SENTIMENT ANALYSIS ON UP-RELATED COMMUNITIES IN FACEBOOK AND TWITTER USING HADOOP

RAHMAT PETER I. DABALOS

May 2016

BIOGRAPHICAL SKETCH

Born on June 20, 1995 in Davao City, Philippines, Rahmat Peter I. Dabalos is the fourth child of Marianito Dabalos and Ma Lourdes Dabalos. Rahmat's siblings are Marrion Dabalos, Kathrina Lua Khozein, and Zarrin Rose Dabalos. Rahmat together with Marianito, Ma Lourdes, Marrion, and Zarrin is currently living in San Pedro, Laguna.

He finished primary and secondary Education on Casa del Niño Montessori and Science High School where he finished as the top 3 of his class. Currently, Rahmat is in his fourth year in University of the Philippines Los Baños taking up Bachelor of Science in Computer Science. During his free time, he usually plays sports or watch TV series.

ACKNOWLEDGEMENT

To the researcher's family, who has always been there for him, supporting his every up and down, he loves you all and he will always look up to each and every one of you. To his parents, who supports him emotionally and financially, now is the time that he gives back to you. To his siblings; Ate Kat with her husband, Rahan, Kuya Marrion, and Ate Zarrin – he gives his gratitude for helping him prepare for all the obstacles there would be on the way. He is who he is because of the culture in the household, he will always be thankful for being a part of the family.

The researcher is taking this opportunity to thank his SP Adviser, Prof. Lailanie Danila, for continuously giving him the challenge of enhancing his work and guiding him every step of the way. To all his college professors, who saw the potential in him and have continuously helped in developing the skills that he needs in order to complete his academic endeavors, he thanks you and hopes to be able to impart knowledge himself to the future generations of Computer Science students. To his sp supervisor, Joseph Anthony Hermocilla thank you for opening new ideas to him, without his insights on the researcher's topic, he wouldn't be able to comply all that is needed for his SP to be a success. To Professor William Remollo, he sends his deepest gratitude, for without his guidance on paper writing, the researcher wouldn't acquire the skills that he needs.

To acquaintance, friends, colleagues, classmates, blocmates, and teammates, he also gives thank for letting him experience a relationship with different kinds of people. You have all prepared him in becoming ready to face on the challenges of what we call the real world. To Ms. Nikki Mae Dee, who has never given up on the researcher, to whom the researcher gets motivation and encouragement, the researcher is most gratified to have her for the researcher would have not completed the task without her unwavering support and care.

This acknowledgement wouldn't be finish without him sending his most heart-felt and sincere gratitude to the institution of which he has acquired all the skills that he has. To the Institution of Computer Science, and to the University of the Philippines – Los Banos, keep on inspiring people with the burning passion to keep on excelling in the various fields of Computer Science.

Lastly, he would like to thank the Lord, from whom he asks for guidance and help when all seems to be hopeless. This is all for the people who believed in him. .

RAHMAT PETER I. DABALOS

ABSTRACT

The rapid increase of the data present in the world can be attributed to the advent of social media, wherein, discourses almost about everything ranging from personal to societal issues are being tackled. There comes the problem of handling and making sense of Big Data. To be able to solve this problem, huge amounts of processing power must be utilized, this is possible through the use of the Cloud Computing as the platform for such studies. Sentiment Analysis is the study of the sentiments, opinions, and moods expressed in written language. It is inside of the wide spectrum of Natural Language Processing, the field in which creation of systems and tools are done to recognize patterns and relationships within written text as it is cumbersome to do it manually. With that, analysis, classifications, and predictions of the meanings in written language is possible with little human involvement on huge amounts of data present in social media.

LIST OF TABLES AND FIGURES

Figures:

Tables:

Appendix:

SENTIMENT ANALYSIS ON UP-RELATED COMMUNITIES IN FACEBOOK AND TWITTER USING HADOOP

INTRODUCTION

Background of the Study

As of today, the world is seeing the rise of technology. We are fortunate to witness the coming of the digital age in the world. Because of this, human life is now becoming dependent on the functionality of modern technology. There is no reason to doubt that the data humans have stored reached a ridiculously large amount of number. According to an article by Lucas Mearian, Humankind has stored more than 295 billion gigabytes (or 295 exabytes) of data since 1986 [1]. To add to that, this number is not constant, this is rising as we speak in an exponential rate. With that in mind, computer scientists will have a hard time dealing with this magnitude of data. In this paper, the venue used for gathering of data would be the social networking sites such as Facebook and Twitter, concentrated on the sites and pages used by UP students in expressing their sentiments. Facebook pages such as, Narinig ko sa UP, Elbi Files, Diliman Files, just to mention some. These pages contain a huge volume of data that are varied and valuable making it a big data problem. Cloud computing is a way to get more computing power when it is needed. Cloud computing makes use of several computers to provide a service as supposed to providing it using only one computer. These computers are allocated equally according to the resources that are required. Processing huge amounts of data can approximately take k x n faster to finish compared to processing it in a single computer,

