2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing

The Evaluation of The Public Opinion

A Case study: MERS-CoV infection Virus in KSA

Anis Zarrad
Department of Computer Science and Information Systems
Prince Sultan University
Saudi Arabia
azarrad@psu.edu.sa

Abdulaziz Aljaloud
Department of Computer Science and Information Systems
Prince Sultan University
Saudi Arabia
abdulaziz.jaloud@gmail.com

Izzat Alsmadi
Computer Science Department
Boise State University
USA
izzatalsmadi@boisestate.edu

***Abstract*— Opinion Mining and Sentiment Analysis are active research trends in natural language processing and data mining. Recently, this research has been extended outside the computer science area to cover other areas such as social science, political science, and business. The explosion of social media such as social networks, Blogs, Twitter, and forums has created unprecedented opportunities for data mining research community. Analyzers can study and analyze users' opinions, attitudes, and emotions about news or social events. Big data focuses on the intelligent analysis of a large amount of data that is typically collected from several different sources. Our focus in this work is to address new challenges raised by combining Apache Hadoop as a big data platform with an opinion mining approach to make a decision we often seek based on collected data from the opinions of people. We presented a case study about MERS virus in KSA to evaluate our proposed approach. A discussion of available dataset and results are also provided.**

Keywords—component; Big Data; Hadoop; Opinion mining; sentimental analysis; MERS-CoV infection Virus, Social Networks

I. INTRODUCTION

Nowadays, the growing importance of Social media coincides with the growth of big data technologies. To this end it is now possible to store a large amount of data being exchanged among people through social networks beyond traditional database systems [1]. McKinsey Global Institute estimates that consumers around the world stored more than six Exabyte (one Exabyte equal to 1,048,576 terabytes) of new data on devices such as PCs and notebooks which is equivalent to more than 4,000 times the information stored in the US Library of Congress [2]. The question "What people think about specific subject" has always been a challenge for decision-maker? For example in our real life you may ask your friends opinions about vote election, best smart device available in the market, Job opportunities etc… Such opinions may be incorrect or/ and untrusted, if the collected data are insufficient and not significant.

Due to the rapidly increasing usage of social networks, Big data will become ubiquitous in the future. This research thesis is directed towards big data collection to determine the appropriate information in order to perform accurate and valuable analysis about humans' opinions, sentiments, and to avoid faulty decisions. We aspire to take advantages from the available huge amount of information exchanged each day in social networks Twitter.

Big data evolved recently with the information overload from the Internet, social networks, etc. Big data tried to collect and gather data from different related sources. The goal is to have a better global vision for a particular issue and be able to make findings based on studying this huge amount of data. Collected data can be structured and unstructured.

In this paper we used Hadoop [16], an open source Apache based data system to collect a large amount of Tweets from users in Saudi Arabia. Other sources might be configured later on like Facebook and LinkedIn. Hadoop is being used widely in companies like Yahoo, Facebook etc. [7]. Our challenges in this work are collecting and analyzing big data to study users opinions about specific subject.

Opinions might be positive, negative or neutral. This makes the mission for the analysis a quite difficult. A Sentiment Analysis, also known as Opinion mining will be presented in this paper, which is the process of identifying positive and negative opinions, emotions, and evaluations [8]. This type of analysis is important for decision maker process to avoid any misunderstanding and provide valuable results.

In this work we present a case study covering a hot topic about MERS-CoV infection Virus [15] in order to analyze people opinions in Saudi Arabia. To address this issue, we

collected huge amount of data using big data technology and try to answers the following questions related to satisfaction: available preventions methods proposed by the ministry of health. Statistics Truthiness, and services offered by the ministry of health. Results can be mapped to different dimensions; for instance, we can find out satisfaction and or non-satisfaction in certain period of time and certain region.

