Sentiment and Stock Price Analysis of the Apple Company Using Tweets

Natural Language Processing Group Project

Jiaye Tang
Computer Science Faculty
Dalhousie University
Halifax NS Canada
jy661114@dal.ca

Xi Rao
Computer Science Faculty
Dalhousie University
Halifax NS Canada
xi.rao@dal.ca

Kris Lin
Computer Science Faculty
Dalhousie University
Halifax NS Canada
Kris.Lin@dal.ca

ABSTRACT

As social media plays a significant role in daily life, emotions and sentiments on those platforms could have a significant impact on people's general opinions and cognition toward products and companies. Even though there are no unified conclusions to the question that whether public sentiments lie a relationship to stock price changes, it is still worthwhile to analyze the effect of sentiments. Sentiment analysis would help companies gain insights from social media platforms and find out if public opinions would cause the stock price to change or be affected by sentiments. Besides, studies in this field could also help investors to collect more information about companies, avoid losses and act accordingly. The project only investigated the sentiments in tweets and the stock price change of Apple Inc. from 2015 to 2019. The goal of this project is to explore sentiment distribution and find out patterns. Moreover, we would like to discover the relationship between sentiment and stock price changes. Popularity, which is the sum of the number of retweets, likes, and comments, was added to exclude less important information. We applied the NLTK toolkit to preprocess tweets, including removing user ID, URL, stopwords, tokenization, lemmatization, etc. VADER and AFINN dictionaries were used to tag the words into "Positive", "Negative", and "Neutral" sentiment. We also found out the stock price gain had a slight correlation with the sentiments. Additionally, sentiments had a stronger correlation with stock close value and a slight negative correlation with stock volume. Meanwhile, we evaluated the performance of two machine learning models, the Naive Bayes classifier and the Support Vector Machine (SVM). The results showed that SVM had a better performance than Naïve Bayes Classifier with both AFINN and VADER sentiment classified label.

1. INTRODUCTION

Stock price analysis is an important topic in the financial industry, whether in building forecasting models or conducting corresponding business activities. From past studies and research, reasons that cause the changes in stock price can be various, for example, the relationship between volume and stock prices. As social media plays a significant role in daily life, emotions and sentiments on those platforms could have a significant impact on people's general opinions and cognition toward products and companies. Because of that, it raises interest in researching the relationship between sentiments and market fluctuations. According to the Efficient Market Hypothesis (EMH), it states that stock market prices are largely driven by new information and follow a random walk pattern (Mittal, Goel, 2011).

However, sentiment analysis on social media is still facing a lot of challenges. Short text, misspellings, uncommon grammar constructions, hashtags, and pictures affect the correctness of results. Some researchers reported that sentiments from social media have no relation between web buzz and stock prices

(R. F. 2001), while other researchers have reported either weak or strong predictive capabilities (Mao, et al. 2015).

Even though there are no unified conclusions to this problem, it is still worthwhile to analyze the effect of sentiments. Because social media has become so common even being an information source that people cannot ignore it. Analyzing investors' sentiments is critical for both companies and investors themselves. Sentiment analysis would help companies gain insights from social media platforms and find out if public opinions would cause the stock price to change or be affected by sentiments. Companies can make strategies to amplify the confidence of investors based on the analysis. Besides, studies in this field could also help investors to collect more information about companies, avoid losses and take action accordingly.

To specify our research, the project only investigated the sentiments in tweets and the stock price change of Apple Inc. from 2015 to 2019. Twitter, a popular social media platform for all-age users, plays a significant role in daily life, and Apple Inc. is one of the most influential companies. The goal of this project is to explore sentiment distribution and find out patterns. Moreover, we would like to discover the relationship between sentiment and stock price change. To exclude less important information, we added a feature, popularity, which is the sum of the number of retweets, likes, and comments. The threshold of popularity is 50. We applied the NLTK toolkit to preprocess tweets, including removing user ID, URL, stopwords, tokenization, lemmatization, etc... VADER and AFINN dictionaries were used to tag the words into “Positive”, “Negative”, and “Neutral” sentiment. Also, we found out the stock price gain had some correlation with the sentiment while the correlation was insignificant. The sentiment had a stronger correlation with stock close value and a slight negative correlation with stock volume. Then we evaluated the performance of two machine learning models, Naive Bayes and Support Vector Machine (SVM) with the AFINN and VADER labelled data. The results showed that the Neutral sentiment is well predicted, followed by the positive sentiment classification, and the negative sentiment prediction is very low, which means in the respect of negative sentiment classification, the lexicon-based approach: VADER and AFINN have a relatively large difference with the supervised machine learning approaches Naïve Bayes classifier and SVM. Depending on the results of evaluation metrics, SVM achieves a better score than the Naïve Bayes classifier. The metrics of the prediction based on the VADER label and AFINN label have no significant difference.

2. RELATED WORK

Sentiments could affect decision-making since most people usually search for reviews before making decisions. Because sentiment analysis is popular, lots of research has been done in this field, especially in terms of Twitter data. The following are previous studies that have contributed to sentiment analysis in the past few years.

Some common approaches of sentiment analysis available in the literature include the subjective lexicon approach, the machine learning approach, and the hybrid approach (D’Andrea et al. (2015) compared the pros and cons of different approaches. From their research, the main advantage of machine learning is its flexibility. Models can be trained and built for specific purposes and contexts. On the flip side, the main advantage would be its low applicability to new data since labelling new data can be costly or prohibitive. The lexicon-based approach does not need any prior training. It uses a predefined list of words, where each word is associated with a specific sentiment. Since the lexicon is widely and generally used for all purposes, it is hard to create a unique lexical-based dictionary to be used for different contexts. In the Hybrid method, a sentiment lexicon is constructed using public resources for initial sentiment detection and then sentiment words as features in the machine learning method. Our study is to analyze tweets and find out the relationship between stock price and tweets of Apple Inc., and we do not need a unique model. Thus, we chose lexicon-based methods for our solution.

Based on previous studies, VADER (Valence Aware Dictionary and Sentiment Reasoner) dictionary is one of the most common and suitable lexicons for analyzing sentiments of tweets because of its better articulation in online media settings (Elbagir, Yang, 2019). VADER also examines other linguistic and grammatical varieties, such as punctuation, capitalization, and the utilization of emoticons (Alsaeedi, Khan, 2019). Another study compared the difference in sentiment analysis between using punctuation marks and not (Oad, et al. 2019). Based on the study, they found that the Exclamation (!) and Question (???) marks increased the Positive and Negative Polarity scores and decreased the Neutral polarity score. In our study, we use dictionaries, VADER and AFINN to analyze tweets and try to contribute some findings.

Furthermore, in recent years, a large number of supervised learning approaches have been used to do the sentiment classification, in paper (Kristiyanti et al., 2018), the authors collected the data of Twitter public comments of the West Java Governor Candidate, preprocess the data with tokenization, n-grams and stemming, and compare the two classifiers: SVM and naïve Bayes classifier on the sentiment analysis. Both the classifiers obtained reliable results and some of the results are affected by n-grams and Indonesian language text where the major data is in English. In our project, after labelling the dataset with VADER and AFINN, we will also perform the sentiment analysis with supervised learning approaches: Naïve Bayes and SVM classifiers and compare the two classifiers' performance.

3. PROBLEM DEFINITION AND METHODOLOGY

Here we present the framework of our project in Figure 1.