where k is the number of computers to be used. It allows scalability and parallel computing, it can easily handle a growing amount of work just by using more computers to process it [2].This kind of platform will be very useful for a big data problem. MapReduce is a programming concept that deals with processing data over a distributed and parallel cluster or cloud computing. This is a concept that deals with the two basic functions on the large data that will be stored in the distributed file system, Map() and Reduce(). Map(), which takes the input data and also run by parallel computers, will process it and produce a list of key value pairs, which is used by the Reduce() function to integrate the key value pairs and produce the result. This kind of programming paradigm is supported by Hadoop, which is an open source framework that allows the processing of big data by using Hadoop Distributed File System (HDFs) and the MapReduce Function. [3]. Sentiment analysis or opinion mining is an area of study in the field of Natural Language Processing. It is directly involved in extracting the mood or opinion of the subjective elements found on a text. This is done by classifying the text according to the opinions expressed in it, whether it is neutral, negative, or positive [4].This technique is actually used by companies in order to know their customer's feedback and suggestions. This technique will be used to process the data that will be gathered from the pages in Facebook and posts in Twitter in order to come up with the sentiments of the students regarding issues that affect them.

Statement of the Problem

According to Facebook in their September 2015 report , their daily active users is 894 million on the average [5]. In this 894 million, the majority of these people are

studying in the university. Which is only logical for the reason that students have easy access to the Internet since, nowadays it is the largest source of information. And it is where students usually spend their time when not in class. There are already studies that are present that is dealing with analytics in social media. But this research's focus will be limited. The goal of this research is to find a consensus among UP students regarding issues they face. Surveying them would be a very tedious task since there are at least 50,000 students studying in the university. so this paper is proposing to develop a tool for gathering the data from Facebook and Twitter and apply sentiment analysis on it on a Hadoop cluster.

Objectives

The main objectives of this study were the following:

- To utilize social media, specifically Facebook and Twitter, as a source of data for sentiment analysis.
- To provide a tool which will make the processing of data in social media effectively and efficiently through the use of Hadoop, and
- To examine, analyze, and report the sentiments of University of the Philippines students to issues that they face in the university or in the country.

Significance of the Study

The tool that was developed was used to examine, understand, and interpret the remarks of the students of the University of the Philippines in issues that matter to them.

Which in turn can be a good source for examining the behavior, awareness, and the way these students think of issues that is of importance to each individual and to the society.The result of the tool can be used in many ways, the staff of the university will be aware of the sentiments of the students at the current time and know what are the needs of the students and how to fulfill them.

Scopes and Limitations

This study produced a tool that will only be used to analyze the posts and tweets of UP students in the community pages in Facebook and Twitter on top of a Hadoop Cluster.HDFS, Flume, Hive and MapReduce were the only modules that was used from the Hadoop framework. The results from running this tool are available to everyone and not only limited to students of UP. With this, the people can easily see the general mood of the UP community with regards to certain issues.

REVIEW OF RELATED LITERATURE

As of today, everything is now being digitized, which means that the data that we have are getting larger. This will cause a problem since the tools and techniques that we use are not designed in handling massive amounts of data. Most of the time, these data are very hard to manipulate and manage. This is the reason for the rise of Big Data in the industry. Being able to handle big data will change everything from human activities up to the industry such as the sciences, government, financing, entertainment, leisure, and many more. The definition of Big Data comes from its characteristics: the Three Vs namely- Velocity, Variety, and Volume. As of now, there is an ongoing debate if data requires all of these three characteristics or is it enough for data to have one characteristic to be classified as Big Data. Velocity is defined as how fast the data is generated. In social media sites such as Facebook, velocity can clearly be seen because a person can post anytime he/she wants. There are millions of pages and users in Facebook, with that capability to post at anytime, data generated from Facebook in form of posts comes in very fast. Variety refers to the differences in the records, which means that the data gathered are in different kinds or values. Another example can be seen in Facebook, looking at the interactions in Facebook, the posts, likes, and comments generate datasets that are different from each other. Lastly, Volume is the characteristic that is associated with the size of the data that is generated .It deals with how the large data sets are stored and processed. When dealing with big data it is usually in terabytes or petabytes [6]. Data Mining on the other hand, is the process in which we model data. One model that would be relevant to this project is Machine Learning. Machine Learning is in the field of

Artificial Intelligence that aims to make sense of a given data. “According to Leskovec, Rajraman, and Ullman: Machine learning practitioners use the data as a training set, to train an algorithm of one of the many types used by machine-learning practitioners, such as Bayes nets, support-vector machines, decision trees, hidden Markov models, and many others” [7]. Machine Learning is done by training a data set in order to recognize the pattern from already processed data that can be used to predict the outcome of the next inputs. From a raw data, by using Machine Learning, we will be able to classify that given data and predict future events. When dealing with Big Data, it is very important that the supporting platform where it is ran is scalable and supports parallel computing to ensure success in data analysis. Cloud computing is the infrastructure that can support it together with Hadoop, a class of distributed data-processing platforms. Furthermore, cloud computing is also effective in addressing the storage needed for performing big data analysis [8].A single piece of hardware even if high-end will always be outperformed by several dedicated hardware running in parallel. There is an ongoing rise on the data that are generated on social media sites, it can be explained by the increase in human interactions in the internet, and because of that many people are getting interesed in doing researches dealing with Big Data analysis. Big data is not only classified according to its size, but also with the relationship of the data with other data. ”A very important property within social media is the connectedness of entities. There are hidden structures as a result of purposes behind actors, individual psychological states, their comments, interactions such as conversations, and shared semantics” [9]. A data funneling process is done by Eugene Ch’ng for the purpose of managing the data through the use of scalable open source architecture for gathering data from Twitter that is

