Analyzing Hashtags Using Sentiment Analysis

Daniel Forgosh, Bemnut Nuru, and Brandon Ubiera
Computer Science
Hood College
djf7@hood.edu, bu2@hood.edu, bn3@hood.edu

May 10, 2019

Abstract

Tweets with specific hashtags from Twitter have been collected using a Twitter developer account. The hashtags that were collected are Facebook, Amazon, Apple, Netflix, and Google. Each hashtag will be its own data set. Stock data has been collected that correspond to each hashtag for the corresponding dates. Sentiment analysis was performed on all of the Twitter data sets to determine if each tweet is positive, negative, or neutral. A database was built to hold all of the tweets and stock data. A website user interface (UI) was built to view the data from the database and allow users to choose different criteria, and the website UI will dynamically generate graphs. These graphs can then be used to analyze the data to find interesting results about Twitter, Twitter users, as well as companies, companies' stocks and items users tweet about.

1 Introduction

Twitter is a large social network platform that allows its users to create short posts called "tweets." A user can use a "hashtag" (#) to indicate a topic that their tweet is about. By using hashtags, tweets that contain the same topics can be collected to create different data sets. Although tweets can be random because they are created by users of a social network platform, there is potential for tweets to provide information that can be used for many purposes by analyzing social patterns [1]. The tweets can be used to promote certain advertisement of companies and they can be used to spread information from the news. Additionally, by performing sentiment analysis on the tweets will add additional information to contribute and be analyzed. There would be a goal in observing how the tweets that focus on company advertisement would be possible in affecting the stock prices.

2 Problem Statement

Users of Twitter create a massive pool of data that can be used for an endless number of purposes. A company can analyze the tweets associated with their company to see if there are any patterns that could indicate changes the company should make. Users might tweet about a specific item/product. If a company has a product that is suddenly gaining popularity on social media, they can expect to soon have an increase in sales; they can prepare for this by increasing inventory of the specific product.

Analyzing the sentiment of tweets can do more than just help companies with marketing. It can help with predicting many topics of all varieties, such as stocks[3, 4]. By collecting large tweet data sets and analyzing them, interesting information can be found.

3 Objectives

Analyzing Twitter sentiment data can help to reveal patterns or changes in trends of how Twitter users feel about certain companies or products. This can be interesting because if there is a change in a trend, a real-world event may be able to be found that would have caused this change in how Twitter users feel about the company or product.

By analyzing Twitter sentiment data along with corresponding stock market data, it can be determined if Twitter sentiment data has a relationship to the stock market data. This can be useful for trying to predict the stock market. Predicting the stock

market is difficult because the price of stock is based on the demand of the stock.

If any changes in trends of the Twitter sentiment data are found and it is related to the stock market data, it can be determined if the change will cause the stock to move up or down in value. This is continuing to help predict the stock market.

4 Related Work

4.1 Twitter Sentiment

Twitter is a large social network platform that allows its users to create short posts called "tweets." The sentiment of tweets can be processed by using one of many data science techniques; one method to process the sentiment analysis of tweets is a library called TextBlob. Many data scientists use sentiment of tweets to help with prediction of many topics.

A major data science topic is stock market prediction. There have been multiple studies linking the use of sentiment of tweets to helping to predict the stock market [2, 3, 4]. Another topic that can benefit from sentiment of tweets is political elections. The sentiment of tweets that contain political candidate names can help to predict the winner of a political election [5]. Very different topics can all benefit from using sentiment of tweets by being able to have accurate predictions.

4.1.1 Impacts on Twitter Sentiment

Since Twitter is a social media platform, data gathered from it can be considered random because the data is users. There could be spikes in the data that could be caused from many reasons [6]. A possible reason is real-world events. Real-world events can cause emotions within a community of any size. It the date of a real-world event can be identified, it can be determined if there is a spike in the data for that specific date [7].

4.2 Website UI Visuals

Websites can be designed in different ways. Some website developers create websites from the ground up using code; this would consist of using HTML, CSS, JavaScript, PHP, JQuery, and other possible scripting languages.

