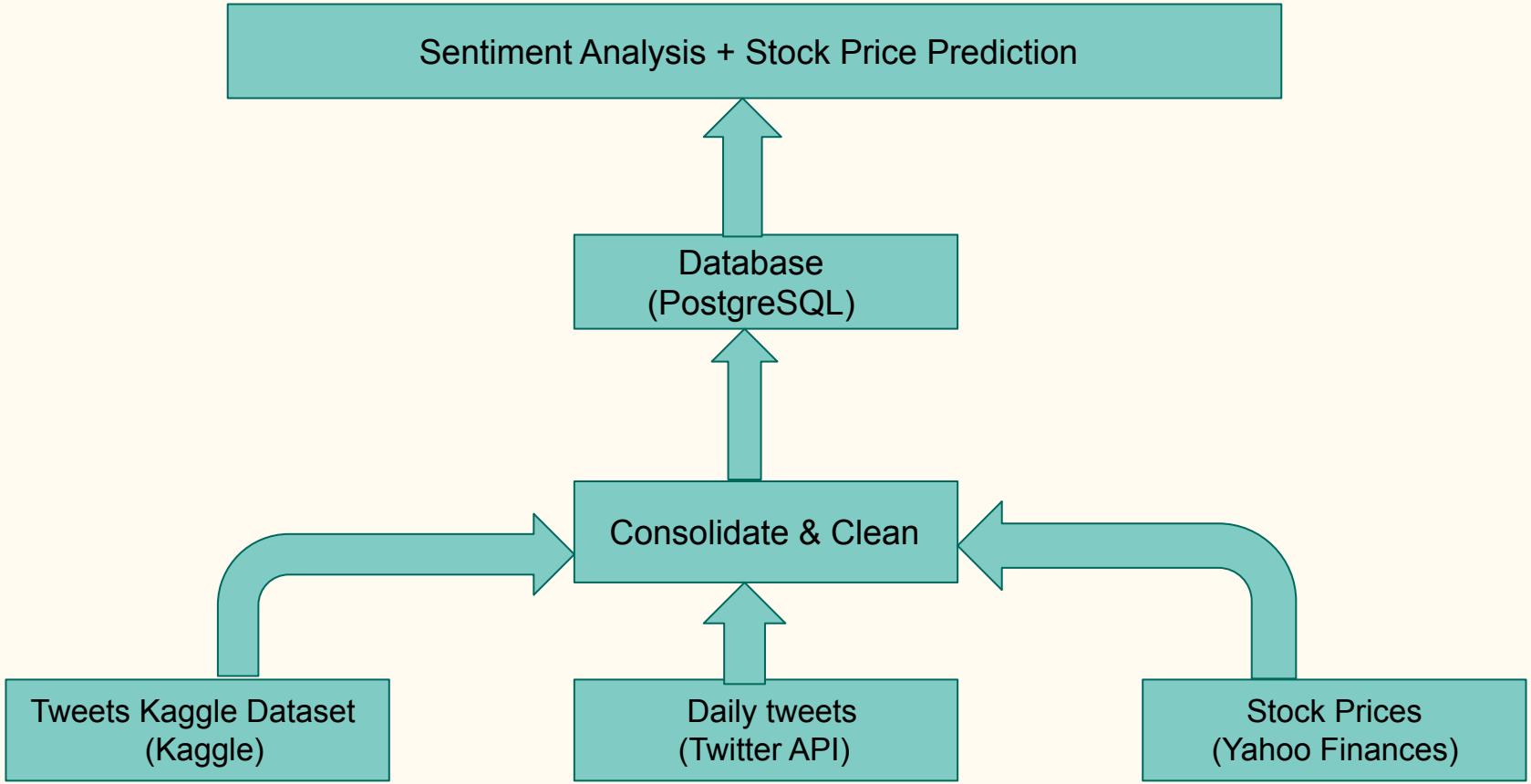
Analyzing the Effects of Twitter Sentiments on Stock Prices

Dataset creation details

- Fetched daily tweets using a python library called “**Tweepy**” which streamed Tweets from their official API
- Incorporated **Kaggle dataset** with company relevant tweets from **2015 - 2020**
- Collected **Stock prices** from “**Yahoo Finance library**” for the same time period to aid historical data analysis and stock price/trend prediction.
- Did some **preliminary data cleaning** like duplicate, special chars, URL removal, handling missing values
- To **ensure** that we have **quality data** we filtered out the tweets based on tags like verified, hashtags, search_query and retweets
- Finally **stored/consolidated** the data from different sources on **PostgreSQL DB** for further analysis.