meaningful for social science researches [9]. Sentiment analysis is one of the most researched topic in the field of Natural Language processing. And because of it, there already exists numerous approaches to extract the emotion being expressed by using the subjective elements found in the text. The two most common approaches are machine learning and lexicon based. Where machine learning approaches are reliant on the selection and extraction of the appropriate set of features in order to detect sentiments. Lexicon-based approaches relies on the sentiment of each lexicon by comparing it with a collection of already known phrases, terms and idioms develop the meaning [10]. There are already existing applications that are used in order to conduct sentiment analysis on social media sites. An example would be Sentbuk, or SENTimental FaceBUK, it is an app developed for Facebook users. It performs sentiment analysis on users walls by classifying each of the sentences first as positive, neutral, or negative and then the overall sentiment will be calculated."The proposed classifier follows a lexicon-based approach, using a dictionary of words annotated with their semantic orientation (positive/negative emotional polarity) and detects additional language, features such as positive interjections(i.e laughs), negative interjections, emoticons, misspells, part of speech tagging or negation(polarity shifter)" [11]. What is great about the paper is that the tool that was used was very accurate in classifying the polarity of the text with 96% success rate. In order to measure the accuracy, the tool was ran and a human manually classified the texts without knowing the results of the classifier [11].

METHODOLOGY

Theoretical Framework

1) Data Source: Data that were collected from:

    A. Facebook

        - Narinig ko sa UP
        - The Elbi Files
        - The Diliman Files
        - University of The Philippines Los Baños
        - University of The Philippines

    B. Twitter: the '#' was used in order to search for tweets that are related to the topic of interest. The Following hashtags were searched:

        - UPsaHalalan
        - UP Halalan
        - UPLB
        - PoeVisitsUPLB
        - AyokongMagmahal
        - Duterte UPLB
        - DuterteVisitsUPLB
        - FacultyCenter
        - MDSUPLB
        - MiriamUtakAtPuso
        - Stephen Villena

- UPFight
- UPLBWalkout
- AbuSayyaf
- BigasHindiBala
- Duterte
- Pilipinas2016
- PilipinasDebates2016
- Presidentiabledebates
- PhilippinesElections2016
- Philippines Elections
- Philippines Elections 2016
- NoToMarcos
- Halalan2016
- PhilippinesElections
- PHVote

2) Data Collection: In order to collect the data from the given sources the following tools were used:

A. RFacebook- A package in R that implements Graph API was used to gather data from public pages and groups in Facebook.

B. Twitter- A package in R that was used to gather data from Twitter through searching for keywords.

C. Apache Flume- Apache Flume is a tool that ingests data for collecting, aggregating and transporting large amounts of streaming data from various servers, in this case from Twitter, to HDFS.

D. Apache Hive - In order to make sense of the data gathered by flume, Hive was used as it supports reading, writing, and managing data stored in a distributed storage such as HDFS through SQL. The architecture of the HDFS, Flume, and Hive is seen in Figure 2. Twitter API acted as the server in which tweets were coming from and then Flume receives it and dumps it into the HDFS while Hive was used for querying.

3) Training: Before processing the tweets, each word in the dictionary that has subjective meaning were classified as positive or negative. A sample data set from this source: https:www.cs.uic.edu/~liubFBSsentiment-analysis.html#lexicon was used for better results. As it is necessary to have Filipino words, the file retrieved from the source was translated in translate.google.com to Filipino, and added to the initial dataset which only contained English words. And for the dataset for the Naive Bayes' classifier, 11,020 tweets and posts that were close to the subject of the data were picked via crowdsourcing in order to prevent a biased dataset, and another data set, which was used in Mr. Regalado's research, was added [13].

4) Pre-processing: Because the data that were gathered were surely composed of many unnecessary data, at this stage, the dataset was filtered so that, only the relevant data was used in the next processes. A script was implemented to remove irrelevant columns and trailing spaces, so that the data were clean. The data extracted on each post or tweet were: id, username, and text.

5) Classifying: The data gathered were classified by two different algorithms using the same input data, and both were also implemented through the MapReduce function, which is unique in Hadoop. Map divides the tasks into smaller tasks and when all of the Map tasks are done, Reduce will combine the results.
    - Dictionary-based: For this algorithm, every word in a post or tweet will be checked if it is in the negative or positive dictionary. If it is in the negative dictionary, the word will be scored -1, 1 if positive, while 0 if it is not in both of the dictionary. Afterwards, the score of each word in the post or tweet was added to get the overall sentiment of the data. The implementation of the Dictionary Based classifier as MapReduce can be seen in Figure 3.
    - Naive Bayes: In order to compute the sentiment of each of the post or tweet in this second classifier, Naive Bayes with Laplace smoothing was used. The difference between this and the former is in the way it gives score to the whole post or tweet. The implementation of the Naive Bayes Classifier on MapReduce can be seen in Figure 4.
6) Storage: The storage for the data that were processed is the Hadoop Distributed File System(HDFS) in a Multi-node cluster in Peak Two Cloud.

