Sentimental Analysis of Social Media Tweets Using Leading Economic Indicators of Recession

by

Howard A Hamburger

Dr. Daniel Ayala-Rodriguez, Advisor

Masters project submitted in partial fulfillment of the
requirements for the Master of Science in Data Science degree
in the College of Arts and Sciences of
Lewis University

Abstract— This project is an investigation on whether Twitter sentiment derived from President Donald Trump's tweets can be used to predict future expansion or contraction in the United States economy using monthly, weekly or daily changes in leading economic indicators that have predicted previous recessions. The Copper-Gold-Ratio, Yield Curve Flattening and The Conference Board's Economic Index were used as the leading economic indicators to determine an expansion or contraction of the economy on any given date the market was open and matched that date with a tweet. The packages that were used to manipulate tweets were Natural Language Processing algorithms, such as RE and the Natural Language Toolkit. TextBlob was also used to measure Sentiment Analysis and HDFS MapReduce to measure frequency count on significant tweet words. Results showed that tweets produced by a powerful person in the world like the President of the United States might be used to predict future expansion or contraction in the economy if certain words or phrases are used often enough and consistently.

1. Introduction

1.1 Motivation

Since President Donald J. Trump has been inaugurated as President of the United States, he has made it a habit to consistently express attitudes and beliefs about important current events by posting several tweet messages a day on Twitter, and they have raised public awareness because of their content. Mr. Trump also wants to take credit for the booming economy that has expanded quickly with heavy growth since he became President. What would the significant consequences be if Mr. Trump's words really do have an influence on something as significant as the U.S. economy? It would be very appropriate to use sentiment analysis on President Trump's tweets to capture his moods and feelings on a day-to-day basis and relate them to the daily expansion and contraction of the U.S. economy. It is possible that a tweet with negative sentiment sent on a certain day could cause a daily contraction of the economy and one with positive sentiment could cause a daily expansion of the economy, or vice versa. A good way to determine whether the economy expands or contracts on a specific business day is to use reliable time-related leading economic indicators that in the past have predicted recessions in U.S. history.

Using key words or phrases that are most frequently used in tweets and daily, weekly or monthly changes in three reliable economic recession indicators that have historically predicted recessions, this paper is going to explore whether positive or negative tweets from the Donald Trump Tweet Archives from January 20, 2017 to January 31, 2018, (@realdonaldtrump, 2018) as determined by a Twitter Sentiment Classification Algorithm, correlate in any way with the expansion and contraction of the United States

economy. If some correlations do exist, it could be demonstrated that the tone and sentiment of the words in Donald Trump's tweets as well as certain words or phrases contained in the tweet, can have an influence on an expansion or contraction of the United States economy. It is even possible that one can predict which direction the economy will go either on a daily, weekly or monthly basis. A definition of Twitter and the various methods used to prepare the tweet messages for sentiment analysis follows.

1.2 TWITTER

Twitter is a well-known online social media platform with over 328 million monthly active users as of the first quarter of 2017 where users can read and post short 140-character messages called tweets as well as follow other users. (Twitter - Statistics & Facts, 2018) Twitter is also a platform where users can express and spread their opinions and attitudes about a topic of interest and there can be very strong sentiment within many of the messages sent. Cleaning up a message so that it can be analyzed for its content can be a challenge. In most tweets, language used by users is very informal with punctuation, misspellings, slang, new words, URLs and abbreviations all of which need to be removed. (Singh & Kumari, 2016) Splitting of attached words also needs to be noticed when cleaning. Natural Language Processing (NLP) is a skill used in data analytics to facilitate the interactions between computers and natural language so that large amounts of natural language data can be processed. Several packages or tools within the Python programming language library exist to help with this processing and are very helpful to help with the cleansing of the tweet. The ones used in this project are Regular Expression (RE), the Natural Language Toolkit (NLTK) and TextBlob.

1.2.1 RE - REGULAR EXPRESSION

RE is a library used for much of the preprocessing of non-alphanumeric characters, including special characters like emoticons. This library takes patterns in the words of tweets and performs specific operations on them based on where the characters are in the orientation of the word. For instance, one might want to remove a character at the very beginning or at the very end of the word. (Python Software Corporation, 2018) This project used RE to find all non-alphanumeric characters and replace them with blank spaces.

1.2.2 NATURAL LANGUAGE TOOLKIT (NLTK)

A tool that is helpful with dealing with insignificant words in the message is NLTK. This toolkit can be

used for removing stop words and to create lemmatization. Stop words are words like “the”, “a”, “is”, or “of” that are used frequently in the English language. Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. (TextMiner, 2014) For example, the words “softly” and “softer” would be grouped together and analyzed as “soft”, and “parsed” and “parses” would be grouped together and analyzed as “parse”.

1.2.3 TEXTBLOB

TextBlob is a Python library for processing textual data and is used consistently in common NLP tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more. It is built on top of NLTK that performs stopwords and lemmatization. (Loria, TextBlob Tutorial: Quickstart, 2018) For the cleaning of tweets in this paper, TextBlob was used for tokenization of the words.

1.3 SENTIMENT ANALYSIS

Another focus of this paper was to determine the sentiment of a Trump tweet to see if positive or negative sentiment in his tweets are correlated in any way to a particular direction the economy takes in one day. The goal of sentiment analysis is to determine the user of social media’s sentiment about a topic by classifying the tweet into categories of feelings and emotions such as positive and negative or neutral. Sentiment analysis has two ways of analyzing text. They are lexicon analysis and machine learning analysis. Lexicon analysis calculates the polarity of a document from the semantic orientation of words and phrases in a document and machine learning analysis involves building models from labeled training datasets in order to determine the orientation of a document. (El Alaoui, Gahi, Messoussi, Chaabi, & Todoskoff, 2018) This paper used the sentiment analysis utility in TextBlob using the polarity portion of property called sentiment that has a named-tuple of the form Sentiment (polarity, subjectivity). (PlanSpace, 2015) This sentiment property is able to evaluate both the polarity of the tweet, whether it be positive, negative or neural, and the subjectivity of the tweet or how objective or subjective the tweet is. This project does not use of the subjectivity portion of this property.

1.3 MAPREDUCE

This paper used an implementation called MapReduce to count the number of unique words in all the tweets and how frequently each of them is used so that appropriate word or phrases that are commonly used in tweets can be searched for to find significant correlations. MapReduce is an implementation for

processing and generating extremely large data sets using a parallel or distributed algorithm on a cluster. It is composed of a map procedure that breaks the objects down into smaller parts, executes a map function on each part, then gives all the results to a sorter. The sorter then sorts similar objects together, in this case words from the dataset. These results are then sent to a reduce method to merge all results into one, in this case a word frequency. (Lakshminarayanan, Acosta, Green II, & Devabhaktuni, 2014) MapReduce is efficient because it runs its various tasks in parallel or on distributed systems depending upon the size of the dataset.

2 Discussion of Related Work

This section will discuss other research not directly related to the creation of this project. Section 2.1 will discuss other research from various fields of study that used Twitter sentiment analysis in some way. Because a lot of data in the area of economics was used in this project, Section 2.2 will describe the leading economic indicators for a recession, how they were determined to be leading economic indicators and the formulas and calculations that determine whether the change of an economic indicator moved in a positive or negative direction.

2.1 Review of Literature On Sentiment Tweet Analysis

Twitter has become a very significant social media tool since 2007 with billions of tweets being streamed online on a daily basis and thus can be easily used for a variety of analytic studies. This paper focuses on how sentiment analysis of tweets created by an influential person like the President of the United States could have a causal effect on economic activity while other studies have used sentiment analysis to monitor human behavior on other specific events or topics. Many have used twitter sentiment analysis on stock market trends on specific stocks like the Apple Corporation. (Sharma, 2013) In the area of journalism, researchers had a goal to find tweets about newspaper reports and measure their popularity by forming a training set using tweets that contained a link to the news and the content of the same news article. The tweet was then preprocessed using morphological analyzers to remove unnecessary words and symbols to create a tweet with only content words. (Demirsoz & Ozcan, 2017)

Many researchers around the world have used sentiment analysis to collect information on human emotions from political elections. One research team used a sentiment analysis approach to classify tweets related to the 2016 U.S. election. A tweet was analyzed based on a user's opinion in real-time and the words that express that opinion were extracted. The polarity of these words was added to a dynamic

dictionary containing these words based on a selected set of hashtags related to a given topic. The tweets were then classified under several classes by introducing new features that strongly fine-tune the polarity degree of a tweet. Good accuracy in detecting positive and negative classes and their sub-classes was achieved. (El Alaoui, Gahi, Messoussi, Chaabi, & Todoskoff, 2018) Another research team complied dictionaries for sentiment and entity detection for the Greek language of tweets that were related to the Greek referendum of 2015 and subsequent legislative elections. Volume analysis, sentiment analysis, sarcasm analysis correction and topic modeling of the tweets were performed with results showing there was strong anti-austerity sentiment and that users of the social media were critical of the way the European Union and Greece handled the political situation. (Antonakaki, et al., 2017) Finally, the sentimental analysis of tweets related to a Legislative Assembly election of 117 members in Punjab, a state in India was researched. They gathered significant implications of messages that were sent over the duration of the election. (Singh, Gupta, & Singh, 2017)

Another field that has used sentimental analysis on human emotions is the area of healthcare. A group of researchers in their study looked at the sentiment dynamics of tweets and news publications on the issue of Ebola. They collected 16,189 news articles from 1006 different publications and 7.1 million tweets with the query term “Ebola” or “Ebola virus”. Analysis included the preprocessing of the tweets and news publications to get the significant word content of both media. To do this they used the N-gram LDA topic modeling technique and the adoption of entity extraction and entity networks. They also introduced the concept of topic-based sentiment scores. Results showed that tweets were more time oriented with less variance than news articles and that the news articles focused more on events like people and location. They concluded that news articles and tweets were distinguishable in terms of content coverage and sentiment dynamics. (Kim, Jeong, Kim, Kang, & Song, 2016)

