Analysis of ZIKA Virus Tweets: Could Hadoop Platform Help in Global Health Management

Radmila Juric
HSN, Norway
rju@hbv.no

Inhwa Kim
Samsung, UK
nina.k@samsung.com

Hemalatha Panneerselvam,
HSN Norway
883554@hbv.no

Igor Tesanovic
Inceptum, Croatia
Igor.tesanovic@inceptum.hr

Abstract

This paper investigates possibilities of enhancing everyday decision making in global health management, by looking at the power of twitter data and the use of big data platforms in order to collect and interpret excessive amounts of information generated in a short period of time. We use the scenario of the ZIKA virus because it has triggered a massive response through tweets and retweets. Our goal is to find out a) if we can make sense of twitter data in a global health scare and b) if information available on Twitter could help in the management and containment of the spread of the virus. The results of manual content analysis of selected tweets has been juxtaposed with the results of the manipulation of the same tweets through the Hadoop platform. We wanted to know which approach should be used for addressing public concerns about the ZIKA virus and answer a) and b) at the same time. Both approaches have their advantages and drawbacks. Therefore this paper should be used as an overview of options available for public health organizations, when they need to manipulate social media data in situations where we need to manage health on a global scale.

1. Introduction

The ZIKA Virus has been portrayed as a mysterious illness with devastating effect and rightly or wrongly proclaimed as the next major global health crisis [1]. We have never experienced such an upheaval of worried and alarming *information exchange* on mosquito-born diseases for two reasons. Firstly, social media have already proved to be an effective mode of exchanging information across the world [2], [3] and in cases of humanitarian crises and public health scare, they are still bringing, fairly quickly, relevant data for making informed decisions on how to respond to such crises [4] [5] [6]. Secondly, globalization changes the way diseases are spread around the world [7] and mosquito-born diseases may not necessarily be limited to poor and tropical countries. Modern borders between countries and our excessive travelling around the Globe, do not protect against the spread of infectious diseases, and the lack of investment in research on tropical diseases [8] has resulted in global panic and debates on whether ZIKA is really our next global health threat [9] [10].

In this research, we look at the problem Public health organizations, and institutions involved in Global health face, if they have to decide which actions to take to answer global health issues. We primarily look at data and information accessible by these organizations and search for methods of processing the data to enable informed decision making.

We have had a long term interest in investigating the power of Twitter in information exchange in communities affected by health [11] [12] and other humanitarian crises [13]. Twitter exceeded expectations in terms of disseminating valuable and correct information to the world [14] [15] and it has become an invaluable source of easily accessible data for anyone. Twitter delivers live information, and illustrates views, opinions and reactions of individuals and organizations, which concerns us all. There are no universal methods, which would guide any organization or individual on how to manage and interpret the meaning of information available on Twitter. We can use text mining [16], apply analytical and statistical tools [17, 18], and interpret sentiments through text clustering [19,20] and keywords produced by topic modelling [21]. However, twitter data maybe extremely noisy and it is almost impossible to predict what the content of tweets would be, even if we follow a strictly defined event on Twitter. These are clear problems if we wish to use Twitter data at the time of health scares or crises. Research communities talk about *making sense* of twitter data [22, 23], which could be feasible through either new data management models or new software technologies.

In order to illustrate our own approach to *making sense of Twitter data* and interpreting the meaning of it in a particular situation, we have to find:

a) a health crisis which is currently of interest to Global Health and

URI: http://hdl.handle.net/10125/41553
ISBN: 978-0-9981331-0-2

b) research questions which have to be answered, in order to make decisions or assess the situation in the health crises.

For a) above, our choice of focusing on the ZIKA virus was obvious. However, we were particularly motivated to look at Tweets with #zika hashtag, because our previous research, on the role of Twitter data in the spread of E-bola, before and after the vaccination [11, 12], delivered interesting results. Therefore our experience of trying to *make sense of Twitter data* in E bola cases, helped us to formulate our method for understanding what was tweeted in relation to the ZIKA.

For b) above, we specify which research questions are to be answered and why the answers to these questions are of interest to Public or Global Health organizations or even individuals. We have decided to reuse some of our research questions from the analysis of Twitter data in E-bola crises for two reasons. The questions are universal in terms of their applicability to any health scare or crisis. They are easily converted into a set of methodological steps, which help to *make sense of twitter data* through their categorization, manual content analysis or any similar type of data processing.

These issues indicate that we have to be in a position to filter relevant live Twitter data (as in a)), read individual tweets, categorize them and perform their content analysis in order to answer research questions from b). This would produce the most accurate answers to any question we may have because the manual content analysis is performed on individual Twitter data by a human being.

However, the speed of generating Twitter data, and the type of data Twitter disseminates, signal that we might need to process it using Big Data technologies [24, 25, 26]. *Making sense of twitter data for ZIKA virus crises* could really be supported by the technology if we process Twitter data according to the rules and characteristics typical of Big Data [27] . This involves the automated collection of Twitter data through a Big Data platform, such as Hadoop [28], and querying the collected data, according to the mechanisms available in the platform, for the purpose of answering the same research question as in b) above. This is a more appealing option for anyone interested in processing user-generated data because it eliminates human intervention at the level of individual tweets and could provide answers to questions Public or Global health organizations may have at the time live tweets are generated and collected.