Dataset Creation Flowchart



Model - TextBlob

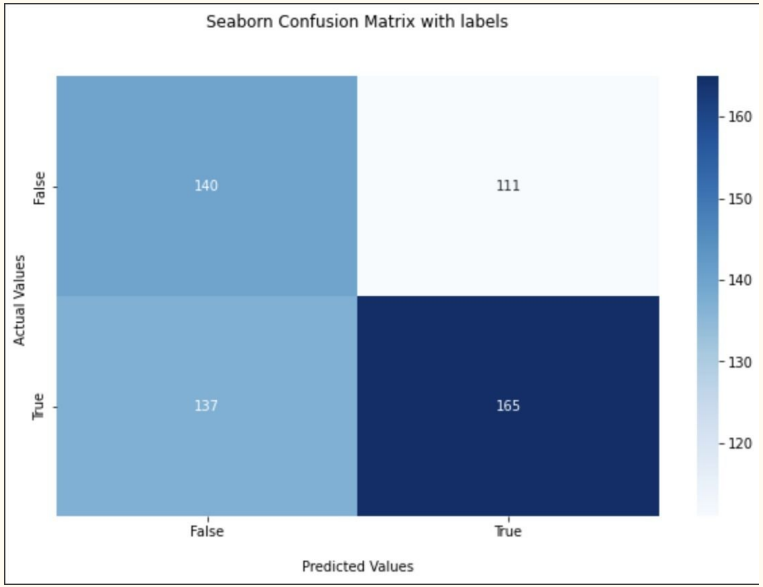
- Uses Natural Language ToolKit(NLTK)
- Supports complex analysis and operations on textual Data
- Calculates Polarity and Subjectivity of a sentence
- Polarity lies between [-1,1] and Subjectivity lies between [0,1]
- Accuracy: 0.508

  For example: We calculated polarity and subjectivity for “I do not like this example at all, it is too boring”. For this particular example, polarity = -1 and subjectivity is 1, which is fair.

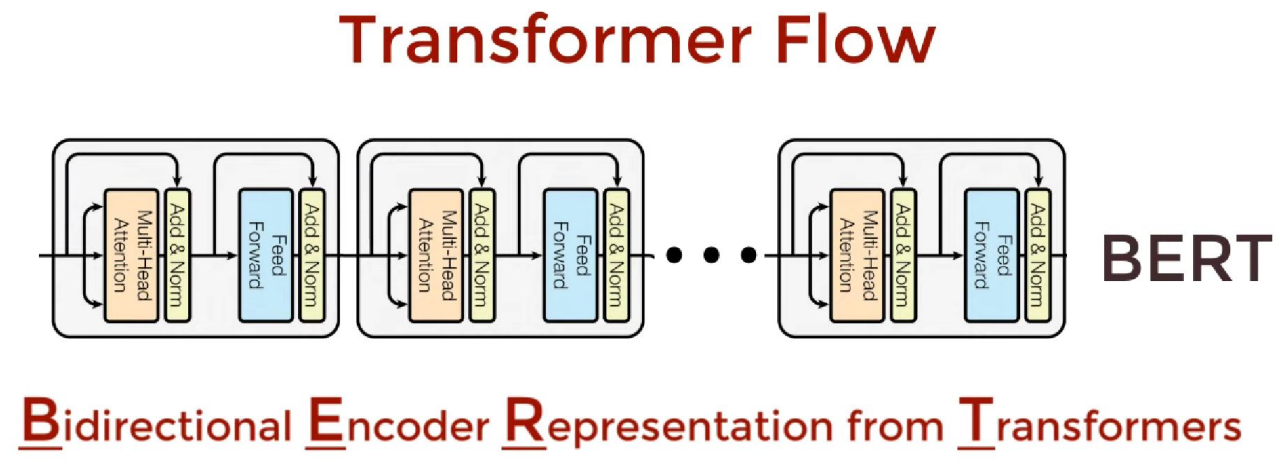
  However, for the sentence “This was a helpful example but I would prefer another one”. It returns 0.0 for both subjectivity and polarity which is not the finest answer we’d expect.

Model - VADER

- VADER (Valence Aware Dictionary for sEntiment Reasoning) is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion
- Lexical approaches look at the sentiment category or score of each word in the sentence and decide what the sentiment category or score of the whole sentence is
- The main advantage of lexical approach lies in the fact that we do not need to train a model using labeled data, since we have everything we need to assess the sentiment of sentences in the dictionary of emotions
- Accuracy: 0.553