Another paper looked at the importance of social media and interaction integration in an existing Smart Clinical Decision Support System (CDSS) that monitors health conditions for patients with chronic diseases. They extracted keywords, concepts and sentiments from patient’s tweets data, trajectory analysis to identify activities based on semantic tags and email analysis to find communication trends in daily routines of patients. The results of the combined data helped health practitioners to better understand the behavior and lifestyle of patients so they can make better decisions about treatment. (Fatima, et al., 2015)

2.2 Description of Leading Economic Indicators

This project made use of three leading economic indicators of recession to determine whether the economic expanded or contracted on a market business day. They are the Copper-Gold Ratio, Yield Curve

Flattening and the Conference Board's Leading Economic Index. In an FEG 2017 Investment Forum presentation, Jeffrey Gundlach, an American investor from Double Line highlighted that the copper-gold ratio is a powerful coincident indicator of the 10-year Treasury yield, a component used to determine Yield Curve Flattening. (Gundlach, 2017) This is due to the fact that the copper-gold ratio and the real yield and inflation components of the 10-year Treasury are highly correlated with global growth. (Cooper, 2017) The Conference Board's Leading Economic Index is also a good indicator of a predicting a recession and is used extensively to analyze many different aspects of any world economy. This index has dipped below zero, an indication of the economy moving towards a recession, prior to seven of the past eight recessions "with few false alarms." (Vlastelica, 2018) It has also been used as a forecasting model for the past five presidential elections to predict votes. (Erikson & Wlezien, 2016) There are ten variables for the Conference Board's Leading Economic Index: 1) the Average Weekly Hours, Manufacturing, 2) the Average Weekly Initial Claims for Unemployment Insurance, 3) the Manufacturers' New Orders, Consumer Goods and Materials, 4) the ISM® Index of New Orders, 5) Manufacturers' New Orders, Nondefense Capital Goods Excluding Aircraft Orders, 6) Building Permits, New Private Housing Units, 7) Stock Prices, 500 Common Stocks 8) Leading Credit Index™, 9) Interest Rate Spread, 10-year Treasury Bonds Less Federal Funds and 10) the Average Consumer Expectations for Business Conditions. (Board, The Conference Board Leading Economic Index, 2018) Each of these leading economic indicators will be discussed in detail below along with descriptions of each of the indicators within the Conference Board's Leading Economic Index.

2.2.1 The Copper-Gold Ratio

The Copper-Gold Ratio is a good indicator of economic expansion and contraction because copper prices tend to rise during an expanding global economy and gold prices tend to rise when there is a contracting global economy. Copper is a malleable metal with anti-bacterial properties and is often used in making pipes for plumbing and is even used to make certain medical devices. (Cooper, 2017) During an expanding economy, industries like construction produce more things that require the use of copper. As more copper is being utilized, the concept of supply and demand is applied; there is an increase in demand for copper and the price for copper goes up. Gold is a less useful metal than copper and is usually a safe-haven investment. Its prices tend to rise when investors are fearful about a contracting global economy. A rising price in gold thus signals a contracting economy, investor fears of a contracting economy or both. (Regenstein, 2018) The Copper-Gold Ratio is simply the market price of copper on a given day divided by the market price of gold on a given day so when there is an increase in the price of copper and a decrease in the price of gold, the ratio increases, a sign of an expanding economy, and when the price of

copper goes down and the price of gold goes up, the ratio decreases, a sign of a contracting economy.

2.2.2 Yield Curve Flattening

Yield Curve Flattening is a concept that suggests a recession may be imminent when the yield curve inverts. This happens when short-term bond yields like the 2-year U.S. Treasury Bonds move higher than yields for longer-term bonds like the 10-year U.S. Treasury Bonds. (Vlastelica, 2018) Increased inflation expectations also tends to play a part in the rise of the yield on 10-year Treasury Bonds during an expansion. As inflation increases, prices for U.S. Treasury bonds decrease resulting in increased yields. (Regenstein, 2018) Basically, Yield Curve Flattening occurs as two-year U.S. Treasury Bonds increase and ten-year U.S. Treasury Bonds decrease. (Yang, 2012)) The yields percentages approach each other in numbers and the curves flatten out. The Yield Curve Flattening data is thus calculated using the daily business day 2-year U.S. Treasury Yield percentage and the daily business day 10-year U.S. Treasury Yield percentage. If the 2-year yield percentage goes up and the 10-year yield percentage goes down, the economy is likely to be contracting. (Yang, 2012) All other possibilities where the 2-year yield percentage goes down and the 10-year yield percentage goes up or both go up or both go down would indicate that no recession is imminent so it is fair to say the economy is potentially growing.

2.2.3 The Conference Board's Leading Economic Index

There are ten variables in the Conference Board's Leading Economic Index. The variables such as average weekly hours, new orders, consumer expectations, housing permits, stock prices, and the interest rate spread, which make up most of the current variables in the index, are variables that tend to shift direction in advance of the business cycle. They also are reported early in the month and help executives make decisions about their company. For these reasons, they get the lion's share of the attention. (The Conference Board, 2001)

The Conference Board's Leading Economic Index uses a diffusion index to measure the proportion of the components that contribute positively to the index. A value of 1 is considered to be a positive contribution indicating an expanding economy, a value of 0 is considered to be a negative contribution indicating a contracting economy and a value of 0.5 is considered to be little change in the index from the previous report. Each component of the diffusion index is a value based on a percentage change from one report on a variable in the leading economic index to a previous report on the same variable. The

Conference Board uses is a rather simple formula to determine the value. They use the amount the percentage change either rises or falls to see if a variable in the diffusion index increased, decreased, or had no change. Percentage changes that rise more than 0.05% are given a value of 1, percentage changes that change less than 0.05% in a positive or negative direction are given a value of 0.5, and percentage changes that fall more than 0.05% are given a value of 0. The 10 values are then summed. That sum is then divided by 10 and then multiplied by 100. A final value over 50 will indicate an expanding economy while a value under 50 will indicate a contracting economy. (Board, The Conference Board Leading Economic Index, 2018)

The following subsections describe each of the ten variables in the Conference Boards' Leading Economic Index.

2.2.3.1 Average Weekly Hours, Manufacturing

This variable measures the average hours worked per week by production or factory-type workers in manufacturing industries. The source is the U. S. Department of Labor's Bureau of Labor Statistics. Also known as "factory hours", this component tends to lead the business cycle because employers usually adjust work hours before increasing or decreasing their workforce. (The Conference Board, 2001)

2.2.3.2 Average Weekly Initial Claims for Unemployment Insurance

This variable measures the average number of first-time filings of new claims for unemployment compensation on a weekly basis. The source is the U.S. Department of Labor. The number of new claims filed for unemployment insurance is typically more sensitive than either total employment or unemployment to overall business conditions, and this variable tends to lead the business cycle. Because initial claims increase when layoffs rise, this variable is inverted when included in the leading index. (The Conference Board, 2001)

2.2.3.3 Manufacturers' New Orders, Consumer Goods and Materials

This variable tracks orders for goods that are primarily used by consumers using monetary dollar values found Census Bureau Data. The Conference Board uses various price indexes from the industry and a chain-weighted price index formula to compute the variable and adjusts the dollar amounts for inflation. New orders are good indicators of lead production because they directly affect the level of both unfilled

orders and inventories that companies monitor when making production decisions. (The Conference Board, 2001)

2.2.3.4 ISM INDEX OF NEW ORDERS

This index monitors the amount of new orders made by customers and is based on a monthly survey conducted by Institution for Supply Management (ISM). As a part of a diffusion index, its value reflects the number of participants reporting increased orders during the previous month compared to the number reporting decreased orders. When the index has a reading of greater than 50 it is an indication that orders have increased during the past month. This variable tends to lead the business cycle and is reported early in the month. (Board, Description of Components, 2018)

2.2.3.5 MANUFACTURERS' NEW ORDERS, NONDEFENSE CAPITAL GOODS EXCLUDING AIRCRAFT ORDERS

This variable tracks orders received by manufacturers in nondefense capital goods industries using monetary dollar values found Census Bureau Data. The Conference Board uses various price indexes from the industry and a chain-weighted price index formula to compute the variable and adjusts the dollar amounts for inflation. It is the producers' counterpart to Manufacturers' New Order, Consumer Goods and Materials. This variable is in the index because new orders of capital goods in particular lead production and the business cycles. (The Conference Board, 2001)

2.2.3.6 BUILDING PERMITS, NEW PRIVATE HOUSING UNITS

This variable measures the monthly change in the number of permits for building private houses that were authorized by local permit issuers. The source is a survey on residential construction produced by the U.S. Census Bureau. The number of residential building permits issued is an indicator of construction activity, which typically leads most other types of economic production. (The Conference Board, 2001)

2.2.3.7 STOCK PRICES, 500 COMMON STOCKS

This variable also known as the "S&P 500", tracks the change in prices of common stocks traded on the New York Stock Exchange. Increases and decreases in this stock index can have an effect on the mood of investors and the changes in interest rates, both of which are good indicators for future economic activity. (The Conference Board, 2001)

2.2.3.8 LEADING CREDIT INDEX™

This index is consisted of six financial indicators: a 2-year Swap Spread, LIBOR 3 months less 3 months Treasury-Bill yield spread, Debit balances at margin account at broker dealer, AAII Investors Sentiment Bullish percentage less Bearish percentage, a Senior Loan Officers C&I loan survey which deals with bank tightening credit to large and medium companies, and Security Repurchases from the Total Finance-Liabilities section of Federal Reserve's flow of fund report. All of these financial indicators look forward to the state of the economy so they are good indicators for future economic activity. (Board, Description of Components, 2018)

2.2.3.9 INTEREST RATE SPREAD, 10 YEAR TREASURY BONDS LESS FEDERAL FUNDS

This variable is similar to a Yield Curve Flattening and is a difference between long-term and short-term rates. It is a simple measure of the slope of the yield curve. It is constructed using the 10-year Treasury bond yield rate and the Federal funds rate. This variable is a good indicator of monetary policy and general financial conditions, because it rises when short-term rates are relatively low and falls when short-term rates are relatively high. When short-term rates are higher than long-term rates, this variable becomes negative and strongly indicates a recession. (The Conference Board, 2001)