Some website developers create websites using responsive HTML5 frameworks; some examples of these frameworks are Bootstrap, Foundation, and Skeleton. By using a responsive HTML5 framework, it can make a website look visually more appealing using pre-designed styles.

Some website developers create websites using a content management system (CMS). A couple examples of CMSs are WordPress and Drupal. A CMS is a software that is designed to hold a large amount of content and help to create a website. A CMS can implement pre-designed website themes.

In addition, website developers can use open source libraries to help implement other parts of a website. An example of when an open source library can be used is to display data/graphs. There are open source libraries that can do this for the website developer such as Google Charts and RGraph.

4.3 Web UI Search

The online query processing is a component that will allow different graphs with new interesting data to be presented. Query words are associated with data set names, which will comprise a dictionary like structure, allowing for fast lookups through the database [9]. The presentation of the data will be most meaningful to the user performing the search will be dependent on the criteria options. The selection of criteria presented to the user has to be compatible and able to properly look through the database by matching with the stored data sets. Query searches will be primarily related to a corresponding companies stock within the database. Stock queries will include: high, low, open, and close. Other searches that are not related to stocks will include interesting company data, and twitter comparisons of positive and negative changes.

5 Methodology

5.1 Twitter Data Collection

There is an API that Twitter provides that allows developer account holders to query and download tweets [10]. Although Twitter provides this API, there are restrictions. The Twitter API will only allow 350 calls per hour for each Twitter developer account [11].

Twitter stops their users from making long posts by placing a restriction on how many characters a post can contain. Twitter increased a tweet maximum character length from 140 characters to 280 characters in 2017 after conducting a test with the

longer tweet length. Although Twitter has doubled the tweet maximum character length, tweets do not provide a complete data set. Due to using an incomplete data set creates a new research field within data mining because less data is available to be analyzed [12].

5.2 Sentiment Analysis

There are different versions of sentiment analysis. A main form of sentiment analysis is where text is processed into three categories: positive, negative, and neutral. Sentiment analysis will not always be correct due to complexity of language and different writing styles [13]. When performing sentiment analysis, there are two different ways to identify sentiment of text. The first technique is to identify the sentiment of each word.This technique might work, but words do not usually act alone; words are used with other words to provide meaning. The second technique is to identify the sentiment of a string of words. How words are associated with each other is typically more important than the words themselves which makes the second technique better for most scenarios [14]. To accomplish the sentiment analysis, the package TextBlob will be used using Python.

5.3 Final Implementation

After the conceptual designs are written and checked, the implementation can begin. Python is going to be used for a way to collect the data of certain tweets. On the provided department server, PHP and MYSQL would be used for the database of the collected tweets found in Twitter and to display the database data on the website user interface. The Web UI will be created using BootStrap as well as standard web development languages such as HTML, CSS, JavaScript, and JQUERY. The Web UI will have availble criteria option based off the collected data withing the database and present the data of the selected criteria using a JQUERY graph library called RGraph [15].

6 Discussion

The overall goal of this project is to find interesting data within data taken from Twitter by relating tweets to company stock data. This will be accomplished by using the Twitter API to gather data from Twitter and an API of stock data related to the hastags collected. Finding the sentiment for each tweet will be performed using sentiment analysis. A database to hold all of the tweets will be created using MYSQL (Methodology 5.3). To display the data from this database and to generate dynamic graphs, a website user interface will be created using standard web development languages. The web UI will be essential for visualization as well as querying specific parts of the stored data. In the future, a larger selection of hashtags, corresponding stock data, more extensive criteria, and a larger database can be collected for a better look at twitter behavior and preferred company items.