Therefore we have two approaches.

In the content analysis, we should have the content of every tweet read and analyzed by the authors through its categorization. This would require that tweets are initially filtered, possibly with software tools, but the results of filtering should be manually checked. It is important to note that the content analysis of filtered tweets is performed without software tools, i.e. manually, with significant author involvement.

On the other side, we have an automated analysis of filtered tweets through Hadoop and its components. This is performed according to the rules available within the platform and cannot include the categorization of tweets as in the manual content analysis.

We are not in a position to advocate which approach is better for Public and Global health organizations, because the answers are not simple and straight-forward. However, we wish to exchange our experience of using both:

(i) manual content analysis of filtered tweets,
(ii) processing the filtered tweets through Big data technologies, such as Hadoop,

in order to compare these two approaches and see if we are able to answer the same research questions. We apply both approaches to Tweets which

- have **#zika** in their body and
- were generated in a over one week in April 2016.

Therefore, the purpose of this paper is twofold:

(A) It illustrates options and the exact steps any individual or organization might use in order to *make sense* of Twitter data in a particular scenario and

(B) It highlights the benefits and drawbacks of using an automated analysis of twitter data through Hadoop.

The choice of using Hadoop was obvious. It is currently a very popular platform in the Big Data world [29, 30, 31, 32]. It has been included in curricula of quite a few university degrees and the amount of peer-reviewed papers on Hadoop exceeds the number of publications of any other Big Data platform [33,34,35,36]. Hadoop extends the statistical computing and graphics capabilities of R language and the number of support groups and forums available on the Internet for Hadoop users is significant.

In section 2 we define the methodology used in this research and specify research questions, by explaining the rationale behind them. We wish to address our main goal and find out if Twitter data did help in managing the problems associated with the spread of the ZIKA virus. In Section 3 we define our steps for collecting relevant tweets. Live twitter data had to be filtered and collected according to our problem domain and therefore we should define a mechanism of selecting tweets, which are relevant to answering our research questions from section 2. In section 4 we illustrate the categories of tweets which have been defined for answering research questions and performing our manual content analysis. In section 5 we look at the process of inserting filtered Twitter data (from section 3) to Hadoop and querying it through HIVE components of Hadoop. We have to check if Hadoop can answer the same research questions as in section 4. In conclusions

we interpret and compare the results of both approaches (i) and (ii) and comment on the feasibility of using the results of this research in real life scenario.

2. Methodology and Research Questions

In this research we wanted to find out if:

> *Twitter was used for disseminating relevant information on the ZIKA virus, and helped to understand how to manage its spreading across the World*.

Such a generic question, which is similar to the one we used for measuring the response of international communities in E-bola crises [11,12], requires answers to many other questions. For example, we would wish to know:

- How many tweets give "facts" about the ZIKA virus in terms of being trustworthy and referring to verifiable information within the tweet body;
- How many professional bodies, which are expected to be involved in the management of the spread of the ZIKA virus did tweet and give professional advice to the population and other professional bodies;
- What is the truth about the ZIKA virus in terms of our understanding if it is a serious danger to human health;
- If we could exchange information on symptoms of and treatments for the ZIKA virus, relevant to the management of its spreading.

The bullets above should constitute the basic set of questions we have to answer if we wish to *make sense of twitter data*, regardless of the approach we use: either (i) or (ii). Furthermore, we might be in a position of not being able to answer these questions completely through the collected tweets. It is very difficult to predict what people will tweet. For example, do we have to know exactly which information is the most relevant for answering our questions: ZIKA virus symptoms, advice, or treatment? Would the person who is infected by the virus tweet? Do we really need to know facts about the ZIKA virus created by healthcare professionals? Would tweeted information from communities affected by the ZIKA virus have a bigger impact on our response to the spread of virus across the Globe? Could twitter data educate the affected communities? Could we alert the WHO more efficiently through Twitter about the ZIKA virus spread? They might have been well prepared and organized for managing the spread of THE ZIKA virus, but experience problems when delivering help?