Model - roBERTa

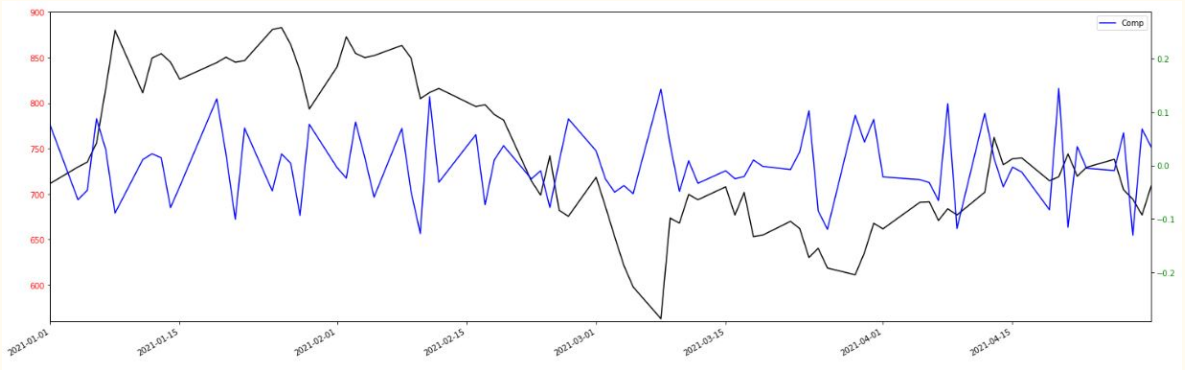


Problems that can be solved by BERT

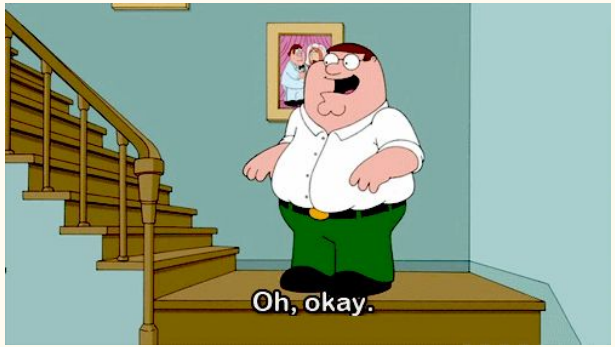
- Neural Machine Translation
- Question Answering
- Text Summarization
- Sentiment Analysis

Model - roBERTa

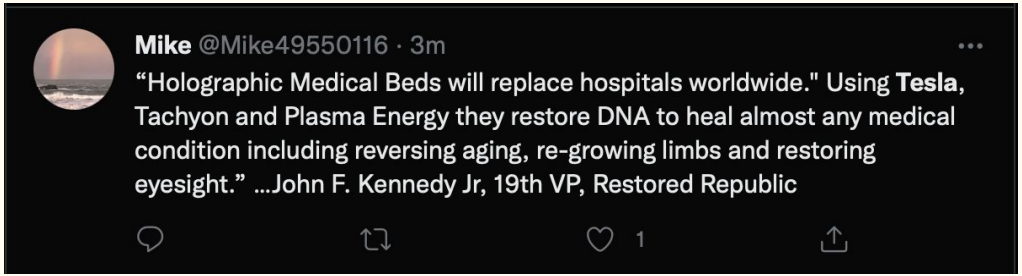
- Roberta is language model developed by Meta that improves on Bidirectional Encoder Representations from Transformers, or BERT, the self-supervised method released by Google in 2018
- BERT is a revolutionary technique that achieved state-of-the-art results on a range of NLP tasks while relying on unannotated text drawn from the web, as opposed to a language corpus that's been labeled specifically for a given task
- RoBERTa builds on BERT's language masking strategy, wherein the system learns to predict intentionally hidden sections of text within otherwise unannotated language examples
- Accuracy: 0.6



About our results ->



- More time can be spent fine tuning the hyper-parameters of roBERTa to get better results
- However, there is a ceiling on our results
  - Stock movements are inherently stochastic
  - We couldn't manually label the tweets ourselves and train a model
  - Pre-trained models were not specifically trained for tweets with a financial focus
  - Only 5000 tweets a day were collect
  - Furthermore a lot of tweets in our scrapings were tweets that were completely unrelated to our aim



Future Scope

- We have investigated the **relation** between **public mood** as measured from a large scale collection of tweets from twitter and the **Stock market values**. Our **results show** that **public mood can be moderately captured** from the large-scale Twitter feeds by the means of natural language processing techniques employed in this project.

- **Future work** regarding this study would include using the model on **different stock markets across the world.** Furthermore, using a data range of more than 10 years may provide more accurate results. Additionally, analyzing the models in **different economic situations** such as booms or recession may allow us to better see the productivity of the models.

- **Extracted sentiments** may be **biased** because **not all the people who trade** in stocks **share** their opinions on twitter. **Stocktwits** is a financial communication platform designed **solely** for sharing ideas and **insights of investors, entrepreneurs and traders.** The current study can be **extended** by incorporating stocktwits data.