The main contribution of this works are as follow: First, a cooperative architecture is designed and implemented between Hadoop and social network Twitter to deal effectively with big data and to collect /store data. Second, analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written Arabic language using an efficient analysis methodology that cover all kind of user behaviors. Third, presents a case study to cover a hot topic in Middle East region (MERS-CoV infection Virus) to evaluate our proposed approach.

The paper is organized as follows. The next section presents related works. Section III provides a brief introduction of the proposed architecture, data collection method using Hadoop, and describes the analysis methodology. Section IV highlights the case study about MERS-CoV infection Virus in KSA and analysis results. The last section concludes the paper..

II. Related Works

In the area of social network analysis a large number of publications have been conducted to evaluate users' opinions. Several approaches of wide applicability have been developed; ranging from Business or marketing, to news or political and social subjects.

In this section we limit ourselves to social subjects that use Big data technologies since it represents the main focus of this study. Kim et al [1] presented an approach to evaluate customers' preferences on certain products based on social networks analysis and users' comments or posts. The proposed approach uses Hadoop big data, to collect about 600,000 Twitter comments from one month period. Big data configuration architecture includes also in addition to Hadoop, HIVE, Chukwa, R, Twitter4J, etc. Different nodes and processing environments are evaluated. Different keywords are used to collect the data. Hannanum Java based morphological analyzer is used to process that data into sentiments. Polarity sentiments produce a class of three possible results or class labels: Positive, negative or neutral. We followed a similar approach in our paper. However, the detailed steps in the experiment are different in our case where the context in our case (i.e. public opinion on an infectious disease) is not product or marketing oriented.

Anjaria et al [2] evaluated several classification algorithms for sentimental analysis. Authors presented two case studies: US presidential election 2012 and Karnataka state election in 2013. Support Vector Machine (SVM) classification algorithm showed best results in terms of prediction accuracy. Another approach was proposed by Zhang et all [3] to detect spam in social networks.

Algorithms are used to detect whether a Tweet is a duplicate or not. Authors collected a large number of Tweets from many users. Users are also classified into 5 classes: Users, robots, information aggregators, marketing accounts and others. In [4] authors evaluated the effect or the relation of the location and the Tweet. They tried to study the evolution of different topics in different locations where focus is giving for volatile or hot topics.

III. Proposed Approach

This section presents the proposed approach used for analyzing a huge amount of Arabic Tweets that are related to the recently discovered MERS-CoV infection virus in Saudi Arabia. However it can be extended to cover other subjects and to deal with other languages such as English , French etc…

The below sub-sections will explain in details the components of our proposed approach which include: data collection, data cleansing and pre-processing, data analysis and results' presentation, these components are sequentially illustrated in figure 1.

A. Twitter as a Data Source

Twitter is an online microblogging service that allows users to expose their thoughts in 140 characters. Saudi Arabia is the country in the region with the highest percentage of active Twitter users among its online population [5]. In 2012, The Global Web Index addressed that 51% of population in Saudi Arabia are active Twitter users; this indicates a huge growth in active usage [6]. Because of the popularity and the huge amount of data created in Twitter, it was selected as the main source and input for this work analysis.

B. Data Collection

A program is designed and implemented to capture all possible Tweets that contain certain words related to our case study i.e. MERS-CoV. Initially, the program was seeded with a set of all related keywords, then we did a frequency analysis on regular bases to discover new trends that directly or indirectly point to our subject, for example, the hashtag #وزارة_الصحة (Ministry of Health), was commonly tagged with Tweets that talk about MERS-CoV.

This program is subjected to capture all real-time streams i.e. any Tweet posted globally in real-time that match our keywords. Also, it is capable of crawling old posted Tweets that precede the first run date or Tweets posted after the program gets stopped accidentally.

The integration with Twitter is done indirectly by an open source tool [7] which is used as an interface to simplify the complicated integration process introduced in the newly Twitter REST API v1.1.