It is difficult to answer all the questions we may have. We could find out exactly which questions are likely to be answered by looking at the content of the tweets and therefore approach (i) always pays off. We also have to be very careful with the methodology we use for (i) and (ii) because it can determine the feasibility of answering questions we may have. In order to address the problem of answering the research questions for (i) and (ii) above we defined our own methodology, consisting of the following tasks:

1. Collecting relevant Tweets – we have to filter tweets in order to make our content analysis feasible. In other words, filtering tweets relevant to #zika will ensure that we can perform a manual content analysis in a reasonable period of time,
2. Specifying the steps in tweet filtering – the choice of automated tool and keywords used in filtering should ensure that we collect ALL relevant tweets in a certain period.
3. Defining the categories of filtered tweets – this would enable us to answer research questions.
4. Performing a manual content analysis on 29,000 filtered tweets – we must assign manually a category to each individual tweet.
5. Answering our research questions – answers are based on the results of the manual content analysis.
6. Using filtered tweets from tasks 1 and 2 above (BEFORE their categorization) and inserting them into a repository of the Hadoop's storage system – we have to find out how Hadoop processes the same tweets, collected in tasks 1 and 2.
7. Formulating and performing SQL queries upon Hadoop's repository – we use the same filtered tweets as in the content analysis, but perform SQL queries using Hadoop's component HIVE.
8. Answering our research questions through the analysis of answers to SQL queries through HIVE.
9. Comparing the results of Steps 5 and 8.

3. Collecting Relevant Tweets

We have long term experience of deploying software tools for collecting relevant tweets prior to their analysis. The description of options we may have for collecting ZIKA virus tweets is outside the scope, and we refer readers to [11][12] for the explanation of the value of our method for collecting tweets.

The method consists of the following three steps:
Step 1: Automated filtering of live tweets according to a *key word* of our choice. We used Tweepy open source tool and modified the underlining Python code to filter tweets, which contain **#zika.** There are two important decisions we made at this step. Firstly, we needed a simple tool, which would allow us to experiment with the type of key word(s) we can use in order to make sure that we will collect tweets relevant for our analysis.

Very expensive tools might be more sophisticated and powerful, but they will remove from us the power of choosing keywords or their combinations in the filtering. Secondly, our numerous experiments in the past outlined that keyword **#ZIKA**, placed within a filtering tool will collect the widest range of tweets, which could be used for both (i) and (ii) from the introduction. Therefore the combination of Tweepy and Python proved to be the best possible combination of tools for filtering tweets with a chosen keyword **#zika**.

Step 2: Determining the dates for and amount of Tweets which are to be collected. We had to predict the max number of tweets we wish to analyze in order to answer the research questions. For performing our selected tweet analysis through Hadoop, the number of tweets is irrelevant, i.e. we could take any number of them when using the platform. However the manual analysis of collected tweets by a human being might not be performed consistently, at the time where Tweets are generated, if we collect an excessive number of them. Humans do not have unlimited concentration for performing categorization of excessive number of tweets in a short period of time. Therefore we have decided that, for the sake of our experiments, we will analyze tweets from 2 consecutive days in April (18th and 19th of April 2016) and perform manual and Hadoop analysis of more than 29,000 Tweets in parallel. The chosen dates are arbitrary. In our tweet filtering throughout one week in April 2016, there was no significant difference in the number of tweets generated on any of these days. Our manual inspection of all the tweets collected between the 15th and 24th of April did not reveal any particular anomaly or discrepancies in the content of the tweets. Each day we filtered a minimum of 13,000 tweets. The were mostly written in English, Spanish in Portuguese. In our content analysis we translated non English tweets.

Step 3: Streamlining filtered tweets to a spreadsheet which will allow us to perform the content analysis and feed the Hadoop platform. This step was extremely important for the former, because we wanted to perform manual content analysis in a reasonable period of time. The streamlining of tweets to the spreadsheet was done by using delimiters in Python code which separated the content of each tweet from any additional information within it. Therefore our final result of streamlining live and relevant tweets into a spreadsheet has structured format: each column in that spreadsheet contains information cut from the collected tweets. We need information on the owner of the tweet/retweet, which owner is retweeted, what the body of the tweet / retweet is and similar, all clearly separated in the columns

4. Answering Research Questions through Tweet Categorization and Manual Content Analysis

Our manual content analysis of collected Tweets is based on their categorization.

It is important to note that our tweet categories are dictated by

- Research questions we defined in section 2 and
- Previous experience of using the same categories for answering similar questions in the case of the spread of E-bola [11.12.13].

Table 1 gives an overall count of tweets and retweets. Table 2 shows how often URLs are used in the body of tweets and how many URLs we can find in tweets. Table 3 gives a number of # available within the body of each tweet (except #zika).

Tweet categories are in the first column of Table 4. Most of the categories are self-explanatory. For example, FACTS are tweets, which contain information which is verifiable. This means that in our content analysis the authors read every individual tweet from the pool of 29,000 tweets and checked if its content is verifiable. This includes visiting all URLs from the tweet's body (if they existed), reading all tweets referenced in the body of the tweet using #. If a known and reputable public organization is the owner of the tweet, it was very likely that the tweet would be categorized as FACT. However, many individuals and unknown people also generated tweets, which were easily verifiable.

OPINIONS are tweets which are not verifiable, but still carry information which is relevant to #zika. Some opinions were easy to detect because they clearly express the individual's views or perception of events related to the ZIKA virus.

URL category belonged to tweets, which use only URLs in their content, which is similar to HAS-TAG-ONLY tweet category. Both of these categories might be interpreted as FACTS, because they might be verifiable through URLs and #. However, they rarely contain too much of any other type of text, except # and http://. This is why we wanted to categorize them separately. A detailed explanation of possibilities in manual tweet categorization is given in [11].