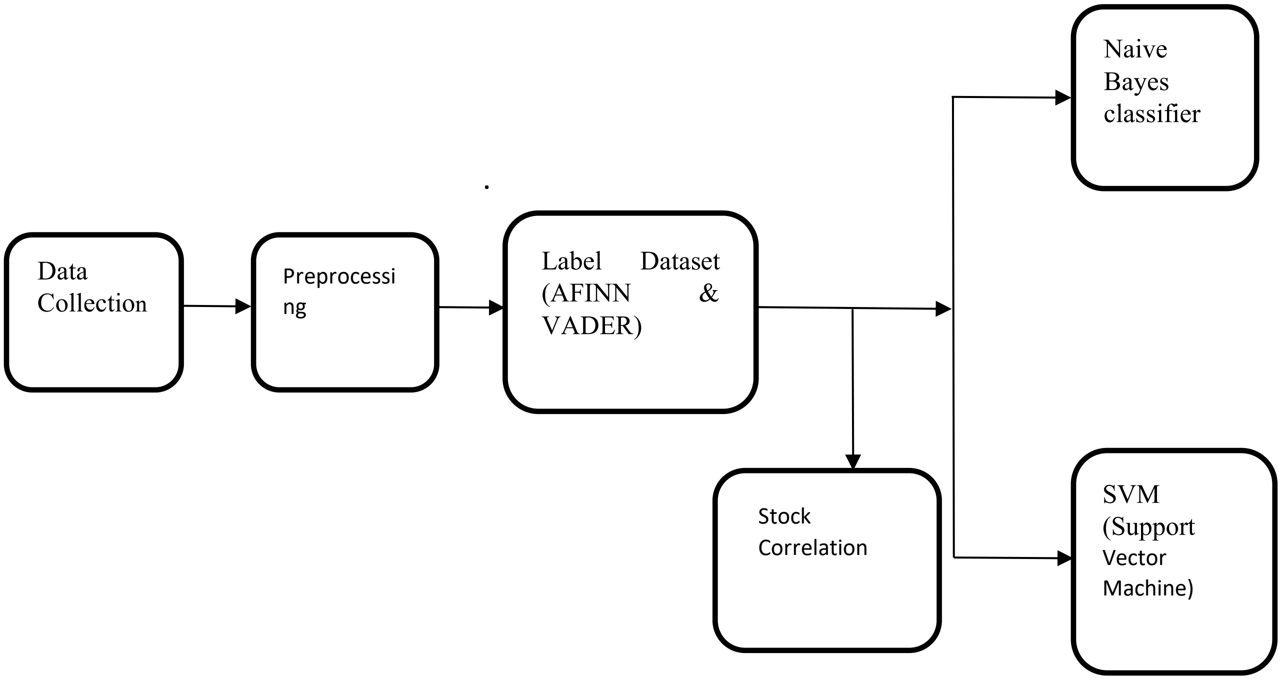


Figure 1. Framework of the proposed structure

3.1 Data Collection

Our study analyzes the sentiments of Apple Inc. on tweets and explored how they related to stock price. Two datasets we used are the tweets dataset and the stock price dataset. The tweets dataset is from Kaggle (Tweets about the Top Companies from 2015 to 2020), it includes three files, Tweet.csv, Company_Tweet.csv, and Company.csv. The original tweets dataset includes tweets that were written about five companies, Amazon, Apple, Google, Microsoft, and Tesla. Those tweets were extracted by appropriate share tickers of each company, and the period was from 2015 to 2019. It contains over 3 million unique tweets with related information such as tweet id, author of the tweet, post-date, the text body, and the number of comments, likes, and retweets of the tweets matched with the related company.

The stock price dataset is from Kaggle (Values of Top NASDAQ Companies from 2010 to 2020) which contains daily OPEN, CLOSE, VOLUME, HIGH, and LOW values of Amazon, Apple, Google, Microsoft, and Tesla companies as tagged by dates.

3.2 Data Preprocessing

3.2.1 NLTK toolkit

NLTK is a free open-source Python package that provides several tools for building programs and classifying data. NLTK is suitable for linguists, engineers, students, educators, researchers, and developers who work with textual data in natural language processing and text analytics. (Elbagir & Yang, 2019) We will use the NLTK toolkit for data preprocessing.

3.2.2 Preprocessing

First, we transformed date data format, combined files, and filtered stock price of Apple Inc. from 2015 to 2019. Post_date data under tweets was in form seconds data, so we converted it to a day format. Since only Company_Tweet.csv includes a ticker symbol, we merged Company_Tweet.csv and Tweet.csv files through the same feature, tweet_id.

Next, we supposed retweets, likes, and comments numbers have an impact on Twitter, which will influence the stock price of the company. Thus, we added a new column “popularity” to the dataset. Popularity is the sum of the retweets, likes, and comments numbers. From Figure 1, the popularity is mostly between 0 to 50. We assumed that tweets with a popularity of less than 50 have a minimal impact on Twitter and the company. Thus, we dropped the rows whose popularity is less than 50, and we had 5834 data points. The threshold of popularity is 50. We used this dataset as our main dataset in the later sentiment analysis.

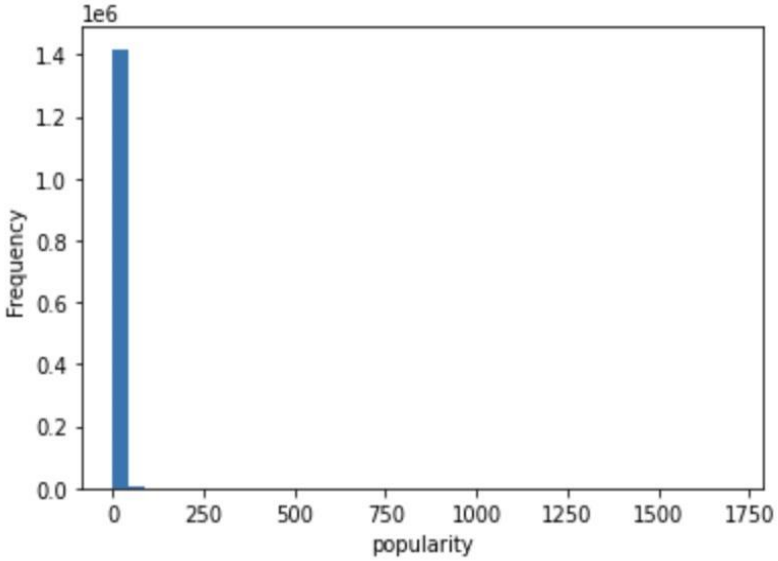


Figure 2. The histogram of the popularity of tweets

To acquire more accurate results, we used Python and NLTK toolkit to transform texts. Tweets usually contain some unnecessary information, such as URLs, punctuations, usernames etc... However, we believe that words in hashtags contain some information, so we did not remove the hashtags. In our dataset, the top five popular hashtags words are #apple, #bitcoin, #trading, #stocks, and #crypto. Those hashtags did contain some information.

The text transformation steps included:

- Removed username;

  The username starting with '@' is useless for obtaining the sentiments of the whole text.

- Removed RT beginning words;

  The old-style retweet texts have no essential information about sentiments.

- Removed URLs;

  URLs are links of websites and have no valuable knowledge for getting the sentiments.

- Removed punctuations;

  Punctuations are symbols like commas, quotes, apostrophes, and so on. In case they will add noise to sentiment analysis, we remove the punctuation.

- Removed single numbers;

  Single numeric terms in the tweets seem to have no relationship with sentiments.

- Removed stopwords;

  Stop words are a set of commonly used words in a language. Examples of stop words in English is "a", "the", "is", "are" etc. (Ganesan, 2019) These words are neutral and inappropriate for sentiment analysis.

- Tokenization

We used the word_tokenize() method in the NLTK module to separate each word in the tweet list.

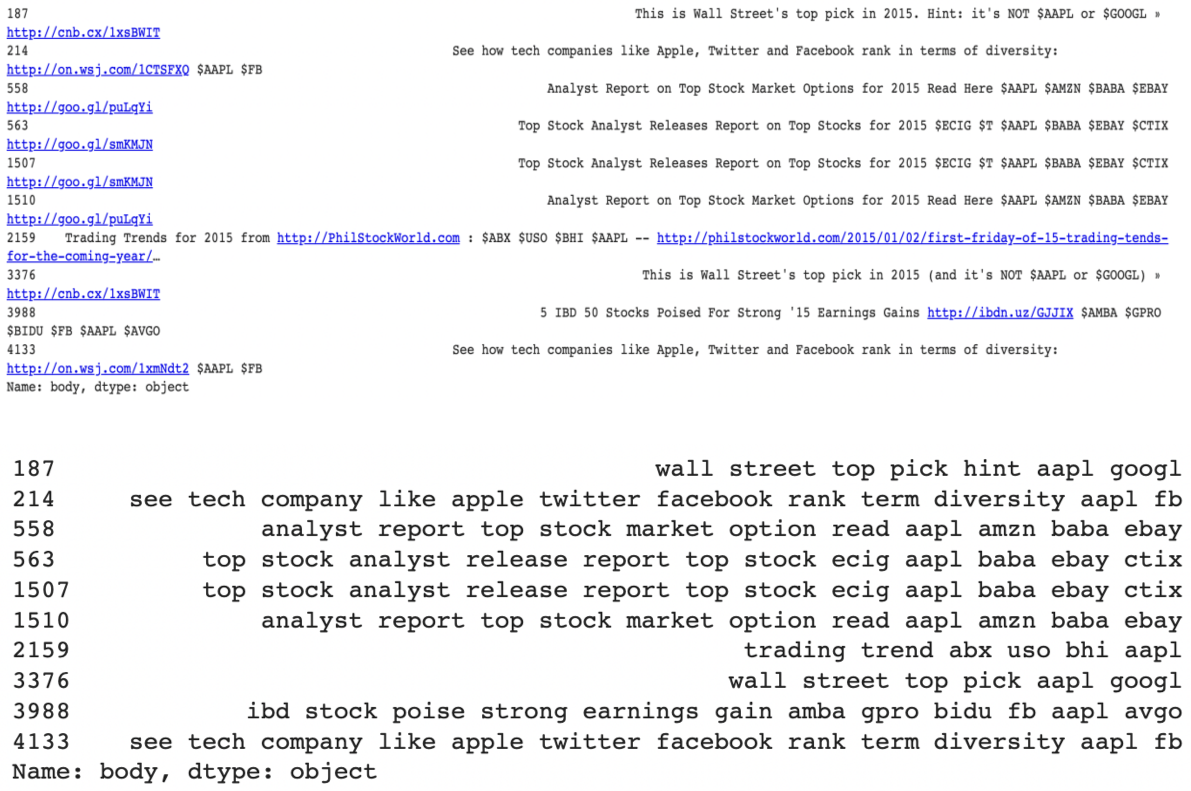
- Lemmatization

  Lemmatization is a linguistic term that means grouping together words with the same root or lemma but with different inflections or derivatives of meaning so they can be analyzed as one item. For example, to lemmatize the words “cats,” “cat’s,” and “cats’” means taking away the suffixes “s,” “’s,” and “s’” to bring out the root word “cat.” (TechSlang, 2022)

- Pos-Tagging

  Part-of-speech tagging is the automatic text annotation process in which words or tokens are assigned part of speech tags, which typically correspond to the main syntactic categories in a language (e.g., noun, verb) and often to subtypes of a particular syntactic category which are distinguished by morphosyntactic features (e.g., number, tense). (Clarin, 2022) Here we use pos tagging to tag the tokens so that the lemmatization could use the contextual information the taggers provide to choose the appropriate lemma.

