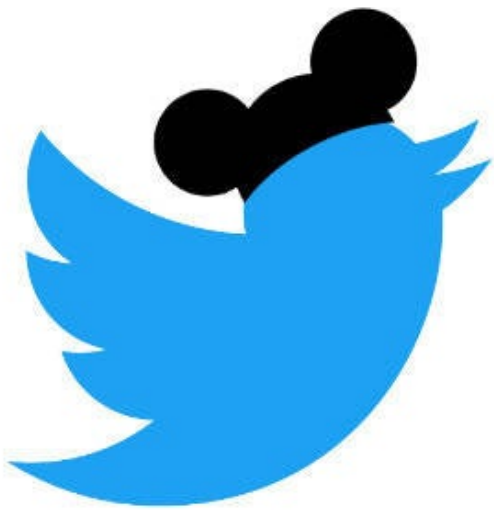
Alexander M. Atzberger
Kroon, mw. dr. A.C.
Big Data, Group 2
28 May 2020

Predicting Twitter Sentiment with Machine Learning



Introduction

"I only hope that we never lose sight of one thing – that it was all started by a mouse." - Walt Disney

The Walt Disney Company has come a long way from its origin: nobody has predicted a small scale entertainment company from rural Kansas City, Missouri turning into one of the biggest media companies 100 years later. Walt Disney's corporate portfolio nowadays studio entertainment, direct-to-consumer streaming platforms, theme parks and multiple media networks (Seymour, 2020). The projections for 2020 were ambitious, with initial plans to open new Star Wars theme parks, launch a new iteration of Marvel movies and the production of much more. Fast forward to May 2020: the global pandemic COVID-19 brought "business as usual" to a grinding halt an inevitably forced most of the world into a stay-at-home situation. For Disney that meant that 75% of their business has been shut down and halted for the past three months. How does their most valuable asset, their customer, talks about them in these times? What is the Groundswell talking about one of the largest entertainment organisations in times of crisis?

With a total of 6,7M followers on Twitter (As of 25 May 2020), Disney is a closely monitored organisation across the world. For communication professionals within the organisation, it is crucial to periodically monitor the sentiment that surrounds their organisation and get external, unobtrusive feedback as to what the people think of the organisation. Research has indicated that positive sentiment towards an organisation positively affects attitude towards an organisation (Liu et al., 2017), better brand recall (Mostafa, 2013) and brand perception (Cambria, 2016). This becomes even more compelling when being able to predict Sentiment with the help of Machine Learning.

Further with regards to Twitter, research has shown that online user generated content has become a crucial source that poses critical influence on customers' brand perception, brand reputation, purchase decision making, and profitability (Browning et al., 2013, Vermeulen & Seegers, 2009)

The emergence of improved computational power and the ability to analyse large datasets with the help of natural language processing with Machine Learning has been a development which compares to a revolutionary theme park: with little personnel labour, it is possible to develop an algorithm that learns and evaluates selected text material (Langley, 2011). The refinement of NLP techniques and the ability to process large data has been at the forefront at the data driven approach that organisations use to make more educated guesses about their consumers.

Further, on a more experimental level, research has shown a meaningful long-term correlation between investor sentiment and the stock price of US companies (Fang & Huang, 2018). Another study has indicated the impact of quarterly financial statements of public organisations on the short-term movement of the stock market (CITATION). Even though no connection between public sentiment and short-term stock market movement, it would be interesting to explore the relationship between the two.

The following research question and sub-questions have emerged from the following paragraphs:

RQ1: Can we predict consumer sentiment with the help of Machine Learning for the Walt Disney Company?

RQ1a: What is the general sentiment of Twitter followers of the Walt Disney Company?

RQ1b: Does the aggregated sentiment of Twitter followers correlate with the stock data movement of the Walt Disney Company?

Overview of Analytic Strategy

To test the research questions above, three initial phases were conducted in the research project, a scraping phase, a wrangling phase and the analysis phase.

Scraping Phase

A total of 12,485 tweets that included the hashtag #disney were scraped with the help of the official Twitter API. Due to limited access of its functionalities, it was possible to scrape recent tweets with a daily limit of 1,600 tweets. For this analysis, tweets from 13 to 22 May were utilised and scraped daily and saved as a .csv file. To limit duplicate tweets in the dataset, retweets were excluded from the data collection.

