# A CASE STUDY ON VERACITY IN TWITTER DATA USING OIL COMPANY RELATED TWEETS

By

KAMMARPALLY PRASHANTH

Bachelor Science in Computer Science

Jawaharlal Nehru Technological University

Hyderabad, India

2013

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
July, 2015

A CASE STUDY ON VERACITY IN TWITTER DATA
USING OIL COMPANY RELATED TWEETS

Thesis Approved:

Dr. K. M. George

Thesis Adviser

Dr. Nohpill Park

Dr. Eric Chan- Tin

Name: KAMMARPALLY PRASHANTH

Date of Degree: JULY, 2015

Title of Study: A CASE STUDY ON VERACITY IN TWITTER DATA USING OIL COMPANY RELATED TWEETS

Major Field: COMPUTER SCIENCE

Twitter is a powerful real-time micro-blogging service and it is a platform where users provide and obtain information, called tweets, at a rapid pace. Because of the volume, velocity, and unstructured nature of tweets, Twitter data can be viewed as big data. In this thesis we study the veracity of tweets using oil industry related tweets. Previous research has shown that most of the tweets posted on twitter are truthful. But the same platform (Twitter) is also used often to spread misinformation intentionally or unintentionally. There is no definitive measures to determine the veracity of tweets based on the tweets themselves. So there is a need for better mechanisms to measure levels of accuracy from tweets.

In this thesis, we propose three measures to estimate the veracity/accuracy of topics based on analysis of tweets. They are topic diffusion, geographic dispersion, and spam rate. We collect tweets associated to topics. Using the tweets we compute the measures and estimate the veracity of topics. Reliable geographic dispersion data was not available in our data set and hence it is not used in validation process. To validate measures, we verity the tweeted information using official data. For this study we streamed oil industry data. Several topics were identified for our analysis. In the case of each topic, tweets unrelated to the topic are considered noise. After noise elimination, tweets are classified according to company names, then the proposed measures are computed. The results are compared against the verification results. In majority of cases, the estimates of veracity of topics by the proposed measures are confirmed by the verification results.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER I

## Introduction

Micro-blogging service Twitter has become one of the most important social networking sites in the present generation [5]. As per www.statista.com , it was estimated that there are 236 million active twitter users worldwide, among them 28% users are from United States, "with more than 140 million active users and over 340 million messages posted per day, Twitter has become one of the most influential media for spreading and sharing breaking news, personal updates and spontaneous ideas" [5]. It also allows people to post their experiences and opinions online. This information can be useful for making informed decisions. The information could be useful to businesses and policy makers alike. However, people can intentionally or unintentionally spread false information. Social media can also facilitate the spread of unverified information [9]. For example, the spread of unverified information that turned out to be false could be far reaching during and after a disaster [1]. The spread of unverified information has negative consequences. It confuses people and interferes with the discovery of useful information and if the information is false, it may lead to misbeliefs, which are difficult to change.

As mentioned previously, nearly 340 million tweets are being posted on twitter every day. Twitter is used by a wide variety of users, of which a large portion – 46% of

active users - corresponds to mobile users. Tweets (messages posted in Twitter) can be published by sending e-mails, and sending SMS text-messages directly from smartphones using a wide array of Web-based services. Therefore, Twitter facilitates real-time propagation of information to a large group of users. This platforms makes it an ideal environment for the dissemination of breaking-news directly from the news source and/or geographical location of events.

For instance, in an emergency situation, some users generate information either by providing first-person observations or by bringing relevant knowledge from external sources into Twitter. In particular, information from official and reputable sources is considered valuable and actively sought and propagated.

Significant amount of research papers based on Twitter data are available in the literature. They span predicting crime rate [19], information spreading [26], and NFL (National Football League) [8]. Public sentiment analysis in twitter data is employed for prediction of a company's stock price movements [21], US primary elections [18], Box Office collection [6] and more.

All the papers mentioned above focus on predictive analytics using twitter data. Twitter data can be viewed as part of the big data ecosystem. They take the veracity of data (one of the important V's of big data) for granted. Due to the volume, velocity, and variety of tweets, Twitter data fall into the category of big data.

"In a recent statistics, IBM estimates that every day 2.5 quintillion bytes of data are created - so much that 90% of the data in the world today has been created in the last two years. It is a mind-boggling figure and the irony is that we feel less informed in spite of having more information available today. The surprising growth in volumes of data has badly affected today's business. The online users create content like blog posts, tweets, social networking site interactions and photos. And the servers continuously log messages about what online users are doing" [45]. The social media sites Facebook, twitter etc. contribute to online data via user posts. The area of Big Data studies these data. Big data is characterized by several V's denoting their properties. Several researches based on the characteristics of big data have been published. However, few were focused on the characteristic Veracity. Some of the commonly mentioned V's [13] are defined below in table 1:

- Volume
- Variety
- Velocity
- Veracity
- Validity
- Volatility

Table 1: Definitions of V's in Big data [13]

| Characteristic | Definitions |
|---|---|
| Volume | The data size will be large could be in TB's |
| Variety | It extends the structured data, including unstructured data of all varieties: text, audio, video, posts, log files etc. |
| Velocity | It will be used when streaming in to the enterprise in order to maximize its value to the business. |
| Veracity | Big Data Veracity refers to the biases and noise in data. Veracity in the data analysis is the biggest challenge when compared to the other characteristics. |
| Value | Value starts and ends with business use case and defines the analytic application of the data. |
| Validity | Validity defines the data correctness and accuracies. |
| Volatility | It defines how long the data is valid and how long should it be stored. |

There are some papers that address the veracity of Twitter data. Our study focuses on the veracity of Twitter data related to oil industry. Oil as it plays a significant role in the world economy today since nearly two thirds of the world's energy consumption comes from the crude oil and natural gas [46]. At present, oil accounts for 40% of total energy consumption in the United States. According to an article of U.S. Energy Information Administration, crude oil prices have plunged 60% since mid-June as global production exceeded demand [46] leading to significant revenue shortfalls in many energy exporting nations. To our

knowledge, no twitter data veracity study relating to the oil industry is available in the literature. The primary objective of this study is to evaluate the accuracy of twitter data. We propose three measures for estimating veracity of topics defined by keywords present in tweets. We selected oil industry related data for our study. As oil plays a significant role in the finance of individuals, accuracy of information propagation related to the oil industry in the social media is important to society. This and the lack of research are reasons that we chose oil industry as the data domain for our study on big data veracity.

As mentioned above, the focus of our study is on the veracity property of big data using Twitter data as it falls into the category of big data. The current work consists of three key parts:

(a) Data collection: Extracting live Twitter data and pre-processing and storing it in a file system.

(b) Data Classification: Classification of the data based on keywords using a java program.