Development Tools

1) Peak Two Cloud(P2C): Big Data problems requires a platform that can support parallel and distributed computing. The Peak Two cloud provided the infrastructure needed in order to create a Multi-node Hadoop cluster. Four virtual

instances of ubuntu-14.04-server-amd64 with 512mb ram were used from the cloud, one will be the Master Node and three others were slave nodes.The creation of each instance is discussed in this article: http://srg.ics.uplb.edu.ph/projects/peaktwo-cloud/peak-two-cloud-resources/user-guide

2) Apache Hadoop: Hadoop is an open-source framework that was used to store the data via the Hadoop Distributed File System, and contains the libraries that were needed to create the Map-Reduce function to be ran on a Multi-node cluster.
    i. Create 4 instances from Peak-Two Cloud
    ii. Download hadoop-src in https://hadoop.apache.org/releases.html
    iii. Extract the tar.gz file and copy it to /usr/local/hadoop/
    iv. Configure hadoop-env.sh: JAVA HOME should be set
    v. Each of instance's IP address should be set to each nodes' /etc/hosts file
3) Graph API: Graph is an Application Programming Interface developed by Facebook that allows developers to get data from Facebook such as, posts, comments, photos, etc through an access token that has an expiration time.In order to have an interface for Facebook's Graph Api and to be able to gather data from Facebook, the package 'RFacebook' in R was used.
4) Twitter API: Twitter has an Application Programming Interface used to gather data from Twitter.The interface used for it is 'Twitter', a package in R.
5) Apache Flume: Apache Flume was used collect, aggregate, and ingest large amounts of data from Twitter and store it directly to HDFS in a distributed manner [14].

6) Apache Hive: Apache Hive was used to read, write, and manage the data gathered by Flume through SQL [15].In this study, hive wass used to structure the data gathered by flume, as it was initially in JSON format. After putting it into a table in Hive, it was dumped as a file wherein the only columns left were id, username, and text. This was faster than any other method as it supports distributed clusters. The query for dumping can be seen in Figure 4. The result of this SQL query was a text file composing of tweets with unique text, screen name and id. 'regex replace' was used to remove carriage return and line breaks in the text column.

Non-Functional Requirements

1) Performance Requirements: The status of the system was monitored, system crash, hang, or error are detected, and the performance according to the efficiency and integrity of the system was made available by Hadoop's built in web app hosted in port 50070 of the masternode.
2) Safety Requirements: This is in terms of the data, the storage was always backed up in case of failures in the Hadoop cluster, it is fail safe since the architecture's main goal is to avoid this from happening.
3) Scalability Requirements: Because this is a Big Data problem, the tool should be able to handle increasing amounts of data and still process it efficiently, further exploration of the configurations of Hadoop properties unlocked Hadoop's built in capability to expand.

RESULTS AND DISCUSSIONS

With regards to the platform that was used, Hadoop Distributed File System, the final specifications can be seen in Figure 6.

The capacity or the storage of the whole cluster is 58.96 GB, this was the result of allocation of space from the 21GB storage of each instance to the cluster. Another thing to look at is the Number of Total Datanode Volume Failures, which is 0B. It means that there was no recent failure in the system.

The first objective of this project was to utilize Facebook and Twitter as a source of data for Sentiment Analysis. The first table shows the total number of tweets and post gathered on different types of implementation.

The total number, 89021, was a substantial amount because the main focus is sentiment analysis, and it is limited only to topics that are relevant to students studying in the University of the Philippines. This number is already sufficient for analysis since patterns and relationships hidden in big data can already be found without having to go through all of it. According to Michael Berry, Patterns and relationships reveal quickly and adding more data would not drastically change the result [16]. Out of the total amount of data, 65% was used as input data. After running the two types of classifiers, Dictionary based and Naive Bayes, on the gathered data, the results is in Table 2.

Some examples of the actual results of the classifiers are listed in the appendix to be able to visualize how it works. Sample tweets and posts are in the first column, and the result of the Dictionary Based and Naive Bayes in the next columns. Examples provided show that there were a number of correctly classified messages for both of the classifiers, instances also when only one was correct, and when there was no correct of the two.

The number of negatively classified tweets and posts were in majority as compared to positive and neutral in both of the classifiers - 45% in Dictionary Based while 64.6% in Naive Bayes.

Through this agreement between the two classifiers, it can be said that the general mood of the people when it comes to UP-related topics and issues were leaning towards the negative. Furthermore, this means that within the discourse of the topics gathered, many were negatively affected and used Twitter and Facebook as avenues for them to vent out their frustrations and dissatisfaction.

The next set of results, as seen in Table 3, was representative of the data in its middle or average value.

The mean, -1.7061, for Dictionary based and 0.6184 for Naive Bayes, were the average values of the results of the tests. The typical score of each post or tweet was close to those numbers and predictions can be made that most of the tweets or post were close to those numbers. The means for the two classifiers' mean, which were in negative interval, can explain as to why there were more negatively classified tweets than positive.

The median, 0, and 0.7050 was the value at which half of the results were greater than or less than that number.

The mode, which is 0 and 0.6442, were the value that had the most number of frequencies in the results.

The standard deviation which explains the average distance of the values to the mean can explain the distribution of the results. Compared to the Naive Bayes classifier, the standard deviation of 5.0517 for the Dictionary Based was not high since the result of each was not restricted to a range of values. While on the Naive Bayes, where each result

is restricted to the interval 0 to 1, the standard deviation of 0.3409 can be interpreted as high. Meaning, the results in the first classifier averagely were close to the mean, and on the second, the average values of the results were far to the mean. To know more about the distribution of the data, look at the plot in Figure 7.

