**Preprocessing and Cleaning Datasets**

First, we are preprocessing the data for stock market. We are creating a new csv file which has merged the company names to its shortform which is being used in the dataset.   
Now “merged\_data.csv”, we convert the column ‘day\_date’ to a proper datetime format and filter out the year 2020 from it. Also filter and separate out data for companies like Microsoft (MSFT) and Tesla (TSLA), and drop rows with missing values. This is then saved as two separate .csv files: “msft\_2020.csv” and “tsla\_2020.csv”.  
  
  
Next, we preprocessing the social media tweets dataset.  
Again we start by merging the companies with tweets dataset. Then converting ‘post\_date’ column to a proper datetime format. The tweets are then filtered out based on the year being 2020 and companies being Microsoft (MSFT) and Tesla (TSLA). These are then saved into separate .csv files: “msft\_tweets\_2020.csv” and “tsla\_tweets\_2020.csv” .  
  
Now we clean the text/posts uploaded by users in both “msft\_tweets\_2020.csv” and “tsla\_tweets\_2020.csv”. After loading the datasets, renaming some of the columns for easy access, any duplicate or missing values is dropped/removed. After this, text preprocessing is done, wherein the text is converted to lowercase, URL’s removed, mentions and hashtags are removed, numbers are removed. But at the same time it keeps some of the punctuations and tokenizes the text. This cleaned data is then saved as “cleaned\_msft\_tweets.csv” and “cleaned\_tsla\_tweets.csv” .

Now we move on the preprocessing and cleaning the news articles dataset.   
The process is the same for both the Microsoft and Tesla news articles.   
First, we load the dataset and then convert the “date” column to a proper datetime format. Then we filter out the year 2020 and drop rows with missing values. Then the text preprocessing function applies, wherein the text is converted to lowercase, removes URLs, removes mentions and hashtags, removes punctuations, removes numbers and removes extra spaces. This cleaned dataset is then saved as “cleaned\_msft\_articles.csv” and “cleaned\_tsla\_articles.csv”.   
  
  
**Sentiment analysis**

Again the process remains same for both Microsoft and Tesla. First we load the cleaned datasets. We start with sentiment analysis of tweets using VADER. and then do the sentiment analysis for news articles using TextBlob. This data is then stored into new .csv files : “sentiment\_msft\_tweets.csv” , “sentiment\_msft\_articles.csv” , “sentiment\_tsla\_tweets.csv” and “sentiment\_tsla\_articles.csv”