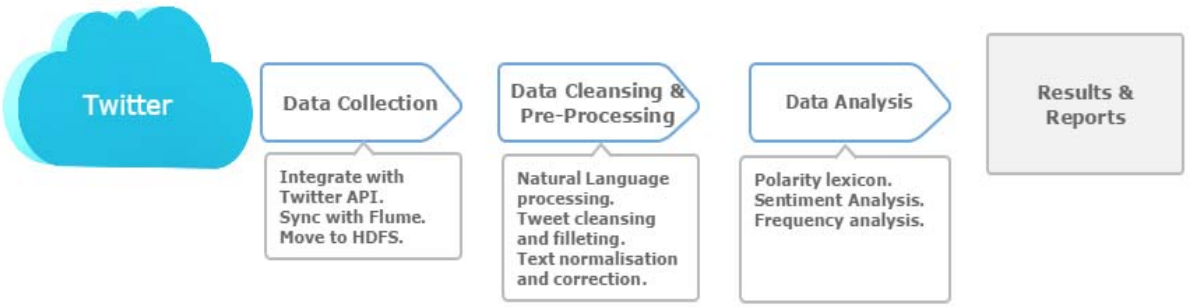


Fig.1. The system architecture for Arabic tweets opinion mining

Figure 2 shows how the captured Tweets get stored into Hadoop File System (HDFS) using Apache Flume. Flume is "a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data" [8]. In our program, Flume is responsible for moving and aggregating all received Tweets to pre-defined locations through a channel to HDFS.

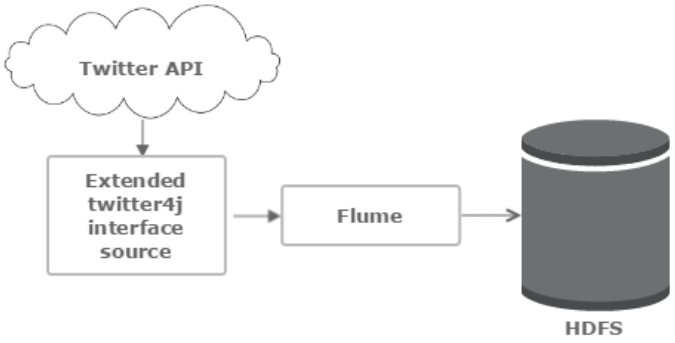


Fig.2. Dataflow from Twitter to HDFS

Around 1,500,000 Tweets were collected in three months. All Tweets information were archived and not discarded like Tweet's location, mentions, hashtags… etc. As Hadoop is capable of dealing with a high volume of data, this amount of information is kept for different types of analysis in future.

C. Data Cleaning and Pre-Processing

Text pre-processing is an essential phase due to the existence of Arabic language challenges addressed in [9]. For a number of reasons, Arabic language requires a complex natural processing, so we developed a tool that takes care of most of the challenges and prepare the text for the final analysis phase as shown in the below steps.

1. **Data Cleaning**: In this step we remove all kinds of unwanted text from Tweets that are not important for our analysis which include: URLs, @usernames, #hashtags, special characters and non-Arabic words.
2. **Remove Arabic Diacritics and Extensions**: Some Tweets contain diacritics even though they are not in a Modern Standard Arabic (MSA). In this task we remove all diacritics and extensions in a Tweet.
3. **Remove Repetition of letters**: Repetition of a letter in a word may show an emphasis on a certain emotion, for example 'كفى أرجووووووكم' (Please stop). This happened frequently in social media, so in this case we try to detect and remove the repeated letters.
4. **Text Correction**: Hunspell API [10], an open source spelling checker that supports Arabic was used to look for any word that is misspelled then correct it.
5. **Normalization**: This is a very important phase to unify the form between words in Tweets and the words in the lexicon dictionary. This is done by the below steps:
   - ✓ Replacing any occurrences of أ, آ, ى and إ with ا.
   - ✓ Replacing any occurrences of ؤ with و.
   - ✓ Replacing any occurrences of ئ with ي.
   - ✓ Replacing any occurrences of ة with ه.