7 Results

7.1 Data

7.1.1 Twitter Data

Apple Twitter data was collected from 10/21/2018 - 1/23/2019. Apple had 87207 positive tweets, 21125 negative tweets, and 182467 total tweets. Amazon, Facebook, Google, and Netflix Twitter data were collected from 10/22/2018 - 1/23/2019. Amazon had 223511 positive tweets, 47856 negative tweets, and 438623 total tweets. Facebook had 99098 positive tweets, 33964 negative tweets, and 238320 total tweets. Google had 78195 positive tweets, 21102 negative tweets, and 199036 total tweets. Netflix had 96737 positive tweets, 29313 negative tweets, and 203726 total tweets.

7.2 Stock Data

Amazon, Apple, Facebook, and Google stock data was collected from 9/24/2018 - 2/15/2019. Netflix stock data was collected from 9/28/2018 - 2/22/2019. There are 100 days of stock for each company; all Twitter data that was collected has a date that corresponds to a stock day that was collected.

7.3 Dynamic Graphs

Due to there being numerous possibilities for combinations of which specific attributes to be in each graph, only certain relevant graphs will be shown in the Analysis section. To see other graphs, visit the web UI listed in the cover page.

8 Analysis

8.1 Twitter Sentiment

8.1.1 Positive/Negative Tweet Comparison

There are always more positive tweets than negative tweets for all technology companies within the data we have collected, but there is this one instance where Facebook has one day, 11/12/2018, where there were more negative tweets (1674 tweets) than positive tweets (1571 tweets). A likely reason for there to be a large amount of negativity towards Facebook is there was a Facebook outage on 11/12/2018 making Facebook inaccessible for a considerable amount of time.

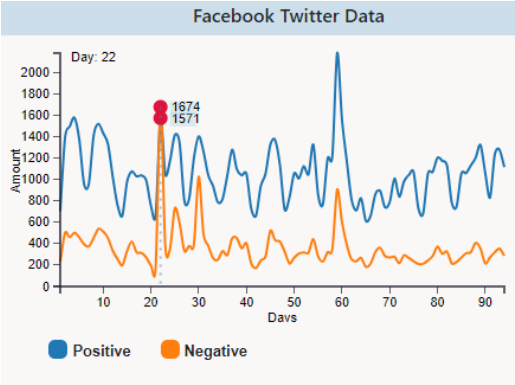

Figure 1: Facebook Twitter data

8.1.2 Tweet Relationships

All attribites for tweets: positive, negative, and total tend to have the same shape within the data we have collected. All attributes go up or down together. Total is the most extreme in its value changes. Positive is medium in its value changes. Negative is small in its value changes.

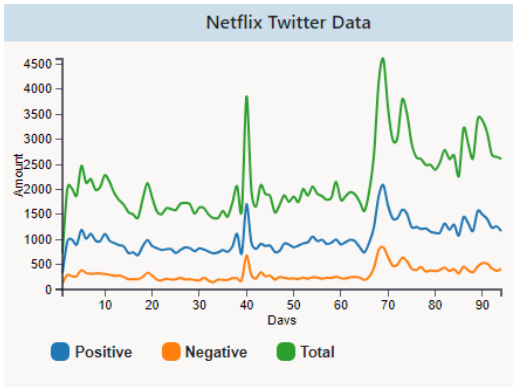

Figure 2: Netflix Twitter data

8.2 Tweet and Stock Comparisons

8.2.1 Apple Keynote Impacts

Apple had a keynote where they introduced the iPad Pro, MacBook Air, and Mac Mini on 10/30/2018 (day 10). Due to this event, there is a large value of tweets on 10/30/2018 (day 10).

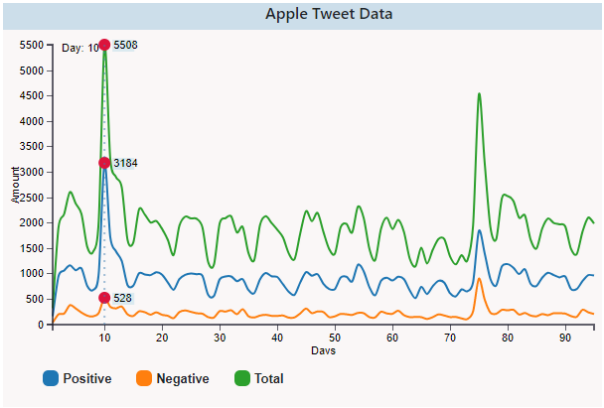

Figure 3: Apple Twitter data

The days trailing Apple's keynote, Apple's stock have a major spike with a max on 11/2/2018 (day 30) in the volume, the number of sales of the stock.