Looking at the distribution of the 1st classifier, it is evident that the standard deviation was small hence, the steep bell curve. What it means is that most of the results gathered have results that were close with each other. The outliers, which were mostly negative explains that there were tweets and posts that were very negative in nature but were not too many.

The distribution in the second classifier support the claim that the difference between the scores of tweets and posts were huge since the bell curve is flat, meaning the results were spread out. The curve which was also skewed to the right gives explanation to the majority of negative results.

In order to have a better visualization of the results, word clouds seen in figures 9 to 12 were to show the most frequently used words in the results of the classifications.

'Love, 'gusto', 'best', were the most used words in the tweets and posts that the Dictionary Based classifier has classified as positive while 'hindi, 'di, 'tanga' were for the negatively classified.

With these results, the effectivity of the classifiers to classify that the tweets and posts into positive and negative can be qualified due to the fact that the words that were widely used in the classes were words that really imply that kind of sentiment.

With regards to the accuracy of both the classifier, a survey composed of 100 randomly sampled tweets and posts were checked by 3 individuals. This was taken from

95% confidence interval with 9.79% margin of error. This was conducted before running the classifier and the agreement among the 3 people within each of the item in the sample were used as the correct sentiment for each tweet or post. Tables 4 and 5 are the confusion matrices of the two classifiers.

Precision is the proportion of messages correctly identified as belonging to a certain class(Positive/Negative/Neutral) among all cases of which the classifier claims that they belong to that class. While, Recall is the proportion of cases correctly identified as belonging to a class among all cases that truly belong to that class .Furthermore, The first classifier, which was the Dictionary Based, had better precision and recall in classifying the tweets and posts as negative or positive compared to the Naive Bayes. And the accuracy, which was the measure of how correct the classifier was, the Dictionary Based classifier proved to be the better of the two.

CONCLUSION AND FUTURE WORK

This research was able to conduct a Sentiment Analysis on data gathered from Facebook and Twitter through the use of Hadoop as the File System, in which two classifiers were created; namely Dictionary Based and Naive Bayes. Both of the classifiers implemented MapReduce, a programming model associated for processing and generating large amounts of data. Although the Dictionary Based classifier was more accurate than the Naive Bayes, the two classifiers yielded similar results which showed a higher percentage of negatively classified tweets and posts.

For future works, it would be worth exploring on the topic of creating a more balanced dataset containing Filipino and English words that will be able produce more accurate results. Using other tools and techniques used in Natural Language Processing is also suggested as the problem of Sentiment Analysis is a complicated field.

REFERENCES

[1] L. Mearian, “Scientists calculate total data stored to date: 295+ exabytes,” Feb. 2011. [Online]. Available: http://www.computerworld.com/article/2513110/datacenter/scientists-calculate-total-data-stored-to-date–295–exabytes.html

[2] G. E. (2015) What is cloud computing? [Online]. Available: http://asia.pcmag.com/networking-communications-softwareproducts/2919/feature/what-is-cloud-computing

[3] R. M. (2015, March) Map reduce definition. [Online]. Available: http://searchcloudcomputing.techtarget.com/definition/MapReduce

[4] C. Bhadane, H. Dalal, and H. Doshi, “Sentiment Analysis: Measuring Opinions,” Procedia Computer Science, vol. 45, pp. 808–814, 2015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1877050915003956

[5] “Company Info | Facebook Newsroom.” [Online]. Available: http://newsroom.fb.com/company-info/

[6] N. Sheikh, “Big Data, Hadoop, and Cloud Computing,” in Implementing Analytics. Elsevier, 2013, pp. 185–197. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/B9780124016965000116

[7] A. R. Leskovec, Jure and J. Ullman, “Data mining,” Cambridge: Cambride University Press, 2014.

[8] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, “The rise of big data on cloud computing: Review and open research

issues," Information Systems, vol. 47, pp. 98–115, Jan. 2015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0306437914001288

[9] E. Ch'ng, "The Value of Using Big Data Technologies in Computational Social Science," Aug. 2014. [Online]. Available: http://arxiv.org/abs/1408.3170

[10] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, "Sentiment analysis: A review and comparative analysis of web services," Information Sciences, vol. 311, pp. 18–38, Aug. 2015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0020025515002054

[11] J. Martin, A. Ortigosa, and R. Carro, "SentBuk: Sentiment analysis for elearning environments," in 2012 International Symposium on Computers in Education (SIIE), Oct. 2012, pp. 1–6.

[12] "About technologies: Analyse Tweets using Flume, Hadoop and Hive." [Online]. Available: http://www.aboutechnologies.com/2014/12/analysetweets-using-flume-hadoop-and.html

[13] R. V. J. Regalado and C. K. Cheng, "FeatureBased Subjectivity Classification of Filipino Text." IEEE, Nov. 2012, pp. 57–60. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6473695

[14] "Welcome to Apache Flume Apache Flume." [Online]. Available: https://flume.apache.org/

[15] "Apache Hive TM." [Online]. Available: https://hive.apache.org/

[16] C. Stedman, "Analytical models in big data environments often best left small." [Online]. Available:

http://searchbusinessanalytics.techtarget.com/feature/Analyticalmodels-in-big-data-environments-often-best-left-small