  Figure 3 shows the first 10 tweets content before and after preprocessing.



```
187                                                                                    This is Wall Street's top pick in 2015. Hint: it's NOT $AAPL or $GOOGL »
http://cnb.cx/1xsBWIT
214                                                          See how tech companies like Apple, Twitter and Facebook rank in terms of diversity:
http://on.wsj.com/1CTSFXQ $AAPL $FB
558                                                                        Analyst Report on Top Stock Market Options for 2015 Read Here $AAPL $AMZN $BABA $EBAY
http://goo.gl/puLqYi
563                                                                    Top Stock Analyst Releases Report on Top Stocks for 2015 $ECIG $T $AAPL $BABA $EBAY $CTIX
http://goo.gl/smKMJN
1507                                                                   Top Stock Analyst Releases Report on Top Stocks for 2015 $ECIG $T $AAPL $BABA $EBAY $CTIX
http://goo.gl/smKMJN
1510                                                                       Analyst Report on Top Stock Market Options for 2015 Read Here $AAPL $AMZN $BABA $EBAY
http://goo.gl/puLqYi
2159    Trading Trends for 2015 from http://PhilStockWorld.com : $ABX $USO $BHI $AAPL -- http://philstockworld.com/2015/01/02/first-friday-of-15-trading-tends-
for-the-coming-year/…
3376                                                                                    This is Wall Street's top pick in 2015 (and it's NOT $AAPL or $GOOGL) »
http://cnb.cx/1xsBWIT
3988                                                                     5 IBD 50 Stocks Poised For Strong '15 Earnings Gains http://ibdn.uz/GJJIX $AMBA $GPRO
$BIDU $FB $AAPL $AVGO
4133                                                         See how tech companies like Apple, Twitter and Facebook rank in terms of diversity:
http://on.wsj.com/1xmNdt2 $AAPL $FB
Name: body, dtype: object
```

```
187                                         wall street top pick hint aapl googl
214     see tech company like apple twitter facebook rank term diversity aapl fb
558              analyst report top stock market option read aapl amzn baba ebay
563          top stock analyst release report top stock ecig aapl baba ebay ctix
1507         top stock analyst release report top stock ecig aapl baba ebay ctix
1510             analyst report top stock market option read aapl amzn baba ebay
2159                                              trading trend abx uso bhi aapl
3376                                             wall street top pick aapl googl
3988            ibd stock poise strong earnings gain amba gpro bidu fb aapl avgo
4133    see tech company like apple twitter facebook rank term diversity aapl fb
Name: body, dtype: object
```

Figure 3. First 10 raw and preprocessed tweets

Lastly, to better analyze the correlation between the stock price and the tweet sentiments, we added one column to the stock dataset to show the gain in the stock price each day.

The value of the gain = close value of stock price - open value of stock price

3.3 Labeling Data

The lexicon-based approach was applied to label the dataset, this technique calculates the sentiment orientations of the whole document or set of a sentence(s) from the semantic orientation of lexicons. The dictionary of lexicons can be created manually as well as automatically generated (Pragnya, 2022). We chose VADER and AFINN dictionaries to label tweets into three categories, “Positive”, “Negative”, and “Neutral”. VADER scores range from -1 to 1, and AFINN scores range from -5 to 5.

4. EXPERIMENT AND RESULTS

4.1 Distribution of Sentiments

Applying VADER and AFINN scores respectively, we could illustrate the normal distribution of sentiment scores for the tweets as the Figure depicted. VADER scores have a relatively regular normal distribution, whereas AFINN scores are not obvious enough.

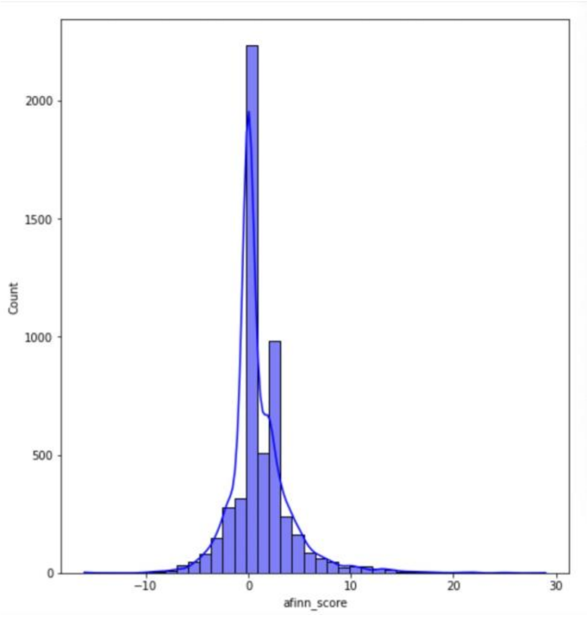


Figure 4. Distribution of compound AFINN score

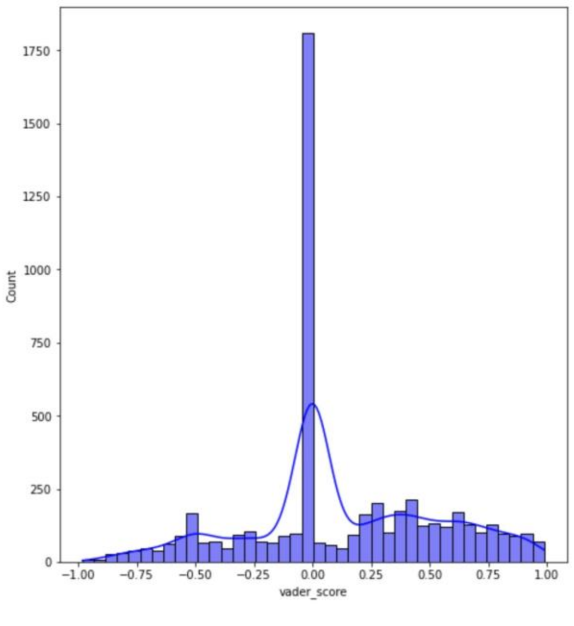


Figure 5. Distribution of VADER score

For VADER scores, we set values -0.25 and 0.25 as the threshold so that the values larger than 0.25 were labelled positive (1), the values smaller than -0.25 are labelled negative (-1) and the values between -0.25 and 0.25 were labelled neutral (0). For AFINN scores, we set the threshold value -1.0 and 1.0 to classify the tweets as neutral (0) when the score values were between -1.0 to 1.0, and the text whose values are more than 1.0 and less than -1.0 were considered as positive (1) and negative (-1) respectively.

As a result, we can have the sentiment count plots for these two dictionaries in Figure 5 and Figure 6. For both VADER and AFINN, the neutral sentiment took the largest proportion, compared with positive and negative. Specifically, positive ones were almost twice as much as negative ones in VADER scores. Additionally, we found that VADER classified more negative and positive sentiments, and less neutral sentiments than AFINN did.

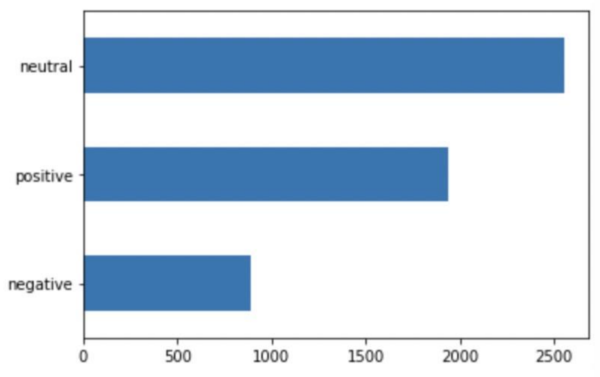


Figure 6. Count plot of VADER sentiments

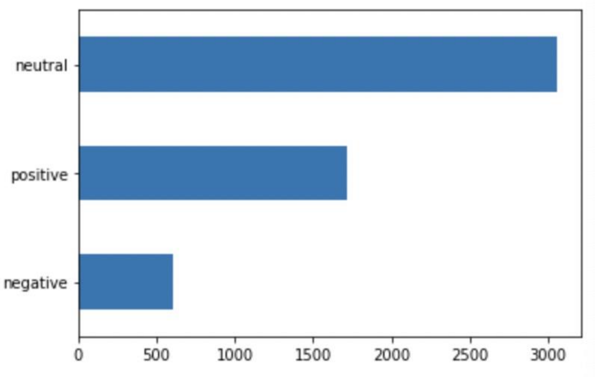


Figure 7. Count plot of AFINN sentiment

By comparing the distribution of AFINN and VADER sentiments over 5 years, we found that the proportion of positive sentiments categorized by both AFINN and VADER increased in 2018 and 2019, as we observed the total number of tweets increased in these two years. Consequently, the number of tweets with positive sentiments about Apple company increased in 2018 and 2019. However, the number of neutral sentiments remained relatively stable, the proportion accordingly decreased in 2018 and 2019 given the total number increased. Even though the negative sentiments increased accordingly as well, they still took the least part among these three sentiments. All in all, from 2015 to 2017, neutral sentiments took the largest proportion, while in 2018 and 2019 positive sentiments were in place of the position.

