock from Yahoo Finance.

Validation

The data, that will be used in this project is not labeled. To verify how the model that I am going to build performs, I will take a sample from the analyzed tweets and verify manually how accurate the extracted sentiment is.

The dataset includes number of likes and retweets, that data can also be used to as a verification of the sentiment of the tweet.

Visualizations

The visualizations will start from basic statistics about the dataset: number of tweets per year, most used words, most mentioned topics, how sentiment of tweets changes over the years, etc.

The second part of the visualizations will aim at answering the main research question. To achieve that the stocks of the most mentioned companies will be compared to what Elon Musk has said about them. Also, the extracted sentiment will be compared to his stock(Tesla). This will be direct comparison of how his opinion influences people's actions.