Table 1. Collected tweets and retweets

Tweets	16010
Retweets	13819
Total	29829

Table 2. How many collected tweets have URL in their body?

URL Count in each tweet	Num of Tweets	% of Tweets
0	6563	22.62%
1	15809	53.32%
2	6860	23.63%
3	149	0.42%
4	3	0.01%
Total	29829	100.00%

Categories SYMPTOMS, TREATMENTS, PREVENTION and GUIDELINES are self-explanatory.

Table 3. How many tweets have additional #

# Count in each tweet	Num of Tweets	% of Tweets
0	20945	70.22%
1	5264	17.65%
2	2099	7.04%
3	1002	3.36%
4	232	0.78%
5	187	0.63%
6	21	0.07%
7	47	0.16%
8	18	0.06%
9	3	0.01%
10	1	0.01%
11	1	0.01%
12	1	0.00%
Total	29829	100.00%

Table 4. Tweet categories for 18/19 April 2016

Category	Count	% of Tweets
FACT	18708	62.72%
OPINION	6410	21.49%
CAMPAIGN	203	0.68%
USED	0	0.00%
URL	236	0.79%
QUESTION	737	2.47%
PREVENTION	776	2.60%
TREATMENT	168	0.63%
N/A	820	2.75%
GUIDELINES	611	2.05%
HASH TAG ONLY	42	0.14%
DONATION	15	0.05%
IMAGE	42	0.14%
PRODUCT	212	0.71%
PRODUCT- COURSE	60	0.13%
PRODUCT- LECTURE	58	0.20%
PRODUCT PROJECT	0	0.00%
SYMPTOMS	60	0.20%
VIRUS ALERT	671	2.25%
Total	29829	100.00%

Category N/A (not applicable) applies to tweets which are not related to #zika: we could not find explanations why Tweepy selected these tweets. USED categories were tweets where **#zika** has been used to promote businesses (not related to ZIKA) or personal believes. The CAMPAIGN category defined tweets which will campaign for funds, donation or any kind of help offered to communities affected by ZIKA. IMAGE category contained tweets which refer to images, and QUESTION category contained tweets which pose questions about ZIKA.

Some tweets gave us additional information which can not belong to any other category. We named them PRODUCTS. We defined that LECTURES/PROJECTS/SEMINARS were advertised as more specific type of PRODUCT related to #zika.

5. Answering Research Questions through Hadoop

In order to analyze our twitter data through Hadoop, we had two options. The first one was to use FLUME in Hadoop in order to collect live tweets and feed HIVE tables with them. Therefore FLUME and HIVE would be responsible for tweet filtering, i.e. they will replace the role of Tweepy and Python used in the method described in Section 3. Due to numerous discussions on Hadoop forums on the complexity of the procedure of using FLUME, HDFS, Hive and Oozie in Hadoop, it has become obvious that we might not be able to collect the same number and type of tweets in Hadoop as we did through our filtering procedure described in Section 3.

The second option was to

- use the collected tweets from section 3, streamlined into a spreadsheet document,
- create an SQL like table using SELECT command in HIVE in order to feed the spreadsheet document through HDFS into HIVE and
- perform SQL like queries to answer the same questions as in section 4.

This option proved to be the safest way of measuring if Hadoop can win the competition with manual content analyses of live tweets. We were able to confirm that tasks 6 and 7 from the methodology were feasible.

Table 5 gives our own collection of results of HIVE queries, which are related to the categories in Table 4, defined in the manual analysis of the tweets from Section 4. In other words we were trying to see if we can mimic what we did in our manual content analysis in terms of using the same tweet categories for analyzing live Twitter data. However in order to find out to which category each tweet might belong, in Hadoop we have to use keywords which are inputted in SQL-like HIVE queries. We used only two options in this task.

- We could use exact words such as "PREVENTION", and "CAMPAIGN" in SQL queries and SQL would return the number of tweets which contain these words in their body.
- We could use a combination of words such as DRUG, TREATMENT, MEDICINE, FIGHT, VACCINE connected with the OR logical operator in order to get tweets which we then categorize as "talking about treatments".

The last two rows of Table 5 give a number of tweets with a minimum of one URL in the body of a tweet and the number of tweets with at least one extra # in addition to #zika. The results in these two rows can not be compared with similar categories in Table 4 because in these two cases we counted different things. However, the rows are somehow counterparts to Tables 2, 3 from section 4. In Hadoop we were not able to retrieve data which would generate information from Tables 2 and 3.

Table 5: SQL query results for categories of tweets

Potential Category	Count of Tweets	% of Tweets
Prevention	176	0.59%
Treatment	913	3.06%
Symptoms	74	0.24%
Campaign	53	0.17%
Virus alert	518	1.73%
Minimum one URL in the body	16555	55.4%
At least one # except #ZIKA in the body	3487	11.6%

Our preliminary comparison of the results of Hadoop queries, run through HIVE, and counts of tweet categories defined in section 4, have revealed that it was a straight forward task for Hadoop to obtain counts of various parts of the contents of selected tweets, if the SQL like command in HIVE supported it. In other words, the basic of SQL-like queries supported by SQL in HIVE did run smoothly and produced the same results as in our manual analysis of tweets from Section 4.