Wrangling Phase

In a second step of preprocessing the tweets, the data then was gathered was then imported into one large Pandas data-frame and split up into the columns "Date" and "Tweet". The time stamps were reduced to year, month day from and the string of the tweet was shortened to remove the username.

Analysis Phase

The last step of the analysis itself consisted of three phases, the sentiment analysis, the Machine Learning algorithm and the experimental correlation between the sentiment and the stock market data.

As the first step, a sentiment analysis was conducted for two purposes: to get an overview of the initial sentiment, but more importantly to categorise the tweets into sentiment categories. The initial sentiment score between -1 and 1 was used to create three categories, negative, neutral and positive to utilise for the training and implementation of the machine learning algorithm.

In the second step, the dataset was split into a training, validation and testing dataset. The machine learning algorithm includes three classifier, the Naive Bayes, Logistic Regression and SVM. This way we are able to compare how more simplistic (Naive Bayes) compare against more elaborate classifiers such as the Logistic Regression and the SVM in the prediction of Twitter Sentiment. The data was tested with both a Count and TFID vectorisers and compared which of the two was more accurate. For both vectorisers, English stop-words were removed to single out any potential words that could bias the results. In a second step, the ROC/AUC measures were used to determine the most suitable cut off point of each model. The most suitable classifier in the end was used on the testing data, with the ROC/AUC measures adjusted to find the optimal cut off point for the final model. The final vectoriser and classifier were then saved for future use.

In a last step, the Twitter sentiment was aggregated based on date and compared to the stock movement data of the same time frame. The correlation and a correlation heat map indicate the quality of the relationship of the sentiment.

Script

The project entails three Jupyter notebook files, the first one being a script for the collection of the data, the second notebook including the script for preprocessing the data and the third for analysing the data. Output documents include the raw data file, the master data frame with added sentiment scores and classification, the saved Machine Learning Script, the Disney Stock, its plot and a correlation plot.

For the analysis the following packages were utilised:

- Matplotlib, Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

- NLTK, Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."

- Numpy, Oliphant, T. E. (2006). *A guide to NumPy* (Vol. 1). Trelgol Publishing USA.

- Pandas, McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference, 445*(1), pp. 51–56).

- Scikit Learn, Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(10), 2825–2830.

- Tweepy, Roesslein, J. (2020). Tweepy: Twitter for Python! *URL:* https://Github.Com/Tweepy/Tweepy.

Conclusion

Overall, the results of the analysis yield interesting insights surrounding the Twitter data on the Walt Disney Company during the time of the pandemic COVID-19. Although there appears to be an overall neutral sentiment and no correlation between stock price data and twitter sentiment, the introduced algorithm is able to detect positive, neutral and negative twitter comments. The implications of the findings will be discussed below.

RQ1: Can we predict consumer sentiment with the help of Machine Learning for the Walt Disney Company?

Yes we can. After running the ML script through three different classifiers, the Naive Bayes, the Logistic Regression and the Support Vector Machine with two different vectorisers, the count and TFID, it became evident that the SVM classifier is the most suitable method with the most balanced precision (.80) and recall (.76) and F1 score (.78). After examining the optimal threshold for the classifier, the recall score improved slightly from .74 to .76.

It was expected that the Naive Bayes would perform least accurate with the count vectoriser, *F1 = .69*. The simple classifier was designed to predict mostly binary outcomes, not three categories. Nevertheless, it serves as a baseline to compare the two more complicated models against. Examining the logistic regression and the SVM with the count vectorisers, it appears that both models are good classifiers, with the SVM classifier having a more balanced precision and recall (LG: Precision: .81, Recall: .74; SVM: Precision: .80, Recall: .75). The latter model has then be used to test the final data. After testing for the optimal threshold with the testing data and the count vectoriser, the F1 score improved by .01 to a final of F1 = .78.

Unfortunately, the ROC/AUC scores are sub-optimal. Of the classifiers, the Logistic Regresssion scored the highest (.25), the final classifier of the SVM scored

.16. This indicates that our model is not suitable to categorise between True Positive Rate against the False Positive Rate at various threshold settings. The model as it is, is unfitting to predict Twitter sentiment.