2.2.3.10 AVERAGE CONSUMER EXPECTATIONS FOR BUSINESS CONDITIONS

This index reflects changes in consumer attitudes concerning future economic conditions and is the only variable in the leading index that is completely expectations-based. A monthly survey by the University of Michigan's Survey Research Center is conducted to collect data on responses to questions relating to economic prospects for the respondent's family over the next 12 months, the economic prospects for the entire country over the next 12 months and the economic prospective for the entire county over the next five years. Each of these expectation responses are then classified as positive, negative, or unchanged. (The Conference Board, 2001)

3 PROJECT DESCRIPTION

This project investigated how the sentiment analysis of tweets and the words used by Donald Trump in his tweets from the period between January 20, 2017 and January 31, 2018 influenced the direction the United States economy takes every business day the commodity markets are open. Figure 1 is a flow

diagram showing an overview of the project. Each tweet from the Donald Trump Tweet Archives was taken one at a time and cleaned with the methods of RE and NLP. This includes stop words and lemmatization of words. This cleaned tweet is then tokenized by the TextBlob method to create a list of individual words. This list was later matched by the date associated with the tweet with a sentiment polarity label and an economic direction label to form a Panda Series Data Frame. Figure 2 shows a flow chart of how the leading economic indicators data were processed to obtain a label of expansion or contraction for each date the commodity market was open for business. This was done using the formulas and calculations as discussed in Section 2. Once all the tweets were all transformed by the packages of RE and NLP and the economic direction labels were established for each date, the tweet's date, sentiment analysis polarity matched with the tweet's date and economic direction label, also matched with each tweet's date, were put into Series lists and joined together into a Python Panda Data Frame. If a tweet was made on a market holiday or weekend, the economic direction label of the next date when the market was open for business was used as the label for that particular tweet. After this data frame was created, it was used to perform SQL-like queries in a Panda Series Data Frame syntax to extract all the tweets from the original Donald Trump Tweet Archives file that contain the word or phrase using words and phrases with high frequency count based on the results from the MapReduce implementation.

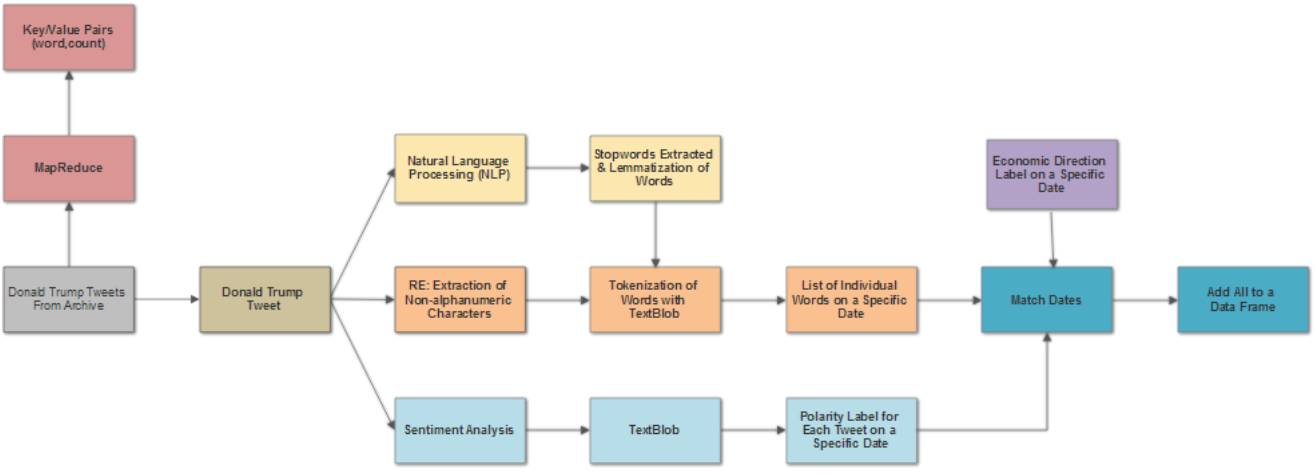

Fig. 1. Flowchart of Overview of Project

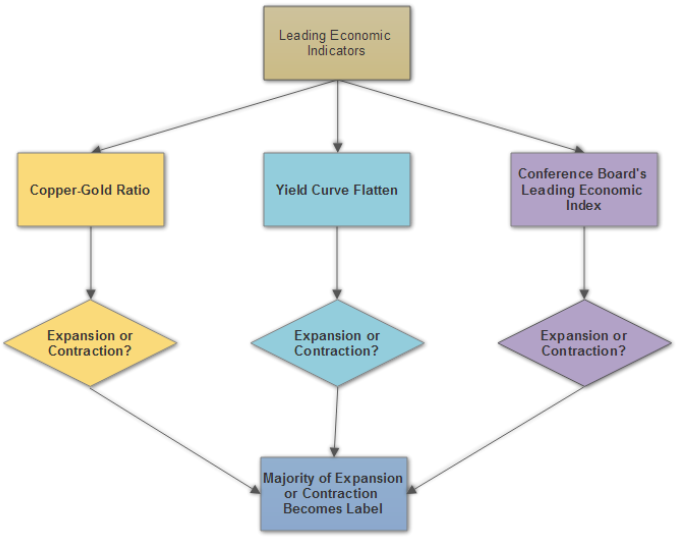

Fig. 2. Flow Chart showing how leading economic indicators were processed

TABLE 1

EXAMPLE OF 6 CATEGORIES OF ECONOMIC DIRECTION AND SENTIMENT POLARITY

Hillary FBI	Sentiment Polarity		
	Positive	Neutral	Negative
Expansion of Economy	30%	10%	20%
Contraction of Economy	20%	10%	10%

Using only the extracted tweets that resulted from the query of a word or phrase, six categories were constructed based on the sentiment polarity and the economic direction label as shown in Table 1.

There are three sentiment outcomes, a positive, a neutral or a negative outcome and two economic labels, "Expansion of Economy" and "Contraction of Economy". Percentages for each category were obtained from the data frame using an SQL-like query that counted the number of occurrences where a specific sentiment polarity and economy label grouped together were identical and dividing that count by the total number of extracted tweets from the word or phrase query. Table 1 shows the percentages for each category using the word phrase "Hillary FBI". Here, 30% of the tweets that were extracted using the query "Hillary FBI" had a positive sentiment and an expansion of the economy label and 20% of them had a negative sentiment and an expansion of the economy label. A pie chart was then constructed using all six categories. Analysis was then done by looking for relationships between the word or phrase and any

salient features in the chart. For example, one slice of the pie could be significantly larger than the rest or one slice that was consistently small or non-existent in most of the charts had a larger than normal slice in a specific chart. Also, if the pie was looked at from an expansion and contraction point of view and divided into two sections, other features could be seen like much more expansion than contraction for a particular word or phrase.

4 Details of the Methodology

The methodology in this project used several packages from the Python 3.6 library and made use of different library attributes or properties of a specific method. Their overall purpose was to step by step manipulate each tweet down to significant words for analysis. First, RE was used which removed non-alphanumeric characters from the tweet. Second, NLTK removed stop words from the tweet and performed a lemmatization of words that are grammatically related. Third, TextBlob, which works on top of NLTK in the Tokenization of the Words property, separated the tweet into individual words. Fourth, once the tweet was down to individual significant words, TextBlob with the property of Sentiment Analysis was used to measure a sentiment polarity of each tweet. Finally, to obtain a set of commonly used words in the tweets to see the effect the sentiment of the tweet had with certain words or word phrases and to see the how that tweet's sentiment affected the economy market on the day that tweet was made, the method of MapReduce was executed on the entire dataset of processed tweets to obtain the frequency count of each significant word.

All of the above methods are described in detail discussing only the details of properties or attributes used in the Python code created to achieve the goals of the project or in the case of MapReduce to obtain the frequency count of all of the tweet words. Section 4.1 will cover the methodology of RE. Section 4.2 will cover the methodology of NLTK, Section 4.3 will cover the methodology of TextBlob using the property of words as related to the Tokenization of tweets. Section 4.4 will cover the methodology of TextBlob using the property sentiment. Section 4.5 will cover the methodology of MapReduce.

4.1 RE – Regular Expressions Operations

The first stage in cleaning up the raw tweet message is to get rid of non-alphanumeric characters. The main focus of the method RE that was used in this paper is to search for these types of characters and to remove them. The patterns RE looks for in the Python code are Unicode strings with 'str' similar to the type 'str' in the Python language. It is common to use the backslash character ('\') to indicate special forms or to allow special characters to be used without invoking their actual meaning. (Python Software

Corporation, 2018)

4.1.1 – Regular Expressions Syntax

RE specifies a set of strings that matches it using a specific function to see if a particular string matches a given regular expression. This paper uses the *sub* and *findall* functions that are explained in Section 4.1.2. Regular expressions can contain both special and ordinary characters. Most ordinary characters, like 'A', 'a', or '0', are the simplest regular expressions; they simply match themselves. One can also concatenate ordinary characters, so last matches the string 'last'. Some characters, like '|' or '(', are special. Special characters either stand for classes of ordinary characters or affect how the regular expressions around them are interpreted. A list of special characters that are used in the project are listed in Table 2 along with their general function. One can refer to RE documentation at https://docs.python.org/3/library/re.html for more details. (Python Software Corporation, 2018)

4.1.2 – Regular Expression functions

The following are the RE functions used in the project to match non-alphanumeric characters in the tweet. There are many more functions that can be used within this package.

re.findall

The *findall* function returns all non-overlapping matches of pattern in string, as a list of strings. The string is scanned left-to-right, and matches are returned in the order found. If one or more groups are

TABLE 2
PARTIAL LIST OF SPECIAL RE CHARACTERS AND THEIR FUNCTIONS
(Python Software Corporation, 2018)

.	(Dot.) In the default mode, this matches any character except a newline.
^	(Caret.) Matches the start of the string
$	Matches the end of the string or just before the newline at the end of the string
*	Causes the resulting RE to match 0 or more repetitions of the preceding RE, as many repetitions as are possible. ab* will match ‘a’, ‘ab’, or ‘a’ followed by any number of ‘b’s.
+	Causes the resulting RE to match 1 or more repetitions of the preceding RE. ab+ will match ‘a’ followed by any non-zero number of ‘b’s; it will not match just ‘a’.
?	Causes the resulting RE to match 0 or 1 repetitions of the preceding RE. ab? will match either ‘a’ or ‘ab’.