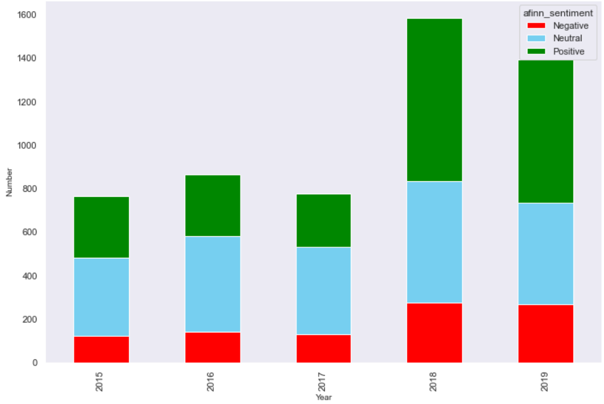


Figure 8. Changes of AFINN Sentiments over 5 years

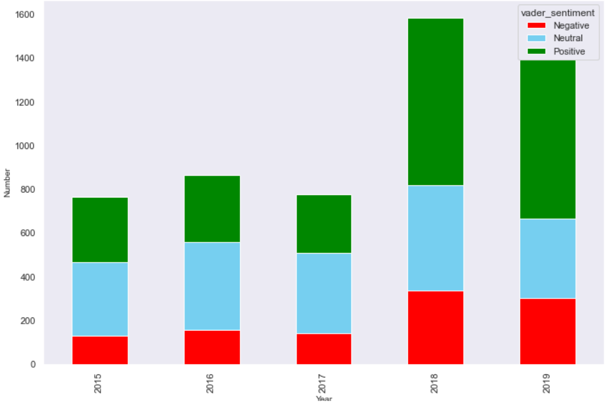


Figure 9. Changes of VADER Sentiments over 5 years

We also tried to find out if the distribution of sentiments has seasonal patterns, as users probably have different sentiments during certain months. According to the stock price changes, the stock price of Apple kept upward since the beginning of 2019, therefore, we chose this year to investigate whether the sentiment had seasonal patterns affecting stock prices. Eventually, we could obtain Figure 9 and Figure 10, which showed that the highest score was observed in March, April, and the end of the year dropped in the middle of the year. One noticeable point is that the distribution of 50% AFINN scores in each month is not apparently different, as AFINN contains fewer words and scored Tweets.

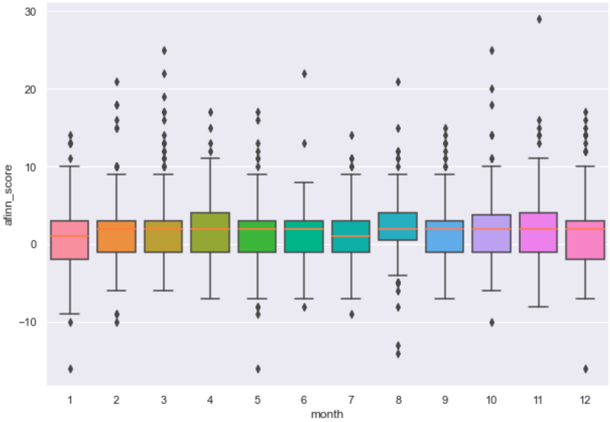


Figure 10. Monthly distribution of AFINN scores in 2019

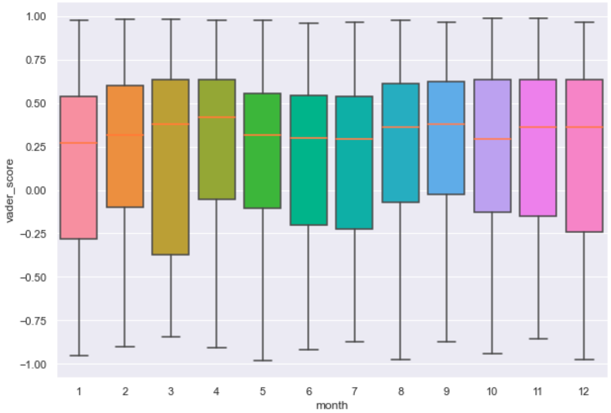


Figure 11. Monthly distribution of VADER scores in 2019

Furthermore, we also investigate the most frequently used words by analyzing the word cloud. As neutral comments have no significant effects, we excluded the neutral text and used wordcloud to show the frequent words of the positive and negative comments. From Figure 10 to Figure 13, those figures are wordcloud for positive and negative words of these two lexicons, respectively. It shows that all the figures contain Apple's NASDAQ ticker symbol: aapl. In positive words, some words like 'good', 'great', 'high', 'strong', and 'best' are frequent. In a negative word cloud, some words such as 'bad', 'don't', 'worry', and 'lose' occur often.

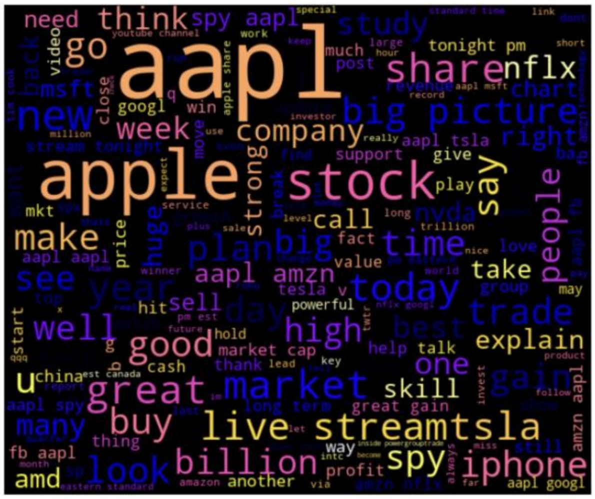


Figure 12. Word cloud for positive words of VADER

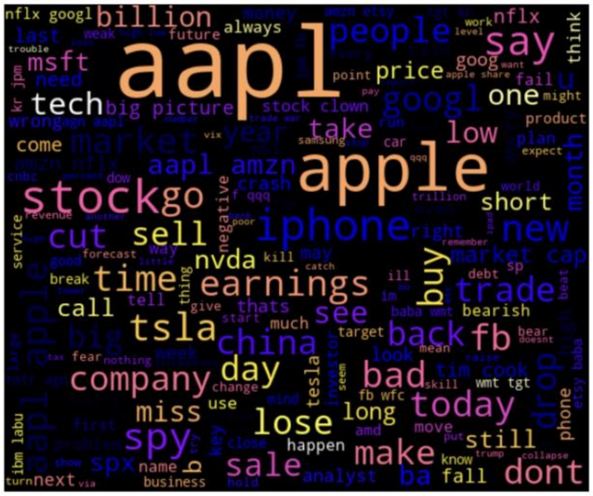


Figure 13. Word cloud for negative words of VADER

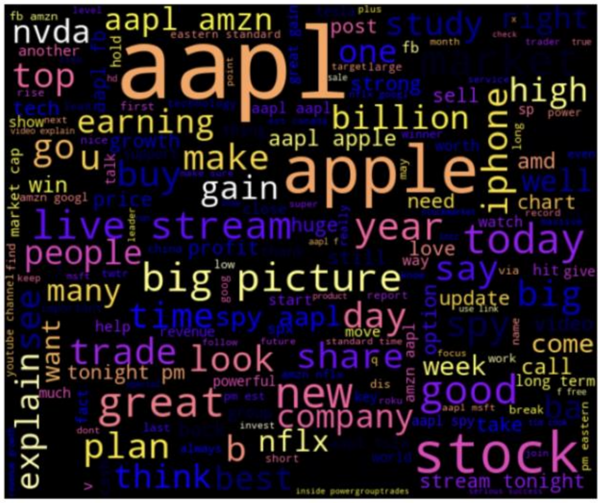


Figure 14. Word cloud for positive words of AFINN

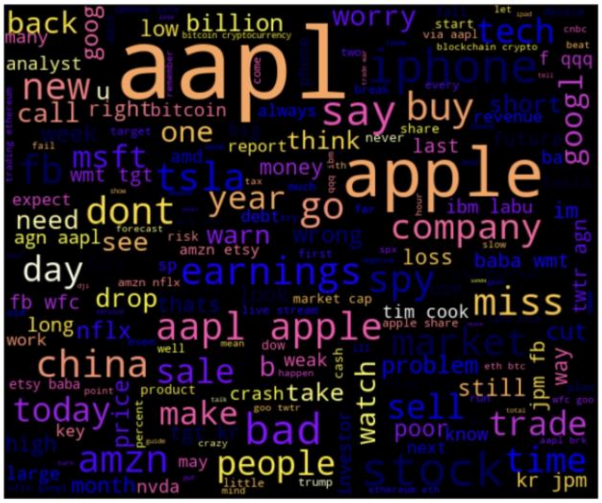


Figure 15. Word cloud for negative words of AFINN

4.2 Correlation between sentiments and stock prices

We tried to find if the sentiment score is correlated with stock price gain, close value and volume. The Spearman correlation test was applied here because we could not consider the distribution of data with this correlation. We grouped the sentiment scores by date and calculated the mean value for each day. By analyzing the daily close/last stock price change in five years, the steep downward trend happened from September 2018 to January 2019. Additionally, since July 2019, the stock price of Apple company kept increasing and reached to highest point in December 2019.