D. Data Analysis

We will present in the below sections, two kinds of analysis using Hadoop, and Sentiment Analyzer and Hashtag Words Frequency Analysis.

1) Arabic Sentiment Analysis

We present an automatic lexicon-based classifier that tries to classify and determine whether a Tweet written in Arabic (MSA or slang Arabic) is holding a positive, negative or a neutral opinion or emotion.

Constructing the polarity lexicon was the big challenge and the main component in this analysis, though; we manually seeded the lexicon by reviewing a collected set of related Tweets with around 1,100 negative words and

850 positive words, the lexicon can contain composite words like 'حسبي الله عليهم'. We then downloaded an English polarity lexicon that is constructed by MPQA [11] and contains 8,222 labeled words in positive, negative or neutral. The lexicon is translated by Google Translate to Arabic then we did a quick review for the translation result before adding it to our lexicon.

A negation list was constructed which contains about 11 words that may emphasize a negative impression when they are combined with positive words, Figure 3 shows how the stop word 'ما' (not) affects the polarity meaning of the phrase.

Fig.3. A phrase affected by the stop word 'ما', (The Ministry of Health performance is not satisfactory)

Our analysis simply depends on the classification method proposed in [12], which is based on computing the number of the occurrences of positive or negative words in a Tweet, this will lead to discover the overall opinion of the Tweet. A Tweet can hold an opinion if it contains one or more negative or positive words, otherwise it is a fact.

And we can say that a Tweet is a positive, if the total number of the positive words occurrences is more than the total for negative words and vice versa.

To formalize this, let's assume that t is a Tweet and T is the collected training set of Tweets, for each t ∈ T:

t is holding an emotion e if POS >0 or NEG >0, otherwise t is a fact.

If POS – NEG > 0 then e is positive.

If NEG – POS >0 then e is negative.

If POS = NEG then e is neutral.

Where

e can be an emotion in {positive, negative or neutral}.

POS is the total number of positive words occurrences in t.

NEG is the total number of negative words occurrences in t.

Algorithm 1 shows how the explained classification method above is implemented. The algorithm took an advantage of Hadoop capabilities of processing; the algorithm can search for up to five consecutive composite words in our polarity lexicon. It also can detect positive words that are affected by a negation word even if they are not consecutive.

Algorithm 1 Evaluate Polarity

```
input is a tweet words W = {w1, … … ,wn}
Output polarity p
set pos to 0
set neg to 0
set neu to 0
for i = 1; i <= sizeof(W) do
    for y = min(sizeof(W) – i, 5) ; y >= 1 do
        set
        set e = checkInLexicon(SW)
        if(e = negative) then
            neg = neg+1
        else if(e = positive) then
            set NW = {w max (i–3,0), … … ,wi–1}
            if(isContainNegationWord(NW) = true)
            then
                neg = neg-1
            else
                pos = pos+1
            end if
        else
            neu = neu+1
        end if
        y = y-1
    end for
    i = i+1
end for

if(pos + neg > 0 ) then
    p = positive
else if(pos + neg = 0 ) then
    p = netureal
else
    p = negative
end if
return p
```

2) Hashtag Frequency Analysis

Hashtags in twitter are words prefixed with '#' that are used to group public messages and discussions [13]. In our work, we analyze all tagged hashtags in the collected set of Tweets because it is important to discover new trends that are related to our targeted subject. For example, a Tweet which is tagged with the hashtag '#لاتبوس_الناقة' (Don't kiss the camel) may not contain any keywords about MERS-Cov, but in fact, it is related to the subject.

This analysis is done in regular basis by finding the top ranking for all hashtags in all collected Tweets in a given range of time. It also finds the frequency for a hashtag over a unit of time; this is useful to track how the hashtag was active in the same period of time.