\	Either escapes special characters (permitting you to match characters like '*', '?', and so forth), or signals a special sequence; special sequences are discussed below.
[ ]	Used to indicate a set of characters. In a set: Characters can be listed individually, e.g. [amk] will match 'a', 'm', or 'k'. Ranges of characters can be indicated by giving two characters and separating them by a '-', for example [a-z] will match any lowercase ASCII letter.
\W	Matches any character which is not a word character. If the ASCII flag is used this becomes the equivalent of [^a-zA-Z0-9_].
|	A|B creates a regular expression that will match either *A* or *B*. An arbitrary number of REs can be separated by the '|' in this way and are tried from left to right. When one pattern completely matches, that branch is accepted. This means that once *A* matches, *B* will not be tested further. To match a literal '|', \\| is used or enclosed it inside a character class, as in [|].
(?(id)yes-pat|no-pat)	Will try to match with yes-pat if the group with given id exists, and with no-pat if it doesn't. no-pat optional and can be omitted.

present in the pattern, the function will return a list of tuples. Empty matches are included in the result. (Python Software Corporation, 2018) Figure 3 shows how the *findall* function is used in the project. The re.findall('(?::|;|=)(?:-)?(?:\)|\(|D|P)', text) is going to find all emoticons in the tweet using the (?(id)yes-pat|no-pat) where ':' is the id for eyes in an emoticon. In the first part (?::|;|=), it will try to find all types of emoticon eyes; either a ':' for a regular eyes open and ';' for a wink or '=' for another type of eyes. (?:-)? will find the emoticon :- and all repetitions of it. (?:\) will find all :\ which is a sad emoticon. \(|D|P)' is going to find mouths that are wide open or with the tongue sticking out.

re.sub

The *sub* function returns the string obtained by replacing the leftmost non-overlapping occurrences of the pattern in the string by the replacement. If the pattern isn't found, the string is returned unchanged. The replacement can be a string or a function. If it is a string, any backslash escapes in it are processed. For example, \n is converted to a single newline character and \r is converted to a carriage return. Unknown escapes such as \& are left alone. Back references, such as \6, are replaced with the substring matched by group 6 in the pattern. In Figure 3, the sub function re.sub('<[^>]*>',",text) will take out any hashtags from the tweet that are delimited by brackets '<', '>' in the tweet and replace it with a double quote ' " '. The ^> inside the [ ] will be treated as a set of characters. The * after the [ ] will match as many repetitions as possible of the '<', '>'. Also in Figure 3, re.sub ('[\W]+', ' ', text.lower()) +\, the \W in [ ] will match any character that is not a word character and replace it with a null value. The '+' will match one or more of the \W. This line of code also lower cases any characters that might be in uppercase. (Python Software Corporation, 2018)

```
def preprocessor(text):
text = re.sub('<[^>]*>', '', text)
emoticons = re.findall('(?::|;|=)(?:-)?(?:\)|\(|D|P)', text)
text = re.sub('[\W]+', ' ', text.lower()) +\
' '.join(emoticons).replace('-', '')
return text
```

Fig. 3. Sample of code from project showing the use of the sub and findall functions from the re library.

4.2 Natural Language Toolkit (NLTK)

The Natural Language Toolkit (NLTK) is being implemented for stop words and for the lemmatization of words. Stop words are common words that generally do not contribute to the meaning of a sentence such as “the” ,“a”, “and” or “of”. This is commonly done in many search engines to save time on indexing. (Perkins, 2014). NLTK comes with a stopwords corpus that contains word lists for many languages. In this case we are using the English language. This corpus is an instance of nltk.corpus.reader in the Python 3.6 Library. It has a words method that will retrieve all the English stop words. In the example here from (Perkins, 2014) in Figure 4, the line of code [word for word in words if word not in english_stops] is the line that will extract all of the stop words in the words list.

```
from nltk.corpus import stopwords
english_stops = set(stopwords.word(‘english’))
words = [“Can’t”, ‘is’, ‘a’, ‘contraction’]
[word for word in words if word not in english_stops]
```

Fig. 4. (Perkins, 2014) Sample of stopword code from the Perkins, 2014 cookbook showing how stopwords is used.

Figure 4 shows a sample from the project that obtains the English stop words from nltk.corpus. This project has a stopwords list that also excludes single letters of the English alphabet. Figure 5 also shows how the project lemmatizes word using word.lemma. lemma is an instance that comes from the nltk.corpus.reader.wordnet

A lemma in nltk.corpus.reader.wordnet document can be created from a <word>.<pos>. <number>.<lemma> where <word> is the morphological stem, <pos> is one of the module attributes like the parts of speech (adjective, noun, verb), <number> is a sense number counting from 0 and <lemma> is the morphological form of interest. Some of the lemma methods used to retrieve related lemmas and return lists of Lemmas include antonyms, hypernyms, instance_hypernyms, hyponyms, instance_hyponyms, member_holonyms, substance_holonyms, part_holonyms, member_meronyms, substance_meronyms, part_meronyms among many others. For more details one can go to the nltk.org website under the corpus.reader.wordnet link. (NLTK Project, 2017)

```
stop = stopwords.words('english')
stop = stop +
[u'a',u'b',u'c',u'd',u'e',u'f',u'g',u'h',u'i',u'j',u'k',u'l',u'm',u'n',u'o',u'p',u'q',u'r',u's',u't',u'v',u'w',u'x',u'y',u'z']

def split_into_lemmas(tweets):
    words = TextBlob(tweets).words
    return [word.lemma for word in words if word not in stop]
```

Fig. 5. Sample code of Stopwords, TextBlob for Tokenization of Words and Lemmatization of Words in the Project

4.3 TextBlob for Tokenization

Tokenization is a basic process of NLP. TextBlob is the package this project uses to separate all of the words in the tweet into individual words. TextBlob is a Python library package that processes textual data and is an API to help work on tasks related to NLP like parts of speech, phrase extraction and sentiment analysis. (Loria, TextBlob Tutorial: Quickstart, 2018) TextBlob's specific purpose with tokenization in this project was combined with the lemmatization and removal of stop words from the tweets. The Python method used in the project, shown in Figure 5, took the unaltered tweet, split the words into tokens or individual words then lemmatized the word and removed a word if it was a stop word. Since the tweet passed into the Python method is already a sentence, the word attribute was used to separate the sentence into words.

4.4 TextBlob for Sentiment Analysis

The TextBlob implementation of sentiment analysis uses a property called sentiment which results in a namedtuple of polarity and subjectivity. The scores for polarity go from -1 to 1. A negative result for polarity indicates negative sentiment while a positive one indicates positive sentiment. In research, a polarity score that falls between 0.05 and -0.05 is considered to be neutral in sentiment. In terms of subjectivity, the number is a float within the range 0.0 to 1.0 where 0.0 is very objective and 1.0 is very subjective. Subjectivity is not used in this project.

4.4.1 How TextBlob Calculates the Polarity and Subjectivity in Sentiment Analysis

Figure 6 shows a sample sentence using the sentiment property and the results of both polarity and subjectivity. When the sentence "not a very great calculation" is put into TextBlob a polarity of about -0.3 meaning it was slightly negative, and a subjectivity around 0.6 resulted. The numbers generated to obtain these scores comes from sloria/TextBlob on GitHub. (Loria, Steve Loria GitHub Page , 2014) The lexicon where the default sentiment calculations are stored within this source code is in a document called en-

sentiment.xml. Each line in this file shows a word it's "pos" or grammatical form like noun or adjective and a "sense" which defines one of the definitions of the word then gives a polarity score, subjectivity score, its intensity and its confidence. Figure 7 shows all the entries for the word "great". To calculate a sentiment or subjectivity score, the average of all the scores under each category is calculated. If that is done to the word "great", the sentiment score is 0.8 and the subjectivity score is 0.75. Negativity is not handled by TextBlob but a word like "not" will multiply the average of "great" by -0.5 and for modifiers like the word "very" the intensity score is used. In the en.sentiment.xml source code the word "very" has an intensity score of 1.3. Now if the negation word "not" is intensified by the word "very", in addition to multiplying by -0.5 for the word "not", the inverse of the intensity score is entered to the equation for both polarity and subjectivity. So for the phrase "not very great", its sentiment score would be -0.5 *(1/1.3)*0.8 ≈ -0.31 which is on the negative sentiment side. Since modifiers have no effect on subjectivity, the subjectivity score would be calculated as (1/1.3) *0.75 ≈ 0.58. These numbers match the scores seen in Figure 6. The words "calculation" and "a" have no sentiment or subjectivity to TextBlob so they are ignored.

```
from textblob import TextBlob

TextBlob("not a very great calculation").sentiment
>>Sentiment(polarity=-0.3076923076923077, subjectivity=0.5769230769230769)
```

Fig. 6. TextBlob Sentiment Analysis example (PlanSpace, 2015)

```
<word form="great" cornetto_synset_id="n_a-525317" wordnet_id="a-01123879" pos="JJ" sense="very good" polarity="1.0" subjectivity="1.0" intensity="1.0" confidence="0.9" />
<word form="great" wordnet_id="a-01278818" pos="JJ" sense="of major significance or importance" polarity="1.0" subjectivity="1.0" intensity="1.0" confidence="0.9" />
<word form="great" wordnet_id="a-01386883" pos="JJ" sense="relatively large in size or number or extent" polarity="0.4" subjectivity="0.2" intensity="1.0" confidence="0.9" />
<word form="great" wordnet_id="a-01677433" pos="JJ" sense="remarkable or out of the ordinary in degree or magnitude or effect" polarity="0.8" subjectivity="0.8" intensity="1.0" confidence="0.9" />
```

Fig. 7. Samples from the en-sentiment.xml file for the word great (Loria, Steve Loria GitHub Page , 2014)

MAPREDUCE

The MapReduce framework was developed for large-scale parallel and distributed processing. The framework is responsible for executing tasks in parallel, dealing with fault tolerance, making sure data transfers are executed smoothly, and making sure the data is load balanced over a cluster of machines usually in a distributed manner. (Lakshminarayanan, Acosta, Green II, & Devabhaktuni, 2014) The data are written to and read from a Hadoop Distributed File System (HDFS). MapReduce is composed of a Map phase

and a Reduce phase where several tasks are executed in the Map phase as well as the Reduce phase in parallel.

