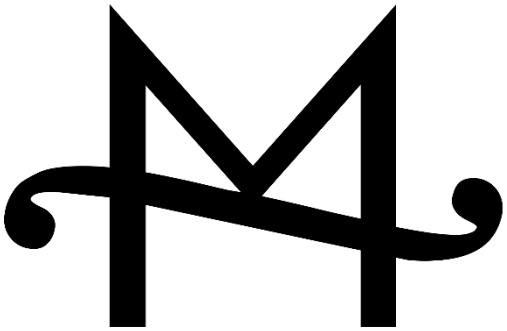
MASA
Malaysian Actuarial Student Association

PPM-016-10-18102018

HERIOT-WATT UNIVERSITY MALAYSIA
1, JALAN VENNA P5/2
PRECINCT 5
62200 PUTRAJAYA



Research Topic:
Impact of Contagious Disease on Different Scale
Field:
News-Based Tweets' Sentiments on the CBOE VIX

Authors:
Matthew Farant | Bernard L | Cheah Eason
Supervisor(s):
Chi Keong Lo

Published by: Malaysian Actuarial Student Association (MASA)
Date of Publication: 17th August 2020

Website : http://www.masassociation.org/
Facebook : http://www.facebook.com/MASAssociation

PPM-016-10-18102018

HERIOT-WATT UNIVERSITY MALAYSIA
1, JALAN VENNA P5/2
PRECINCT 5
62200 PUTRAJAYA

TABLE OF CONTENTS

1 ABSTRACT

The growth of social data is rapidly increasing nowadays. During the global COVID-19 outbreak, many individuals and organizations have posited their viewpoints on social media regarding the current epidemic. Twitter is widely used among the general population and is one of the leading social media platforms in obtaining news and information on a topic. Currently, tweets posted by influential twitter accounts like news accounts are arguably a reliable proxy of the public sentiments' on COVID-19. Therefore, this research intends to make good use of this information to generate insights in general, whether news twitter accounts are practicing fear-mongering and whether it affects the financial market. In this research, we attempt to analyze the sentiments of most-followed news accounts on twitter during the COVID-19 outbreak and find out their impacts on the Chicago Board Option Exchange's Volatility Index (CBOE VIX), or also known as the "fear gauge," one of the indicators of S&P 500 movement. Further correlation and statistical tests have found no correlation between average polarity and CBOE VIX.

PPM-016-10-18102018
HERIOT-WATT UNIVERSITY MALAYSIA
1, JALAN VENNA P5/2
PRECINCT 5
62200 PUTRAJAYA

2 INTRODUCTION AND STATEMENT OF PROBLEM

With the advent of the internet, leading to the explosive growth of social media platforms like Facebook, Twitter, micro-blogs, etc., the rate at which information is being disseminated and retrieved has heightened at a monumental pace [1]. The rise and successes of these platforms, for example, Twitter–to which represents the social media platform our thesis examines–towards rapid informational flow, can be inferred upon high accessibility and user's ease of use in utilizing the platform for a bigger purpose it was intended to be used as [2]. According to Oberlo, Twitter has over 330 million monthly active users, with 500 million tweets sent each day [3]. Of course, it is not just the factual ease of use users have on Twitter or technology in general that led to its success, but also their perception of the ease of use to which other great academic minds have mapped out a positive relationship between both items [4]. The amount at which opinions are shared has given rise to the chance to utilize a dataset of public opinions for analytical applications across a variety of fields [5].

The process of drawing analytical insights from digital forms of opinionated data has been aptly given the term sentiment analysis or also, opinion mining [5]. In a more technical definition, it also relates to the computational study of people's opinions towards a respective target (which can be individuals, entities, issues, and events) along with their attributes. Being an active research area in Natural Language Processing, the practical applications of sentiment analysis has seen its research not just limited to the field of computer science but also towards practical business and societal domain use cases as a whole [5]. Systems are being built in these fields where opinions form a major driver towards behavioural activities.

This paper will explore the application of sentiment analysis in the field of finance, specifically analyzing news-based tweets on Twitter to predict the price movement in the CBOE VIX. In essence, we are subscribing to the Efficient Market Hypothesis in which we conform to the view that markets, in reality, are of the typical semi-strong efficiency form–publicly available information which is inclusive of publicly opinionated data (i.e., tweets)–will be reflected in the price movements of an asset class (i.e., the VIX). This paper will show whether there is a correlation between the polarity of news tweets disseminated by reputable news sources in CNN, BBC and NBC with the index price movements of the CBOE VIX – widely known as Wall Street's "fear gauge"–for the timeframe where the occurrence of COVID-19 had led to detrimental macroeconomic effects globally [6].

3 LIMITATIONS OF STUDY

This paper explores the application of sentiment analysis from news-based tweets datasets in drawing inference on any correlation with the VIX as a representative of public fear in the stock market.

Datasets utilized have demographics limited to English speakers, as the medium of language of news portal is in English. The time frame of the Twitter dataset is also limited to from January 2020 to July 2020, where the COVID-19 began to emerge until it reached its peak infection rate globally. This implies that datasets from other contagious diseases are not considered in this study. We will only use the CBOE VIX as the representation of the S&P 500 options. This research will not take into account other indices outside the US.

4 METHODOLOGY

Description of Dataset

In this research, we will use R to build our dataset by scraping tweets from various accounts on Twitter. We will use the *rtweet* package and standard Twitter API to scrap a maximum of 3200 tweets from each Twitter account observed. This package is used due to its exceptional features compared to other mainstream Twitter-scraping packages (e.g., twitteR, streamR, RTwitterAPI).

We will be focusing on six most significant news accounts on twitter: CNN Breaking News (@CNNbrk), BBC Breaking News (@BBCbreaking), NBC Breaking News (@Breakingnews), Wall Street Journal (@wsj), New York Times (@nytimes), and Bloomberg (@business). The scraping process will generate six different data frames, which the details are shown in this table below:

Account	Time Range (YYYY/MM/DD)	Average Number of Tweets per day	Number of Followers
@CNNbrk	2019/12/07 – 2020/7/29	14	58.4 M
@BBCBreaking	2017/11/21 – 2020/07/28	4	44.8 M
@BreakingNews	2018/09/20 – 2020/07/28	6	9.5 M
@wsj	2020/07/06 – 2020/08/12	84	17.9 M
@nytimes	2020/06/26 – 2020/07/29	94	47.1 M
@business	2020/07/19 – 2020/07/29	291	6.5 M

Therefore, there will be a total of **19,200 distinct tweets** from six different accounts that we will be using during this research. Notice that from the six news accounts above, the first three are breaking-news accounts, and the following three are regular news accounts. The main difference between those two types of news accounts is the average number of tweets per day. Breaking-news accounts have a relatively lower amount of tweets per day, compared to regular news accounts, which is why breaking news accounts show a bigger picture on the sentiment changes from the beginning of 2020 until the mid of 2020. We will only use the breaking-news accounts dataset for the sentiment analysis, later on, to be compared with the VIX. However, we will still use all of the datasets for an Exploratory Data Analysis. The tweets dataset that is generated by *rtweet* will be a data frame object, with 3200 rows and 90 columns. Each column represents the attributes of a particular tweet: tweet id, date of creation, text, etc[1].