IV. CASE STUDY: MERS-COV INFECTION VIRUS IN SAUDI ARABIA

A. Introduction and Settings

The first case of MERS-CoV infection virus was reported in 2012 in Saudi Arabia [15]. MERS-CoV infection developed severe acute respiratory illness complemented with fever, and cough. About 30% of people confirmed to have MERS-CoV infection have died. On May 2, 2014, the first U.S. imported case of MERS was confirmed in a traveler from Saudi Arabia to the U.S. Health Department in Saudi Arabia closely monitor the MERS situation and work with international partners to better understand the risks of this virus and stop the virus spreading, People living in KSA were struggling about the efforts made by the government to protect the populations and the offered services. Big data is used to analysis the opinions. For this a single-node Hadoop 2.2.0 cluster with the HDFS and a pseudo-distributed mode is configured on Ubuntu Linux 14.04.

Hive 0.31.0 is configured with Hadoop, Apache Hive is data warehouse software built on top of Hadoop that is capable of analyzing and querying large datasets stored in Hadoop's HDFS using a SQL-like language called HiveQL [14]. For the Sentiment Analysis, we developed multiple Hive User Defended Functions (UDF) that are used inside Hive queries for extracting opinions.

For Frequency Analysis, we developed a Java MapReduce program that is able to find the most frequent hashtags associated with positive and negative Tweets stored in HDFS.

B. Result Analysis

In three months, we have collected 1,498,513 Tweets for the targeted subject MERS-Cov in May, June and July 2014. Table 1 shows some statistics about the training set.

TABLE I: TRAINING SET

	Total
#All collected set	1,498,513 Tweets
#Duplicated and Spam	326,164 Tweets
#Total unprocessed	103,889 Tweets
#Total related to MERS-Cov	1,068,460 Tweets
#Total hashtags	3,677,142 hashtags
#Total processed words	18,574,436 words

The performed Arabic Sentiment Analysis on the collected data shows that 21,9% of the Tweets are positive, 39.8% are negative and 38.2% are neutral as shown in Figure 4.

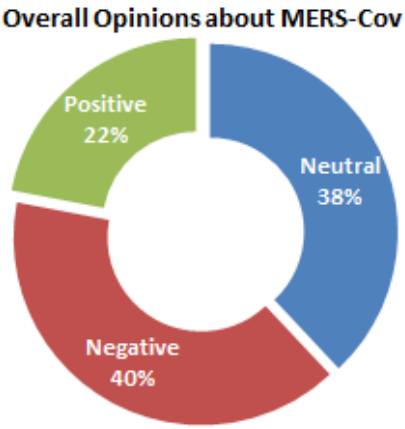
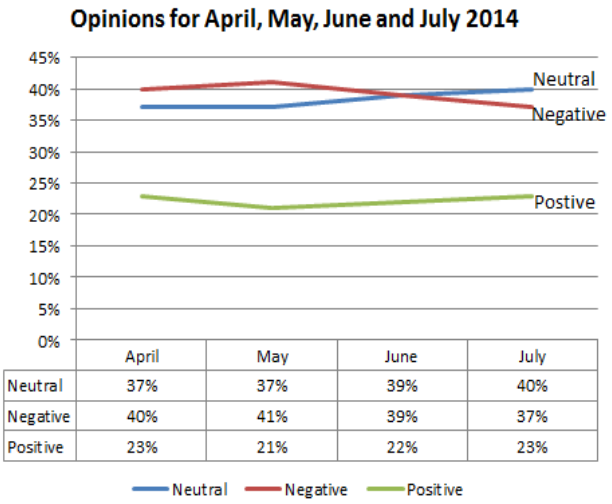


Fig.4. Overall Opinions about MERS-Covfrom the collected dataset

Figure 5 shows the overall opinions about MERS-Cov against the consecutive months: April, May, June and July 2014. Analysis was performed by simple queries using Hive. Fig. 4 shows that there is a slightly increase in the positive and neutral Tweets along with a decrease of the negative Tweets. This may give a clue about a satisfaction over the mentioned months.



	April	May	June	July
Neutral	37%	37%	39%	40%
Negative	40%	41%	39%	37%
Positive	23%	21%	22%	23%

Fig.5. Overlall opinions for the months April, May, June and July 2014.