This project used the implementation of MapReduce to find the counts of all of the significant words in every tweet in the dataset. Figure 8 shows a theoretical example of a MapReduce algorithm that counts words. In this figure, it is dealing with fruit but in the project, it is dealing with individual significant words from tweets. MapReduce first starts on the large list of words by taking the list and splitting it into smaller hopefully equal parts. For example, if the input is a list of words in a text file like [Apple Orange Mango Orange Grapes Plum], the Map Function will distribute the words into two parts [Apple Orange Mango] and [Orange Grapes Plum]. This is done so that these parts can be executed in parallel on different HDFS data nodes which could also be on different machines. This helps cut down on execution time dramatically.

Now the word is prepared for counting. A Map function will create a Key-Value pair and give every word a Value 1 with the word as a Key. This function can also be executed in parallel on different HDFS data nodes. The output produced by a Map function is a set of Key-Value pairs for all the words which could be sitting in different HDFS data nodes. Local Combiners on one machine can do a mini-reduce of similar Keys to help make the next Sort and Shuffle phase more productive. All of these outputs from every machine that performed a Map function is then sent on to a Sort and Shuffle phase where similar values are combined together. The function reduceByKey in Hadoop Spark, used in this project, is a function that can handle such a task. All outputs that have “Orange” as output are combined into one list and a count of them is produced, all output that have “Apple” are combined into one list with a count of them and so on. The result is a list of Key-Value Pairs with the word as a Key and the count as the value.

MapReduce has other components to the system that help keep it productive and running smoothly. A Scheduler ensures that multiple tasks from multiple map or reduce jobs are executed on the cluster and a Synchronization component ensures that the reduce phase does not start until the map phase is done.

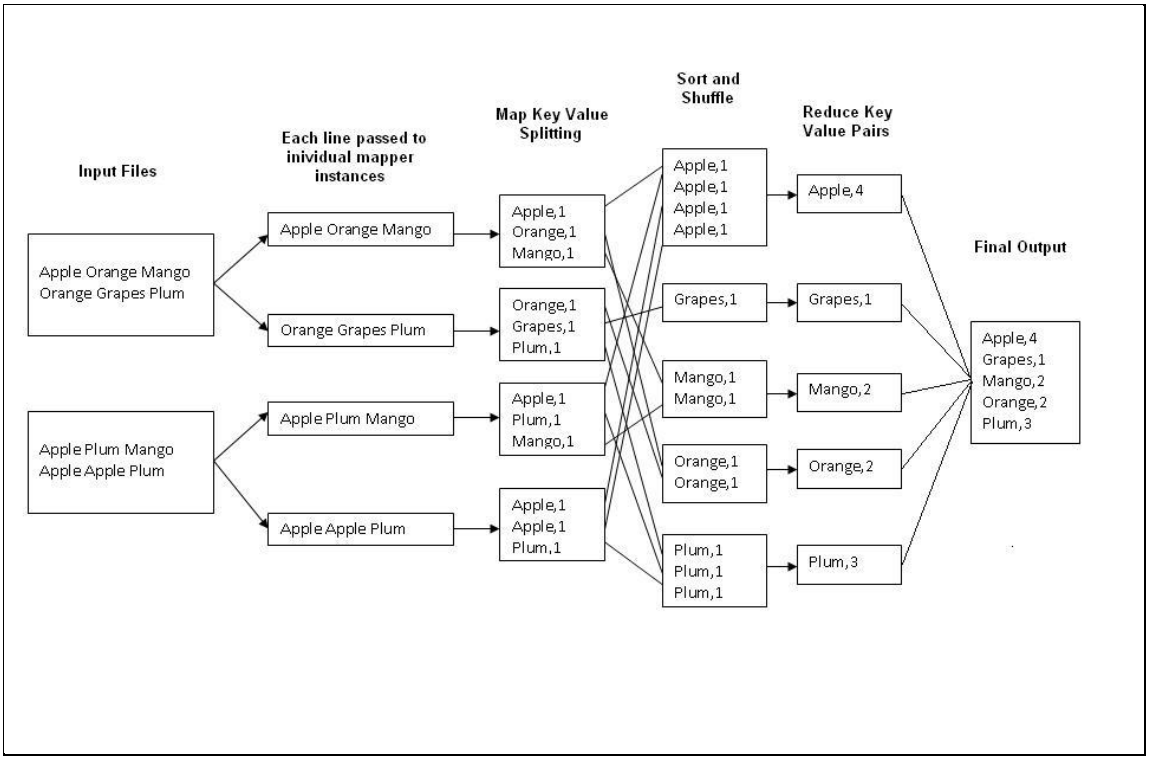

Fig. 8. MapReduce Word Count example (Used with permission)

4 EXPERIMENTS/IMPLEMENTATION DESCRIPTION

The main data structure where the implementation of the project is taken from is an Panda Series Data Frame which includes a series of tweet dates, a series of tweets in their original form, an economic tag for the tweet which is either an “expansion” or “contraction” and a sentiment polarity of ‘1’ if it has positive sentiment, ‘0’ if it is neutral in sentiment and ‘-1’ if it has negative sentiment. The results of the MapReduce Key/Value pairs of words with their frequency counts was compiled into a list from the highest frequency word to the lowest. All words that were only found once in the tweets were discarded. This list was used to choose appropriate words and word phrases to use based on significant events that occurred with President Trump in 2017. Phrases like “fake news”, “make america great”, and “tax cut” were used to search for these words in all of the tweets. A list of tweets was added to a list when the word or all of the words in the phrase were found in the tweet. All of the tweets and their related dates, sentiments and economic direction (expansion or contraction) were all incorporated into a Panda Series Data Frame so that Panda Data Frame queries, similar to an SQL query, of each word or phrase can be performed that calculates the percentages of each category as shown in Table 1. Table 3 shows the list of words and phrases used along with the count of the number of tweets that matched the query, the

percentage of the largest category and the largest category sentiment and economic direction. A sample tweet from the list is included that matches the sentiment value very accurately. This experiment was performed to get a general summary of the sentiment and expansion or contraction labels for each word or phrase that was queried in the program.

TABLE 3
SUMMARY OF WORDS OR PHRASES AND THEIR ATTRIBUTES

	Number of Samples	%	Sentiment	Expansion or Contraction	Tweet Sample
Politics					
healthcare/ obamacare	123	48.0	Positive	Expansion	obamacare premiums are going up up up just as i have been predicting for two years obamacare is owned by the democrats and it is a disaster but do not worry even though the dems want to obstruct we will repeal amp replace right after tax cuts
tax cut	116	56.0	Positive	Expansion	yesterday was a big day for the stock market jobs are coming back to america chrysler is coming back to the usa from mexico and many others will follow tax cut money to employ-ees is pouring into our economy with many more companies announcing american business is hot again
tax cut democrat/ dems	24	54.2	Positive	Expansion	the tax cuts are so large and so meaningful and yet the fake news is working overtime to follow the lead of their friends the defeated dems and only demean this is truly a case where the results will speak for themselves starting very soon jobs jobs jobs
tax cut republican	52	54.5	Positive	Expansion	biggest tax bill and tax cuts in history just passed in the senate now these great republicans will be going for final passage thank you to house and senate republicans for your hard work and commitment
The Economy					
economy	44	54.5	Positive	Expansion	really great numbers on jobs amp the economy things are starting to kick in now and we have just begun don t like steel amp aluminum dumping
make america great	58	86.2	Positive	Expansion	together we will make america great again americafirst https t co ajic3f lnki 58
people job/working	15	73.3	Positive	Expansion	does anybody really want to throw out good educated and accomplished young people who have jobs some serving in the military really
News Media					

fake news	156	60.9	Negative	Expansion	the fake news refuses to talk about how big and how strong our base is they show fake polls just like they report fake news despite only nega- tive reporting we are doing well no body is going to beat us make america great again
foxandfriends	104	35.6	Neutral	Expansion	ivanka trump will be interviewed on foxandfriends
Realdonald trump	84	39.3	Positive	Expansion	rt gopchairwoman realdonaldtrump is the paycheck president learn how the tax bill will put more money in your pocket amp how to contact
Foreign Affairs					
fbi hillary	10	30.0	Positive	Expansion	so general flynn lies to the fbi and his life is destroyed while crooked hillary clinton on that now famous fbi holiday interrogation with no swearing in and no recording lies many times and nothing happens to her rigged system or just a double standard
great meeting	35	80.0	Positive	Expansion	great meeting with a wonderful woman today former secretary of state condoleezza rice usa https t co zuri ic3ywg
north korea	52	38.5	Positive	Expansion	when will all the haters and fools out there realize that having a good relationship with russia is a good thing not a bad thing there always playing politics bad for our country i want to solve north korea syria uk raine terrorism and russia can greatly help
russian/russia collusion	16	43.8	Positive	Expansion	it is now commonly agreed after many months of costly looking that there was no collusion between russia and trump was collusion with hc
Wall	39	38.5	Positive	Expansion	we must have security at our very dangerous southern border and we must have a great wall to help protect us and to help stop the massive inflow of drugs pouring into our country

Once the percentages were obtained for the six categories, a pie chart was constructed for each word or word phrase showing the percentage results for each of the six categories. Pie charts were then grouped by category to compare and contrast the percentages to make conclusions about the sentiment and economic direction of tweets within specific topics as discussed in Section 3.