(c) Data Analysis: Computing several statistics and evaluation measures based on the data

Apache Flume is used for collecting tweets. Flume extracts the tweets using keywords listed in a configuration file. Some of the keywords are Halliburton, BP, Exxon Mobil, Baker Hughes Incorporated, Schlumberger limited, Chesapeake energy, Devon energy, Transocean limited, North Dakota, North Dakota oil well shutdown and Encana. The data fetched from twitter is stored on Hadoop Distributed File System [HDFS] [23]. The obtained data is in "JSON" format so the data can be parsed in two ways either by "Map-

Reduce program [12] or Hive [4]". I have used "Hive" to parse the data. Retrieved data from twitter may contain numerous tweets that are not useful to the analysis that we have performed. We refer to such tweets as noise. We use a java program to eliminate this noise and classify the data. The Java program generates different files containing the tweets of a specific keyword. Following are the set of keywords used:

- Halliburton
- BP
- Baker Hughes Incorporated limited
- Devon energy
- Transocean limited
- Schlumberger limited
- Exxon Mobil
- Encana
- Akastor SA
- Oceaneering International Limited
- Chesapeake Energy
- Chevron
- Conoco Phillips

As our objective is to propose and validate measures to estimate veracity of topics, we partitioned the initial dataset into manageable parts based on company names

listed above. Topics are defined for each company data to perform validation analysis. Some of the reasons for selecting the company names are

- The company names are in the top 100 oil producing companies globally (http://twinengine.com/oilandgas/the-top-100-oil-gas-companies-on-social-media/ ).
- They are involved in some of controversies like Bribery, Deep water horizon oil spill and economic losses.
- There are lots of rumors spreading over the social media about the companies. Some of the rumors are "company is going to cut some thousands of jobs", selling their shares, top politicians are involved in the company profits, main cause of animals' death etc.

"Keyword pair pattern matching" method is used for classifying the tweets, keyword pairs used and the reason for choosing those keywords are mentioned in section 3.3, Data Processing. Once the classification is done, then we perform comparative analysis of the tweets with the data published on the official government websites.

Government sources [27, 28, 30] are used for verification of information propagated in the tweets. We believe this approach is the most trustable approach for verification of contents in tweets.

CHAPTER II

# Literature Review

## 2.1 Related Work

Twitter has attracted a considerable amount of research in recent years. With the rapid increase in utilization of online information storage and social networking sites like Facebook, Flicker, Twitter and LinkedIn the amount of data available is larger than ever before. As indicated by "aci.info" Twitter users collectively tweet almost 300,000 times each moment on twitter [14]. Increase in the utilization of social networking sites has driven numerous social researchers to examine whether specific patterns in the stream of tweets might be able to predict real-world outcomes [10, 24].

Twitter's streaming API has been utilized all through the domain of social networking sites to see how users behave on these platforms. It is utilized to gather information for a variety of topics. Researchers have utilized this data for a wide variety of purposes such as, to predict stock market movements, to predict election outcomes etc. Many methodologies were implemented to analyze the twitter data. Due to the wide use of Twitter data in different exploratory fields, it is important to understand how sub-sample of the data generated can affects the results.

A useful function of social media is to collect and display collective opinion [25]. In Twitter, an indicator of collective opinion is the number of people who have forwarded a tweet. Forwarding of tweets is called retweeting. Twitter displays the total number of retweeting associated with each tweet that has been retweeted. This retweet count signifies the popularity of the associated tweet. For example, Facebook shows the number of 'likes' to indicate the collective liking of articles, photos, posts, communities, pages etc. And also, it gives user ratings of some famous Websites like Amazon, eBay, Netflix, YouTube and other similar websites, and user ratings are aggregated as collective preferences.

Collective opinion is useful because it allows individuals to learn about the majority's opinion and consider options that they would not have considered otherwise. In this way, collective opinion can increase people's knowledge as well as help them make decisions. A study on twitter under crisis response [20] showed that aggregate crowd behavior could help detect false information on twitter and suggested a possibility of building a verification tool.

These reasons have motivated the need for research with data processing that takes advantage of large sets of data from the Twitter. Based on keywords related to oil companies to verify whether data (tweets) generated on these platforms can be reliable or not.

A significant amount of research has been done on extracting, classifying and analyzing the twitter data. Though tweets have simple meaning, just one or two keywords may capture the main theme [47]. Twitter data have been used for prediction in several areas.

Andrew H. Tapia, Kathleen A. Moore, and Marcelo Mendoza have published their work on analysis of tweets in the paper titled "Beyond the Trustworthy Tweet: A Deeper Understanding of Microblogged Data" [1]. The research reported in this thesis is very closely related to their work.

A brief summary of their work is described below:

They have streamed the data based on the keywords: "Trust, Microblogging, Disaster, Twitter, Humanitarian, Relief, and NGO", before and after the disaster and then compared the results to see how well they co-relate, and concluded that microblogged data is useful in some situation where information is limited, especially in emergency situations. In others, such as search and rescue operations, microblogged data may never meet the standards of quality required.

Kathleen M. Carley, and Fred Morstatter have conducted an experiment on "Is the Twitter Data Good Enough? Comparing Data from Twitter's Steaming API with Trusted Data [15, 32]" and concluded that Twitter data is biased. They compared the top hashtags found in the tweets, a feature of the text commonly used for analysis. Also used topic analysis for better understanding the content of tweets and discovered two main types of topics: informational and emotional. Finally, all studies showed that the problem of identifying

topics in geographical Twitter datasets (the data collected from twitter using streaming API using Tweet Tracker with exactly the same parameters(same attributes mention in configuration file)), and they proposed models to extract topics relevant to different geographical areas in the data studies. These models show the topics users discuss, drive their geolocation.

Work in social psychology has recommended that individuals have a strong inspiration to contrast their suppositions and other individuals take after aggregate feeling because of their craving to make right reactions under instability [31]. These studies have shown that conformity takes place in face-to-face environments. Moreover conformity happens in online networking situations. For example an online experiment showed that the number of times a piece of music was downloaded in the past could predict its future popularity when the number of downloads was available to the users. These results suggest that people have a strong tendency to adopt collective opinion on social media. Extracting and detecting of information is always a challenging task for researchers. Collecting the twitter data begins with identifying the topic of interest [48]. Streaming of twitter data based on keywords or hashtags requires use of API's.

Lukoianova, T., & Rubin have published a paper on 'Veracity Roadmap: In Big Data Objective, to see the Truthful and Credible? '[16]. Their paper gives a guide to hypothetical and experimental meanings of veracity along with its practical implications. They explored veracity across three main dimensions: 1) objectivity, 2) truthfulness, 3) credibility – and

proposed to operationalize each of these dimensions with existing tools. We combine the measures of veracity dimensions into one composite index: – the big data veracity index.

The interdependency of opinions makes Veracity an important aspect of microblog ecosystem.

## 2.2 Problem Statement

The overall objective of this research is to study the big data component Veracity in Twitter data. Veracity of data can have skewing effect on the results of data analysis with serious consequences. Therefore understanding veracity is an important problem. We define three indices of veracity. Our approach is different from that of Lukoianov and Rubin [16] who also has proposed three indices as the dimensions of veracity. Their veracity index OTC depends on external validation of tweet contents in the index computation. Our approach is to estimate the veracity from the data itself. Due to unavailability of geographical location of tweets, we are able to verify only two indices.

# CHAPTER III

# METHODOLOGY

The major tasks in experiment design are:

- Data collection,
- Evaluation measures, and
- How accuracy is determined?

## 3.1 Tools Used

### Apache Hadoop

Apache Hadoop [2] is a platform that offers an efficient and effective method of storing and processing large amounts of data. Unlike traditional offerings, Hadoop was designed and built from ground up to address the requirements and challenges of big data. At its core, Apache Hadoop is a frame work for scalable and reliable distributed data storage and processing. It allows for the processing of large datasets across clusters of computers using a simple programming model. At the core of Apache Hadoop are the

Hadoop Distributed File System or HDFS and Hadoop Map Reduce, Hive which provides a framework for distributed processing.

