Aditya Ranade
Dr. David Jeong
COMM173 (Spring 2023)
14 June 2023

Statistical Analysis: Tweets vs. Stock Price and Cross-Checking Human-Labeled Data

INTRODUCTION

- Please set the stage for your topic of choice.

In the past few years, social media has become more prevalent and has become a popular way of communicating and spreading information. Along with the increase of social media, it continues to act as a double-edged sword in the sense that social media can be used to negatively influence or positively influence various people, organizations, companies, groups, and more. Large conglomerates and companies such as Tesla, Apple, United Airlines, Delta Airlines, and more have taken advantage of social media to share their products, offerings, and updates with audiences worldwide. Similarly, customers have used social media platforms to share their thoughts and honest experiences with particular products and services. At times, the feedback can be so harsh or amazing that a company's shareholders may hear about it, and thus make decisions on their holdings of positions within the company. This may lead to the stock price increasing greatly or decreasing at a rapid pace as well.

- What is the context of your topic?

In recent years, the social media platform called Twitter has been in the news for several reasons, particularly because of its recent takeover by Elon Musk, a technology executive and CEO of Tesla Inc. Usage of Twitter has become more controversial since the takeover, and several known celebrities have also left the platform because of its controversies such as Hasan Minhaj. However, back in the 2010s, the platform was growing in popularity and was an active place for people to express their opinions. It is important to note that one of the industries that heavily used Twitter to understand customer sentiments and protect brand credibility was the Airline industry. When a customer would be upgraded, or would lose their bag, or would receive poor customer service, or would be given a shuttle to the gate, they would tweet about it, tagging the airline's account with an '@' while doing so. When a tweet would be something negative, the airline would try to take matters into the corporate level and try to resolve it immediately, so the tweet doesn't cascade, and the brand image and reputation is protected. Similarly, when an airline gets more positive tweets, it is something they may retweet or share so that their brand's loyalty and reputation increase. When looking at tweets in a major quantity, more negative tweets could potentially mean bad news for the company, its reputation, and ultimately, its stock price or valuation (if private). Of course, Twitter may not be used actively nowadays, but it would be important to monitor more widely used platforms such as Facebook, Instagram, and/or

Reddit for such feedback and thoughts/experiences posted by customers.

- Why is this important in [a particular field]?

Analyzing the sentiments of customers' thoughts and experiences on a given service or product is important for the brand to understand the ‘lived experience’ of its customers. Analyzing this at a deeper level can allow a brand to uncover important feedback that can help improve their products and services for their customers, and ensure that mistakes are not repeated. Constantly observing negative feedback and trying to improve on that is also something that will be good for the brand’s reputation as it would allow for more customers to recommend the product to other people. Without analyzing real customer feedback, the brand has no way of knowing how customers are feeling about the product. A lot of negative feedback being given and a brand being oblivious to this is not the best thing as this will heavily affect the company’s reputation as well as to company’s stock price, valuation, and revenue. This idea of analyzing customer feedback on social media platforms can be applied to any field or company that has a product or a service, but it is especially important in the airline industry as it is a service-based industry, and many customers have brand loyalty when it comes to airlines.

- What is the problem [in this particular field or area] that you are solving?

The problem I am trying to solve is to establish a connection between the stock price, revenues, and/or valuation of a company based on the sentiments of customer feedback on a social media platform (in this case, Twitter). I feel that companies may have two vastly different verticals when it comes to finance/business valuation/revenues and social media. My goal is to see if there is a correlation between these two and to see if any cross-functional knowledge may help the company grow and provide a better product/service to its customers.

- What is your research question?

My research question is what is the connection between the sentiments of tweets ‘@’ an airline and the stock price/revenues/profits coinciding with that particular timeframe? For example, if analyzing tweets from Jan 2014 to Apr 2014, we would take a look at the sentiments of tweets in that time range for tweets that tweeted ‘@’ the airline. If we see more negative sentiments, we would expect the earnings/revenues/profits/valuation to decline, and vice versa. To get a more complete picture, it would be important to analyze a controlled timeframe and use that as a basis to see how the earnings/revenues/profits/valuation (in other words, financial aspect) of the company changes based on the sentiments of social media presence of feedback. Not only do I want to understand this, but since I have found and will be using a dataset that is human-labeled for sentiments, I want to find out what the margin of error is between the human-labeled tweets and machine-labeled tweets (using nltk in Python).

CAUSAL INFERENCE

- Introduce the significance and importance of causal inference in statistical understanding in general.

Causal inference in statistical understanding is a very important aspect to consider when performing any type of statistical analysis with a given set of data. Understanding causal inference is important as we are able to go beyond just the statistical measures and correlation between variables to see if there is a cause-and-effect relationship that exists between them. When we conduct correlation analysis for example, we can see how closely associated two variables are, but the correlation analysis alone does not tell us whether the variables have some cause-and-effect relationship between them. With an understanding of causal inference, we are able to go beyond just using numerical understanding to come up with conclusions but rather use a deeper understanding of the variables to make future predictions and decisions.