5 RESULTS

The words and word phrases in Table 3 were divided into four topics: 1) Politics, 2) The Economy, 3) News Media and 4) Foreign Affairs. In general, the U.S. Economy expanded more than it contracted on a day to day market basis from January 20, 2017 to January 31, 2018, the period the data collected for

this project covers. In approximately 80% of the period to period market changes in the leading economic indicators, the economy expanded and in approximately 20% of the changes, the economy contracted. Results in Table 3 and in Figures 9 through 12 showed that positive sentiment in tweets and economic expansion (shaded light blue) have the highest percentages in a pie chart mainly because the U.S. economy did so well in 2017. Results are therefore going to show a lot more expansion in the economy than contraction. It is also clear that Donald Trump and those who helped him write tweets used positive words almost all of the time, which could possibly have an influence on the high percentage of the positive

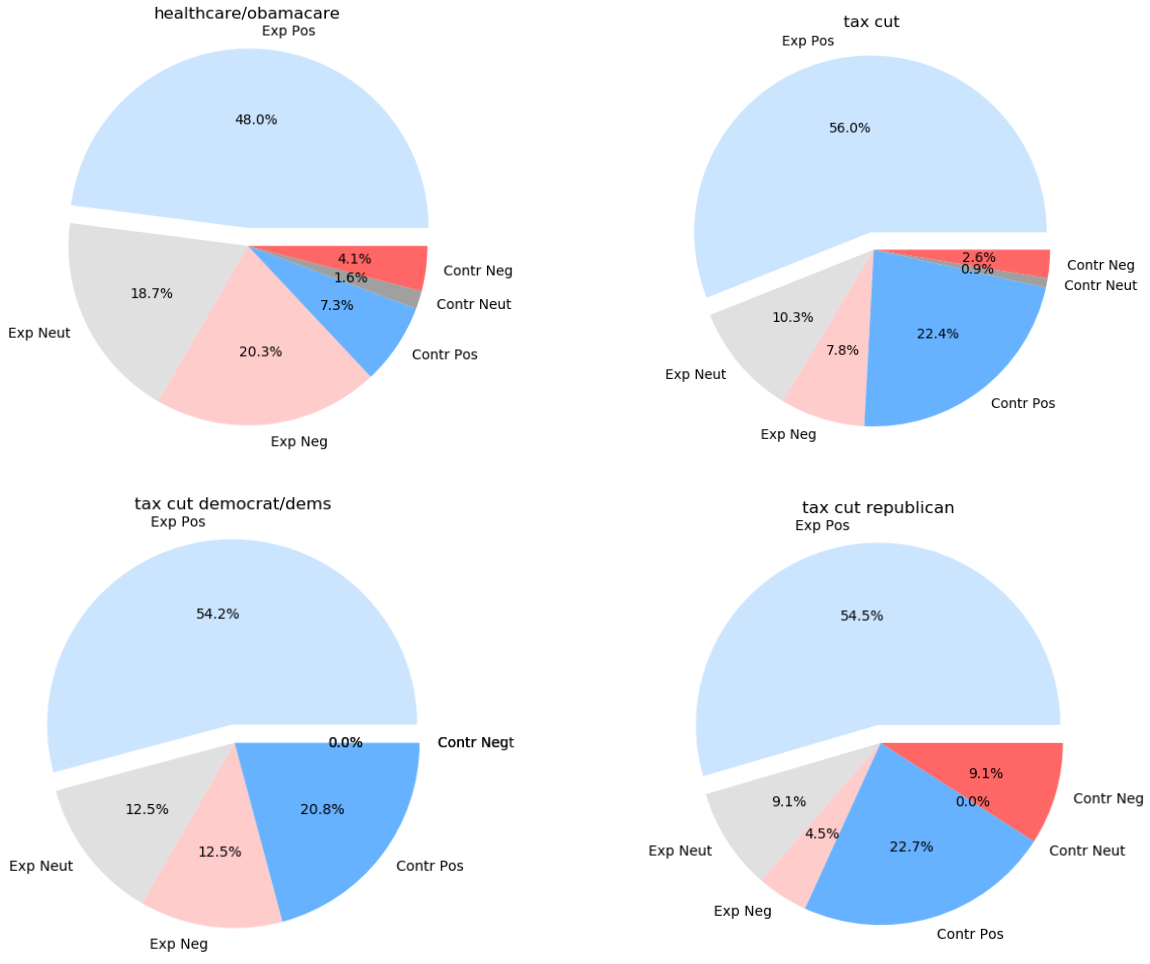

Fig. 9. Pie charts showing percentage results on the topic of politics

sentiment with an expanding economy category. Therefore, in order to make these results more meaningful and significant in the area of how the tweets influence the direction of the economy, it is best to focus on some of the salient features of the pie charts outside of the expansion of the economy category. The most interesting categories on the pie chart would be when the sentiment was positive yet the economy showed a contraction on the same day the tweet was sent (shaded in dark blue) or when the sentiment

was negative in a tweet on one day yet the economy expanded on that day (shaded in pink).

6.1 The Influence of tweets on Politics

Figure 9 shows four pie charts about different topics related to two political events in 2017, the attempted repeal of Obamacare and tax cuts. In regard to the attempted repeal of Obamacare or when healthcare is discussed in a tweet, 20.3% of the total tweets on this topic showed an expansion in the economy when the tweet had negative sentiment. It is possible that using some negative sentiment in tweets on a topic like Obamacare and healthcare, which some people might have a negative viewpoint on, is a politically smart decision when writing tweets about this topic.

In regard to tax cuts, there was a lot of positive sentiment with contraction in the economy in all of the pie charts. This could signal that not all investors agreed that tax cuts were a good idea. A contraction in the economy with negative sentiment when the word "republican" was also inside the tweet appeared to be significant. This category shows 9.1% when the word "republicans" is used in a tweet verse 0% when "democrat" or "dems" was used. It appears Mr. Trump used some negative sentiment when talking about republicans in his tweets. After investigating which tweets contributed to this specific category using the data frame query for that category in the Python program, three tweets were found that showed a consistent theme in all of them. The tweets say that the republicans were working hard or pushing hard to come up with a tax cut plan. It appears that the word "hard" was classified as negative.

The contraction in the economy and positive sentiment were pretty much the same for the word "republican" vs "democrat" or "dems" at 22.7% and 20.8% respectively, indicating that the focus of a contraction in the economy for a small fraction of investors is based on the notion of a tax cut that is not partisan in nature. The expansion in the economy and negative sentiment at 12.5% for when the "democrat" or "dems" word was used in the tweet on this topic was much higher. This just means that the tweets were not as kind to democrats and they were to republicans which is to be expected from a Republican president.

6.2 The Influence of Tweets on The Economy

The topics related to the economy are shown in Figure 10 and include the use of the word "economy", and the topic of jobs and making America great again. Overall, these topics had a very positive effect on the expansion of the economy using positive sentiment most of the time. The topic and Mr. Trump's

main campaign slogan “Make America Great Again” worked very well with 86.2% in the economy expansion with positive sentiment mainly due to the very positive word “great”. The word “great” was also the

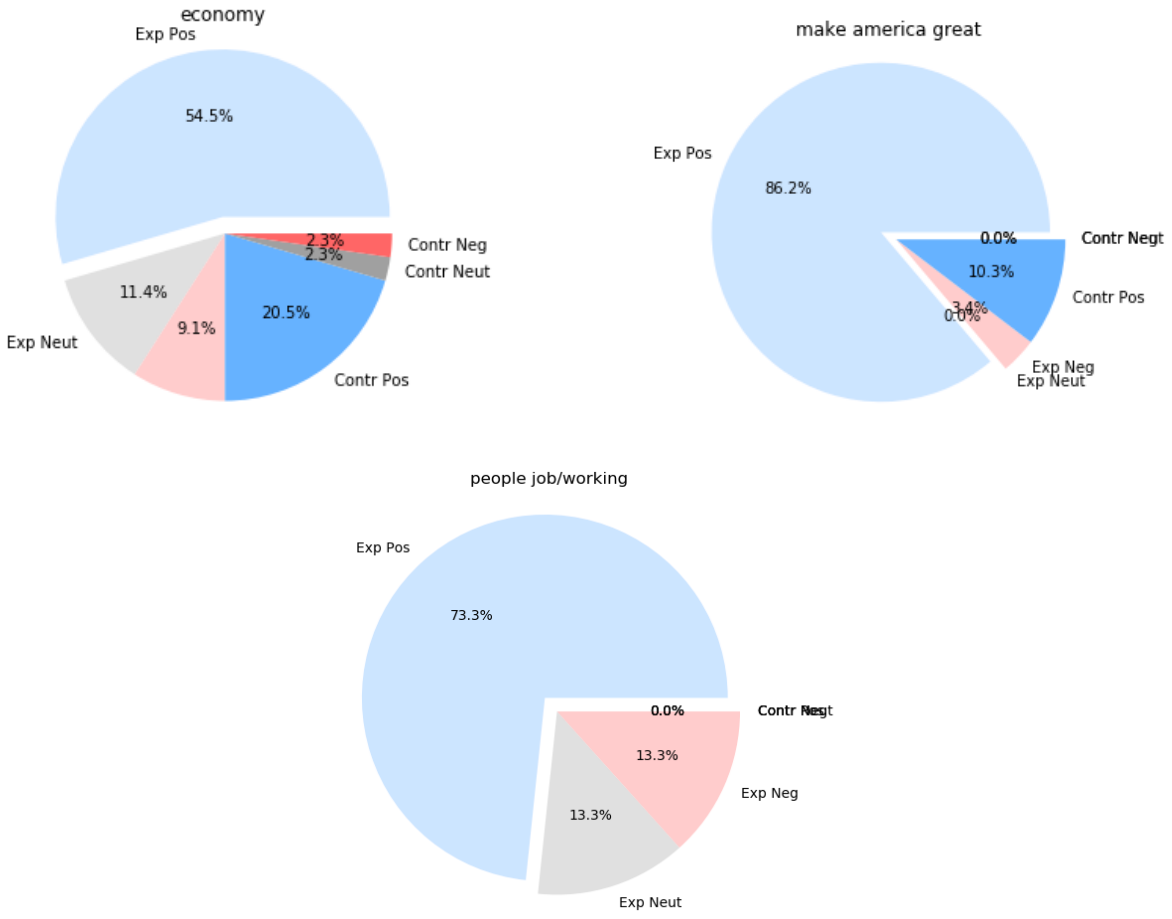

Fig. 10. Pie charts showing percentage results on the topic of the economy. Notice that Expansion of the Economy and Positive Sentiment stand out the most in all of these charts.

most frequently used word in all of the tweets because it was used in this phrase and was also used often in the phrase “great meeting”, perhaps because Donald Trump thought a lot of meetings were great. The word “economy” approximately follows the fact stated earlier that when the expansion and contraction tags of the economic direction in each tweet were evaluated with no sentiment polarity, the economy expanded 80% of the time and contracted 20% of the time. When the word phrase “people job” or “people working” was used in a tweet, the economy expanded 100% of the time which strongly indicates that Mr. Trump’s words about jobs stimulated growth in the economy.