4.2.1 Spearman Correlation

Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables. (Complete Dissertation, 2022)
The Spearman correlation calculation formula is (Liu, 2021):

$$S_{x,y} = 1 - \frac{6\Sigma_{i=1}^{n}(x_i - y_i)}{n\cdot(n^2 - 1)} \tag{1}$$

Where x represents the list of score means, y represents the list of stock values, n is the number of elements, and here are the days from 2015.01.01 to 2019.12.31.
H0 (Null hypothesis): there is no correlation between sentiment and stock.
H1(Alternate hypothesis): there is a correlation between sentiment and stock
The threshold is 0.05, which means if the p-value is larger than 0.05, H0 stands and H1 rejects and if the p-value is less than 0.05, H0 rejects and H1 stands. Figure 15 to Figure 20 are figures of stock price gain, close_value, volume and the sentiment score of VADER and AFINN from 2015 to 2020.

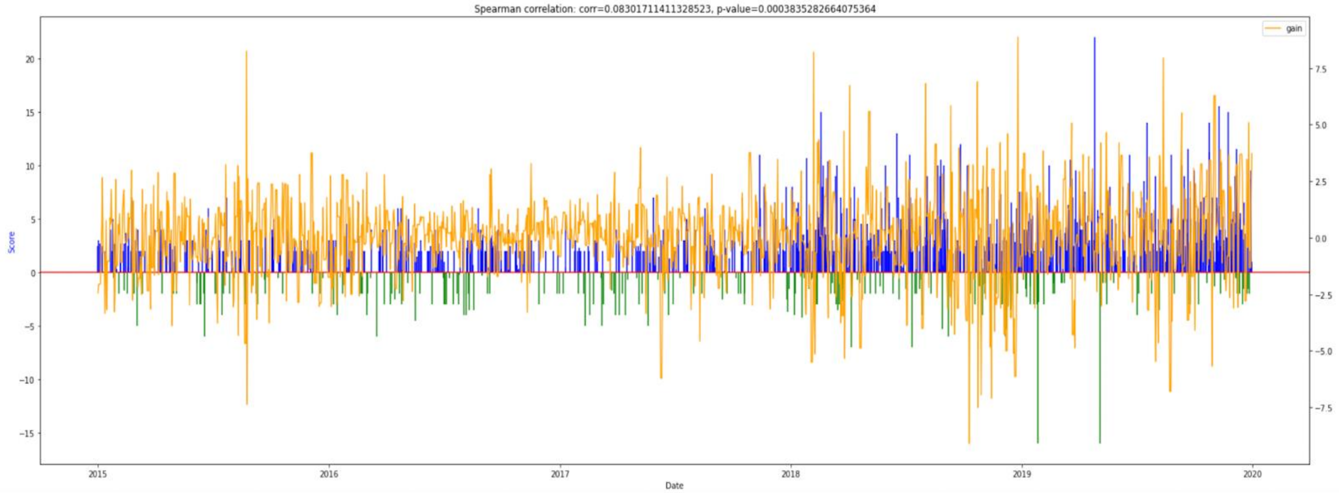


Figure 16. Correlation of stock price gain and sentiment score with AFINN

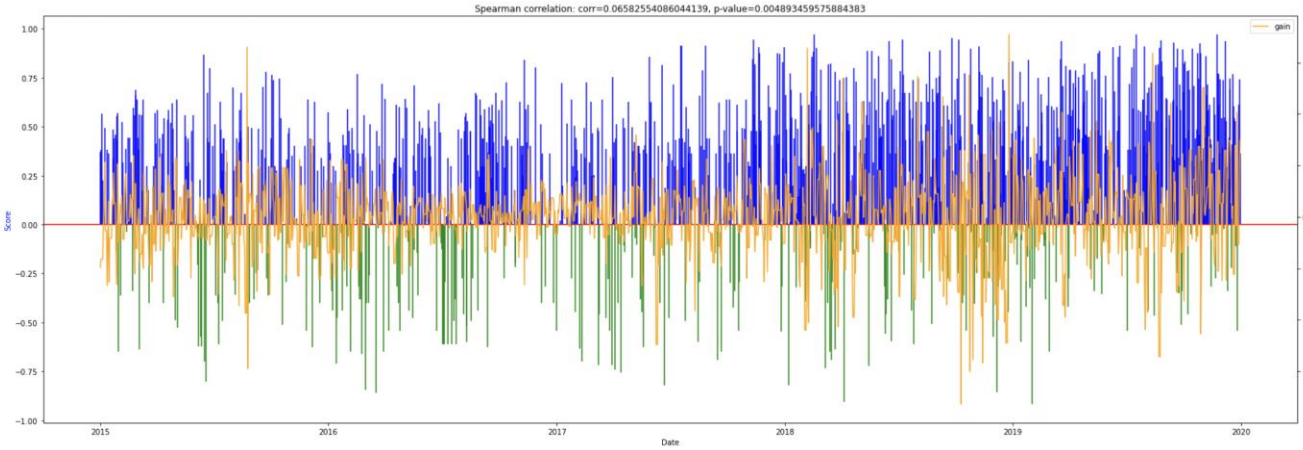


Figure 17. Correlation of stock price gain and sentiment score with VADER

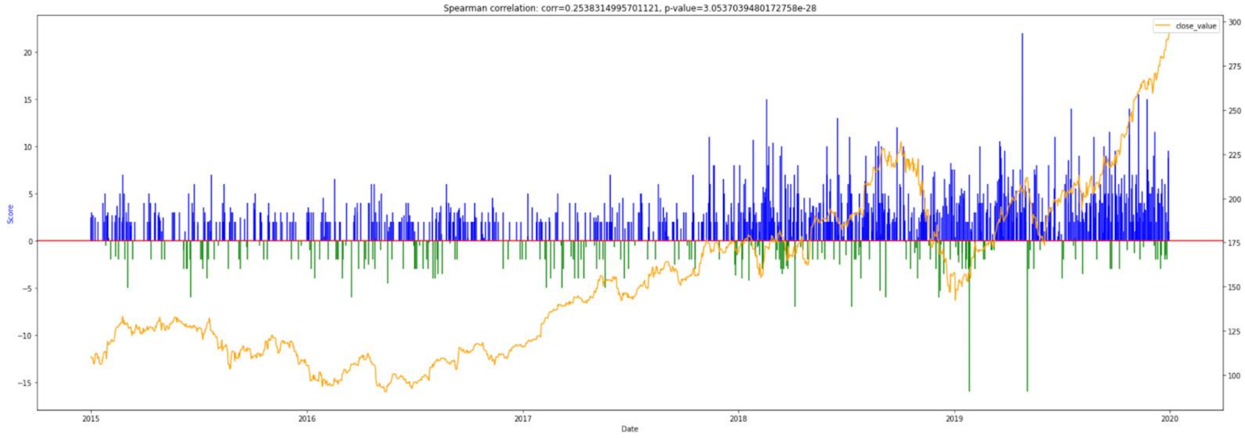


Figure 18. Correlation of stock price close value and sentiment score with AFINN

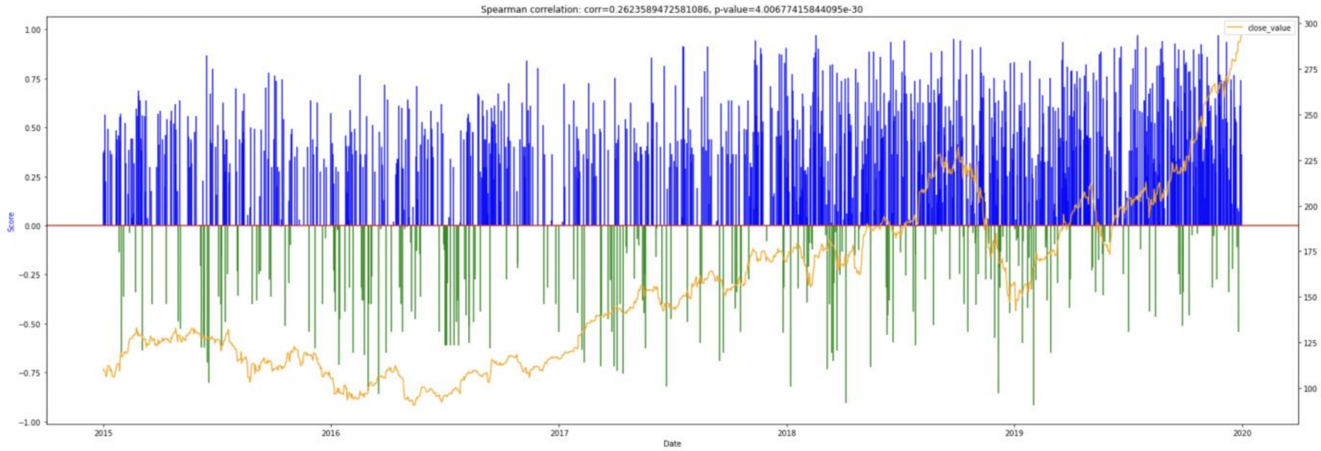


Figure 19. Correlation of stock price close value and sentiment score with VADER

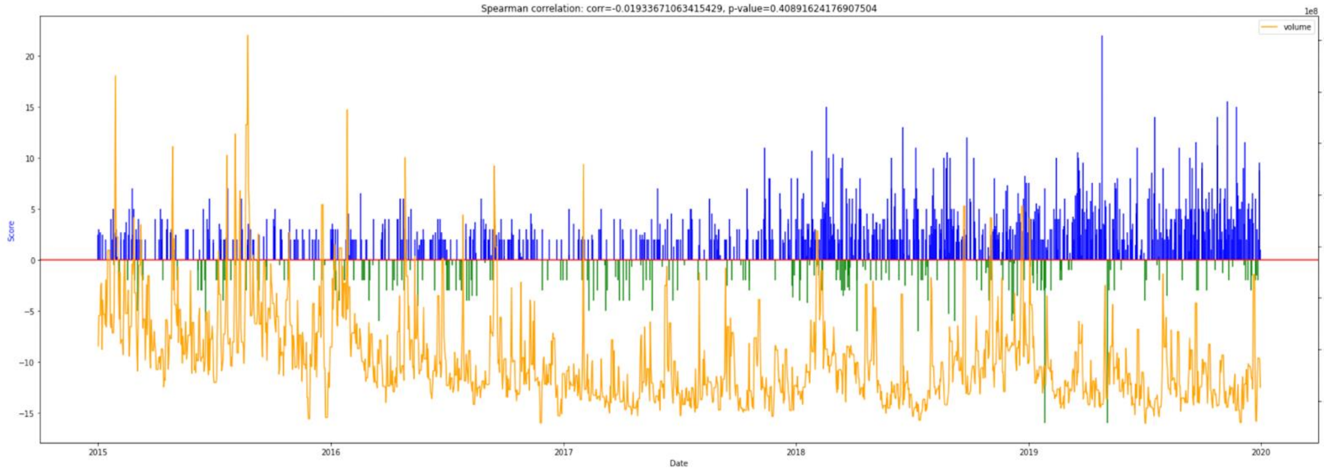


Figure 20. Correlation of stock volume and sentiment score with AFINN

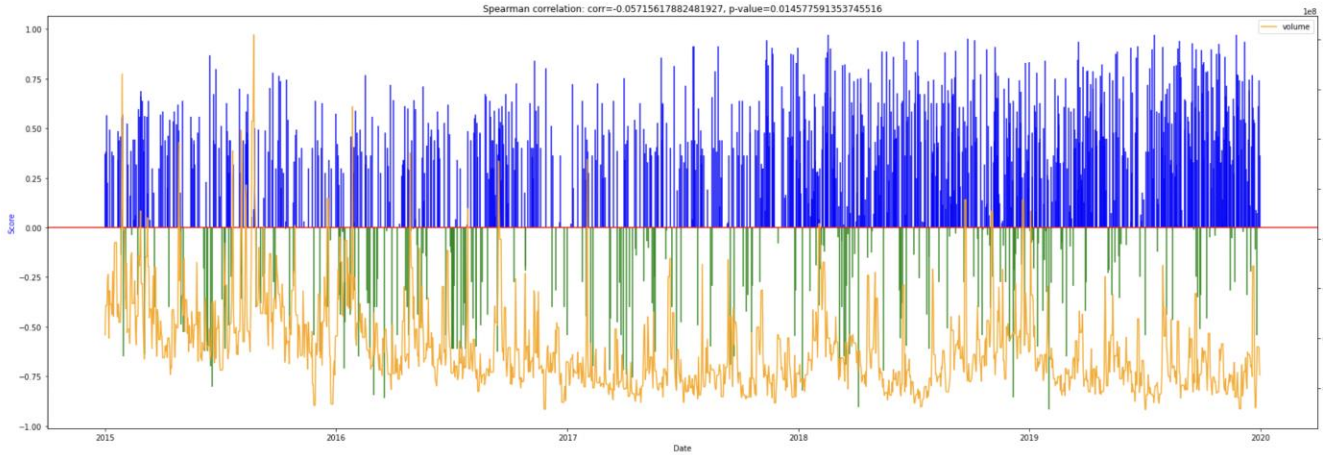


Figure 21. Correlation of stock volume and sentiment score with VADER

Table 1

Spearman’s correlation coefficient and p-value for stock and sentiment

		Stock gain	Close value	Volume
VADER	Correlation coefficient	0.06583	0.26236	-0.05716
	P – value	0.00489	4.00677e-30	0.014578
AFINN	Correlation coefficient	0.08302	0.25383	-0.01934
	P – value	0.00038	3.05370e-28	0.40892

According to the figures and table 1, for stock price gain, the Correlation coefficient of AFINN and VADER is close to 0.05 and the p-value is less than the alpha value: 0.05. Also, the p-value of AFINN sentiment is 0.00038, which is much smaller than 0.05. Based on this observation, it rejects the H0 hypothesis. This indicates that there is a relationship between stock gain and sentiment while the relationship is weak because of the small correlation coefficient. Considering the close value, both the p-values for VADER and AFINN are extremely close to 0 and the correlation coefficient value gets increased. It shows that there is an extraordinarily strong probability to reject the H0 hypothesis and the strength of the correlation between sentiment and close value is much stronger than with the stock price gain. Obviously, stock close value is strongly correlated with the sentiment. While both VADER and AFINN have negative correlation coefficient values. The p-value of VADER is less than 0.05 and the p-value of AFINN is more than 0.05. It indicates that the H0 hypothesis is rejected for VADER sentiment and is accepted for AFINN sentiment. This means that the sentiment produced by VADER has a low inverse correlation with volume and suggests that there is no significant relationship between volume and AFINN sentiment.