FIGURES

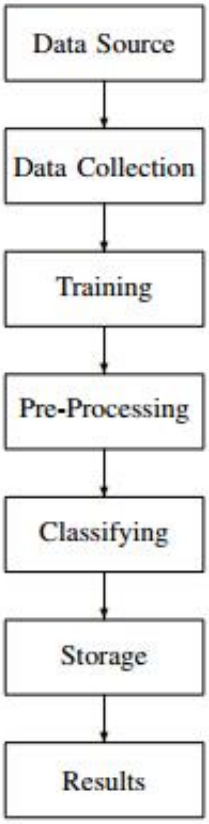

Fig. 1. Theoretical Framework of the Project

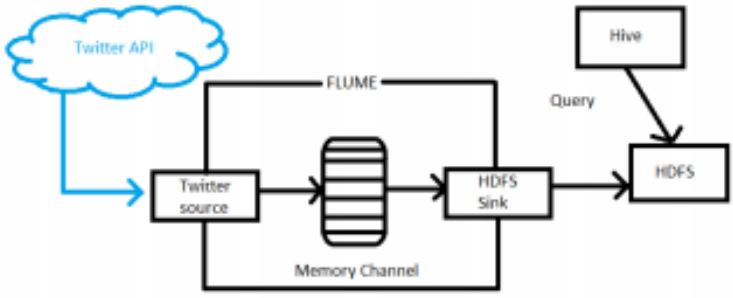

Fig. 2. Architecture of HDFS, Flume, and HivE [12]

Algorithm 1 Map Class

```
1: procedure MAP(key,value,context)
2:     line = Text.toString()
3:     splitted[]= line.split(",", 3)
4:     StringTokenizer s = new StringTokenizer(splitted[2])
5:     while (s.hasMoreTokens() do
6:         Score= checkDictionaries(s.nextToken())
7:         context.write(splitted[2], Score)
8:     }
```

Wherein checkDictionaries() is the function where the current word is checked in both of the dictionaries - negative, and positive.

Algorithm 2 Reduce Class

```
1: procedure REDUCE(Key,values,context)
2:     int sum = 0
3:     while values.hasNext() do
4:         sum += values.next.get()
5:     context.write(key,sum)
```

Fig 3. Dictionary Based Algorithm

Algorithm 3 Map Class

```
1: procedure MAP(key,value,context)
2:     line = Text.toString()
3:     splitted[]= line.split(",", 3)
4:     StringTokenizer s = new StringTokenizer(splitted[2])
5:     while (s.hasMoreTokens() do
6:         pwNeg,pwPos= getWordProb(s.nextToken())
7:         context.write(splitted[2],pwNeg,pwPos)
8:     }
```

Where getWordProb() is a function to compute the probability of each word as negative or positive.

$$P(w|neg) = \frac{(count(w)inNeg) + k}{denom1}$$

$$denom1 = count(totalNeg) + k \times (dicSize + count(newWords)$$

$$P(w|pos) = \frac{(count(w)inPos) + k}{denom2}$$

$$denom2 = count(totalPos) + k \times (dicSize + count(newWords)$$

Algorithm 4 Reduce Class

```
1: procedure REDUCE(Key,values,context)
2:     pmNeg = 1
3:     pmPos = 1
4:     while values.hasNext() do
5:         pmNeg,pmPos = getMProb(values.next.get())
6:     pmMeg = getMsg()
7:     pneg = getProb()
8:     context.write(key,sum)
```

getMProb() is:

$$P(msg|neg) = P(w_0|neg)P(w_1|neg)...P(w_n|neg)$$

$$P(msg|pos) = P(w_0|pos)P(w_1|pos)...P(w_n|pos)$$

getMsg() is:

$$P(msg) = P(msg|neg)P(neg) + P(msg|pos)P(pos)$$

and and, getProb() is:

$$P(neg|msg) = \frac{P(msg|neg) \times P(neg)}{P(msg)}$$

where: k is the smoothing factor, which is 2. Where, P(neg—msg) is the probability of each post or tweet to be negative. Probabilities greater than 0.5, were considered as negative, 0.5 as neutral, otherwise it will be positive.

Fig 4. Naïve Bayes Algorithm

```
1  INSERT OVERWRITE LOCAL
2  DIRECTORY '/home/ubuntu/trial'
3  ROW FORMAT DELIMITED
4  FIELDS TERMINATED BY ','
5  LINES TERMINATED BY '\n'
6  SELECT MAX(id),MAX(user.screen_name),
7  regexp_replace(text,'(\r\n|\r|\n)', '␣') AS text
8  FROM 'tweets' GROUP BY text;
```

Fig 5. Apache Hive Query

10331 files and directories, 10210 blocks = 20541 total filesystem object(s).

Heap Memory used 46.49 MB of 70.27 MB Heap Memory. Max Heap Memory is 966.69 MB.

Non Heap Memory used 43.22 MB of 43.38 MB Commited Non Heap Memory. Max Non Heap Memory is 214 MB.