Table 6: Selection of tweets from UNICEF

Owner of the Tweet (UNICEF)	Count of Tweets
UNICEF Guatemala	2
UNICEF Mexico	2
UNICEF El Salvador	2
UNICEF USA	1
UNICEF Colombia	1
UNICEF Venezuela	2

Therefore we were able to count quickly the number of tweets and retweets, how many tweets mentioned ZIKA treatment, symptoms, diagnosis etc. Hadoop created the same results as in Table 1: total no of tweets is 29829 and there are 13819 retweets amongst them.

Tables 6 and 7 are Hadoop's output which we did not obtain in our manual content analysis, because we did not categorize the owners of tweets and re-tweets, apart from distinguishing individuals from organizations. Table 6 revealed that various branches of UNICEF tweeted (at least once). Table 7 shows that in the forest of various bodies (owners of tweets are NOT individuals) NEWS agencies were leading in their attempts to disseminate information about ZIKA and health organizations are lagging behind.

Table 7: Selection of professional bodies which tweeted

Name of Professional	Count of Tweets
UNICEF	10
NEWS Agencies	1482
Companies with US in their names	11
TRAVEL Agencies or Organizations	34
Organisations with HEALTH in their names	440

It is important to note that in HIVE SQL like queries we used English, Portuguese and Spanish words equally in order to include tweets written in these two languages in our analysis.

We can conclude: it is likely that, to a certain extent, we are able to answer our research questions through either Hadoop or our manual content analysis.

6. Conclusions and Discussions

6.1. Results of Tweet Analysis

The results given in the previous two sections show that our content analysis and SQL queries run through Hadoop give similar answers to our research questions.

If we look at our main goal of the research and ask

> "*if Twitter was used for disseminating relevant information in order to help to manage the spreading of the ZIKA virus across the world*"

then our answer is NO, regardless the way we analyzed tweets. Tweets on symptoms, treatments and guidelines, for managing the ZIKA virus are very rare (their number is significantly small). Furthermore, professional organizations do not tweet sufficiently and it is obvious that majority of tweets are generated by individuals. They are highly present with their FACTS and OPINIONS on the ZIKA virus. Individuals do

contribute towards the dissemination of relevant information on ZIKA and their contribution is as significant as the involvement of professional bodies. However, it is disappointing that professional bodies do not dominate in tweets or retweets. Unfortunately, these are the same results we obtained in 2014/15 when analyzing twitter data related to spread of e-bola in West Afrika.

On the other side, it is good to know that the credibility of tweets is relatively high (FACTS, URLs and #) which is easy to conclude from Table 4, but slightly more difficult to see from the results of Hadoop's queries. Hadoop would require more specific queries to run before we can clearly see the same results. Therefore, there is a good will and attempts amongst individuals and some organizations to give us relevant, true and verifiable information on #zika virus.

What could we conclude from this NO answer?

We think that the lack of tweets issued by professional bodies might be the main culprit for the NO answer. However, it does not mean that the collected data has no value for any Public or Global health organization. The opportunities of querying such a huge pool of information are numerous and any interested party could have learned about "situations" related the ZIKA virus in various locations across the world almost instantly, i.e. as soon as tweets were generated.

Both types of analysis show something unusual. They revealed one of the most striking outcome of our tweet analysis, which we did not expect and which might explain why the answer to our main question is NO. Most of the tweets with #zika hashtag in their bodies are there to express panic, worries, fear and desperation amongst twitter owners in order to alert the whole world to the danger of THE ZIKA virus. This is what we primarily learned from the manual content analysis of 29,000 tweets. We expected higher counts of tweets in all categories in Table 4, BUT simple statements on ZIKA dangers with one URL added to it, issued by disturbed owners of these tweets dominated in our pool of 29,000 tweets. Furthermore, having almost 63% of tweets verified as FACTS and only 21% which were OPINIONs, with an extremely small number of tweets related to the prevention, treatment, and symptoms of and guidelines for managing the spread of the virus, shows that worries of ordinary people dominate in this particular health scare.

6.2. Tweet Categorization versus Hadoop Queries

This paper would be incomplete if we do not address the benefits and drawbacks of our two different ways of analyzing live tweets. In order to find out exactly which approach ((i) or (ii)) proved to be more efficient and promoted to Public and Global health organizations, if they wish to learn from live tweets, we have to find out:

1. If the lack of categorization of tweets, which was a prerequisite for the manual analysis in section 4, may have affected the results we obtained through Hadoop, i.e. is tweet categorization essential in the analysis of tweets?
2. If Hadoop and its SQL-Like query facilities in HIVE will answer all the questions we managed to answer through the Tweet categorization and
3. If Hadoop will offer more results from its queries, which we were not able to obtain in our manual content analyses, This may happen because the tweet categories were either defined in advance or not suitable for this problem domain.