4.3 Sentiment Analysis with Machine-Learning Models

Before constructing and training the classifiers, we split the dataset into 80% training dataset and 20% testing dataset. Then we are able to train the classifiers, in this project we use two classifiers: naive Bayes, and SVM because they are state-of-art approaches for sentiment classification according to (Dashtipour, et al. 2016) (Mohammed, Sawar, 2022). For all the classifiers, we use the same features and dataset.

4.3.1 Feature Extraction

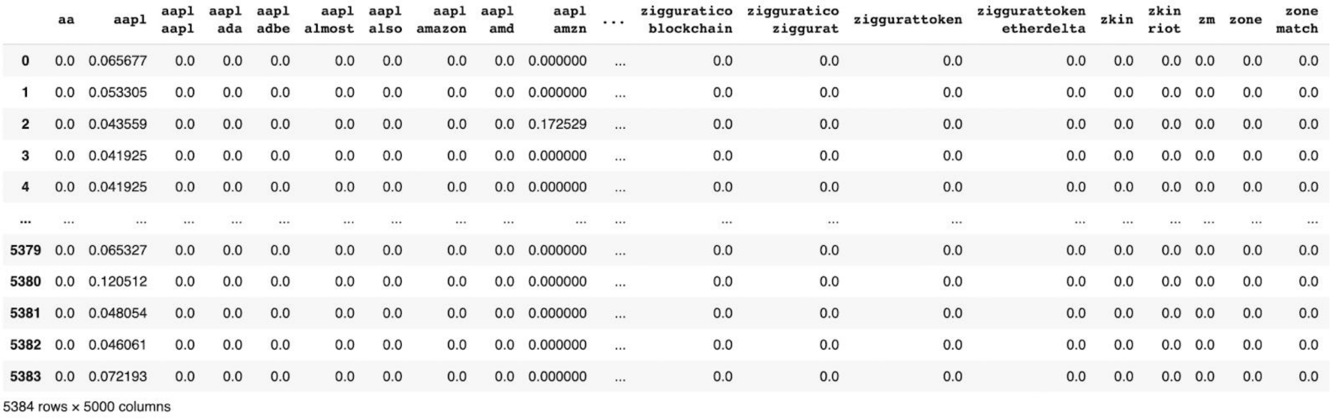
The tweet comments are unstructured data, then we have to do feature extraction before training. Feature extraction is part of the dimensionality reduction process, in which an initial set of raw data is divided and reduced into more manageable groups. So, Feature extraction helps to get the best feature from those big data sets by selecting and combining variables into features, thus, effectively reducing the amount of data. (Chatterjee, 2022) Here we use the extraction approach: IF-IDF associated with N-grams.

4.3.2 IF-IDF

The term frequency-inverse document frequency (also called TF-IDF), is a well-recognized method to evaluate the importance of a word in a document. The term Frequency of a particular term (t) is calculated as the number of times a term occurs in a document to the total number of words in the document. IDF (Inverse Document Frequency) is used to calculate the importance of a term. There are some terms like “isan”, “and” etc. which occur frequently but don’t have importance. IDF is calculated as IDF (t) = log(N/DF), where N is the number of documents and DF is the number of documents containing the term t, et al. 2019) Then the TF-IDF formula of term t is represented as:

$$w_t = tf_t \cdot \log(N / DF_t) \qquad (2)$$

Here we use TF-IDF vectorizer TfidVectorizer() with n-grams to convert the tweet text features into a matrix so that it could be the input for the classifiers. The n-grams range is set (1, 2) and 1 refers to unigram, and 2 refers to bigrams so that one token and a pair of neighbouring tokens are chosen for feature extraction. Figure 15 shows the TF-IDF matrix results with the max features of 5000.



	aa	aapl	aapl aapl	aapl ada	aapl adbe	aapl almost	aapl also	aapl amazon	aapl amd	aapl amzn	...	zigguratico blockchain	zigguratico ziggurat	ziggurattoken	ziggurattoken etherdelta	zkin	zkin riot	zm	zone	zone match
0	0.0	0.065677	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.053305	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.043559	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.172529	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.041925	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.041925	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5379	0.0	0.065327	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5380	0.0	0.120512	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5381	0.0	0.048054	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5382	0.0	0.046061	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5383	0.0	0.072193	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5384 rows × 5000 columns