**HDFS [23]:** HDFS stores individual files in large blocks, allowing it to efficiently store very large of numerous files across multiple machines and access individual chunks of data in parallel, without needing to read the entire file into a single computers memory. Reliability is achieved by replicating the data across multiple hosts. This approach allows HDFS to dependably store massive amounts of data.

**HIVE [4]:** mainly deals with large amount of data (TB's), it is a data warehouse infrastructure build on top of the Hadoop that complies SQL queries as map reduce jobs and run the job in the cluster.

Basically Hive data is organized in four different ways

- Databases
- Tables
- Partitions
- Buckets (clusters)

I have organized my hive data in the form of tables.

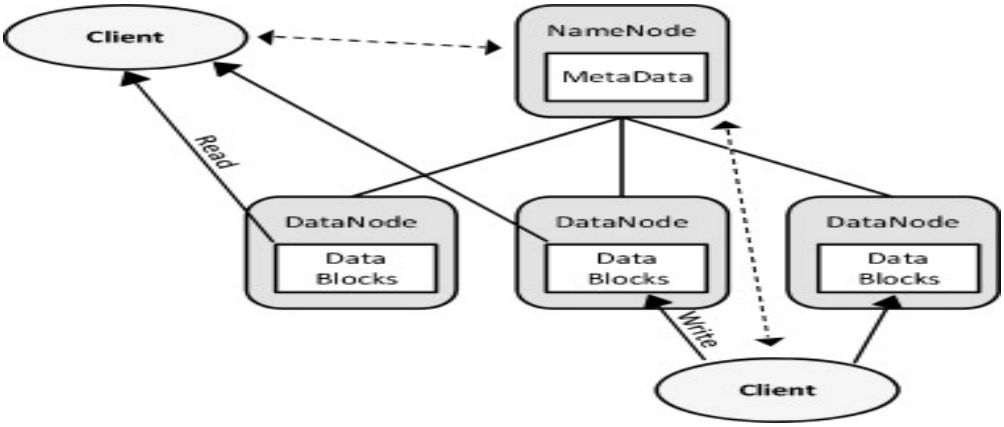


Figure 1: Hadoop File System and Data Communication Architecture [23].

**Apache Flume [3]:** Apache Flume is a tool designed and implemented by Apache Software Foundation. The purpose of Flume software is to collect, aggregates, and move large data files from many sources to a central storage.

**Procedure how Data is Stored on HDFS**

- Twitter data is streamed into Hadoop cluster, using Apache Flume
- To start streaming the data, flume uses 5 main integral components of an agent namely source, sink, channel, Events, Agents
- **Source:** It is the part of the flume that connects to the source of the data and starts sending them to a channel
- **Channel:** It acts as pathway between sources and sinks
- **Sinks:** These is the final stage of the flume data flow
- **Events:** An event is the basic unit of data that is moved using Flume.

- **Agents:** An agent is the container for a Flume data flow.

Flume can be used to retrieve tweets and store in the Hadoop file system. Flume is configured to capture the streaming data from one of the twitter end points. As seen before, flume stores data streamed from various sources into HDFS. The storage functional block of flume is called as HDFS Sink. Its pictorial representation is given in Fig 2.

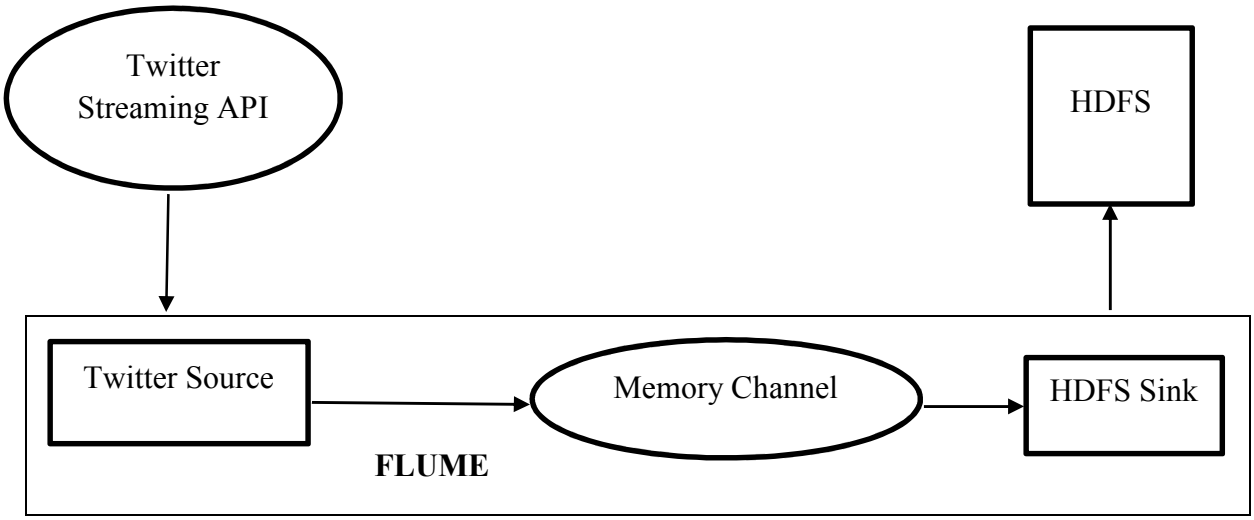


Figure 2: Pictorial Representation of the Flume Agent and data flow [3]

## 3.2 Data Collection

Apache Flume is used to retrieve data from the tweet stream. Twitter is a microblogging website [29] where users read and write millions of short messages on a variety of topics every day Twitter data collection can rely on standardized API services, so till now we have collected **87GB** of data from twitter based on the keywords related to crude oil companies. We examined **769,092** tweets, which were published on Twitter's public message board between 11/07/2014 to 11/26/2014, 02/13/2015 to 02/27/2015 and 05/22/2015 to 06/12/2015. We streamed 6 week of data from twitter. I have collected the tweets using the set of keywords {Halliburton, BP, Baker Hughes Incorporated, Schlumberger Limited, Exxon Mobile, Chesapeake Energy, Ensco plc, Akastor ASA, Devon Energy, Subsea TSA, Oceaneering International, EnCana, Brent Crude, Transocean limited and North Dakota}

**Process of Streaming Twitter Data**

- In order to stream the data from twitter we need to start the flume agent
- Create a twitter application (apps.twitter.com/dev.twitter.com) and provide all necessary details to get the “API Keys”.
- API Keys are used while streaming the data from twitter
- Twitter data is streamed into Hadoop cluster, using Apache Flume
- We need to create a configuration file for the flume agent.

- In the configuration file we need to specify the entire necessary tokens to the attributes and the appropriate keywords. All configuration files are listed in the Appendix
- The data obtained from twitter is in JSON format. Sample tweet data is given in Appendix

## 3.3 Data Processing

Once we have the Twitter data loaded into HDFS, we can stage it for querying by creating an external table in Hive. Using an external table will allow us to query the table without moving the data from the location where it ends up in HDFS. Apache Hive provides an interface that allows users to easily access data in Hadoop using SQL. Hive compiles SQL statements into Map Reduce jobs, and then executes them across a Hadoop cluster. Now that Twitter data is in JSON format, Hive allows us to define how the data is represented on disk. Hive SerDe interface is used to specify how to interpret what we have loaded. SerDe stands for Serializer and Deserializer, which are interfaces that tell Hive how it should translate the data into something that Hive can process.