Configured Capacity:	58.96 GB
DFS Used:	42.87 GB (72.71%)
Non DFS Used:	8.83 GB
DFS Remaining:	7.26 GB (12.31%)
Block Pool Used:	42.87 GB (72.71%)
DataNodes usages% (Min/Median/Max/stdDev):	72.71% / 72.71% / 72.71% / 0.00%
Live Nodes	3 (Decommissioned: 0)
Dead Nodes	0 (Decommissioned: 0)
Decommissioning Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion	0
Block Deletion Start Time	5/15/2016, 1:24:19 AM

Fig 6. HDFS Specifications

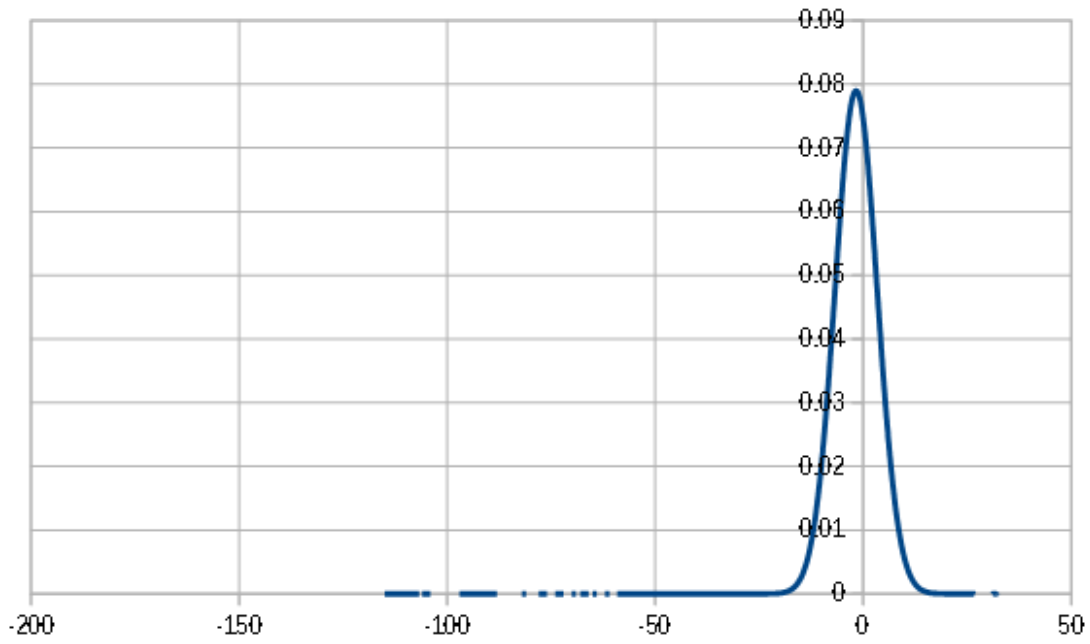

Fig 7. Distribution of First Classifier

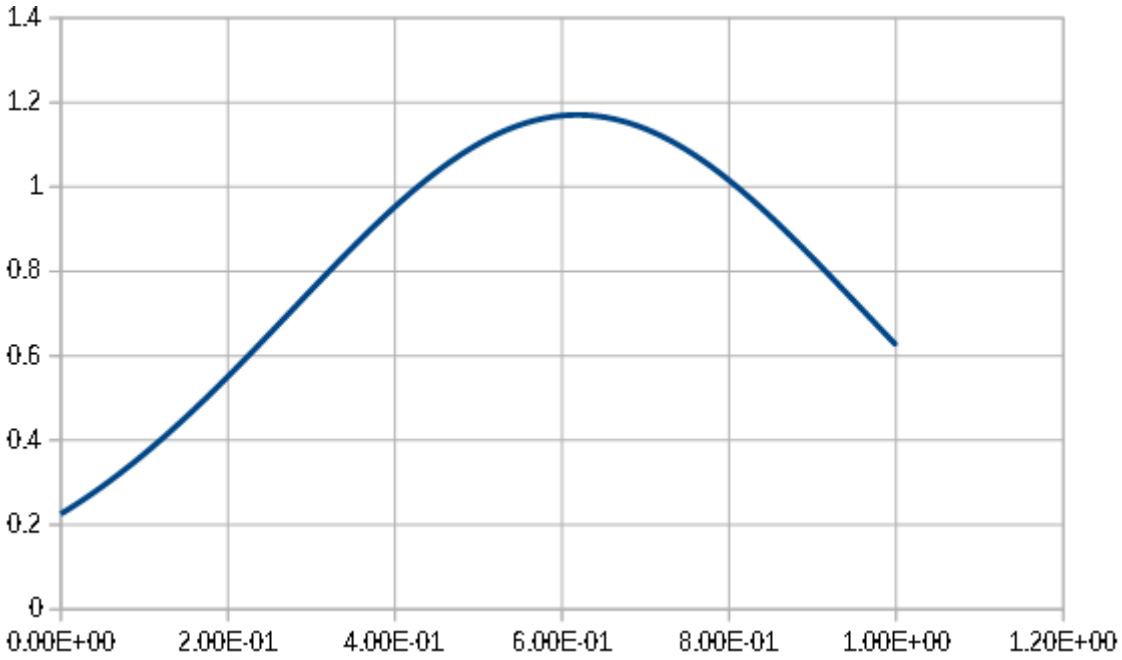

Fig 8. Distribution of Second Classifier

Fig 9. Word Cloud for 1st Classifier Positive Results

Fig 9. Word Cloud for 1st Classifier Negative Results

Fig 9. Word Cloud for 2nd Classifier Positive Results

Fig 9. Word Cloud for 2nd Classifier Negative Results

TABLES

TABLE I
DATA GROUPED ACCORDING TO IMPLEMENTATION:

Source	Total Number'
Rfacebook	37346
twitteR	37299
Apache Flume	14376
Total:	89021