For the hashtag frequency analysis, the total frequency for 9,303 unique hashtags is 3,677,142. Table 2 shows the top 5 hashtags over the evaluated period.

TABLE II: TOP FIVE HASHTAGS

Rank	#Hashtag	Frequency
1	كورونا	1,011,848
2	توعية مجتمع	730,890
3	وزارة الصحة	129,072
4	السعودية	97,789
5	فيروس كورونا	96,576

Figure 6 shows the most 33 frequent hashtags separated across a world cloud chart, bigger words in size, denote a high frequency.

Fig.6. A word cloud chart generated to represent the most 33 frequent hashtags

We applied the time dimension on the hashtag '#فيروس_كورونا' as an example, to measure the activity of the hashtag during the months: April, May and June 2014. Figure 7 shows how the same hashtag was active during the consecutive months.

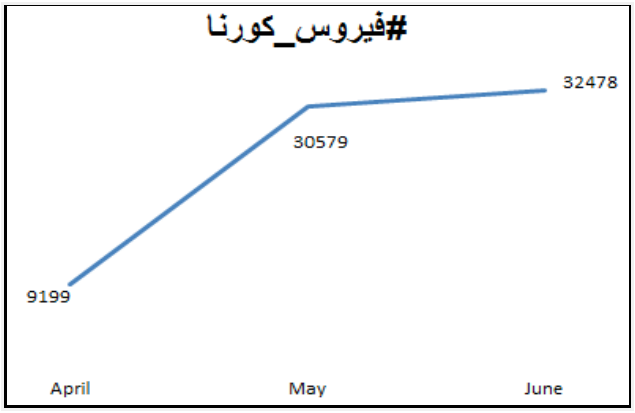


Fig.7. Hashtag activity chart for the hashtag #فيروس_كورونا during the months April, May and June 2014

From the hashtag frequency analysis we drilled down to find the opinion polarity for two of the top 33 frequent hashtags. Figure 8 shows the polarity of the hashtag '#موسم_الحج' (Pilgrimage Season) which is about how the Ministry of Health and Haj will deal with MERS-Cov in the next pilgrimage season. Figure 9 shows the polarity of hashtag '#عادل_فقية' (the new Minister of Health) which is about the performance of the recently assigned minister after the previous minister got dismissed. Efforts of pervious health minister were not appreciated by most of the KSA residents for the last three months.