**Data cleaning and classification based on keyword pair matching:**

Twitter data are always subjected to noise, noisy data are meaningless data. This term has often been used as a synonym for corrupt data. However, its meaning has expanded to include any data that cannot be understood and interpreted correctly by machines, such as unstructured text. Removing the tweets that are noise is an important goal of data cleaning as noise hinders most types of data analysis [49].

In this work, we are concerned with tweets related to a list of companies only. So, we define noise as any tweet that does not contain keyword pairs as described below:

The process and the keyword pairs used to eliminate noise is described below. We compute and report the percentage of tweets that are considered noise.

In this work we define two different sets of keywords. They are company names and words expressing positive/negative sentiments about the company. The sets $R_1$, $R_2$, and $R_3$ represent the set of company names, positive words, and negative words respectively.

$R_1$= {Haliburton, BP, Transocean, Exxon Mobil, Baker Hughes, Schlumberger, Conoco Philipps, chevron, Akastor SA, Oceaneering international, Devon energy, Chesapeake, and Encana}

$R_2$ = {Profits, employment rate increase, stock price increase, share value rise, stock value, and profits gained}.

$R_3$ = {Unemployment, jobs cuts, job reduce, stock price decrease, Gulf oil spill, Bribery scandal, Scam, Nigeria, Iraq Profiteer, Obama, Dick Cheney, Government hands, oil price decrease, falling/decrease of oil prices, Iraq war, Value Act cuts, Goldman Sachs, oil spill, 2010 Gulf oil spill, most vicious Iraq war profiteer, Oil Jitters Stymie, Weakness on Analyst Downgrade, dolphin deaths, Deep water Horizon oil spill, largest marine oil spill, fines, Bid, and stock falls}.

The keyword pairs used for noise elimination are defined as the set

$R = \{(R_1 \text{ X } R_2) \cup (R_1 \text{ X } R_3)\}$

A sample set of keyword pairs are given below

R = {((Halliburton, Profits gained), (Halliburton, employment increase) … (Halliburton, stock value increase), (BP, Profits gained)… (BP, Stock value increase), (Transocean, profits gained) ….. (Transocean, stock value increase), (Devon energy, profits gained) ….. (Devon energy, stock value increase), (Baker Hughes, profits gained) …. (Baker Hughes, stock value increase), (Exxon Mobil, profits gained) ….. (Exxon Mobil, stock value increase), (Schlumberger, profits gained) … (Conoco Philipps, profits gained)…. (Akastor SA, profits gained)) ∪ ((Halliburton, unemployment), (Halliburton, oil spill) …. (Halliburton, bribery), (BP, unemployment), (BP, oil spill) ….. (BP, Iraq profiteer), (Transocean, unemployment), (Transocean, oil spill) ….. (Transocean, Iraq profiteer), (Schlumberger, unemployment), (Schlumberger, oil spill) ….. (Schlumberger, Iraq profiteer), (Devon Energy, unemployment), (Devon Energy, oil spill) ….. (Devon Energy, Iraq profiteer), (Exxon Mobil, unemployment), (Exxon Mobil, oil spill) ….. (Exxon Mobil, Iraq profiteer) …… (Conoco Philipps, unemployment))}

Now we perform data filtering based on the above mentioned keyword pairs with the help of a Java program, the program is written in such a way that, it reads the keyword pairs and matches them against the text attribute present in the tweet, if the keyword pairs are present in the tweet then we write those tweets into a new file. Same program is executed with the different set of keyword pairs with all the companies. Tweets which contain these keyword

pairs will be treated as valid tweets and remaining will be considered as noisy data for our work.

The main reason to select the above keyword pairs are

- Availability of verification information.
- Terms that are viewed as positive when referring to a company, and
- Terms that are used negatively by critics when referring to a company.

Each tweet will contain 56 attributes, among them we identified 4 attributes to be useful for this thesis. The attribute and their definitions are listed below:

**Text:** It gives the description of the tweet

**Geo tags:** These gives geographical location from where the tweet has been tweeted

**Time_zone**: gives the USD time zone tweeted by the user

**Name**: gives the user name of the tweet by whom it was tweeted

## 3.4 EVALUATION MEASURES

The best method to determine the veracity of tweeted information is to verify its accuracy compared with information at official sites. This process could be very time consuming. Furthermore, an official site may not exit for all information. Therefore, it will be beneficial if veracity can be determined from the tweets themselves. This research proposes three measures or indices to evaluate the veracity of tweets. The indices we propose are somewhat similar to the dimensions proposed in [16]. The three indices we propose are indented to measure the spread of information in terms of volume, geographic spread, and the repetition in the volume. The intuitive argument is that information with high volume and high inflation rate that spreads widely could be questionable. Combining all three indices, we propose to associate a degree of veracity to the associated information or topic. The three measure that we propose are:

1. Diffusion Index
2. Geographic Spread Index
3. Spam Index

Diffusion Index:

This measure is used to find, how fast information has spread through Twitter, the information can be either rumor or truth. As per the paper “Beyond the Trustworthy

Tweet: A Deeper Understanding of Microblogged Data Use by Disaster Response [1]", false information has spread faster than truth.

Diffusion index is defined as:

$$\frac{\text{Number of Unique users}}{\text{Total Number of tweets}}$$

Geographic Spread Index:

This index is intended to measures the geographic breadth of the spread of the information. It is defined as:

$$\frac{Number\ of\ Unique\ Location}{Total\ number\ of\ tweets}$$

Due to the unavailability of tweets with accurate location information this measure is not further considered in this research.

Spam Index:

The spam index measures the impact of repeated tweets by the same user. Repeated tweet can be viewed as inflating the diffusion. Intuitively, this phenomenon is similar to spamming which propagates questionable information. The spam index is defined as follows:

$$\frac{\sum_{over\ unique\ users} \frac{1}{unique\ user\ tweet\ count}}{\#tweets}$$

# CHAPTER IV

## EMPIRICAL RESULTS

In this section, we present the results of the analysis performed. As defined in section 3.3, valid tweets are those tweets selected using the keyword pair. The list of companies for which we have data and the number of valid tweets associated to them are shown in Table 2. The valid tweets are divided into verified and unverified groups. A tweet is called *verified* if an official site is found that substantiated the information contained in the tweet. Otherwise it is called an *unverified* tweet. In this research, only government sites (Security Exchange Commission, Department of Justice, and Federal Beaure of Investigation) are considered as official sites. So, for a tweet to be considered verified, the information it provides must be found in one of the three government sites listed above. The counts of verified and unverified tweets associated to the companies are given in Table 2. The tweet counts given in Table 2 include repeated tweets. Corresponding unique counts are given in Table 3. This data is used to validate the veracity indics.

To validate the indices, we have identified *topics* associated to three companies based on data availability. A ***topic*** is defined by a set of keywords. The tweets containg the keywords contribute to the topic.