There is plenty of room for improvement for the presented model, as it does not appear to be the most accurate algorithm to detect sentiment of tweets. More specifically, trying to predict a sentiment score based on a 140 character strings can prove to be difficult. Despite filtering out retweets, a brief glance at the scraped datasets indicates a large number of tweets ends up being simply spam or false advertisements. Future research should examine literature that finds support for different features to predict sentiment and take those into account when running the analysis. Additionally, based on the presented findings it is difficult to say if twitter followers of an entertainment organisation score similar sentiment scores than consumers of a B2B organisation, who most likely use different words, hashtags and ways to interact with their respective organisation. As interesting as it would have been to examine different organisations with multiple features, this would have overextended the scope of the presented research project. Nevertheless, this analysis was an important first step in assessing sentiment of Twitter users and can be utilised as a platform for future more in-depth analysis.

RQ1a: What is the general sentiment of Twitter followers of the Walt Disney Company?

We can conclude that the general sentiment of Walt Disney Companies Twitter followers is rather neutral (Overall Sentiment Score = 0.02). Despite filtering out retweets and wrangling the raw data, the presented Twitter sentiment is difficult to compare to, as the current global pandemic takes an emotional toll on people, yet Disney offering a streaming platform might increase the mood of some fans. Additionally, due to the limitations of the Twitter API, the sentiment of merely 8 weekdays was collected and examined. This leaves little room to examine a change of mood or compare it to less uncertain times. It would be highly interesting if future research would be able to sample a tweets from different point in times and examining them across time. Regardless, the sentiment analysis served on the one hand to classify the tweets for the Machine Learning algorithm, and it gives an interesting perspective that even in times in which Walt Disney utilising 1/4 of their organisational capacity, its followers and fans are not in a negative mood.

RQ1b: Does the aggregated sentiment of Twitter followers correlate with the stock data movement of the Walt Disney Company?

The results of the exploratory correlation between stock closing data and sentiment aggregated per day over the course of the 13-22 May indicates no correlation whatsoever $r(8) = 0.21$, $p = .615$. Despite the meaningful impact of investor sentiment on an organisations stock market , user generated content appears not to have any effect at all. This points to the shortcoming of the time frame of the data collection: the time frame is short and nothing signifi cant happened during these days. Future research might examine the impact of sentiment when large events such as the annual D23 conference, or the opening of a new theme park occur and compare the sentiment to the stock market price.

Overall, the results yield important findings: the machine learning script is a vital attempt (not yet perfect) to predict the sentiment of Twitter followers of the Walt Disney Company. With more access to Twitter data from different times, researchers might be able examine meaningful organisational moments and give an educated outlook, how their customers reacted. This is even more relevant for communication practitioners who conduct regular pulse checks of their fanbase online: sentiment checks with the help of the sentiment analysis and ML algorithm can help map whether customers like or dislike a new product, movie or theme park. The presented script is an outline of what is possible within the emerging world of natural language processing. As Walt Disney said himself, *"it is kind of fun to try to do the impossible".*

References

Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., & De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods & Measures, 8,* 190–206.

Browning, V., So, K.K.F., Sparks, B. (2013). The influence of online reviews on consumers' attributions of service quality and control for service standards in hotels. *J. Travel Tourism Mark. 30* (1–2), 23–40.

Cambria, E. (2016). Affective Computing and Sentiment Analysis. *IEEE Intelligent Systems, 31*(2), 102-107.

Langley, P. (2011). The changing science of machine learning. *Machine Learning, 82* (3), 275–279.

Liu, X., Burns, A. C., & Hou, Y. (2017). An Investigation of Brand-Related User-Generated Content on Twitter. *Journal of Advertising, 46*(2), 236-247.

Fang L., Yu H., Huang Y. (2018). The role of investor sentiment in the long-term correlation between U.S. stock and bond markets. *International Review of Economics & Finance, 58*(2), 127-139.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 40*(10), 4241-4251.

Seymour, L (2020). Cutting losses amid pandemic, Disney closes 'Frozen' on Broadyway. Retrieved from https://www.forbes.com/sites/leeseymour/

2020/05/14/cutting-losses-amid-pandemic-disney-ices-frozen-on-broadway/#13ee0ef216b1

Vermeulen, I.E. & Seegers, D. (2009). Tried and tested: the impact of online hotel reviews on consumer consideration. *Tourism Manage, 30* (1),123–127.