As mentioned in the previous part, we will try to find the correlation between the tweets' sentiment and the CBOE VIX. The VIX's CSV dataset is downloaded directly from the CBOE official website. It will be converted into a data frame with 137 observations, showing the price movement from

[1] The complete dictionary of a tweet data frame object is available on Developer-Twitter's official website: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object

2020/01/02 until 2020/07/17 (excluding non-working days), and five columns: date, open, high, low, and close.

Data preparation

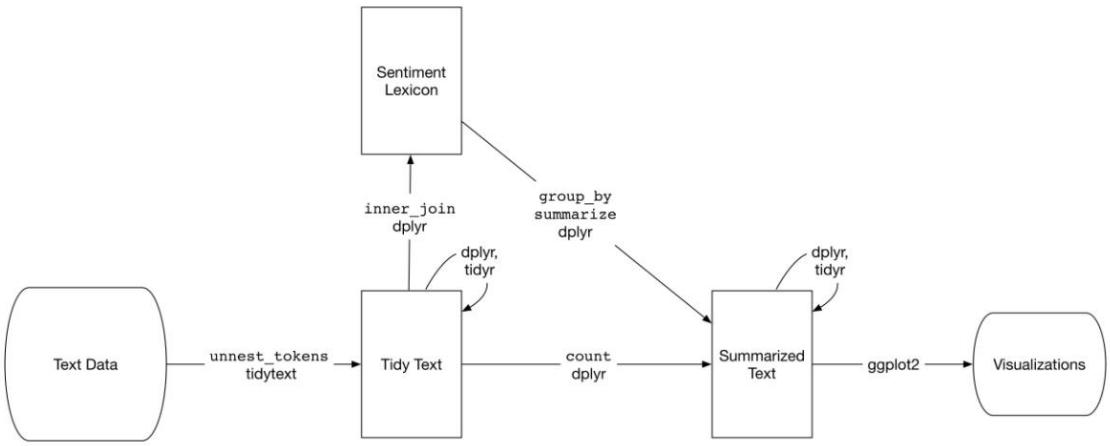
The data preparation and cleaning processes will be mostly done in the tweets dataset. First, we will only filter the tweets that contain coronavirus-related keywords using the grepl function. The keyword includes: "COVID 19", "Coronavirus," "Corona," "COVID," "ncov," "2019-ncov", "SARS-CoV-2", and "lockdown" (case ignored). It is crucial to have a bunch of keywords due to term changes and ambiguation. After we filtered the tweets, we will do a token normalization to clean the text column. According to (Christoper et al., 2008), token normalization is the process of canonicalizing tokens so that matches occur despite superficial differences in the character sequences of the tokens [7]. A token can be a single word, n-grams, sentence, or paragraph. We will do this cleaning process using the *tidytext* and *textclean* packages. This process includes:

- Converting all strings into lowercase
  This is done to prevent discrimination between two same strings with different cases because R is case-sensitive
- Remove contractions in words
  e.g., isn't → is not
- Remove symbols
  e.g. %, $, &, etc.
- Remove URLs
  e.g., t.co links in most tweets.
- Remove all non-words characters.
  e.g.
, \r, \n, etc

In addition, when we do the text mining process, we will also remove all the stop words after the tokenization process. News accounts were chosen to be our dataset due to the absence of misspelled words, internet slangs, and emojis.

Method of Research

After preparing the dataset, this research will be divided into two parts: text mining and modelling. The text mining process will include n-grams analysis, polarity analysis, and emotion analysis that will be conducted in both R and Python. Below is the standard flowchart of a text mining and sentiment analysis in R:



The general workflow for text mining and sentiment analysis. Source: Tidy Text Analysis with R [8]

The first part of the text mining process will be tokenization. Tokenization is the process of splitting a tweet into tokens. We will break a tweet into a single word (unigrams), a pair of two words (bigrams), and a pair of three words (trigrams). The result of this process will later be analyzed and visualized using bag-of-words (word cloud), term frequency, and narrative networks.

For the sentiment analysis part (polarity and emotion analysis), instead of using an available sentiment lexicon, we will use the *sentimentr* package due to its ability to handle valence shifters, compared to other mainstream packages like *qdap* or *syuzhet*. Valence shifters are words that are able to change the sentiment of a whole sentence. It could be a negator, amplifier, and many other types of shifters. Therefore, when it comes to sentiment analysis, this package will examine the sentiment sentence-by-sentence for a particular tweet. However, a single tweet may have multiple sentences, and a single day may have numerous tweets, which is why we will do an aggregation by finding the average sentiment for a single day, for a particular account.

The result of the polarity analysis is binary: positive or negative. In this research, it will also be presented using a value between -1 (indicating a strong negative polarity) until 1 (indicating a strong positive polarity). After aggregating for average sentiments for each day, we will generate a time series data of average sentiment for a single twitter account. On the other hand, the result of emotion analysis is non-binary. A tweet may have different types of emotions, such as fear, optimism, sadness, and so forth. We will find the number of words for each kind of emotion and find the highest emotion type for each twitter account.

After the text mining part is finished, we will have two different time series data: tweet sentiment and historical price movements of the VIX. This research aims to find how these two time series correlate to each other by finding whether there is a significant cross-correlation. Cross-correlation and the statistical tests are conducted using the tseries and stats package in R.

5 LITERATURE REVIEW

According to (Liu, 2012), Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language [9]. Sentiment analysis has been widely researched in the domain of online review sites to generate summarized opinions of users about different aspects of products. Identifying such sentiments from online social networking sites can help emergency responders understand the dynamics of the network, e.g., the primary users' concerns, panics, and the emotional impacts of interactions among members. The advent of social media has created an unprecedented environment for people to share their thoughts with the world. These online platforms like Facebook, twitter are usually the first resort people turn to in times of crisis to voice their opinions and relay other crucial information. But when it comes to detecting sentiments out of this gigantic pool of views, it becomes an arduous task, and doing it manually is practically impossible. However, Twitter posts only contain a few words. Hence, it's easy to scrape.

Mathematical Point of View of Sentiment Analysis

According to the *sentimentr* documentation[2], an augmented dictionary method is used by *sentimentr* due to better results than using a simple lookup dictionary approach that does not consider valence shifters. First, to assign a polarity value to a text object, the algorithm utilizes a sentiment dictionary to tag polarized words. Each paragraph ($p_i = \{s_1, s_2, \dots s_n\}$) composed of sentences, is broken into element sentences ($s_i, j = \{w_1, w_2, \dots, w_n\}$) where w is the words within sentences. Each sentence (s_j) is split into bag-of-words (ordered). All punctuations are removed except for pause punctuations (e.g., commas, colons, semicolons), which are considered a word inside the sentence. We can label these tokens as i,j,k notation as $w_{i,j,k}$. For example, $w_{1,2,3}$ would be the third word of the second sentence of the first paragraph. The term paragraph merely represents a complete turn of talk.