Table 2: Count of valid tweets, verified tweets, and unverified tweets.

| Company name | Number of Valid tweets | Number of verified tweets | Number of unverified tweets |
|---|---|---|---|
| Halliburton | 45,058 | 10656 | 34402 |
| BP | 54, 955 | 49489 | 5466 |
| Baker Hughes | 794 | 0 | 794 |
| Schlumberger | 945 | 0 | 945 |
| Devon Energy | 196 | 0 | 196 |
| Exxon Mobil | 330 | 0 | 330 |
| Encana | 204 | 0 | 204 |
| Transocean | 14977 | 11395 | 3582 |
| Akastor SA | 202 | 0 | 202 |

Table 3: Count of unique tweets

| Company name | Unique tweet count | verified tweet count | unverified tweet count |
|---|---|---|---|
| Halliburton | 79 | 35 | 44 |
| BP | 101 | 80 | 21 |
| Baker Hughes | 5 | 0 | 5 |
| Schlumberger | 5 | 0 | 5 |
| Devon Energy | 9 | 0 | 9 |
| Exxon Mobil | 3 | 0 | 3 |
| Encana | 1 | 0 | 1 |
| Transocean | 41 | 33 | 8 |
| Akastor SA | 1 | 0 | 1 |

Tables 4, 5, and 6 show the list of topics and the measures for three companies. We did not compute measures for other companies as data was not available among the tweets we collected. The tables compare the veracity indices Diffusion index and Spam index against the percentage of verified tweets. Both indices range between 0 and 1. A higher value for the indices indicate a higher level of veracity. Verification results mostly agree with the indices in the case of Halliburton and not in the case of BP and Transocean. Figures 3, 4, and 5 show the relative positions of the topics in the space of Diffusion and Spam indeces.

Table 4: Veracity of Halliburton topics

| Topic | Valid # of tweets | Verified Tweets | Unverified Tweets | Diffusion index | Spam index |
|---|---|---|---|---|---|
| Nigeria Bribery | 921 (2.3%) | 921 (100%) | 0 (0%) | 0.025 | 0.014 |
| oil spill | 12180 (27%) | 4634 (38%) | 7546 (62%) | 0.36 | 0.18 |
| Iraq profiteer | 9482 (21%) | 3021 (31.8%) | 6461 (68.2%) | 0.19 | 0.11 |
| Unemployment | 7014 (15.5%) | 0 (0%) | 7014 (100%) | 0.12 | 0.08 |
| Stock price increase | 4436 (9.8%) | 0 (0%) | 4436 (100%) | 0.16 | 0.06 |
| Government hands to make Halliburton rich | 5467 (12.1%) | 0 (0%) | 5467 (100%) | 0.12 | 0.06 |
| Horizon, Mexico, Deepwater oil spill | 5558 (12.3%) | 2080 (37.4%) | 3478 (62.6%) | 0.07 | 0.035 |

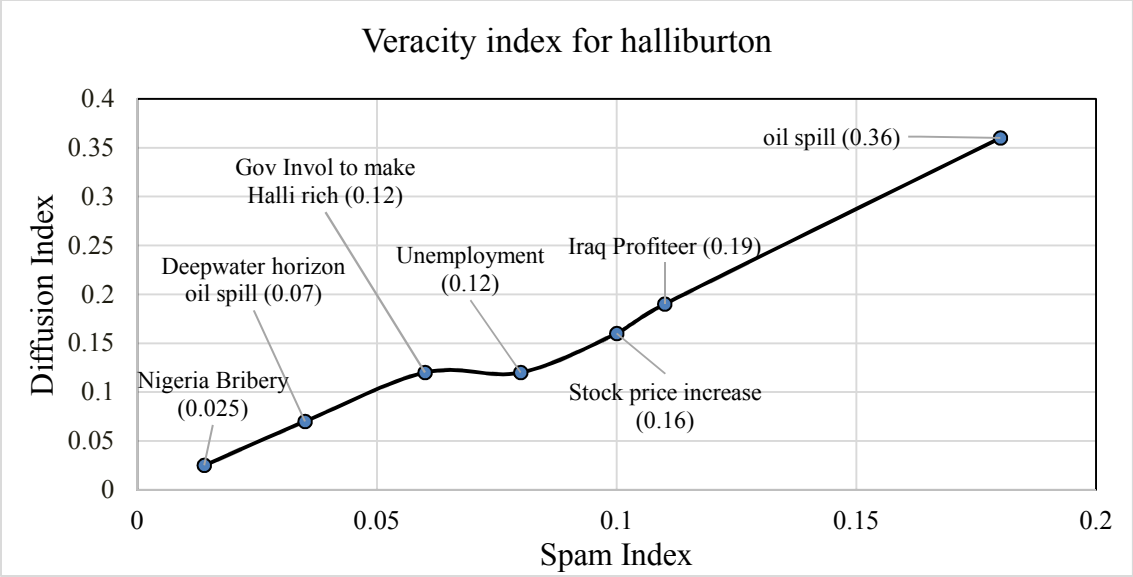


Figure 3: Positions of Halliburton Topics based on Diffusion and Spam indices

Table 5: Veracity of BP topics

| Topic | Number of tweets | Number of Verified Tweets | Number of Unverified Tweets | Diffusion index | Spam index |
|---|---|---|---|---|---|
| Wild life death | 15461 (28%) | 14011 (90.6%) | 1450 (9.4%) | 0.14 | 0.09 |
| Largest oil spill in U.S history | 9601 (17.4%) | 9601 (100%) | 0 (0%) | 0.16 | 0.1 |
| Deepwater, gulf, Horizon oil spill | 20601 (37.4%) | 17841 (86.6%) | 2760 (13.4%) | 0.24 | 0.14 |
| Compensations/fines | 9292 (17.2%) | 8036 (86.4%) | 1256 (13.6%) | 0.12 | 0.07 |

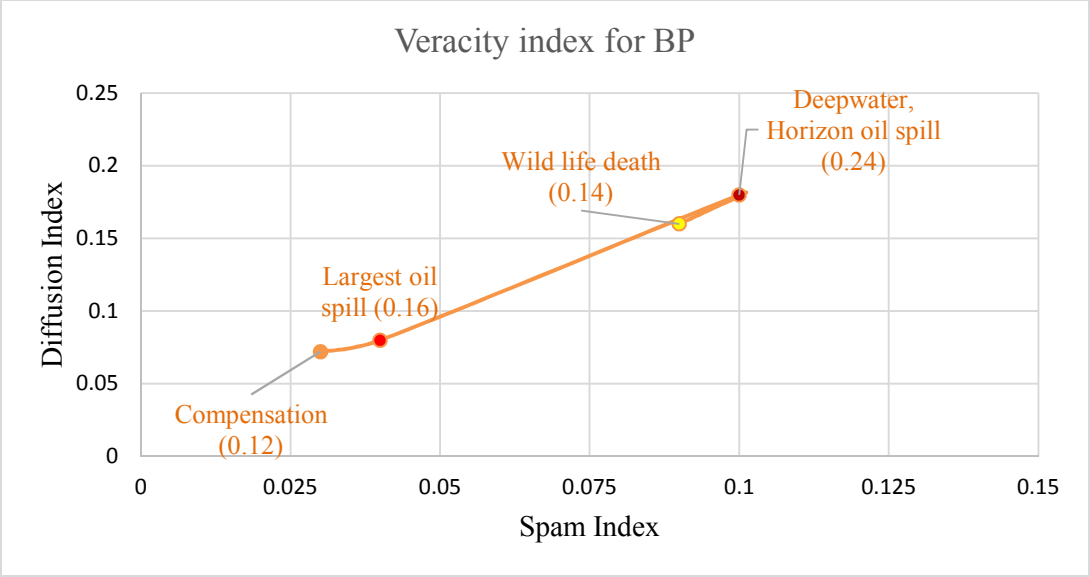


Figure 4: Positions of BP Topics based on Diffusion and Spam indices

Table 6: Veracity of Transocean topics

| Topic | Number of Valid tweets | Number of Verified Tweets | Number of Unverified Tweets | Diffusion index | Spam index |
|---|---|---|---|---|---|
| Settlement with gulf cost | 13944 (93%) | 11295 (81%) | 2649 (19%) | 0.34 | 0.2 |
| Oil price rise | 1033 (7%) | 100 (9.6%) | 933 (90.4%) | 0.2 | 0.12 |