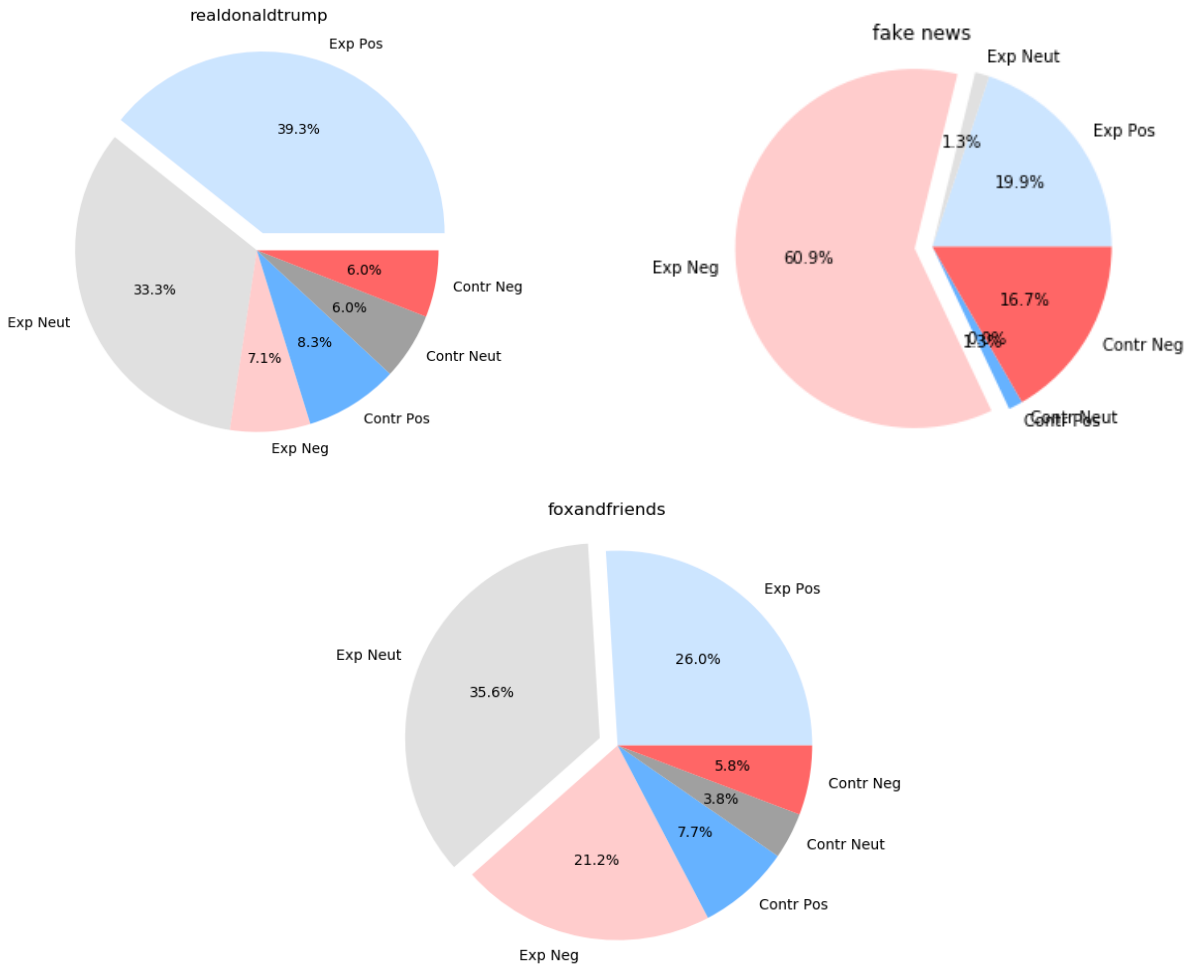

Fig. 11. Pie charts showing the percentage results on the topic of the news media

6.3 The Influence of Tweets about the News Media

Figure 11 shows the pie charts on topics related to the news media. They include "fake news" and two tags used quite often that were considered useful to analyze in the project: "foxandfriends" and "realdonaldtrump". Mr. Trump referred to the phrase "fake news" on the liberal media in his tweets so often that it has been called a common phrase that anyone familiar with Donald Trump's tweets knows quite well. It appears the word "fake" has a very strong sentiment polarity that leans a tweet towards negativity thus very little positive sentiment in the "fake news" pie chart exists. This pie chart follows the general trend of the economy expanding 80% of the time and contracting 20% of time from one market day to another in 2017 so it appears the phrase "fake news" has insignificant effects on the movement of the economy from one period to another. The words "realdonaldtrump" and "foxandfriends" are two phrases that occur often in Mr. Trump's tweets. The Fox and Friends program on the Fox News Network is watched by Donald Trump frequently. He also compliments the people on the show and makes many

references to topics or people that were on the program, probably to get other people to watch it. It appears that tweeting about this program often shows an 82.8% expansion in the economy on the pie chart. This is insignificant as it follows the general trend of the economy expanding 80% of the time and contracting 20% of time from one market day to another in 2017. As for the "realdonaldtrump", it also follows the general trend of the economy 80% of the time and contracting 20% of the time in 2017 plus the categories in the contracting economy side of the pie are all equal. This word also seems to have an insignificant effect on movement in the economy from one period to another.

6.4 The Influence of Tweets about Foreign Affairs

Figure 12 shows all of the topics related to Foreign Affairs and anything that has to do with foreign policy or immigration. One would think "hillary fbi" would have some effect on the economy but it appears that little was tweeted on this subject and it is pretty evenly split among the six categories so it appears to be an insignificant phrase. The phrase "great meeting" has the word "great" in it and refers to any meeting he had with someone so there is no negative sentiment in these tweets. It follows the general trend of the economy 80% of the time and contracting 20% of the time in 2017 so it is also an insignificant phrase. In general, when the phrase "north korea" is mentioned in a tweet, the economy expanded from one market period to another much more often than it contracted. All sentiments are pretty spread out within the sentiment category with more of them being positive. It appears that investors have confidence in Mr. Trump when dealing with North Korea; otherwise the markets would not go up most of the time the topic is mentioned in a tweet.

The topic on Russia and collusion is following the general trend of the economy 80% of the time and contracting 20% of the time in 2017 and most of the tweets have positive sentiment. There appears to be no significant effect on the movement of the economy at least in this time period in 2017. Talk about the wall between the U.S. and Mexican border has always been a hot topic since Donald Trump's campaign for President. The idea of building a wall between the United States and Mexico theoretically could stimulate the economy due to the creation of new jobs; however, the pie chart for "wall" showed an 84.6% expansion in the economy with almost equal slices for positive and negative sentiment, so sentiment about the wall doesn't seem to have much effect on the movement of the economy, at least in 2017.