- Introduce the significance and importance of causal inference in solving the problem you have mentioned in the introduction section.

In the context of this setting, causal inference is a crucial part of understanding the relationship between the financial aspect of companies and how customers' sentiments are on social media. We will use causal inference to potentially understand how negative and positive sentiments can negatively or positively affect a company's financial aspect. This causal inference that is made will help companies make better improvements in the product and service offerings to customers, as well as better decisions on marketing strategies to enhance brand loyalty and reputation and plan better financially to understand profits/losses based on customer sentiments (i.e. if we find that negative sentiments correlate to a lower stock price or decline in revenues, then the company can activate its cash position to make sure they won't have to layoff as many people or can limit hiring).

- Identify the proposed causal argument you will attempt to address in your proposed analysis.

The proposed causal argument in the proposed analysis will aim to understand a causal relationship between the sentiments found in tweets by customers in which an airline is tweeted, and the performance of the financial aspects of the company (stock price, revenues, profits/losses, etc.). What we will look for is that if negative sentiments are more prevalent, then is there a decline in financial performance, or vice versa?

- According to Pearl, what are 3 ways you can make a strong argument of causality in your research question?

When understanding the causality relationship, it will be important to make note of these three particular parameters detailed by Judea Pearl:

1. Confounding Variables: When trying to establish causality between two variables, it is also important to understand the existence of confounding variables, or in other words, a third variable that might affect the relationship between the other two variables you are attempting to establish a relationship between. In the case of finding causality between the sentiments of tweets and the financial performance of a company, it will be important to take note of the other variables that might be present that may affect the relationship such as the market conditions or any large news events. For example, if looking at the

time period of 2020, there was a huge drop in every company's financial performance at the onset of the pandemic. If when trying to establish such a relationship, one does not look at extraneous variables such as this, one may come up with incorrect conclusions that could potentially affect future predictions or decisions.

2. Temporal Order of Events and Cause-Effect Relationships: It will be important to understand the chain of events when analyzing the relationship between the tweets at a particular airline and the stock price, that the financial performance is a result of the tweets. This means that it is important to analyze the tweets and look at the financial performance after the fact. Not during or before the fact. Therefore the step-by-step analysis involves analyzing the sentiment of tweets first before anything else, then followed by the financial performance analysis. This way, a proper causal relationship can be established between the sentiment of the tweets and the financial performance of an airline as a result.
3. Counterfactual Reasoning: It will be important to analyze the results of the analysis to see what actually occurred in the causal analysis versus what would have occurred under different circumstances. For example, if we observe that a large number of negative tweets has a correlation with the financial performance decreasing, we can think of what would have happened if this were not the case for the sentiments (if they were neutral or positive). Using this, we can isolate and understand the effect of social media sentiments directly on financial performance. Again, it is hard to make conclusions here because we are not accounting for any other confounding variables that will, can, and may affect financial performance.

DATA

- What type of data will you need to address this problem?

The data that I will need is some data scraped from Twitter with the tweets that are directly related to an airline, which can be found by finding tweets that tag an airline's Twitter account in them by using the '@' symbol. The dataset would need to include data from various different airlines that includes the tweets and their text content. The time of the tweet (specifically, the date) would also need to be provided. Additional metadata of each tweet such as the location of the tweet would also be beneficial to conduct further analysis and research using.

- Identify the public dataset and provide the link to the dataset within the text of your methodology section as so: I will be using the Big Mac Index in this report.

The dataset that I plan to use is the "US Airlines Twitter (Overtime)" dataset obtained from the following link: https://www.kaggle.com/datasets/thedevastator/sentiment-analysis-of-us-airline-twitter-data. This is a very famous dataset and the credit is given to: https://data.world/socialmediadata, for scraping the dataset. Here is a quick description of the dataset obtained from socialmediadata's post of the dataset: "A sentiment analysis job about the problems of each major U.S. airline. Twitter data was scraped from February of 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service")".

- ○ Please also explain how you would scrape your dataset

This data set would ideally be scraped directly from Twitter using an API, but for the purpose of this project, I am using a pre-scraped dataset that has already been obtained

- Please describe and explain various characteristics of the dataset
    - ○ How many data points?

There are 14,639 tweets in this dataset, and there are 21 columns. Here is a list of the columns obtained from Kaggle: index, unit id, golden, unit state, trusted judgments, last judgment at, airline sentiment, airline sentiment confidence, negative reason, negative reason confidence, airline_sentiment_gold and retweet count. There is also text included for each tweet as well as tweet location and user timezone.

- ○ What are the primary variables in the dataset?