In order to address 1) above, we have to conclude that the manual analysis and predefined categories gave more precise number of answers regarding ZIKA symptoms, treatment, prevention and guidelines. We were able to detect more tweets, which belong to these categories, compared to SQL like queries in Hadoop. In HIVE we relied on pure key word matching which was used in the "LIKE" operator of the HIVE SQL. Therefore the SELECT command in SQL which uses the LIKE operator is case sensitive, reads and compares strings only, and chooses the content of HIVE tables in the results on the basis of exact word matching. The manual analysis was naturally more precise and could include tweets where key word matching was not essential for tweet categorization. Furthermore, we were also not able to detect through Hadoop tweets, which tweets might be opinions. It was impossible to find words, which may appear in tweets of this category, which we could associate with opinions and which can be used in SQL like queries for word matching.

In order to comment on 2) we must outline that it was impossible to run queries in HIVE which needed joining an SQL table with itself, without SQL query optimization. It took too long to run them on 26,000 tweets (we could not wait to see the results). HIVE does not support the full SQL standard (which is expected) and some queries simply will not run in Hadoop. This particularly applies to questions where we wanted to know "*how many tweets contain more than 3 URLs or more than 4 # in their body*". We could probably be able to tweak this deficiency in HIVE by either using a different component of Hadoop for queries or interfering with the automatically generated code by Hadoop's components. In both cases, the time did not allow us to experiment further with Hadoop and we had to accept that SQL in HIVE has its deficiencies.

In order to comment on 3) we have to emphasize that SQL in HIVE gave us a very fast and efficient option of counting ANYTHING we were able to store in Hadoop's HDFS. The idea of using MapReduce and

perform counting of various "words" which may appear in Tweets proved to be extremely valuable [21]. We did not have this opportunity in our manual content analysis of tweets. For example:

- we were able to see, in the results of HIVE queries, that there are a number of professional organizations which belong to UNICEF across the world, which tweeted on the ZIKA virus (Table 6). We were not able to detect this solely through our categorization of tweets.
- SQL was able to search for words and their counts if they appear in the name/part of the name of the twitter owner, which we were not able to detect in our categorization of tweets (Table 7).

6.3. Lesson Learned

The important result of this research is that we were not able to have a clear comparison between (i) and (ii).

Firstly, results in Sections 4 and 5 show how the analysis of the same live tweets may look different. We could not compare the tables in the previous two sections as like-to-like for many reasons:

- If we could not find a particular word, which could be used in SQL like queries in Hadoop in order to see tweet categories then these tweets will not be discovered by Hadoop;
- Some queries were extremely simple to write in HIVE, but at the same time, their results are very difficult to find in our content analysis;
- If we need to investigate the presence of particular sentences (not a single word!) within tweets, then SQL like queries will not help and different technologies must be used.

Therefore all these Tables 1-7 compliment each other. The way we created the results stored within the tables was often dictated by our domain of interest and the type of questions we wish to ask. This is not unexpected: the method of Big Data analysis always depends on questions we expect to be answered through the analysis.

Secondly we used a process for filtering relevant #zika tweets described in Section 3 because it is a prerequisite for our manual content analysis. However, filtering of tweets is not required by the Hadoop platform. There are other ways of feeding Hadoop's HDFS with relevant live tweets, as mentioned in Section 5 (through FLUME components). Our tweet selection has guaranteed that we collected relevant tweets and therefore we did not have to worry if our data given to Hadoop or used in our content analysis was not "clean". However, we could not assess if FLUME would be equally efficient. We conducted no experiments with FLUME because of the complexity of the prescribed process [36].

Thirdly, we have to bear in mind that the success of Hadoop and its components lies in their extremely efficient COUNTING of enormous amount of data, which are important in finding and understanding their role within a Big Data pool. Therefore these counts we receive through efficient query like commands should be sufficient to answer questions we may have. We have to accept that we sometimes might not get all our questions answered because "counts" might not be sufficient to *make sense* of Twitter data and we would need a new approach to the analysis. T summarize: answers to questions we may have using Hadoop are based on counting. In our manual content analysis our categorization of tweets carries the semantics we need for answering the same questions.

6.4. Recommendations

It is difficult to recommend and favor either of these two approaches (i) and (ii) to the live tweet analysis without thinking about issues raised in the previous section. Therefore each organization which is willing to analyze live twitter data at the time it is generated, should be aware of the benefits and drawbacks of each of them. This particularly applies to the use of Big Data technology, without assessing if it does bring benefits at the moment when it is used. The Hadoop platform is an extremely complex and rich set of interconnected software components, which require significant skills in order to use them properly.

One piece of advice is obvious: if Public and Global health organizations need more precision in twitter data analysis, then manual content analysis of tweets has no competitors. A clear picture of the content of the selected 29,000 tweets can not be obtained by running SQL queries in Hadoop. We might not be able to formulate an SQL query for all possible research questions we may have. Furthermore:

- No Big data platform could verify the content of each tweet with the same precision as in a manual content analysis.
- The research questions we formulate directly influence the categorization of tweets in the content analysis and therefore it is unlikely that we will not be able to answer them.