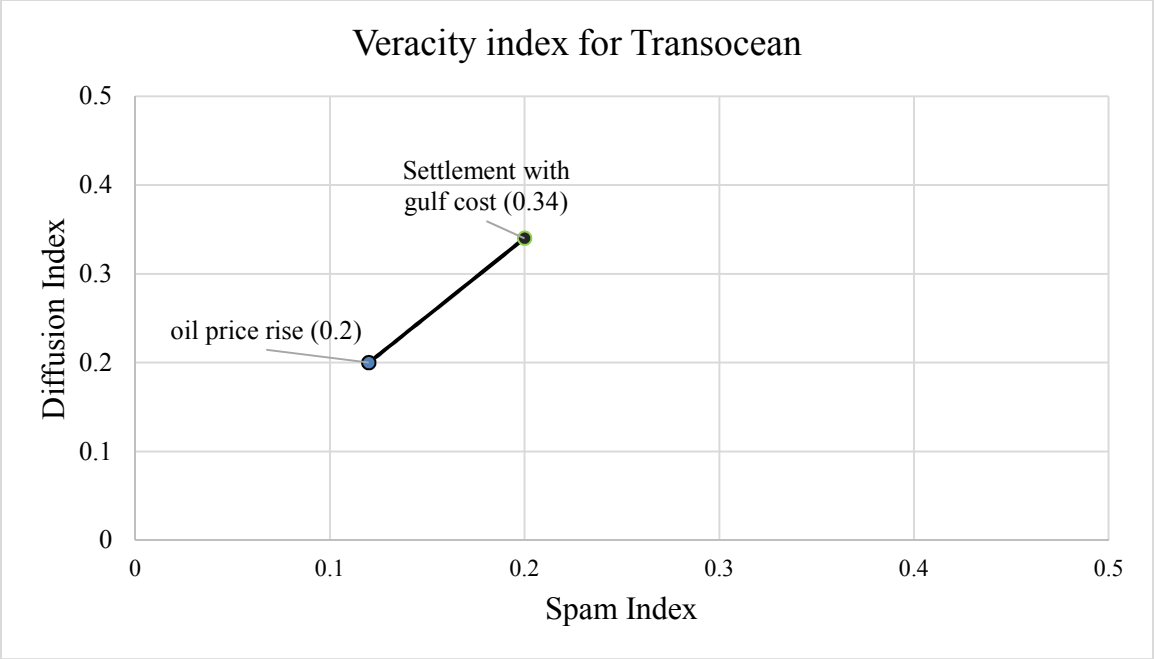


Figure 5: Positions of Transocean Topics based on Diffusion and Spam indices

Based on the data obtained, the analysis indicates that the veracity indices are useful and necessary only if the topics are not based on physical events such as oil spills that are well publicized by reputable news outlets. In such cases, tweets reflect the news and are truthfull. So, the tweets can be trusted and veracity indices are not required. The indices turnout to be false negatives. In the case of Halliburton, the indices are validated by the verified tweet data. This is because all topics and associated tweets are not based on events that are well publicized. For all the topics of BP, the indices turnout to be false negatives. As illustrated in the above figures, the indices could be used to compare the veracity of topics.

# CHAPTER V

## CONCLUSION

Microblogging service (Twitter) has gained significance as an information resource. Therefore, the veracity of information propagated through this medium becomes an important concern. Largescale diffusion of false information is damaging to all aspects of society. In this thesis we propose three indices to measure the veracity of Twitter topics. The veracity of a topic depends on the veracity of contributing tweets. We have validated two indices based on topics related to three companies in the oil sector. Our data collection method did not provide geotag information. Therefore we were unable to validate the third index which is Geographic spread. The indics are not useful or necessary to evaluate topics known to be true. Otherwise, the indices perform well. The indices can also serve as a veracity comparison measure for topics.

Our work is an initial attempt at defining and validating the veracity indices. The limited data available to us prevented us from reaching definitive validation of the indices. Further validation of the indices is proposed as future work.

REFERENCES

[1] Andrew, H. T., Kathleen, A. M., and Nicholas, J. J. Beyond the Trustworthy Tweet: A Deeper Understanding of Microblogged Data Use by Disaster Response and Humanitarian Relief Organizations. Proceedings of the 10th International ISCRAM Conference – Baden - Baden, Germany, May 2013, pp: 770-779.

[2] Apache Hadoop, http://hadoop.apache.org/ [07/14/2014].

[3] Apache Flume, Architecture, Getting Started, http://cwik.apache.org/confluence/display/FLUME/Getting+started [07/14/2014].

[4] Apache Hive, http://hive.apache.org/ [07-14-2014].

[5] Beidou, W., Can, W., Jiajun, B., Chun, C., Weivivian, Z., Deng, C., and Xiaofei, H. Whom to Mention: Expand the Diffusion of Tweets by @ Recommendation on Micro-blogging Systems. WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil, pp: 1331-1340.

[6] Bernardo, A. H., and Sitaram, A. Predicting the Future with Social Media, arXiv: 1003.5699 [cs.CY], March 2010, pp: 1-8.

[7] Castillo, C., Marcelo, M., and Barbara, P. Information Credibility on Twitter. Proceedings of the 20th International Conference on World Wide Web, ACM New York, USA, and ISBN: 978-1-4508-0632-4, 2011, pp: 675-684.

[8] Chris, D., Shiladitya, S., Kevin, G., and Noah, A. S. Predicting the NFL (National Football League) using Twitter, arXiv: 1310.6998 [cs.SI], October 2013, pp: 1–11.

[9] Fang, J., Wei, W., Liang, Z., Dougherty, E., Yang, C., Chang-Tin, L., and Ramakrishna, N. Misinformation Propagation in the age of Twitter: A quantitative analysis of tweets

during Ebola crisis reveals that lies, half-truths, and rumor can spread just like true news. IEEE Computer Science, 2014, pp: 90-94.

[10] Hughes, L. A., and Leysia, P. Twitter adoption and use in mass convergence and emergency events. Proceedings of the 6th International ISCRAM Conference – Gothenburg, Sweden, May 2009, pp: 1-10.

[11] Jim, J., Andrea, H. T., Bajpai, K., Yen, J., and Giles, L. Seeking the Trustworthy Tweet: Can Microblogged Data Fit the Information Needs of Disaster Response and Humanitarian Relief Organizations. Proceedings of the 8th International ISCRAM Conference, May 2011, pp: 1-10.

[12] Jeffrey, D., and Sanjay, G. Map-Reduce: Simplified data processing on large clusters. Communication. ACM, January 2008, pp: 107–113.

[13] Kevin, N. beyond Volume, Variety, and Velocity is the Issue of Big Data Veracity. Inside BIG DATA au. September 12 2013, pp: 1-9, http://www.insideblogdata.com.