The primary variables are the "text" variable which is the main variable in which the tweet text is, the "negativereason" variable where the negative reason is categorized if the tweet ended up being negative, the "tweet_created" variable which tells us the date of the tweet, as well as the "airline_sentiment" variable which contains the human label for whether a tweet was positive or negative.

DATA PREPARATION

- What are some general characteristics of your dataset that may cause bias or other issues in your analysis?

Some issues that may occur from the dataset that may cause bias or other issues would be null tweets (tweets with no text content) after stopwords are removed. We would need to clean the dataset of this and make sure that none of that exists to ensure when we are parsing there are no errors. Another issue that seems to be prevalent is that all the tweets in this famous dataset are from one particular month in one year (February 2015). Because of this, we don't have too much of a wide time range of data to work from. But since we are doing two parts: checking the correlation between sentiments of tweets and stock price as well as checking the accuracy of human labeling of sentiments and sentiment labeling using nltk.

- How would you clean, prepare, and/or pre-process your data in order to avoid issues in your analysis?

The way we would prepare and pre-process the data would first be to check if there are any variables that we will not need in our analysis. For example, we would not need the "tweet_id" or "_unit_id" variables because those are unique to each tweet and add no content to analyze. After doing this, we would look at what other columns have several null values. For example, if there is a column with more than 14,000 null values, we would probably not want to use this in our analysis because it would not yield anything of value to us. After clearing out null values that are appropriate and removing columns that we do not need, we can remove stopwords from the tweets and can proceed with the analysis.

- What type of analyses would you need to conduct to address your research question?

The analyses that I would need to conduct to address my research question would be two-fold. The first part involves checking the sentiment of each of the tweets. We can remove the stopwords and pass each of the tweets through a tokenizer that would tokenize each tweet into words. Each tweet can then be passed through an nltk sentiment analyzer, which can determine whether each token (word) is positive, negative, or neutral. Using this information, we can make the decision on whether a tweet is positive, negative, or neutral. If there is more positive tokens in the tweet, then the tweet is positive, if more negative tokens, then the tweet is negative, and finally if more neutral tokens, then the tweet is neutral. Once we have this info, we can view the percentages for each airline, and see if an airline had more positive, negative, or neutral tweets. After we see this, the next step would be to analyze the company's financial status (primarily, stock price movement) in that particular month, day, or quarter, and see if there is a correlation. We are trying to see if there are more positive or neutral tweets, the stock price go up, and if there are more negative tweets, does the stock price go down. The second part of the analysis would be to take a look at what the tweet was classified as from the nltk sem=ntiment analyzer and compare that to the human-labeled sentiments for each tweet. The goal would be to measure the error (delta) in between the human-labeled tweets and the machine-labeled tweets.

- How would you conduct your analysis?
  - Which variables would you be analyzing?

The variables that we would be analyzing are the "airline_sentiment, airline_sentiment:confidence, negativereason, airline, text, and tweet_created".

  - What types of variables are your variables of interest (e.g., continuous, binary, string).

airline_sentiment: string
Airline_sentiment:confidence: string
negativereason: string
airline: string
text: string
tweet_created: string

CONCLUSION

- What is the general takeaway from this proposed analysis?

The general takeaway from this proposed analysis is that we get a better understanding of how customer outlook, feedback, and customer presence on social media can affect (or not affect) a brand's financial position. The idea is that we will get to know the importance of monitoring such social media posts about a brand and whether it is something that needs to be heavily considered when it comes to how the outlook of a brand's financial status will look like. In addition to this, we will also know how much we can trust a machine. Assuming the human-labeled data in this dataset is accurate, we can think of how accurately the nltk sentiment analyzer from Python analyzes the sentiments in the tweets compared to what humans thought. We can use this information to draw conclusions of what might happen for large datasets and if it would be ok to have a machine-label dataset and perform supervised learning using them, or is it something that should continue to be done by humans.

- What is the *safest* conclusion that you can draw if the hypothetical results of your analysis turned out in the way that you expect it to?

The safest conclusion that I can draw if the hypothetical results of my analysis turned out in the way that I expect it to, then it would be safe to say that if an airline has more positive sentiment than negative sentiment tweets in a given timespan, we would expect to see a better financial outlook in terms of the stock price, for that particular timeframe as well. Similarly, if an airline has more negative sentiment than positive sentiment tweets in a given timespan, we would expect to see a worse financial outlook in terms of the stock price, for that particular timeframe as well. Finally, if the airline has more neutral tweets than both positive or negative tweets, we would have to look at the stock price to make a conclusion for that particular time period. We would expect the stock price to remain in the same range, but not decrease or increase rapidly. It is important to consider that we are not taking into account any confounding variables in this case that may affect the stock price. This is solely trying to draw a conclusion between the social media outlook and the stock price. In addition to this, another conclusion that would be safe to draw is that human-labeled data remains to be more accurate than machine-labeled data. This is because the machine may see more neutral words than anything, and will proceed to label the text of a particular tweet as neutral when in reality it may be positive or negative. There is also a logical conclusion to this: when we are working with machines, we are many times trying to replicate the work of a human. The work of a human in this case may always be better than a machine, however, it would be important to analyze the delta/error between what the human has labeled and what the machine has labeled to see if machine labeling, at least for sentiment analysis, is something that can be used for large datasets.