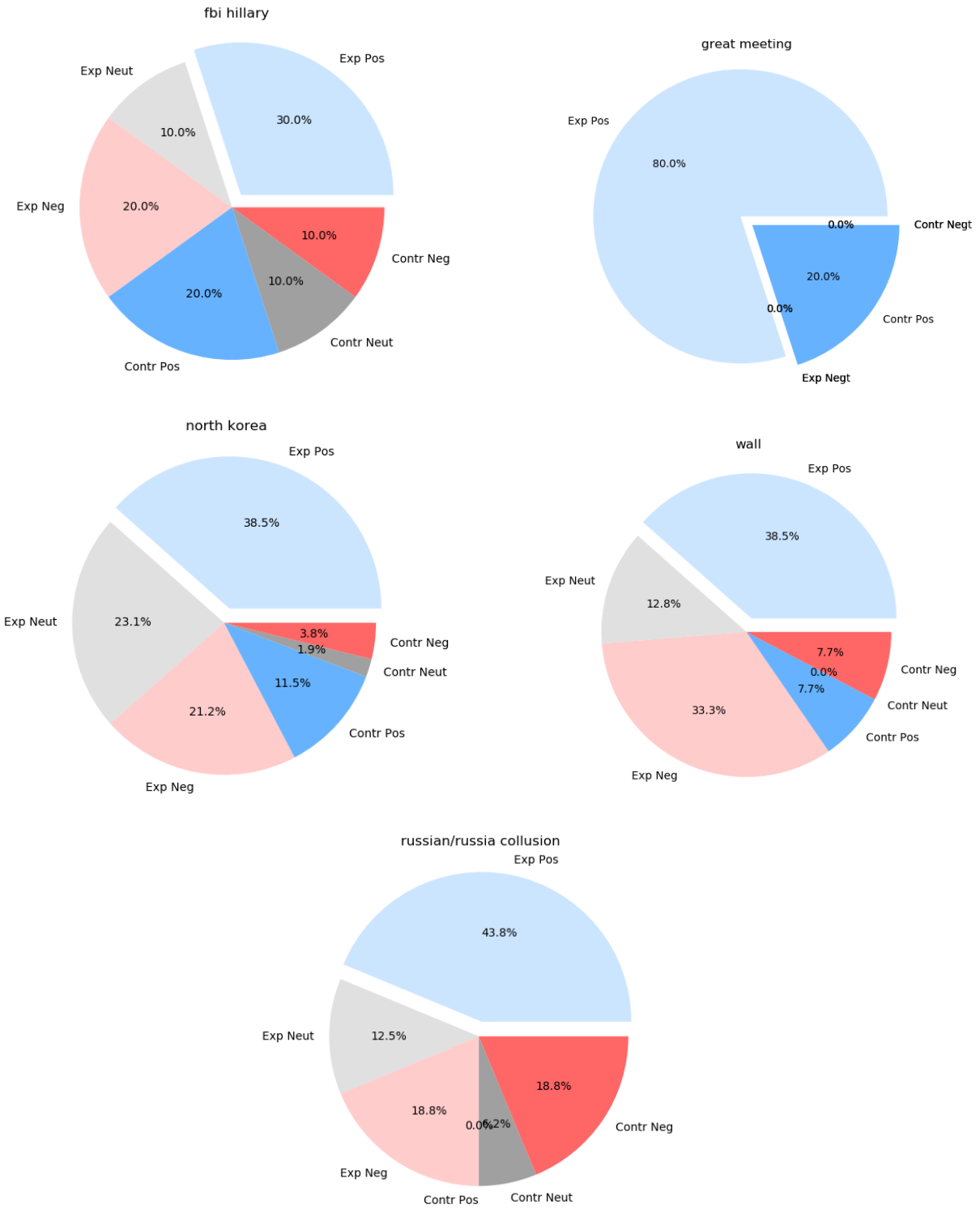

Fig. 12. Pie charts showing the percentage results on the topic of foreign affairs.

7 Discussion of Results

Two salient features stood out in the results, and most of it was due to how the sentiment analysis calculation determined the polarity of the tweet. Many of the words used in the tweets have meanings that can be both negative and positive in nature which would be hard for a computer to determine unless a machine learning algorithm had a large enough training set to work with to differentiate these meanings. It appears that TextBlob Sentiment Analysis was not sophisticated enough in differentiating the positive and negative meanings of the word "hard" because it did not evaluate the words before and after it. The word "hard" has a negative connotation when compared to "soft" but when it is used in the phrase "hard working" (as it often was in the tweets), it supposedly becomes somewhat positive in nature. This is something that should be improved upon in future development of sentiment analysis.

It is wise to keep in mind that the data from this project is within a time period where the economy was expanding rapidly. If the period had leading economic indicators showing a higher percentage of contraction in the economy, results could potentially be very different from what was found in this paper. More recent data could not be included in this project because it wasn't available at the time of this paper. The algorithms and methods used to calculate the economic data can be used on future data by adding to the economy related datasets. Such additions could possibly produce different results especially if more data showing contractions of the economy exist. Longer time periods of data should also help with the reliability of the results, especially if there is a noticeable change in the direction of the economy during Donald Trump's term as President of the United States. For instance, the dataset of all of the tweets made during Donald Trump's presidency would be much more significant than just portions of them. This project can also be used with tweets of future presidents of the United States if they choose to tweet as much as Donald Trump tweets. It probably would be a wise thing to do as long as the President understands the power of positive and negative words and when to use them.

8 Conclusion

The purpose of this project was to see how the sentiment of Donald Trump tweets and the words or phrases he uses in them correlate to a daily expansion or contraction in the United States economy. Based on the above results that cover the period from January 20, 2017 to January 31, 2018, the economy did quite well at expanding in this period. One might argue that Donald Trump can take some credit for the expansion of the economy by using positive words most of the time and negative words when it is appropriate to influence investors in the U.S. economy. Because the data is highly

skewed towards economic expansion, no real conclusions about the predictability can be made until data showing significant contraction in the economy is seen to have similar effects. It is possible that tweets produced by a powerful person like the President of the United States could possibly be used to predict future expansion or contraction in the economy if certain words or phrases are used often enough and consistently, but more data that reflects a contracting economy needs to be investigated before making a definitive conclusion.

REFERENCES

1. @realdonaldtrump. (2018, March 19). *Trump Twitter Archive.* Retrieved March 19, 2018, from http://www.trumptwitterarchive.com/archive
2. Antonakaki, D., Spiliotopoulos, D., V.Samaras, C., Pratikakis, P., Ioannidis, S., & Fragopoulou, P. (2017). Social media analysis during political turbulence. *PLoS ONE, 12*(10). doi:e0186836. https://doi.org/10.1371/journal.pone.0186836
3. Board, T. C. (2018). *Description of Components*. Retrieved from The Conference Board: https://www.conference-board.org/data/bci/index.cfm?id=2160#BCI106
4. Board, T. C. (2018, February 22). *The Conference Board Leading Economic Index.* Retrieved March 19, 2018, from https://www.conference-board.org/data/bcicountry.cfm?cid=1
5. Cooper, J. (2017, 7 28). *The Copper-Gold Ratio.* Retrieved from seekingalpha.com: https://seekingalpha.com/article/4092103-copper-gold-ratio?page=2
6. Demirsoz, O., & Ozcan, R. (2017). Classification of news-related tweets. *Journal of Information Science, 43*(4), 509–524.
7. El Alaoui, I., Gahi, Y., Messoussi, R., Chaabi, Y., & Todoskoff, A. (2018). A novel adaptable approach for sentiment analysis on big social data. *J Big Data, 5*(12). Retrieved from https://doi.org/10.1186/s40537-018-0120-0
8. Erikson, R. S., & Wlezien, C. (2016, October). Forecasting the Presidential Vote with Leading Economic Indicators and the Polls. *PS, Political Science and Politics, 49*(4), 669-672.
9. Fatima, I., Halder, S., Saleem, ,. M., Batool, R., Fahim, M., Lee, Y.-K., & Lee, S. (2015). Smart CDSS: integration of Social Media and Interaction Engine (SMIE) in healthcare for chronic disease patients. *Multimed Tools Appl, 74*, 5109-5129. doi:10.1007/s11042-013-1668-5
10. Gundlach, J. (2017, August 29). *FEG 2017 Investment Forum.* Retrieved March 29, 2018, from http://www.feg.com/hubfs/2017%20Forum/Presentations/FEG%202017%20Investment%20Forum%20-%20Jeffrey%20Gundlach.pdf
11. Kim, E. H.-J., Jeong, Y. K., Kim, Y., Kang, K. Y., & Song, M. (2016). Topic-based content and sentiment and analysis of Ebola virus on Twitter and in the news. *Journal of Information Science, 42*(6), 763-781.
12. Lakshminarayanan, M., Acosta, W. F., Green II, R. C., & Devabhaktuni, V. (2014). Strategic and suave processing for performing similarity joins using MapReduce. *J Supercomput, 69*, 930-954.
13. Loria, S. (2014, July 5). *Steve Loria GitHub Page* . Retrieved from GitHub.com: https://github.com/sloria/TextBlob/blob/eb08c120d364e908646731d60b4e4c6c1712ff63/textblob/en/en-sentiment.xml
14. Loria, S. (2018). *TextBlob Tutorial: Quickstart.* Retrieved from textblob.readthedocs.io: http://textblob.readthedocs.io/en/dev/quickstart.html
15. NLTK Project. (2017). *Source code for nltk.corpus.reader.wordnet.* Retrieved from nltk.org: http://www.nltk.org/_modules/nltk/corpus/reader/wordnet.html
16. Perkins, J. (2014). *Python 3 text processing with NLTK 3 cookbook : over 80 practical recipes on natural language processing techniques using Python's NLTK 3.0.* Birmingham, UK: Packt Publishing.
17. PlanSpace. (2015, June 7). *TextBlob Sentiment: Calculating Polarity and Subjectivity.* Retrieved from planspace.org: https://planspace.org/20150607-textblob_sentiment/
18. Python Software Corporation. (2018). *re - Regular Expression Operations.* Retrieved from docs.python.org: https://docs.python.org/3/library/re.html
19. Regenstein, J. (2018, 2 21). *Copper/Gold and 10-Year Treasuries.* Retrieved from Reproducible Finance.com: http://www.reproduciblefinance.com/2018/02/21/copper-gold-and-10-year-treasuries/

20. Sharma, J. (2013). *Sentiment Analysis on Tweets and their Relationship with Stock Market Trends.* University of Maryland, Baltimore County, Computer Science. Ann Arbor: ProQuest Publishing,.
21. Singh, A. K., Gupta, D. K., & Singh, R. M. (2017). Sentiment Analysis of Twitter User Data on Punjab Legislative Assembly Election, 2017. *I.J. Modern Education and Computer Science, 9*, 60-68. doi:10.5815/ijmecs.2017.09.07
22. Singh, T., & Kumari, M. (2016). Role of Text Pre-processing in Twitter Sentiment Analysis. *Procedia Computer Computer Science, 89*, 549-554.
23. TextMiner. (2014, July 18). *Dive Into NLTK, Part IV: Stemming and Lemmatization*. Retrieved from textminingonline.com: http://textminingonline.com/dive-into-nltk-part-iv-stemming-and-lemmatization
24. The Conference Board. (2001). *Business Cycle Indicators Handbook.* New York: The Conference Board.
25. *Twitter - Statistics & Facts*. (2018). Retrieved from Statista The Statistics Portal: https://www.statista.com/topics/737/twitter/
26. Vlastelica, R. (2018, January 2). *Trouble ahead? What 4 recession indicators say about the economy.* Retrieved March 14, 2018, from https://www.marketwatch.com/story/trouble-ahead-what-4-recession-indicators-say-about-the-economy-2017-12-28
27. Yang, X. H. (2012). Yield Curve Inversion and the Incidence of Recession: A Dynamic IS-LM Model with Term Structure of Interest Rates. *Int Adv Econ Res, 18*, 177-185.