Tokens (words/unigrams) in each sentence ($w_{i,j,k}$) are looked up and compared to an available dictionary of pre-polarized words. Positive ($w_{i,j,k}{}^{+}$) and negative ($w_{i,j,k}{}^{-}$) words are labeled with the value +1 and −1 respectively (or other weighting, if the user provides another sentiment lexicon). These will create a polar cluster ($c_{i,j,l}$) which is a subset of a sentence ($c_{i,j,l} \subseteq s_i, j$).

The cluster of words is pulled from the polarized word and defaults to 4 words before and 2 words after the polarized word to be considered as valence shifters. There are many types of valence shifters such as neutral ($w_{i,j,k}{}^{0}$), negator ($w_{i,j,k}{}^{n}$), amplifier/intensifier ($w_{i,j,k}{}^{a}$), or de-amplifier/downtoner ($w_{i,j,k}{}^{d}$). Neutral words have no value in the model but they affect the word count (n). Then, each of the polarized words is weighted according to the weights from the *polarity_dt* argument and then weighted by the function and number of the valence shifters. Pause locations are labeled (indexed) and utilized in calculating the upper and lower bounds in the context cluster due to the marks that indicate a change in thought. Words prior are not necessarily linked with words after the punctuation marks.

[2] https://github.com/trinker/sentimentr

How the valence shifters affect the weighting:

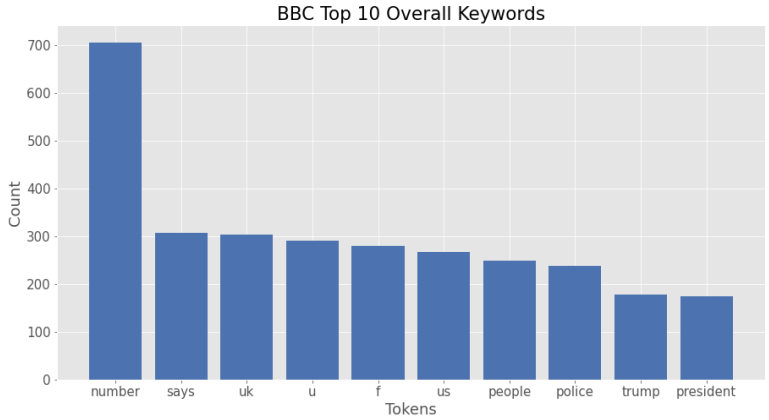
Valence shifters (type)	Example	Effect
Amplifiers	“Very,” “Greatly,” “Truly”	Increase polarity by 1.8
De-amplifiers	“Almost”, “Barely”	Decrease polarity by 1.8
Negation	“Can’t”, “Not”, “No”	Flip the sign of polarity

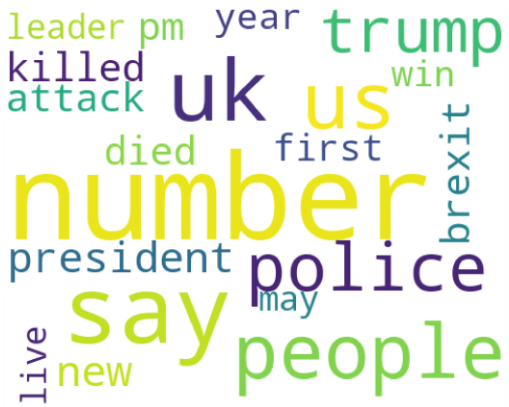
Adversative conjunctions like “but”, “however”, and “although” also affect the weightage of the context cluster. Users may provide a weight to be used with amplifiers or de-amplifiers. Last, the weighted context clusters will be totaled and divided by the square root of the word count, resulting in an unbounded polarity value/score for each sentence. To get the average of all sentences within a paragraph, the user can take the average sentiment score or use an available weighted average.

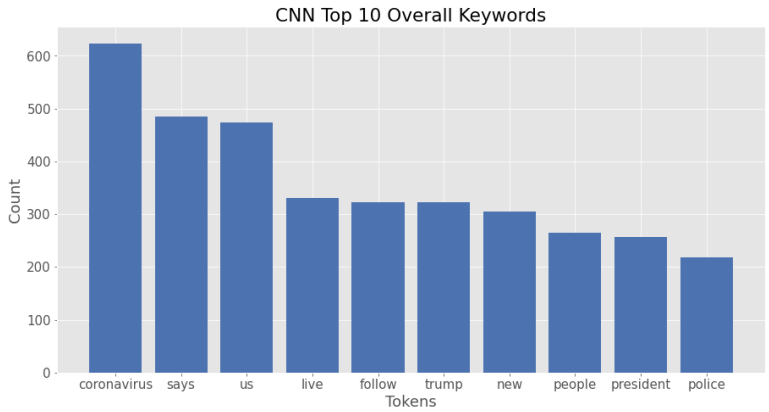
6 RESULTS

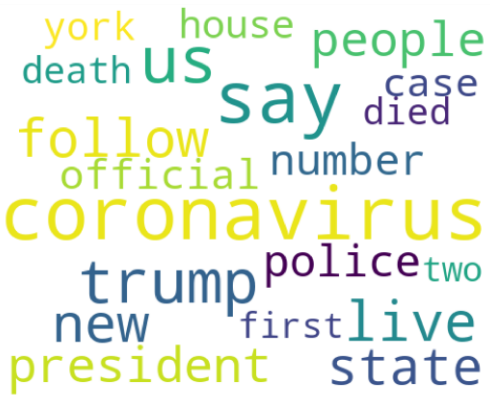
The year 2020 is full of major events around the world such as bushfires in Australia, negative oil prices, US presidential election, and of course the infamous coronavirus outbreak. As a starter, we might want to know what is the most discussed topic by the news twitter accounts this year. One of many ways to find out is by generating a word cloud, made out of a bag-of-words. A bag-of-words is a vector containing frequencies of a word in a collection of sentences (in this case, tweets).

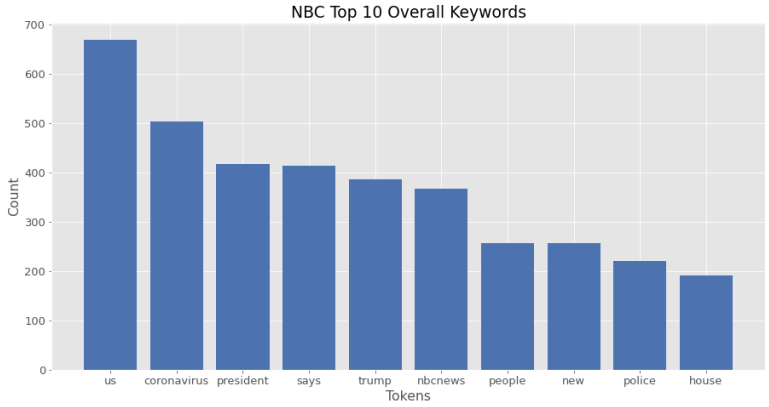
We performed an Exploratory Data Analysis via Python's *pandas* & *WordCloud libraries* regarding most commonly used words across all Twitter accounts we have identified along with an overall aggregated result. The simple Bag-of-Words approach was used to identify the main keywords of our datasets. The following are bar charts tabulating the top 10 most used words overall tweets for a specific account in their specified period along with their respective word clouds.