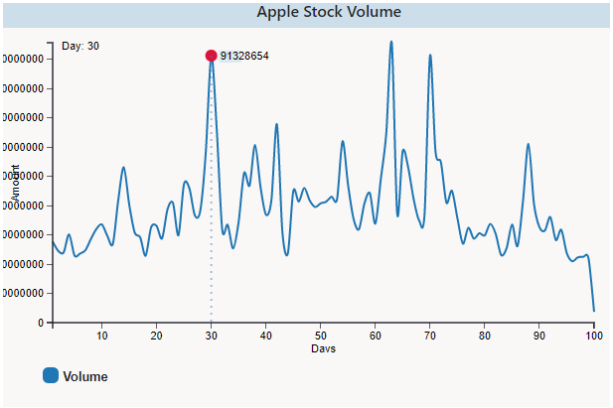

Figure 4: Apple stock volume data

8.2.2 Google Speaker Impacts

Google has a major spike in the value of tweets (4619 tweets) on 12/12/2018 (day 52). This is most likely due to a software engineer from Google, Janet Kuo, being the keynote speaker at KubeCon.

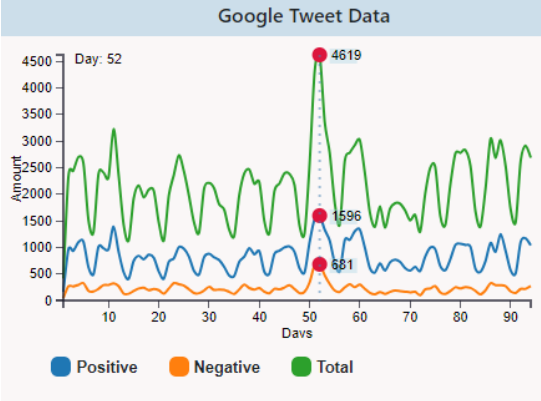

Figure 5: Google Twitter data

Google had a local maximum in their stock value on 12/12/2018 (day 56). The volume of Google's stock was close to a local minimum. This makes sense because if there is a large number of people selling their stock in Google, the value of stock will go down.

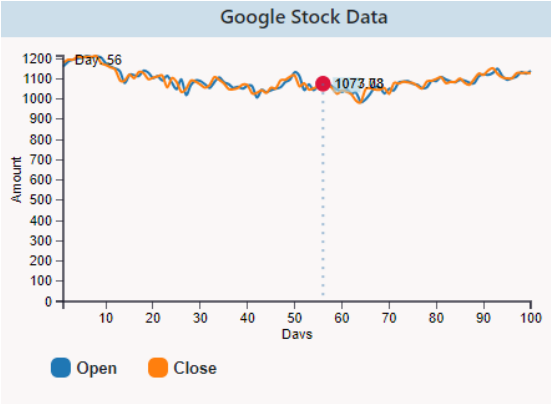

Figure 6: Google stock data

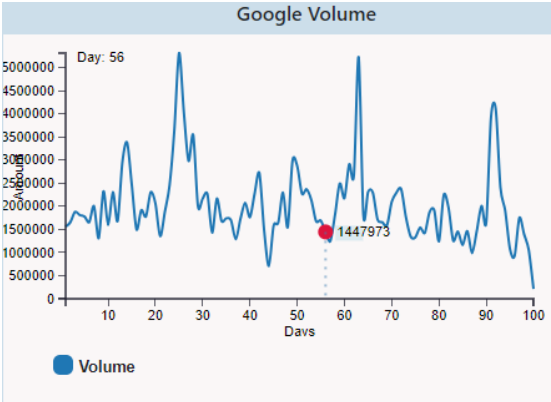

Figure 7: Google stock volume data

References

[1] Maksym Gabielkov, Ashwin Rao, and Arnaud Legout. 2014. *Studying social networks at scale: macroscopic anatomy of the twitter social graph.* In Proceedings of the 2014 ACM international conference on Measurement and modeling of computer systems (SIGMETRICS '14). ACM, New York, NY, USA, 277-288.