- How would the hypothetical results of your analysis contribute to our understanding of causal inference?

The hypothetical results of my analysis would contribute to the understanding of causal inference as we would see if there is a cause-and-effect relationship between the public outlook/view/feedback of a brand and how the brand continues to grow and progress. At the end of the day, the financial position of a company will help its future growth as a better financial position means more cash and more investors, and a better valuation. Through this analysis, we are able to find out how social media can be leveraged in a way to better predict the financial outlook of a company, ultimately, for a company to make better decisions in regards to its financial position, product and service offerings, customer loyalty, and more.

- How would the results of this analysis contribute to your field of interest (e.g., marketing)?

The results of this analysis will contribute to my field of interest in many ways. First of all, these results could potentially be used to train a machine learning model that can use this info to analyze future tweets and data to come up with predictions of how it may affect the stock price. Secondly, it combines both the social media and financial aspects which are both of interest to me. Stock trading is something that I have been doing actively for the past 5 years, and the results of this analysis would be great to see how (if in any way) the tweets are affecting the movement of a company's stock. Obviously, one would need more data and would have to implement a real-time scraper, which may be hard to do in today's time due to the change in Twitter's API and policy.

Future Directions

- Identify 3 possible future directions for this proposal.

There are several future directions that this proposal and project can take, however, there are three that are of the most interest to me:

1. Analyze customer sentiment about airlines on different social media platforms: If possible to do so, it would be interesting to scrape platforms like Facebook, Instagram, Reddit, and others to see how customers are expressing their views on a particular brand. We would scrape the data from here and perform a similar analysis to how we did for Twitter. In addition to this, it might be cool to look at a certain time period but merge together customer sentiments from all the different platforms, which would result to even more data. The primary thing of analysis here will be the "text" portion of a post, the particular airline it is targeting, and when it was posted. This is something that can commonly be found across posts on social media platforms, thus giving us a streamlined way to analyze the data.
2. Look at social media posts of domestic airlines in a different country, and potentially see how they change over time: A direct example of this is that since Tata (an Indian conglomerate) has taken a large dip into the aviation scene in India, domestic travel, in particular, has gotten a lot better and customers are taking to social media to share their happiness and satisfaction with the company's successful changes. Before this, the domestic aviation scene in India was not very good with several delayed flights, missing baggage, and in general, low customer satisfaction. It would be interesting to scrape data from the past 6 years across social media platforms and conduct a similar analysis to that of this project. From this, we would do more of an analysis of the financial statements or balance sheets of the airlines and look at their growth over time with the new product offerings and increase in customer satisfaction.
3. Create a real-time data scraper and train a machine learning model to constantly provide predictions on stock price based on new social media data: The idea here would be to scrape more data from Twitter and other social media platforms as mentioned in the first future direction above and if there is some relation with an airline and its stock price, we would use that information from the past data to feed into the machine learning model, to give us constant updates on how we may expect the stock price to move based on

customer sentiments that it is scraping off the social media platforms. The only problem with this is that it is very computationally expensive to be running a real-time data scraper, and would be technically difficult to do so as well.

Reflection

- Please reflect on your process of learning in this class - where did you start and where are you now?

When I came into this class, I already had a good understanding of several data science concepts as my major was Computer Science with a Data Science emphasis. I knew how to apply several statistical methods to data through code, but mostly in Python. Through this class, I learned how to extensively use the R language to conduct data analysis, and how some operations are more efficient in R than in Python. In addition to this, it was great to see data science being applied to various different platforms such as YouTube, Reddit, Twitter, and more. I had never really experienced the 'scraping' portion of obtaining data as in many cases I used a dataset that had already been scraped, but it was great to see the computational processes that go into data scraping. I feel in the future, I will have a greater understanding of the social aspect of companies and in what ways data science can be applied to improve a company's financial performance.

- What are areas of curiosity in data analytics that you would like to continue to pursue in the future?

The areas of curiosity in data analytics that I would like to continue to pursue in the future are how data analytics can be applied to different settings such as security, social media, and supply chain/operations. In the future, I would also like to enter the restaurant business (franchisees/stand-alone stores) and would like to see how data analytics and data science can be used to optimize the business and make it grow to the maximum level. I believe that the world in the future is going to run on data and that it is really important to understand how data will be used in the future to make predictions and make the lives of humans easier.