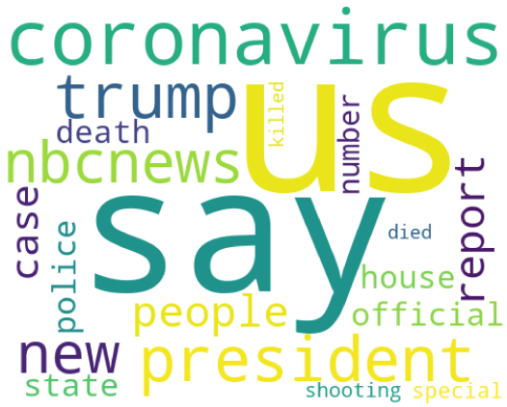


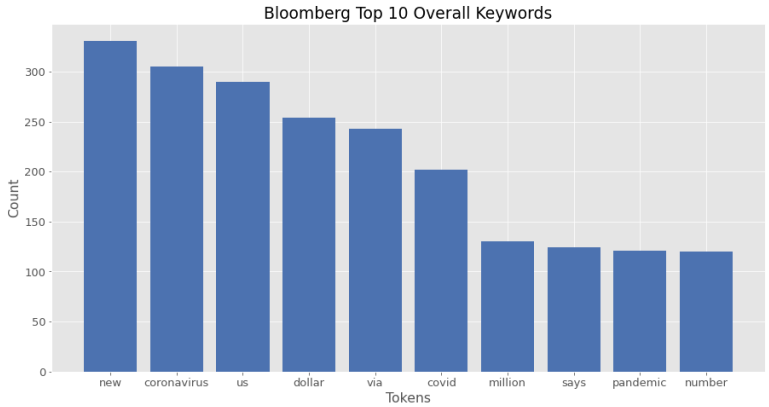




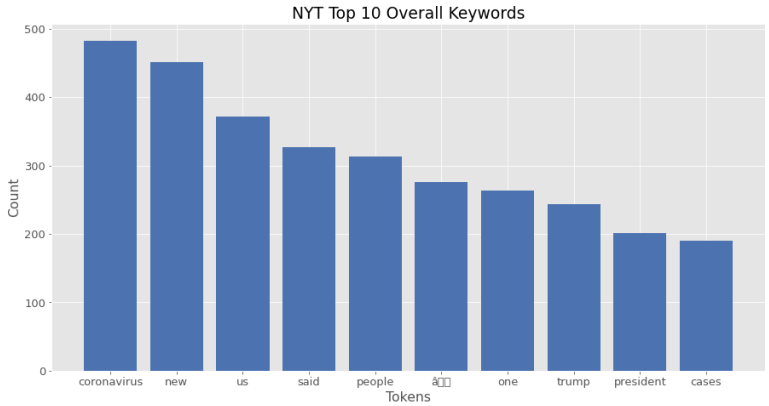



NBC Top 10 Overall Keywords
Count
Tokens
us
coronavirus
president
says
trump
nbcnews
people
new
police
house

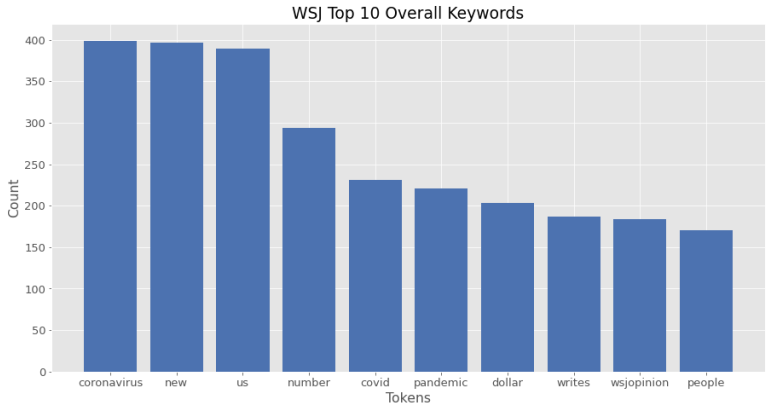

coronavirus
trump
us
death
killed
nbcnews
number
say
case
police
died
report
house
people
official
new
president
state
shooting
special


Bloomberg Top 10 Overall Keywords
Count
Tokens
new
coronavirus
us
dollar
via
covid
million
says
pandemic
number


plan
trump
one
market
via
number
dollar
billion
may
say
new
china
coronavirus
first
year
us
pandemic
government
covid
million

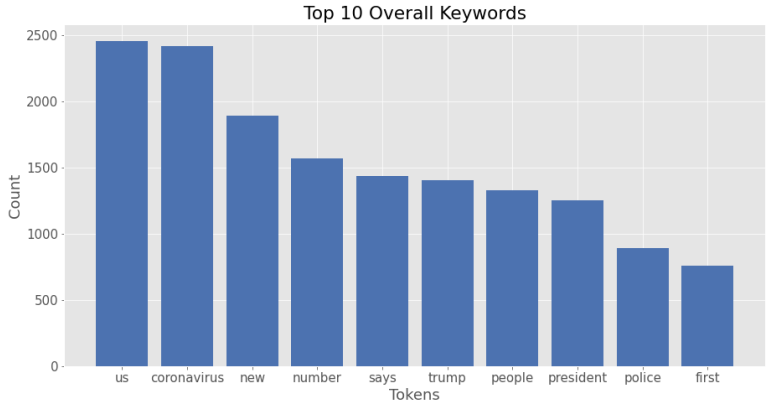

NYT Top 10 Overall Keywords
Count
Tokens
coronavirus
new
us
said
people
one
trump
president
cases

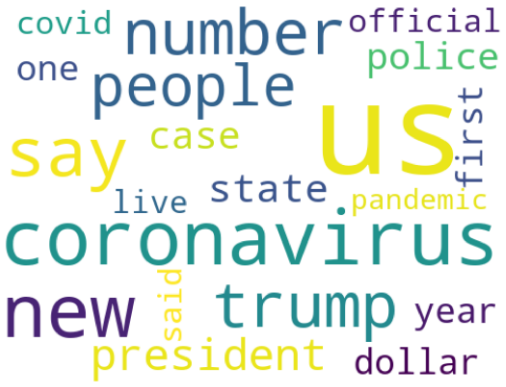

black
people
us
president
state
many
police
new
pandemic
year
york
coronavirus
say
case
trump
one
said
day
time
first


WSJ Top 10 Overall Keywords
Count
Tokens
coronavirus
new
us
number
covid
pandemic
dollar
writes
wsjopinion
people


year
covid
say
look
case
state
write
number
one
people
new
trump
million
president
coronavirus
said
us
dollar
wsjopinion
pandemic