[14] Kevin, S. Blog post The Data Explosion Minute by Minute – Infographic www.aci.info in 2014.

[15] Kathleen, M. C, Fred, M., Jürgen, P., and Huan, L. Is the Sample Good Enough? Comparing Data from Twitter's streaming with Twitter's Firehouse arXiv: 1306.5204 [cs.SI], ICWSM, 2013, pp: 1-15.

[16] Lukoianova, T., and Rubin, V. Veracity Roadmap: Is Big Data Objective, Truthful and Credible? Advances in Classification Research Online, 24(1), doi: 10.7152/acorn. V24i1.14671, 2013, pp: 1-12.

[17] Lewandowsky, S., Ecker, K. H. U, Seifert, C., Schwarz, N. and Cook, J. Misinformation and its Correction: Continued Influence and Successful Debasing. Association for Psychological Science in the Public Interest, 2014, pp: 106-131.

[18] Lei, S., Andranik, T., Timm, O. S., Philipp, G. S., and Isabell, M. W. Predicting Elections with Twitter: What 140 characters Reveal about Political Sentiment. Proceedings of 4th International AAAI Conference on weblog & social media, 2010, pp: 178 – 185.

[19] Gerber, S. M. Predicting Crime Using Twitter and Kernel Density Estimation: Decision Support System (EISEVIER), Journal Article, 2014, Vol: 61, pp: 115-125.

[20] Marcelo, M., Barbara, P., and Carlos, C. Twitter under Crisis: Can we trust what we RT? Proceedings of the first workshop on social media analytics SOMA10, pp: 71-79.

[21] Sang, C., and Sandy, L. Predicting Stock Market Fluctuations from Twitter An analysis of the predictive powers of real-time social media. ACM Transactions on Intelligent Systems and Technology, 2011, pp: 1-24.

[22] Singh, N. Big Data Analytics. International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India, 2012, pp: 19-20.

[23] Shvachko, H., Kiang, S., and Chandler, T. The Hadoop distributed file system. In Mass Storage Systems and Technologies (MSST), IEEE 26th Symposium on, May 2010, pp: 1–10.

[24] Sun - Ki, C., Mohammad – Ali, A., and Huan, L. Real – world Behavior Analysis through social media lens. Proceedings of the 5th International Conference on social computing, 2012, pp: 18-26.

[25] Scott, G., Boyd, D., and Lotan, G. Tweet, Tweet, Retweet: Conversational Aspect of Retweeting on Twitter. HICSS – 43, IEEE: Kauai, HI, January 2010, pp: 1-10.

[26] Tauhid, R. Z., Ralf, H., Gael, J., and David, S. Predicting Information Spreading in Twitter. International Conference on IEEE, January 2010, pp: 1-10.

[27] The United States Department of Justice www.justice.gov

[28] The FBI Federal Bureau of Investigation www.fbi.gov.

[29] Twitter study. Technical Report, pear Analytics, August 2009, pp: 1-5. Source: http://www.quantcast.com/twi4er.com#demographics

[30] US Securities and Exchange Commission's www.sec.gov.

[31] Vasuki, S., Yuko, T., and Toshihiko, M., Toward a Social-Technological System that Inactivates False Rumors through the Critical Thinking of Crowds. 46th Hawaii International Conference on System Sciences, 2014, pp: 1-23.

[32] Vasuki, S, and Luke, H. Making Social Media Research Reproducible. ICWSM Workshop, 2015, pp: 2-7.

[33] U.S. Securities and Exchange Commission: Sec Charges KBR and Halliburton for FCPA violations For Immediate Release 2009-23; Washington, D.C., Feb. 11, 2009, page 1 of 3 http://www.sec.gov/news/press/2009/2009-23.htm.

[34] JUSTICE NEWS Department of Justice, Office of Public Affairs, For Immediate Release, Wednesday, February 11, 2009: KBR LLC and Halliburton Pleads Guilty to Foreign Bribery Charges and Agrees to Pay Criminal Fine of $402 Million, Page 1 of 2 http://www.justice.gov/opa/pr/kellogg-brown-root -pleads-guilty-foreign-bribery-charges

[35] United States District Court Southern District of Texas Houston Division. Securities and Exchange Commission, Plaintiff, vs. Halliburton Company and KBR, INC., Defendants. Civil Action No.: 4:09-399. Dated: February 11, 2009, pp: 1-15.

[36] U.S. Department of Health and Human Science U.S. ENVIRONMENTAL PROTECTION AGENCY, OFFICE OF INSPECTOR GENERAL and Semiannual Report to Congress: Halliburton Pleads Guilty Regarding Oil Spill, April 1, 2013-Septemper 30, 2013, pp: 1-255.

[37] United Nations General Assembly: Human Rights Council, Promotion and protection of all human rights, civil, political, economic, social and cultural rights,. District. General 9 September 2013 GE. 13-16847, pp: 1 -2.

[38] United States Department Attorney's Office, Northern District of Florida, Department of Justice. U.S., Attorney's Office. For Immediate Release, Tuesday, August 27, 2013. USAO-NDFL, page 1 of 2. http://www.justice.gov/usao-ndfl/pr/seven-indicated-fraudlent-bp-oil-spill-claims

[39] United States Department of Justice: Deepwater Horizon (BP) Oil-Spill Fraud: Rokeisha Barrios, New Orleans Woman Charged with Defrauding Gulf Coast Claims Facility, Tuesday, January 31, 2012, pp: 1-258 http://www.justice.gov/criminal-oil-spill

[40] U.S., Securities and Exchange Commission. A Study of the Economic Impact of the Deepwater Horizon Oil Spill, greater New Orleans Inc. Michael Hecht, President and CEO, April 20, 2011 pp: 1-325.

[41] United Sates Department of Justice United Nations, Environmental Security Agency, BP oil spill contributed to dead dolphins, scientist say, citing tissues samples. Dolphins in

the Gulf of Mexico swim through oil at the heights of the BP Deepwater Horizon oil spill in 2010, January 26, 2015, pp: 1-25.

[42] The FBI Federal Bureau of Investigation: U.S. Attorney's Office, Northern District of Alabama: Birmingham Division: Conspirators in Gulf Oil Spill Fund Fraud Sentenced. January 26, 2015, pp: 1-2. http://www.fbi.gov/birmingham/press-release/2015/CIGFFS .

[43] U.S. Securities and Exchange Commission: Sustainability Review 2013, Deep Water. The Gulf Oil Disaster and the Future of offshore Drilling. Report to the President National Commission on the BP Deepwater Horizon oil spill and offshore Drilling. January 2011, pp: 1-2.

[44] U.S. Securities and Exchange Commission: BP will plead Guilty and pay over $525 Billion to settle civil charges by the SEC. Clifford Kraus, Nov., 15, 2012, pp: 1 of 2 http://www.nytimes.com/2012/11/16/business/global/16ihtbp16.html

[45] Hadoop Tutorial, http://hadooptutorials.co.in/.

[46] http://www.eia.gov/todayinenergy/detail.cfm?id=19451

[47] Zhang, X., Fuehres, H. & Peter, A. Gloor. Predicting Stock Market Indicators through Twitter "I hope it is not as bad as I fear". The 2nd Collaborative Innovation Networks Conference - COINs2010, Procedia - Social and Behavioral Sciences, Volume No. 26, 2011, pp: 55–62.

[48] Bong Sug, K. C. Insights from hashtag #supply chain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research. International Journal of Production Economics, Volume No. 165, 2015, pp. 247–259.