Figure 22. TF-IDF matrix results

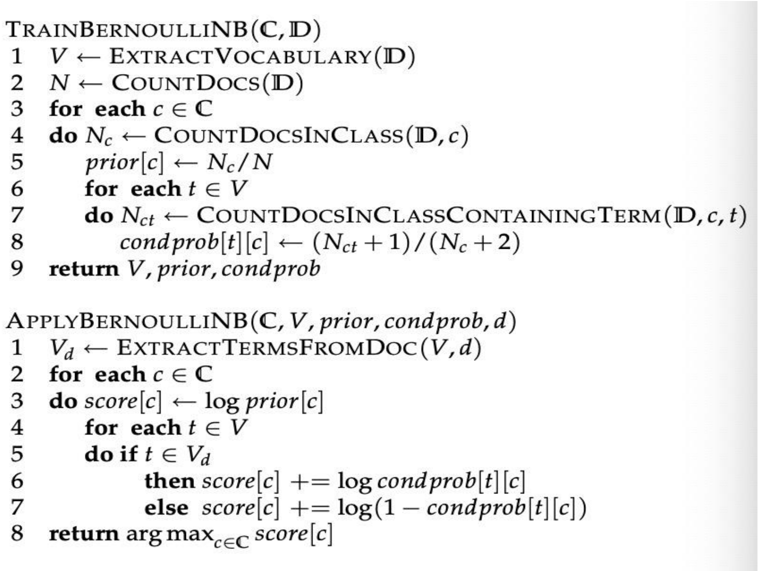
In Figure 15, word aapl appears in the first documents and last five documents, aapl amzn appears in the second document and their TF-IDF values are small, which means they are common words or conjunctions with low weights. The other tokens shown in the figure neither appear in the first five tweets nor the last five tweets.

4.3.3 Naive Bayes

A well-known Bayesian network classifier is the Naïve Bayes’ classifier is a probabilistic classifier based on Bayes’ theorem, considering Naïve (Strong) independence assumption. (Dey, et al. 2016) Bayes's theorem for a data point A of the given class B is:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \qquad (3)$$

We use Bernoulli Naive Bayes in this project. It is a classifier that works efficiently on the binary concept when the items appear or not, unlike Multinomial NB, Bernoulli NB does not notify the frequency of the term. It does not manipulate the same multinomial process where the term frequencies are considered by the multinomial approach. In contrast, the Bernoulli NB approach is only beneficial in determining the presence of a term in the text under consideration. (Ressan, et al. 2022) The algorithm for Bernoulli naive Bayes is:



```
TRAINBERNOULLINB($\mathbb{C}, \mathbb{D}$)
$V \leftarrow$ EXTRACTVOCABULARY($\mathbb{D}$)
$N \leftarrow$ COUNTDOCS($\mathbb{D}$)
for each $c \in \mathbb{C}$
do $N_c \leftarrow$ COUNTDOCSINCLASS($\mathbb{D}, c$)
    $prior[c] \leftarrow N_c / N$
    for each $t \in V$
    do $N_{ct} \leftarrow$ COUNTDOCSINCLASSCONTAININGTERM($\mathbb{D}, c, t$)
        $condprob[t][c] \leftarrow (N_{ct} + 1)/(N_c + 2)$
return $V, prior, condprob$

APPLYBERNOULLINB($\mathbb{C}, V, prior, condprob, d$)
$V_d \leftarrow$ EXTRACTTERMSFROMDOC($V, d$)
for each $c \in \mathbb{C}$
do $score[c] \leftarrow \log prior[c]$
    for each $t \in V$
    do if $t \in V_d$
          then $score[c] \mathrel{+}= \log condprob[t][c]$
          else $score[c] \mathrel{+}= \log(1 - condprob[t][c])$
return $\arg\max_{c \in \mathbb{C}} score[c]$
```

Figure 23. Bernoulli Naïve Bayes Algorithm (C. D, et al. 2008)

4.3.4 Support Vector Machine (SVM)

“Support Vector Machine” (SVM) is a supervised machine learning algorithm that can be used for both classification and regression challenges. However, it is mostly used in classification problems. (Ray, 2017) The best hyper-plane is the one which has the largest distance to the closest points in the training dataset for each class. In this project, we aim to find this kind of hyper-plane to classify the sentiments.

4.3.5 Classifier Result Evaluation

To compare the classifiers on sentiment analysis, we use precision, recall, f1- score and accuracy as evaluation metrics. Precision here measures the rate of correctly predicted tweets of all the positive tweets

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

The recall measures the percentage of the correctly predicted tweets of all the true positive tweets and the false negative tweets.

$$Recall \; = \frac{TP}{TP + FN} \tag{5}$$

F1-score uses both precision and recall.

$$F1 - score \; = \frac{2\cdot(\Pr ecision \cdot Recall)}{\Pr ecision + Recall} \tag{6}$$

Accuracy measures the rate of total correct classified tweets of all the dataset.

$$Accuracy = \frac{TP \; + TN}{TP + TF + FP + FN} \tag{7}$$

TP, FP, FN, and FP are four categories from the confusion matrix named True Positives, False Positives, False Negatives, and False Positives. True Positives and True Negatives represent the data that are perfectly predicted, and False Positives and False negatives represent the classified data that is predicted into the wrong classification.

Table 2 shows the precision, recall, f1-score, and accuracy of the naive Bayes classifier and support vector machine using the VADER label with TF-IDF matrix as input.

Tabel 2
Metrics comparison of sentiment classification for different models with VADER label

		Precision	Recall	F1-score	Accuracy
Naïve Bayes classifier	Negative	0.31	0.53	0.39	0.70
	Neutral	0.88	0.68	0.76	
	Positive	0.58	0.78	0.67	
Support Vector Machine	Negative	0.43	0.65	0.52	0.75
	Neutral	0.84	0.73	0.78	
	Positive	0.73	0.81	0.77	

Table 3 is the sentiment classification metrics using the AFINN with TF-IDF matrix as input.

Table 3
Metrics comparison of sentiment classification for different models with AFINN label

		Precision	Recall	F1-score	Accuracy
Naïve Bayes classifier	Negative	0.24	0.45	0.31	0.67
	Neutral	0.86	0.68	0.76	
	Positive	0.48	0.69	0.57	
Support Vector Machine	Negative	0.37	0.79	0.50	0.77
	Neutral	0.90	0.76	0.82	
	Positive	0.69	0.81	0.75	

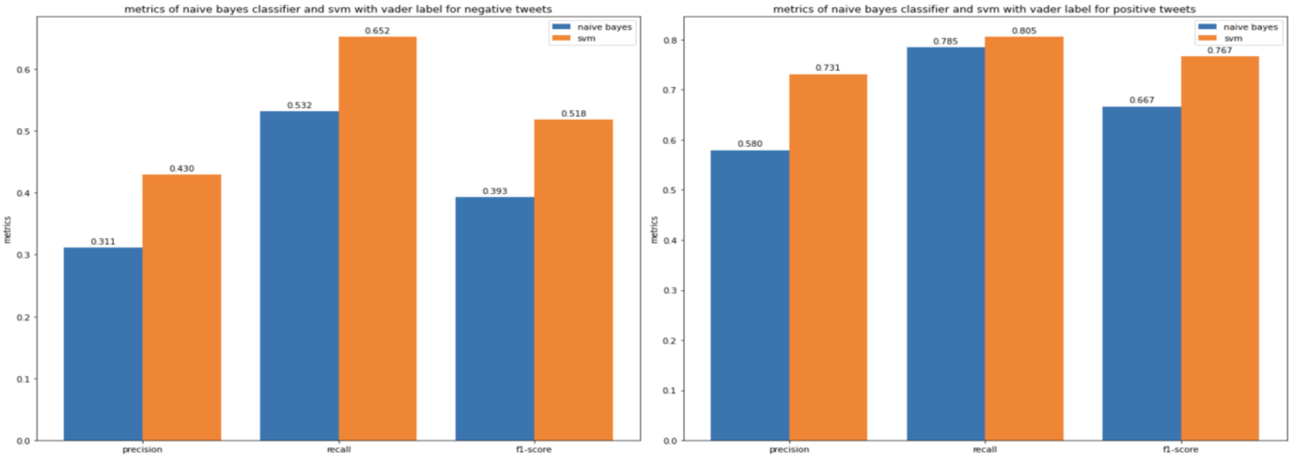


Figure 24. Metrics performance of Naïve Bayes Classifier and SVM with VADER label for negative tweets and positive tweets