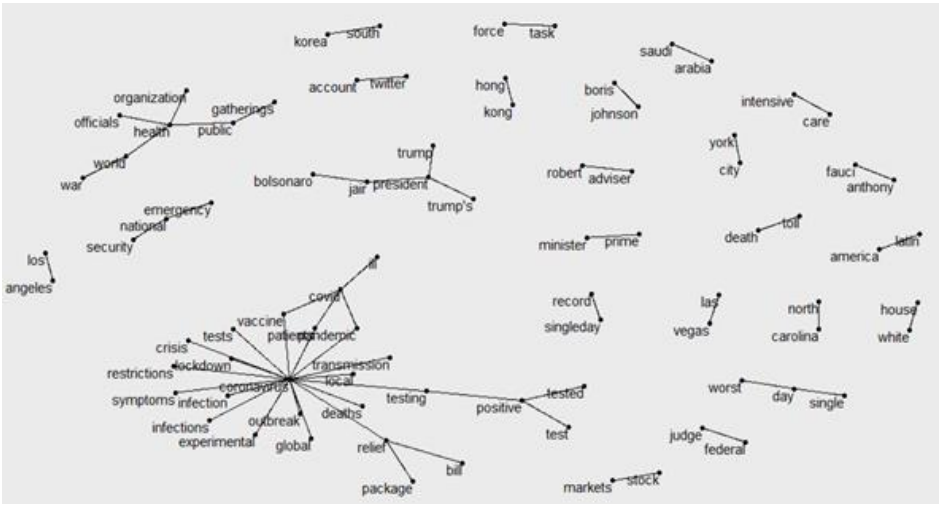
The following two visualizations show the overall aggregated result across the 6 accounts.



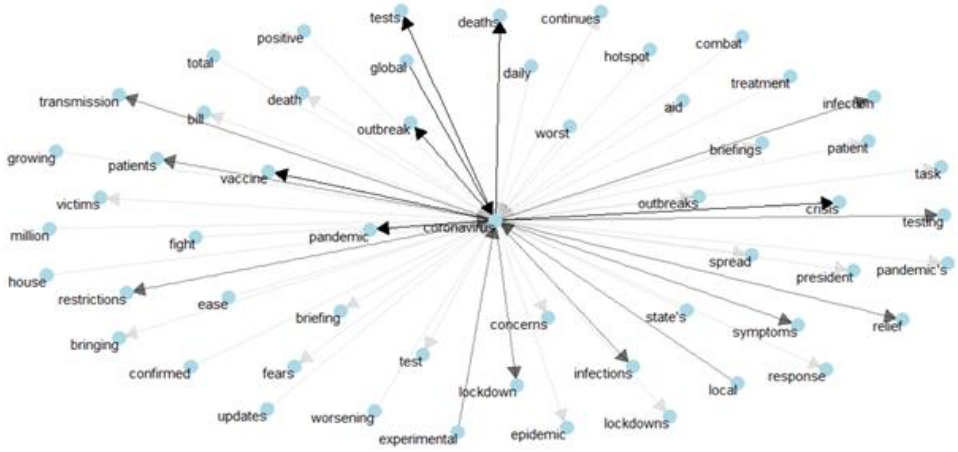


The word clouds above show the most used unigrams in various news accounts. Larger size of word means higher frequency of occurrences. Even though the dataset hasn't been filtered for coronavirus keywords, we can already see that the coronavirus outbreak has become one of the most discussed topics besides any other major events this year. We can see that aside from the 'coronavirus' keyword being one of the most used words in tweets across the six accounts, terms that relate to the United States such as "us," "trump" and "dollar" are also frequently recurring. Thus, the motive of performing this section of an EDA adds support to a consideration towards our hypothesis where the tweets can reflect the price movements of the VIX, which is based in the United States.

Continuing this analysis, a bigram analysis (a pair of two words) will also be conducted to all six twitter accounts. Instead of using a table to show the most-frequent bigrams, a narrative network will be generated. A narrative network is a graph that shows the relationship between words. It contains three variables: from, to, and weight. "from" indicates the node (point) an edge (arrow) is coming from. "to" indicates the node an edge is going towards. "weight" is a numeric value associated with each edge. The length of the edge is randomized with pre-determined seed.



The graph above shows the narrative network of all 6 news twitter accounts. It is already filtered for the bigrams with more than two occurrences. However, some of the graphs are already obvious like "Los-Angeles", "hong-kong," "prime-minister," and so forth. Hence, a filtering process will be conducted again to filter only bigrams that contain the word "coronavirus."

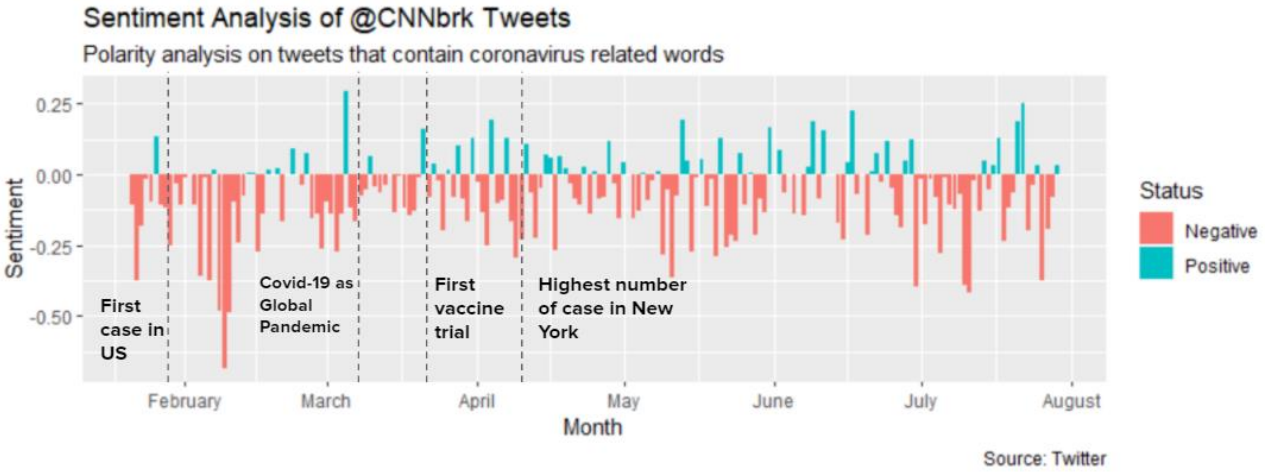


After the graph is generated, we can see which bigrams have the highest number of occurrences by looking at the edges' thickness. According to the graph above, the word "coronavirus" is always linked with the word "deaths", "outbreak," "crisis," "pandemic," "vaccine," and "tests." Most of the words mentioned just now are relatively negative words. However, we can't quickly conclude that most news twitter accounts are practicing fearmongering during this pandemic. The next part about polarity analysis will give more insight into how the sentiment changes over time.

Polarity Analysis

The polarity analysis is conducted on three breaking-news accounts: CNN Breaking News, BBC Breaking News, and NBC Breaking News. The result of the analysis is a time-series data showing the average sentiment for each day from January to July for each account. One of the obstacles of doing a polarity analysis on a COVID-19 tweets data is the confusion on ambiguous words' sentiment. For example, "positive" and "negative". As we know, in the medical lexicon, "positive" (test results) usually has a negative sentiment, vice versa. Other than those two words, words like "top," "new," and "high" which are considered as positive words, could have a different meaning if it is paired with negative words (e.g., death). Unfortunately, *sentimentr* failed to categorize some words as valence shifters. Therefore, we need to detect tweets that contain ambiguous words and adjust the polarity manually. Other than changing the polarity manually, this issue could be solved by using an available medical lexicon. However, we couldn't find any open-source medical lexicon for this research.