[49] Xiong, H., Pandey, G., Steinbach, M., & Kumar, V. Enhancing Data Analysis with Noise Removal. IEEE Transactions on Knowledge and Data Engineering, Volume No. 18, 2006, Issue 3, pp. 304-319.

APPENDIX I – Configuration file for streaming Twitter data

TwitterAgent.sources = Twitter

TwitterAgent.channels = MemChannel

TwitterAgent.sinks = HDFS

TwitterAgent.sources.Twitter.type = edu.cs.okstate.CustomFlumeSource.TwitterSource

TwitterAgent.sources.Twitter.channels = MemChannel

TwitterAgent.sources.Twitter.consumerKey = qRth6LxfXs2kKsE1uMahwjIxD

TwitterAgent.sources.Twitter.consumerSecret =
vl32VcubcK9EtuxBM5aTEsRQvFu1bg1ms2X0qcl7KHVNUKTThd

TwitterAgent.sources.Twitter.accessToken = 349038207-
IQdclqAq4Yl6RU2XcblplX7A1zhBwqwUa9SnwR9s

TwitterAgent.sources.Twitter.accessTokenSecret =
HYvBBuhLg6bSvuQlj3at1J3Y7cJQ7neIbAPW0sEmqFNMW

TwitterAgent.sources.Twitter.keywords = Halliburton, Baker Hughes Incorporated, Schlumberger Limited, Transocean limited, ensco pic, National oilwell varco, Akastor ASA, technip SA, subsea7SA, oceaneering international, Exxon mobil, chesapeake

energy, devon energy, BP, Encana, platts market data-oil, crudes, Brent crude, Bunkers, Tankers, european petroleum swaps, asian petroleum swaps, volatility in oil industry, American petroleum institute gravity, NYMEX, geopolitical factors, opec, crude oil futures price, tight oil,shortage of oil, volatility, long-range dependence oil prices, oil well shutdown, crude oil shutdown

TwitterAgent.sinks.HDFS.channel = MemChannel

TwitterAgent.sinks.HDFS.type = hdfs

TwitterAgent.sinks.HDFS.hdfs.path = hdfs:
//namenode:54310/user/hadoopusr/flume_data/Prashanth_tweets

TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream

TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text

TwitterAgent.sinks.HDFS.hdfs.batchSize = 100

TwitterAgent.sinks.HDFS.hdfs.rollSize = 0

TwitterAgent.sinks.HDFS.hdfs.rollCount = 0

TwitterAgent.channels.MemChannel.type = memory

TwitterAgent.channels.MemChannel.capacity = 1000000

TwitterAgent.channels.MemChannel.transactionCapacity = 100

APPENDIX II – Sample JSON data

{"filter_level":"medium","retweeted":false,"in_reply_to_screen_name":null,"possibly_sensitive":false,"truncated":false,"lang":"en","in_reply_to_status_id_str":null,"id":534420687662743552,"extended_entities":{"media":[{"sizes":{"thumb":{"w":150,"resize":"crop","h":150},"small":{"w":340,"resize":"fit","h":255},"medium":{"w":600,"resize":"fit","h":450},"large":{"w":640,"resize":"fit","h":480}},"id":534420687385927680,"media_url_https":"https://pbs.twimg.com/media/B2qkyKBCIAAGW6z.jpg","media_url":"http://pbs.twimg.com/media/B2qkyKBCIAAGW6z.jpg","expanded_url":"http://twitter.com/BuddhaKatWX/status/534420687662743552/photo/1","indices":[112,134],"id_str":"534420687385927680","type":"photo","display_url":"pic.twitter.com/HvVsQxNPU7","url":"http://t.co/HvVsQxNPU7"}]},"in_reply_to_user_id_str":null,"timestamp_ms":"1416250795652","in_reply_to_status_id":null,"created_at":"Mon Nov 17 18:59:55 +0000 2014","favorite_count":0,"place":null,"coordinates":null,"text":"2:00 PM 22F RH: 68% DP: 13F Rain: 0.000 in. Wind: N/0MPH Gust: 1 BP: 29.75in/Rising slowly #ketteringohwx #ohwx http://t.co/HvVsQxNPU7","contributors":null,"geo":null,"entities":{"trends":[],"symbols":[],"urls":[],"hashtags":[{"text":"ketteringohwx","indices":[91,105]},{"text":"ohwx","indices":[106,111]}],"media":[{"sizes":{"thumb":{"w":150,"resize":"crop","h":150},"small":{"w":340,"resize":"fit","h":255},"medium":{"w":600,"resize":"fit","h":450},"large":{"w":640,"resize":"fit","h":480}},"id":534420687385927680,"media_url_https":"https://pbs.t

wimg.com/media/B2qkyKBCIAAGW6z.jpg","media_url":"http://pbs.twimg.com/media/B2qkyKBCIAAGW6z.jpg","expanded_url":"http://twitter.com/BuddhaKatWX/status/534420687662743552/photo/1","indices":[112,134],"id_str":"534420687385927680","type":"photo","display_url":"pic.twitter.com/HvVsQxNPU7","url":"http://t.co/HvVsQxNPU7"}],"user_mentions":[]},"source":"<a href=\"http://www.weather-display.com\" rel=\"nofollow\">Weather Display Tweet<\/a>","favorited":false,"in_reply_to_user_id":null,"retweet_count":0,"id_str":"534420687662743552","user":{"location":"Kettering, OH","default_profile":true,"profile_background_tile":false,"statuses_count":49495,"lang":"en","profile_link_color":"0084B4","profile_banner_url":"https://pbs.twimg.com/profile_banners/1496017094/1407807410","id":1496017094,"following":null,"protected":false,"favourites_count":155,"profile_text_color":"333333","verified":false,"description":"Weather conditions for Kettering & SE Montgomery County Ohio every 30 minutes from our Ambient WS-2080.","contributors_enabled":false,"profile_sidebar_border_color":"C0DEED","name":"Kettering Ohio WX ⛅","profile_background_color":"C0DEED","created_at":"Sun Jun 09 16:27:33 +0000 2013","default_profile_image":false,"followers_count":233,"profile_image_url_https":"https://pbs.twimg.com/profile_images/486582903720382465/iL9wcXx9_normal.jpeg","geo_enabled":true,"profile_background_image_url":"http://abs.twimg.com/images/themes/theme1/bg.png","profile_background_image_url_https":"https://abs.twimg.com/images/th

emes/theme1/bg.png","follow_request_sent":null,"url":"http://bkwxcam.ddns.net","utc_offset":-18000,"time_zone":"Eastern Time (US & Canada)","notifications":null,"profile_use_background_image":true,"friends_count":230,"profile_sidebar_fill_color":"DDEEF6","screen_name":"BuddhaKatWX","id_str":"1496017094","profile_image_url":"http://pbs.twimg.com/profile_images/486582903720382465/iL9wcXx9_normal.jpeg","listed_count":13,"is_translator":false}}

VITA

PRASHANTH REDDY KAMMARPALLY

COMPUTER SCIENCE

Master of Science

Thesis: A CASE STUDY ON VERACITY IN TWITTER DATA USING OIL COMPANY RELATED TWEETS

Major Field: Computer Science

Biographical:

Education:

Completed the requirements for the Master degree with a major in Computer Science at Oklahoma State University, Stillwater, Oklahoma in July, 2015.