[2] Arijit Chatterjee and Kendall Nygard. 2017. *Predicting Stock Close Price Using Microsoft Azure.* In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (ASONAM '17). ACM, New York, NY, USA, 749-757.

[3] Man Li, Chi Yang, Jin Zhang, Deepak Puthal, Yun Luo, and Jianxin Li. 2018. *Stock market analysis using social networks.* In Proceedings of the Australasian Computer Science Week Multiconference (ACSW '18). ACM, New York, NY, USA, Article No. 19.

[4] Andrew Huang. 2016. *Data mining twitter to model stock price movement.* In Proceedings of the 3rd Multidisciplinary International Social Networks Conference on SocialInformatics (MISNC, SI, DS 2016). ACM, New York, NY, USA, Article No. 32.

[5] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. *A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle.* In Proceedings of the ACL 2012 System Demonstrations (ACL '12). ACM, New York, NY, USA, 115-120.

[6] Yuexin Mao, Wei Wei, Bing Wang. 2013. *Twitter volume spikes: analysis and application in stock trading.* In Proceedings of the 7th Workshop on Social Network Mining and Analysis (SNAKDD '13). ACM, New York, NY, USA, Article No. 4.

[7] Ema Kušen, Mark Strembeck, Giuseppe Cascavilla, and Mauro Conti. 2017. *On the influence of emotional valence shifts on the spread of information in social networks.* In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '17). ACM, New York, NY, USA, 321-324.

[8] Rahman Tashakkori and Zachary Andrews. 2018. *A Team Software Process Approach to*

Database Course. In Proceedings of ACMSE'18 (ACMSE'18). ACM, New York, NY, USA, 1-8.

[9] Stelios Paparizos, Alexandros Ntoulas, John Shafer, Rakesh Agrawal. 2009. *Answering web queries using structured data sources.* In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (SIGMOD '09). ACM, New York, NY, USA, 1127-1130.

[10] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, Robert C. Miller. 2011. *Tweets as data: demonstration of TweeQL and Twitinfo.* In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data (SIGMOD '11). ACM, New York, NY, USA, 1259-1262.

[11] Changhyun Byun, Hyeoncheol Lee, and Yanggon Kim. 2012. *Automated Twitter data collecting tool for data mining in social network.* In Proceedings of the 2012 ACM Research in Applied Computation Symposium (RACS '12). ACM, New York, NY, USA, 76-79.

[12] Jiliang Tang, Yi Chang, and Huan Liu. 2014. *Mining social media with social theories: a survey.* ACM SIGKDD Explorations Newsletter. ACM, New York, NY, USA, Volume 15 Issue 2, December 2013 20-29.

[13] Mahnaz Roshanaei and Shivakant Mishra. 2014. *An analysis of positivity and negativity attributes of users in twitter.* In Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (SIGMETRICS '14). ACM, New York, NY, USA, 365-370.

[14] Fotis Aisopos, George Papadakis, and Theodora Varvarigou. 2011. *Sentiment analysis of social media content using N-Gram graphs.* In Proceedings of the 3rd ACM SIGMM international workshop on Social media (WSM '11). ACM, New York, NY, USA, 9-14.

[15] Lihua Hao, Christopher Healey, and Steve Hutchinson. 2013. *Flexible web visualization for alert-based network security analytics.* Proceedings of the Tenth Workshop on Visualization for Cyber Security. ACM, New York, NY, USA, 33-40.