CNN Breaking News



The bar plot above shows the time series of average tweets polarity from January to July (183 days). The plot above is generated from 877 distinct tweets containing the coronavirus keywords. Red bars indicate a negative polarity and blue bars indicate a positive polarity. According to the plot above, we can see that most of the strong negative sentiment tweets are gathered at the beginning of the year (January - March) as people are still afraid and didn't know about how dangerous the coronavirus is yet. Since the global pandemic announcement, positive tweets such as lockdown measures and social distancing practices are starting to grow. Also, since the first vaccine trial in the US, news accounts start to discuss vaccines and build optimism at the end of the outbreak.

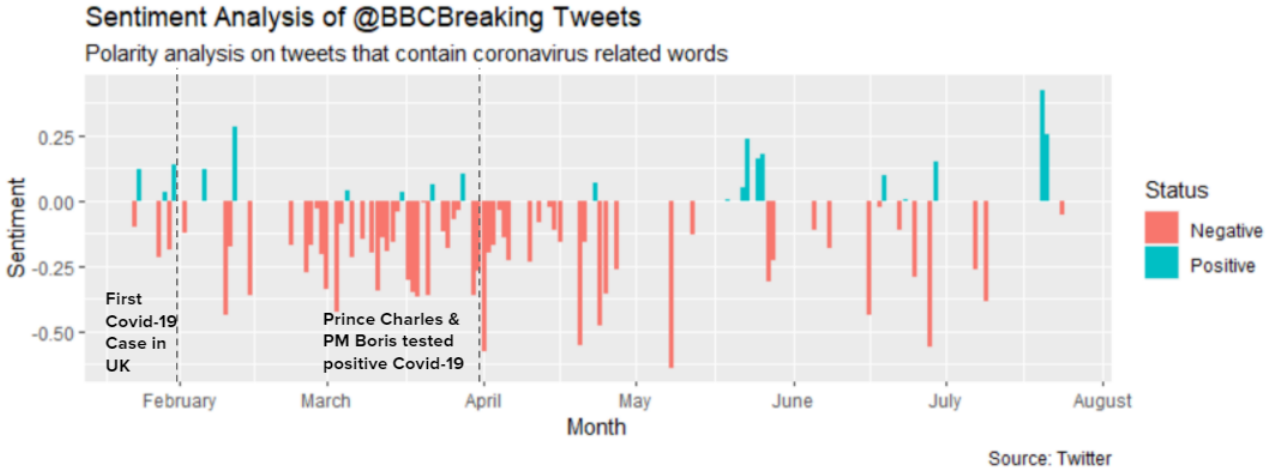
Sample of positive tweets:

- "*israel has further tightened coronavirus restrictions including limiting gatherings to people and closing all leisure and entertainment venues effective sunday morning*"
- *"the fda approves the use of a new onsite test that could detect coronavirus in approximately minutes*"
- *"british health care workers are taking part in a global clinical trial to test the effectiveness of antimalarial drugs chloroquine and hydroxychloroquine against coronavirus"*

Sample of negative tweets:

- *"jcpenney files for bankruptcy the covid crisis is the final blow to a ... year old company struggling to overcome a decade of bad decisions executive instability and damaging market trends"*
- *"dr anthony fauci says social distancing measures appear to be working a but stresses that the coronavirus pandemic is still a very serious situation follow live updates"*
- *"the global death toll from the wuhan coronavirus is at least people surpassing the number of fatalities from the deadly sars outbreak follow live updates"*

BBC Breaking News



The bar plot above shows the time series of average tweets polarity from January to July (89 days). The plot above is generated from 159 distinct tweets containing the coronavirus keywords. Red bars indicate a negative polarity and blue bars indicate a positive polarity. As we know, BBC is a news company based in the UK. Therefore, we will match some major events with the plot above. We can see that the patterns are nearly the same as CNN Breaking News' sentiments. Most of the negative sentiment tweets are gathered when Prince Charles and PM Boris Johnson are tested positive for coronavirus (around March 20th). In the same month, the government also imposed a lockdown measure and stay-at-home notice.

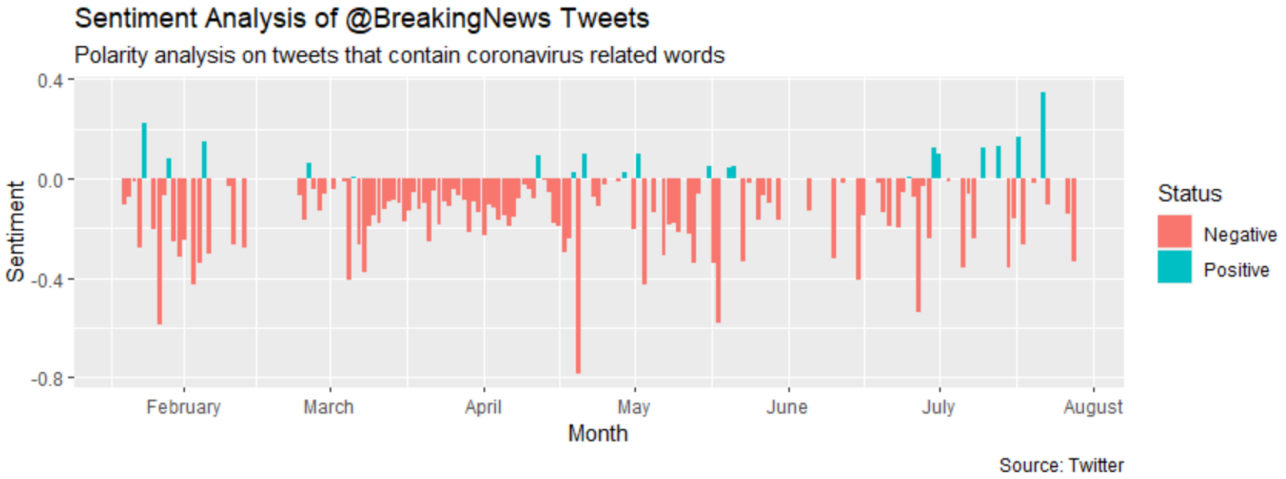
Sample of positive tweets:

- *"coronavirus vaccine developed by oxford university appears safe and trains the immune system key early trials show"*
- *"president trump considering imposing quarantine on new york as coronavirus cases there increase to more than..."*
- *"eu leaders agree covid economic recovery package with a ... bn in grants and loans after fourth night of talks"*

Sample of negative tweets:

- *"us jobless rate rises to as coronavirus pandemic devastates the economy million jobs lost in april"*
- *"uk government declares coronavirus a serious and imminent threat to public health"*
- *"coronavirus death toll in italy overtakes china's after rising by ... to ... "*

NBC Breaking News



The bar plot above shows the time series of average tweets polarity from January to July (139 days). The plot above is generated from 612 distinct tweets containing the coronavirus keywords. Red bars indicate a negative polarity and blue bars indicate a positive polarity. As we know, NBC is also based in the US like CNN. Therefore, there are no labels for major events, as they are already mentioned in the previous CNN plot. Notice that most of the negative sentiment tweets are gathered between March and mid-April, when the coronavirus cases are rapidly increasing, especially in New York.