However, Hadoop offers something important which was not feasible to perform in manual content analysis: the advantage of FAST counting of the occurrences of any important word(s) in live tweets, which would give a different insight on the content of twitter data. In such cases, do we really need the precision secured by a manual content analysis? Precision does not always play an important role in the analysis of user generated data and might not be always essential when deploying Big Data analysis [36].

It is also interesting to note that the time needed to obtain results shown in in section 4 (manual content analysis of 29,000+ tweets) was much shorter than the time we needed for managing our questions through Hadoop's queries. This is not a criticism of the technology. Solutions offered for managing Big Data are still in their infancy and we hope that obstacles we faced when using Hadoop will be removed soon

Finally, the deployment of our experiments in real life should take into account that our research team is interdisciplinary. It consists of researchers, students and professionals from different disciplines, which proved to be essential in choosing and using analytical tools and Big Data technologies, interpreting twitter data and answering the questions we may have. Therefore it is assumed that anyone interested in *making sense* of twitter data would deploy inter-disciplinary teams when managing excessive amount of live data through Big Data technologies.

We could not find published papers, which could be compared to or support the process and results of our live tweet analyses. At the time of writing, there were no publications, which investigate the suitability of Hadoop's platform in cases of health scare. Manual content analyses of live twitter data are extremely rare and often not practical: they are time consuming and difficult to perform. Therefore we could not find related work which could fit this paper.

However, we might trigger a few discussion points:

- Should Public Health organizations use Big Data technology in the analysis of live data in decision making, in spite of the complexity of using the technology on an ad-hoc basis, i.e. when a public health scare appears? Big Data technology is deployed on complex platforms and it requires significant expertise and resources to process data, which makes them difficult to use in general.
- How much precision do we need in the analysis of enormous amounts of data in healthcare? If we can process millions of live data in a very short period, would this be more important, than having a relatively small number of data processed with a very high level of precision? What do we trade-off by performing either of these two options?
- Which messages might Public Health organizations pass to software developers involved in the creation of languages, storage systems and retrievals techniques, which dominate in Big Data platforms? Which type of big data analysis is needed?

We are currently testing the power of Hadoop's FLUME in tweet filtering, i.e. we are replacing steps 1-4 of our methodology with the automated selection and analysis of tweets through Hadoop.

10. References

[1] The Zika Virus, TIME, May 16, 2016, pp. 22-21.

[2] D. Dumbrell, and R. Steele, "#wordlhealthday 2014: The Anatomy of a Global Public Health Twitter Campaign", In Proc. of the 48th Hawaii International Conference on System Science, HI, USA, January 2015, pp. 3094-3103.

[3] J. Hou, G. Xiong, D. Fan, and T.R. Nyberg, "Modeling and analysis of information dissemination mechanism of social media", In Proc. of the 2012 Int. Conference on Service Operations and Logistics, Suzhou, China, 2012, pp.377- 382.

[4] S. E. Halse, A. T., A. Squicciarini, and C. Caragea, "Tweet Factors Influencing Trust and Usefulness during Both Man-Made and Natural Disasters", In Proc. of the ISCRAM 2016 Conference, Rio De Janeiro, Brazil, 2016.

[5] C. Chew and G. Eysenbach, "Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak", PLOS ONE. DOI: 10.1371/journal.pone.0014118. 2010, Available from http://www.ncbi.nlm.nih.gov/pubmed/21124761

[6] Y. Xie, Z. Chen, Y. Cheng, K. Zhang, A. Agrawal, W. Liao, A. Choudhary, "Detecting and Tracking Disease Outbreaks by Mining Social Media Data", In Proc. of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, 2014, pp. 2958-2960.

[7] WHO, "Globalization and infectious diseases: A review of the linkage", 2014. Available from http://www.who.int/tdr/publications/documents/seb_topic3.pdf

[8] WHO, "Investing to Overcome the Global Impact Of Neglected Tropical Diseases", WHO Report, 2015, available from http://apps.who.int/iris/bitstream/10665/152781/1/9789241564861_eng.pdf

[9] R. McKie, "Zika virus could be bigger global health threat than Ebola, say health experts", In Observer, 30 January 2016, Available from https://www.theguardian.com/world/2016/jan/30/zika-virus-health-fears

[10] S.Tavernise and D.G. McNeil, "Zika Virus a Global Health Emergency, W.H.O. Says", New York Times, 1 February 2016, Available from http://www.nytimes.com/2016/02/02/health/zika-virus-world-health-organization.html?_r=0

[11] C. Everiss, J. Feny, and R. Juric, "Ebola Crisis: An Investigation Into Levels of Communication Following Vaccination". In Proc. of the 20th International Conference on System Design and Process Science, Texas, USA, 2015, pp.474-483.