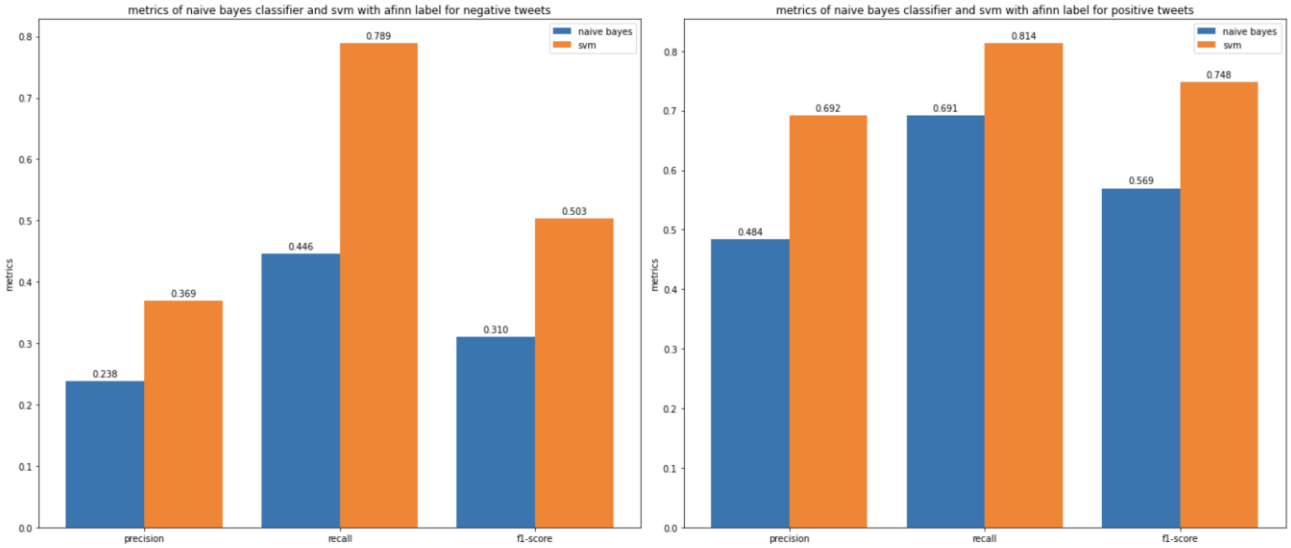


Figure 25. Metrics performance of Naïve Bayes Classifier and SVM with AFINN label for negative tweets and positive tweets

According to tables 2 and 3, and Figures 22 and 23, the overall performance of SVM is better than Naïve Bayes Classifier in terms of precision, recall, f1-score and accuracy while all the classifiers have a bad precision on predicting negative comments, and the performance of neutral and positive tweets prediction are almost good for all the classifiers. Also, there is no noticeable difference between the performance of the classifiers with the AFINN label and the VADER label. The accuracy of the Naïve Bayes classifier is slightly higher and the accuracy of SVM is slightly lower with the VADER label. Also, the gap of the metrics with AFINN label between the Naïve Bayes classifier and the SVM is larger than that with the VADER label.

5. CONCLUSION AND FUTURE WORK

In this study, we conducted sentiment analysis on tweets and discovered how different sentiments would affect the stock price. We used the NLTK toolkit and regex expressions to preprocess the tweet contents and applied lexicon-based approaches AFINN and VADER to label the dataset. We created some visualizations to analyze the distribution of sentiments, frequencies of sentiments, sentiments change over time, seasonal patterns, and word clouds. And we performed the sentiment classification through the Naive Bayes classifier and the Support Vector Machine classifier and evaluation metrics to compare the classifiers. VADER behaves better than AFINN. Also, SVM achieved a higher accuracy than the Naive Bayes classifier. The reason might be the assumption of the Naïve Bayes classifier that a feature is unrelated to other features, and this ignores the contextual relationship between the tokens in the tweets' comments. Then we plotted the images of the stock price gain, close value, and volume with the sentiment score and calculated their correlation coefficient and p-value to analyze their correlation. We found the stock price gain had some correlation with the sentiment while the correlation was very weak, and the sentiment had a stronger correlation with stock close value and a weak inverse correlation with stock volume. And here we only researched the correlation between stock and sentiment. However, there are many factors that would affect the stock price, which could result in a deviation in our experiments. In our future work, since the classification of the negative comments is bad, which affects the whole accuracy, we plan to focus on the improvement to identify and classify the negative emotions. We could use the unsupervised learning approach like K-means to label the dataset and compare it with the lexicon-based approaches to see if it could make up for some deficiencies or utilize the unigram, bigram or n-gram models before the lexicon-based approaches and see how the models influence the sentiment polarity. We could also use some other feature extraction techniques like word2vec and deeper models as well.

REFERENCES

Anshul Mittal, Arpit Goel. (2011) Stock Prediction Using Twitter Sentiment Analysis. Standford University, CS229.

Tumarkin, & Whitelaw, R. F. (2001). News or noise? Internet message board activity and stock prices. Financial Analysts Journal, 57, 41–51.

Mao H, Counts S, Bollen J (2015) Quantifying the effects of online bullishness on international financial markets (No. 9). ECB Statistics Paper

Alessia D'Andrea, Fernando Ferri, Patrizia Grifoni , and Tiziana Guzzo (2015) Approaches, Tools and Applications for Sentiment Analysis Implementation. Article in International Journal of Computer Applications

Shihab Elbagir and Jing Yang (2019) Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment. Proceedings of the International MultiConference of Engineers and Computer Scientists 2019

Abdullah Alsaeedi and Mohammad Zubair Khan (2019) A Study on Sentiment Analysis Techniques of Twitter Data. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2, 2019

Ammar Oad , Imtiaz Hussain Koondhar, Pinial Khan Butt, Huang lei1, Aneel Oad, Mansoor Ahmed Khuhro and Sajida Raz Bhutto (2021) VADER Sentiment Analysis without and with English Punctuation Marks. International Journal of Advanced Trends in Computer Science and Engineering Volume 10, No.2, March - April 2021

Bowman, M., Debray, S. K., and Peterson, L. L. (1993) Reasoning about naming systems. ACM Trans. Program. Lang.

Ding, W. and Marchionini, G. 1997. A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

Fröhlich, B. and Plate, J. (2000) The cubic mouse: a new device for three-dimensional input. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Tavel, P. (2007) Modeling and Simulation Design. AK Peters Ltd., Natick, MA.

Sannella, M. J. (1994) Constraint Satisfaction and Debugging for Interactive User Interfaces. Doctoral Thesis. University of Washington.

Forman, G. (2003) An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3.

Brown, L. D., Hua, H., and Gao, C. (2003) A widget framework for augmented interaction in SCAPE. In Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology.

Yu, Y. T. and Lau, M. F. 2006. A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions. J. Syst. Softw.

Spector, A. Z. (1989) Achieving application requirements. In Distributed Systems, S. Mullender, Ed. ACM Press Frontier Series.

Kavita Ganesan. (2019) What are Stop Words? AI Foundations, NLP Concepts, https://www.opinosis-analytics.com/knowledge-base/stop-words-explained/.

Kaggle, Tweets about the Top Companies from 2015 to 2020, https://www.kaggle.com/datasets/omermetinn/tweets-about-the-top-companies-from-2015-to-2020?resource=download&select=Tweet.csv.

Kaggle, Values of Top NASDAQ Companies from 2010 to 2020 https://www.kaggle.com/datasets/omermetinn/values-of-top-nasdaq-copanies-from-2010-to-2020?select=CompanyValues.csv

TechSlang. (2022) What is Lemmatization? A short definition of Lemmatization. https://www.techslang.com/definition/what-is-lemmatization/

Clarin. (2022) Part-of-speech taggers and lemmatisers. https://www.clarin.eu/resource-families/tools-part-speech-tagging-and-lemmatisation

Swayanshu Shanti Pragnya. (2022) VADER (Valence Aware Dictionary and sentiment Reasoner) Sentiment Analysis. https://medium.com/mlearning-ai/vader-valence-aware-dictionary-and-sentiment-reasoner-sentiment-analysis-28251536698. MLearning.ai

Finn Årup Nielsen. (2011) A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718 in CEUR Workshop Proceedings 93-98.

Himanshu Lohiya. (2018) Sentiment Analysis with AFINN Lexicon. https://himanshulohiya.medium.com/sentiment-analysis-with-afinn-lexicon-930533dfe75b

Boost Labs. (2014) Word Clouds & the Value of Simple Visualizations. https://boostlabs.com/blog/what-are-word-clouds-value-simple-visualizations/#:~:text=A%20word%20cloud%20is%20a,the%20more%20important%20it%20is.

Dashtipour, K., Poria, S., Hussain, A. et al. (2016) Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques. Cogn Comput 8, 757–771.

Hamza, Mohammed; Gupta, Sawar. (2022) A Comparison of Sentimental Analysis Algorithms on Twitter Data Using Machine Learning. TechRxiv. Preprint.

Sampriti Chatterjee. (2022) What is Feature Extraction? Feature Extraction in Image Processing. Great Learning. https://www.mygreatlearning.com/blog/feature-extraction-in-image-processing/

Ravinder Ahuja, Aakarsha Chug, Shruti Kohli, Shaurya Gupta, Pratyush Ahuja. (2019) The Impact of Features Extraction on the Sentiment Analysis. Procedia Computer Science. Volume 152.

Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, Sweta Tiwari, (2016) Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier", International Journal of Information Engineering and Electronic Business (IJIEEB).

Murtadha B. Ressan, Rehab F. Hassan. (2022) Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets. Indonesian Journal of Electrical Engineering and Computer Science.

Manning, C. D., Raghavan, P., & Schütze, H. (2008) Introduction to information retrieval, Cambridge: Cambridge University Press.

Sunil Ray. (2017) Understanding Support Vector Machine (SVM) algorithm from examples (along with code), https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

Complete Dissertation. (2022) Correlation (Pearson, Kendall, Spearman). https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman/

D. A. Kristiyanti, A. H. Umam, M. Wahyudi, R. Amin and L. Marlinda. (2018) "Comparison of SVM & Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-

2023 Based on Public Opinion on Twitter," 2018 6th International Conference on Cyber and IT Service Management

Shihab Elbagir and Jing Yang. (2019) Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment. Proceedings of the International MultiConference of Engineers and Computer Scientists

Hui Liu. (2021) Chapter 7 - Description methods of spatial wind along railways. Wind Forecasting in Railway Engineering.