Sample of positive tweets:

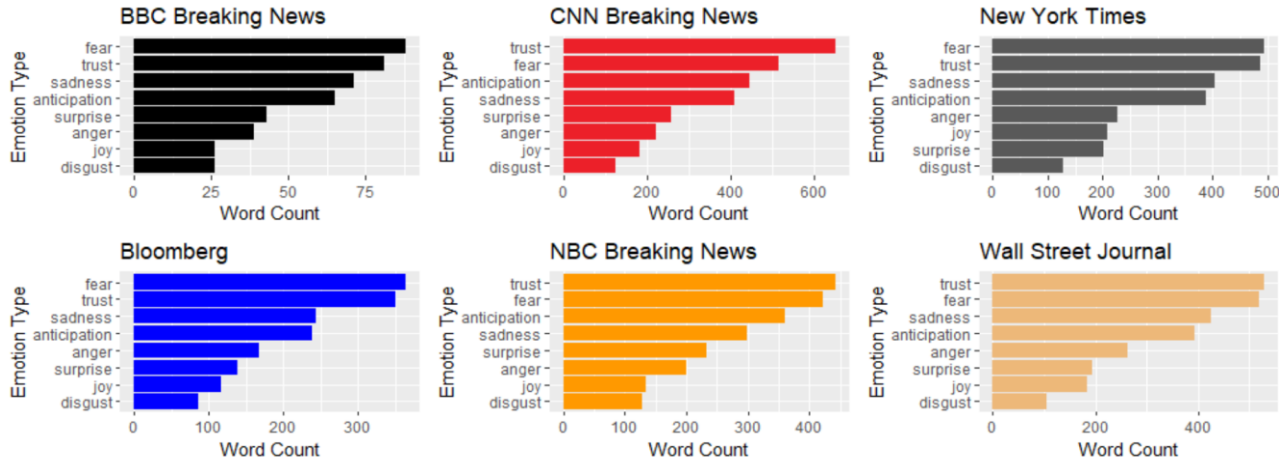
- *"nbcnews white house releases new coronavirus guidance including the recommendation to avoid gatherings of people"*
- *"president trump says he respectfully asks media and politicians to not incite panic around the coronavirus this is something that is being handled professionally"*
- *"president trump has signed the coronavirus aid bill into law white house says"*

Sample of negative tweets:

- *"china's coronavirus death toll climbs to as the government scrambles to contain the outbreak"*
- *"spanish coronavirus death toll reaches after recording fatalities in hours the highest number in the country's outbreak so far spain has also reported a record cases of the virus"*
- *"britain's prince charles has tested positive for the coronavirus and has mild symptoms but otherwise remains in good health clarence house says"*

Emotion Analysis

To see whether news accounts are practicing fearmongering amid the pandemic, we need to address the type of emotion for each tweet instead of just doing a binary polarity analysis (either positive or negative). The emotion analysis generated this bar plot below, showing the most dominant type of emotion for each account.



From the graph above, we can see that 3 out of 6 news accounts are dominated by "fear" tweets. On the other hand, the rest are dominated by "trust" tweets.

Sample of fear tweets:

- *"the us must get control of the covid pandemic or risk seeing deaths skyrocket well into the multiple hundreds of thousands the association of american medical colleges warns"*

Sample of trust tweets:

- *"the white house and senate reached a historic trillion stimulus deal to jolt the economy struggling through the coronavirus pandemic"*

Other than these six news accounts, the practice of fearmongering is considered to be ubiquitous during this outbreak. Mass media companies usually use this technique to gain more ratings or viewers. However, this practice might influence how investors behave, as news can change one's risk profile. The next part will show how the sentiments in news twitter accounts correlate with the VIX.

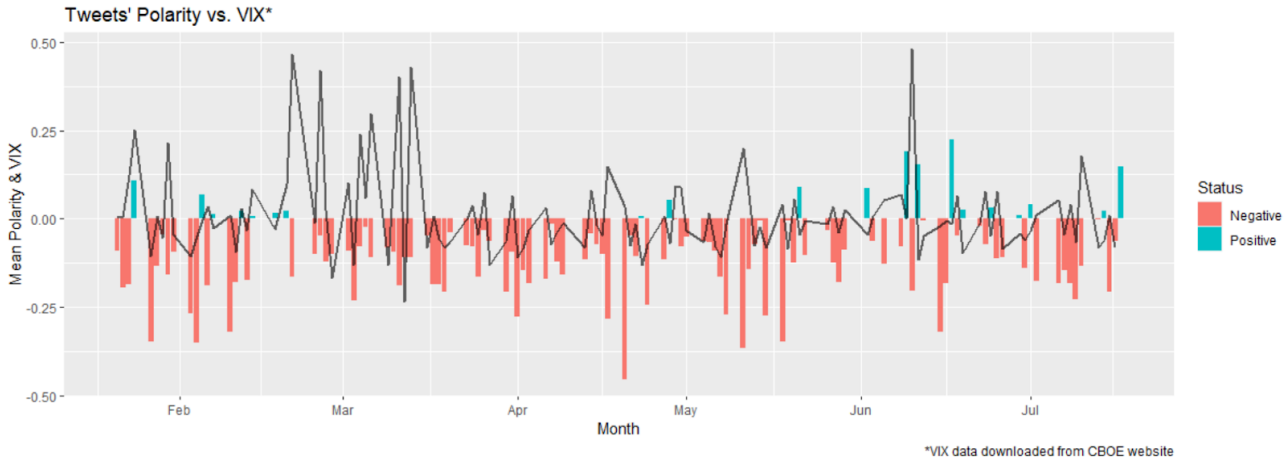
Average Sentiments and VIX

In this part, we are going to find the correlation between average polarity and VIX changes. The average polarity is the combination of all polarities from @CNNbrk, @BBCbreaking, and @Breakingnews tweets. The VIX changes are the percentage change of VIX's close price that expressed mathematically as the formula below:

$$\Delta VIX = \frac{VIX.Close_t - VIX.Close_{t-1}}{VIX.Close_{t-1}}$$

The plot below visually shows the movement of both time series, average polarity of three news accounts combined and the VIX changes. As we know from the previous plots, the bar plot shows the polarities and the

line plot shows the VIX changes. However, we can't simply conclude whether there is a correlation between the time series merely by looking at the plot below, which is the reason why proper statistical tests need to be conducted.