[12] R. Juric and I. Kim, "Can Twitter Transform Communities Affected by e-bola", In Proc. of the 20th International Conference on System Design and Process Science, Texas, USA, 2015, pp. 506-511.

[13] K. Pettai, R. Juric, and A. A. A. Bechina, "The Power of Microblogging in Disseminating Information in Humanitarian Crises: A Study of Nepalese Earthquake", In Proc. of the 20th Int. Conference on System Design and Process Science, Texas, USA, 2015. pp.184-191.

[14] MJ. Paul and M. Dredze, "You Are What You Tweet: Analyzing Twitter for Public Health", In the Proc. of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 2012, pp.265-272.

[15] G. Armour, "Communities Communicating with Formal and Informal Systems: Being More Resilient in Times of Need", Bulletin of the American Society for Information Science and Technology, 36(5), 2010, pp 34-38.

[16] D. Godfrey, C. Johns, C. Sadek, C. Meyer, and S. Race "A Case Study in Text Mining: Interpreting Twitter Data from World Cup Tweets", University Report, 2010, Available at http://meyer.math.ncsu.edu/Meyer/PS_Files/CaseStudyInTextMining.pdf

[17] G. Kilpatrick, "10 Awesome Twitter Analytics and Visualization Tools", In Twitter Tips and Tools, 13 June 2015, Available from http://twittertoolsbook.com/10-awesome-twitter-analytics-visualization-tools/

[18] C.X. Lin, B. Zhao, Q. Mei, and J. Han, "PET: a statistical model for popular events tracking in social communities", In Proc. of the 16th ACM SIGKDD, Washington, DC, USA, 2010, pp. 929-938.

[19] D. Chakrabarti and K. Punera, "Event summarization using tweets", In Proc. of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 2011, pp.66-73.

[20] F. Liu, Y. Liu, and F. Weng, "Why is "sxsw" trending? Exploring multiple text sources for twitter topic summarization", In Proc. of the Workshop on Language in Social Media, Portland, Oregon, USA, 2011, pp.66-75.

[21] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet Allocation", The Journal of Machine Learning Research 3, 2003, pp.993-1022.

[22] D. Laniado and P. Mika, "Making Sense of Twitter Data", ISWC 2010, Volume 6496 of the series Lecture Notes in Computer Science, 2010, pp 470-485

[23] S. Tan, Y. Li, H. Sun, Z. Guan, X. Yan, J. Bu, C. Chen, X. He, "Interpreting the Public Sentiment Variations on Twitter, In IEEE Transactions on Knowledge and Data Engineering, 6(1), September 2012, pp. 1-14.

[24] V. N Gudivada, R. Baeza-Yates, V. V. Raghavan, "Big Data: Promises and Problems", Computer, The IEEE Computer Society, March 2015, pp.20-23.

[25] S. Kaisler, F. Armour, J. A. Espinosa, W. Money, "Big Data Issues and Challenges Moving Forward", In the Proc. of the 46th Hawaii International Conference on System Science, HI, USA, January 2012, pp.995-1004.

[26] M. N. Manu and K. R. Anandakumar, "Current Trends in Big Data Landscape", In 2015 IEEE Int. Conference on Computational Intelligence and Computing Research (ICCIC), December 2015.

[27] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics, International Journal of information Management, 35(2), April 2015, pp.137-144.

[28] HADOOP, Available from http://hadoop.apache.org/

[29] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial", IEEE Access, Volume 2, 2014, pp. 652-687.

[30] Y. Huang, X. Lan, X. Chen, and W. Guo, "Towards model based Approach to Hadoop Deployment and Configuration", In the Proc. of the 12th Web Information System and Application Conference, January 2015.

[31] T. B. Murdoch, A. S. Detsky, "The Inevitable Application of Big Data to Health Care", JAMA, 309 (13), 2013, pp. 1351-1352.

[32] H. You and D. Wang, "Research and Implementation of Massive Health Care Data Management and Analysis Based on Hadoop", In Proc. of the 4th Int. Conf. on Computational and Information Sciences, 2012, pp. 514-517.

[33] S. Thakur and M. Ramzan, "A systematic review on cardiovascular diseases using big-data by Hadoop", In Proc. of the 6th Int. Conf. on Cloud Systems and Big Data Engineering (Confluence), January 2016, pp. 351-355.

[34] R. Ranjan and R. Misra "Epidemic disease propagation detection algorithm using MapReduce for realistic social contact networks", In proceedings of IEEE International Conference on High Performance Computing and Applications (ICHPCA), December 2014, pp. 1-6.

[35] M. A. Cifci, D. C. Ertugrul, and A. Elci, "A Search Service for Food Consumption Mobile Applications via Hadoop and MapReduce Technology, In Proceedings of the IEEE 40th annual Computer Software and Application Conference (COMPSAC), Volume 2, June 2016, pp. 77-82.

[36] A. B. Patel, M. Birla, and U. Nair, "Addressing Big Data Problem Using Hadoop and MapReduce", In NIRMA University International Conference on Engineering (NUiCONE), December 2012.