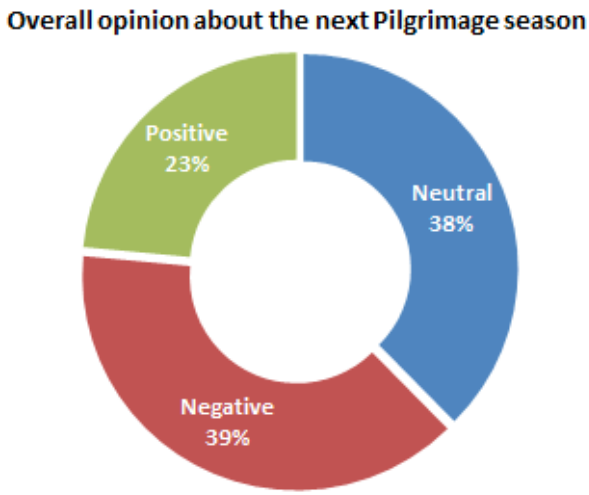


Fig.8. Pilgrimage Season Hashtag

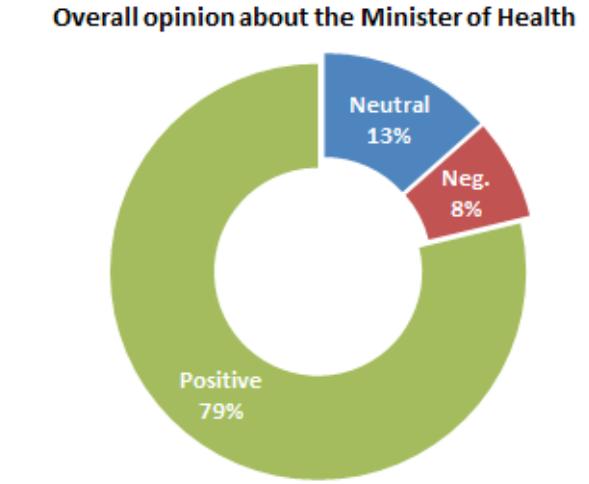


Fig.9. Minister of Health Hashtag

As shown in Figure 9, people showed positive opinion form the new minister of health who was assigned recently.

V. CONCLUSION

The value of information collected from social networks has been the subject of many studies in different fields. Most of the existing approaches used standard database systems to collect and analyze data. In this paper, we collected social data related to an infectious virus/disease that spreads in KSA (MERS-CoV) in 2014. A dataset of 1,500,000 Tweets is collected and analyzed.

Big data techniques are also used to process the large data collected. We showed some focused Hash tags such as "Pilgrimage" or "New prime minister". Those are considered two significant milestones in the course of the event. The approach can be easily extended to include several different subjects. We think that such studies can provided a simple automated method to evaluate public opinions. Output of such studies can be very helpful for decisions makers.

ACKNOWLEDGMENT

The authors would like to express their thanks to Prince Sultan University and Boise State University for supporting this work.

REFERENCES

[1] J. S. Kim, M. H. Yang, Y. J. Hwang, S. H. Jeon, K. Y. Kim, I. S. Jung, C. H. Choi, W. S. Cho, and J. H. Na, "Customer Preference Analysis Based on SNS Data", 2012 Second International Conference on Cloud and Green Computing, pp. 106-113, 2012.

[2] M. Anjaria, and R. Reddy Guddeti, Influence Factor Based Opinion Mining of Twitter Data Using Supervised Learning , COMSNETS 2014: pp. 1-8, 2014.

[3] Q. Zhang, H. Ma, W. Qian, and A. Zhou, "Duplicate Detection for Identifying Social Spam in Microblogs", IEEE International Congress on Big Data, pp. 52-61, 2013

[4] M. Saravanan M, D. Sundar, and S. Kumaresh, "PROBING OF GEOSPATIAL STREAM DATA TO REPORT DISORIENTATION" IEEE Recent Advances in Intelligent Computational Systems (RAICS), 2013

[5] Marcello Mari, "Twitter usage is booming in Saudi Arabia", 2012.

[6] The Global Web Index, "The Fastest Growing Social Platform", 2012.

[7] J. Maximilian, K. Gjergji and N. Felix,"Analyzing and Predicting Viral Tweets", 13–17, 2013.

[8] R. Sanjay, "Big Data and Hadoop with components like Flume, Pig, Hive and Jaql", 2013

[9] F. Ali and S. Khalid, 'Arabic Natural Language Processing:Challenges and Solutions', 2009.

[10] P. Tommi, L. Krister, "Building and Using Existing Hunspell Dictionaries and TEX Hyphenators as Finite-State Automata".

[11] MPQA Subjectivity Sense Annotations, http://mpqa.cs.pitt.edu/lexicons/subj_sense_annotations/.

[12] N. Mohammed.N. Al-Kabi, I. Alsmadi, H. Amal Gigieh and H A. Wahsheh, "Opinion Mining and Analysis for Arabic Language", 2014.

[13] T. Oren, R. Ari, "What's in a Hashtag? Content based Prediction of the Spread of Ideas in Microblogging Communities",2012.

[14] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, R. Murthy, "Hive – A Warehousing Solution Over a Map- Reduce Framework," 2009.

[15] T. Jaffar, A. Abdullah and M. Zaid, "Middle East respiratory syndrome novel corona (MERS-CoV) infection", 2013.

[16] T. Sanjay, K. Monika, , "A review on Hadoop", 2014.