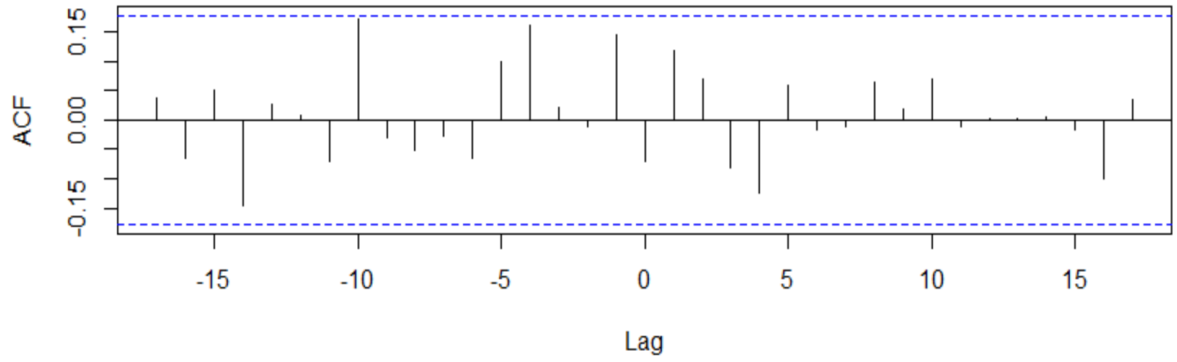
First, we are using the time series from January to July and see whether there is a significant correlation between the two time series. Before finding the cross-correlation, some statistical tests are needed to make sure that both of the time series are stationary. To test stationarity, we will use the Anderson-Darling test and KPSS test. Below are the results of the statistical tests:

Type of test - time-series data	P-value	Interpretation
ADF Test - Mean Polarity	0.01	Stationary
ADF Test - VIX Changes	0.01	Stationary
KPSS Test - Mean Polarity	0.0724	Stationary
KPSS Test - VIX Changes	0.1	Stationary

The statistical tests showed that all of the time-series data are a stationary time series. Hence, we continue to find the cross-correlation between the two time series. Cross-Correlation Function (CCF) is preferred due to its ability to identify lags between the correlation of two time series.

PPM-016-10-18102018
HERIOT-WATT UNIVERSITY MALAYSIA
1, JALAN VENNA P5/2
PRECINCT 5
62200 PUTRAJAYA

Using the available time series and ccf function in R, we generated the cross-correlogram below



The blue dotted line shows the upper bound and the lower bound for the Confidence Interval. The result above shows that with 95% Confidence Interval, there is **no significant** relationship between the news tweets' sentiment and the VIX.

7 CONCLUSION

In conclusion, there is no significant correlation between the average polarity of CNN Breaking News, BBC Breaking News, and NBC Breaking News tweets and the CBOE VIX. Although many news companies are accused of practicing fear-mongering in social media (Twitter) and most of the sentiments of news-based tweets are negative, the sentiment of their tweets are not necessarily the primary determinant of investor behaviour. This research has a numerous of limitations. Therefore, we would like to give some suggestions to future researchers, such as follows:

- Gather more data from various sources other than Twitter:
  Newspaper headlines, article headlines, Facebook posts, Instagram posts, and any other mass media that might also influence investor behaviours. Different weights are also required due to the difference in impact to the society or investors with different types of risk profile.
- Use public tweets instead of news-based tweets:
  - Public tweets with emojis can be treated as a training data. With an available training and testing data, machine learning algorithms will be able to assign the polarity accurately.
- Beware of lurking variables:
  There might be a third variable that influences both public sentiment and VIX—for example, the coronavirus cases itself, public policy, lockdowns, etc. The lurking variable might result in a spurious correlation.
- Fearmongering might influence other things than the VIX (e.g., consumer's behaviour and producer's behaviour)

8 APPENDICES

Rstudio version: 1.3.959
Session info:
R version 4.0.0 (2020-04-24)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 18363)
Matrix products: default
locale:
LC_COLLATE=English_Indonesia.1252
LC_CTYPE=English_Indonesia.1252
LC_MONETARY=English_Indonesia.1252
LC_NUMERIC=C
LC_TIME=English_Indonesia.1252

attached base packages:
stats, graphics, grDevices, utils, datasets, methods, base

other attached packages:
igraph_1.2.5, ggraph_2.0.3, scales_1.1.1, ggpubr_0.3.0, lmtest_0.9-37, zoo_1.8-8, tseries_0.10-47, lubridate_1.7.8, sentimentr_2.7.1, textclean_0.9.3, tidytext_0.2.5 ,forcats_0.5.0, stringr_1.4.0, dplyr_0.8.5, purrr_0.3.4, readr_1.3.1, tidyr_1.1.0 ,tibble_3.0.1, ggplot2_3.3.1, tidyverse_1.3.0, rtweet_0.7.0

Full Codes and Datasets
All codes and datasets are uploaded in our GitHub repository
https://github.com/matthewfarant/MASA_RI
All datasets are provided in the "datasets" folder. File with "full" prefix shows full twitter dataset of an account without cleaning & filtering process. File with "clean" prefix shows cleaned & filtered Covid-19 twitter dataset. File with "nf" suffix shows a fully cleaned twitter dataset without filtering (Covid-19 keywords) process.

9 GLOSSARY

Confidence Interval
A confidence interval, in statistics, refers to the probability that a population parameter will fall between two set values for a certain proportion of times.

CSV
Comma Separated Value.

Dataset
A collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer.

Exploratory Data Analysis
Exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

Fear-mongering
The action of deliberately arousing public fear or alarm about a particular issue.

Lexicon
the complete set of meaningful units in a language.

Natural Language Processing
A subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data

Packages
Packages are collections of functions and data sets developed by the community

Scraping
A technique in which a computer program extracts data from human-readable output coming from another program.

Seed
A random seed is a number or other value that is generated by software using one or more values. For example, hardware information, time, or date are different examples of values that help generate a random value used by a program or encryption.

S&P 500
Standard & Poor's 500 Index is a market-capitalization-weighted index of the 500 largest U.S. publicly traded companies.

Stationary

Stationarity means that the statistical properties of a process generating a time series do not change over time.

Stop words

stop words are words which are filtered out before or after processing of natural language data (text).

String

a string is traditionally a sequence of characters, either as a literal constant or as some kind of variable.

10 REFERENCES

[1] Boyd, D. and Ellison, N., 2010. Social network sites: definition, history, and scholarship. *IEEE Engineering Management Review*, 38(3).

[2] Gil, P., 2020. *What Is Twitter? And How Does It Work?*. [online] Lifewire. Available at: https://www.lifewire.com/what-exactly-is-twitter-2483331 [Accessed 15 August 2020].

[3] Oberlo.com. 2020. 10 Twitter Statistics Every Marketer Should Know In 2020 [Infographic]. [online] Available at: https://www.oberlo.com/blog/twitter-statistics [Accessed 15 August 2020].

[4] Schepers, J. and Wetzels, M., 2007. A meta-analysis of the technology acceptance model: Investigating subjective norm and moderation effects. Information & Management, 44(1).

[5] Liu, B., 2012. Sentiment Analysis And Opinion Mining. San Rafael, Calif.: Morgan & Claypool.

[6] Edwards, T. and Preston, H., 2017. A Practitioner's Guide to Reading VIX. [online] (201). Available at: https://www.spglobal.com/spdji/en/education-a-practitioners-guide-to-reading-vix.pdf [Accessed 15 August 2020].

[7] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

[8] Silge J, Robinson D (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS*, **1**(3).

[9] Liu, B., 2012. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), pp.1-167.