TABLE II
RESULTS OF CLASSIFIERS:

Classifier	Positive	%	Neutral	%	Negative	%
Dictionary	9696	16.7	22218	38.26	26160	45.04
Naive Bayes	20165	34.72	399	0.68	37510	64.6

TABLE III
MEASURES OF CENTRAL TENDENCY:

Classifier	Mean	Median	Mode	Standard Deviation
Dictionary based	-1.7061	0	0	5.0517
Naive Bayes	0.6184	0.7050	0.6442	0.3409

TABLE IV
CONFUSION MATRIX FOR 1ST CLASSIFIER

	Positive	Negative	Neutral	Overall Class:	Recall
Positive	TP=21	FNe=10	FNu=3	34	61.76%
Negative	FP=11	TNe=41	FNu=7	59	69.41%
Neutral	FP=2	FNe=1	TNu=4	7	14.29
Truth Overall	34	52	14	100	
Pre	61.76%	78.85%	28.57&		

Overall Accuracy = 66%

TABLE V
CONFUSION MATRIX FOR 2ND CLASSIFIER

	Positive	Negative	Neutral	Overall Class:	Recall
Positive	TP=21	FNe=13	FNu=0	34	61.76%
Negative	FP=20	TNe=38	FNu=1	59	64.41%
Neutral	FP=1	FNe=6	TNu=0	7	0
Truth Overall	42	57	1	100	
Precision	50%	66.67%	0		

Overall Accuracy = 59%

APPENDIX A: Sample Results from the Two Classifiers

Sample	Dictionary Based	Naïve Bayes
Buti na lang nakita ko si Kuya na may hawak na FREE HUGS na banner sa C-Park kagabi. Sobrang down na kasi ako saka sobrang rough ng week ko, kaya at least napagaan niya loob ko dahil lang sa hug na yun. Sorry kung di ako pumayag sa picture kasi di ko na talaga kaya ngumiti. Sayang naman yun kung baka requirement lang siya sa isang subject mo. Drama ko, pero thank you pa rin :) - Free Hugs,2014, CEAT!	**Positive**	**Positive**
RT @NataSupernova: Good is the new cool.,Talking about Elbi in the ”beauty&fashion” category of #M2020 tomorrow morning	**Positive**	**Positive**
This guy should be the Pambansang OA ng Pilipinas - ang OA lang sobra BPI Julia Vargas #halalan2016 #PiliPinas2016	**Negative**	**Positive**
ICYMI: Palace denies hand in demolition job vs Duterte	**Negative**	**Negative**
Grabe na talaga ang pulitika sa Pilipinas. Namumudmod ng pera si Poe at Roxas sa Payatas dahil marami kaming mga mahihirap na kapitbahay	**Negative**	**Negative**
May ka apartment akong taga UP law, cute niya, petite. Wala lang crush ko siya. Momol naman tayo minsan katok ka lang sakin ;))-admudude,1111, Im outside UP Diliman	**Negative**	**Positive**
Dahil lagi akong umuupo sa likod at malaki yun katawan ng guy na nasa unahan ko, hindi kita ng prof ginagawa ko.So ayun. Puro FB at selfies lang pag may klase. :3Gusto kong ipost sa Wall ko kaso makikita ng mga kapatid ko, taz di na ako good example, so anonymous na lang. :) 1 Selfie per Class 2006 Law	**Negative**	**Positive**
Nagpost ang friend ko ng recent screenshot ng website ng CRS. Nakaemphasize sa pic: Encoding of Ineligible Students for First Semester AY 2011-2011 now availableFRIEND1: talagang 2011-2011 yun siguro. kasi wala na daw tayo ng 2012. hahahahahaFRIEND2: parang ewan lang eh, si crs nakikiend of the world din! XDKasabay sa topic na to ang di pa pagsasara ng preenlistmentFRIEND 3: May 32 palang daw ngayon, bukas pa yung June 1. =))*Ngayon ko lang napansin yung 2011-2011 =))))	**Negative**	**Negative**
Kung hindi kayang ma-control ni Duterte ang fans niya eh baka hindi niya rin magawa ang Peace and Order na pinapangako niya	**Negative**	**Negative**
Maligayang Araw ng Mga Puso mga ka-TDF! Tandaan natin: Maaaring walang forever, pero mayroong always. :)	**Positive**	**Positive**
sobrang ganda mo, every time nakikita kita buo na araw ko, kada punta ko sa engg umaasa akong makita kakaso nung isang araw nakita kita may kasamang iba :(loverboy,2014,Engineering	**Positive**	**Positive**

I really just need to tell this ang bigat na sa puso eh.To whom it may concern,Hello! And fvck you with feelings. :)Grabe sa tinagal kong nagsestay sa campus, ngayon lang ako naka-experience na no receipt no key. Like whut the fuck is that?! Ilang sem na ko nag-aaral dito sa elbi and this was the first time to see a policy like that. Shit lang. Kung dati pupunta ako sa dorm, check in, bayad later. Unlike ngayon. NO RECEIPT NO KEY policy ngayon.Kung pinayagan lang ako mag-apartment ng mother ko eh masaya na ko ngayon. Eh wala eh. Baka ma-aksidente daw ako ng wala sa oras.And another thing, please wag mag-aannounce ng deadline na sakto sa holiday or walang pasok tapos hindi imomove. Got it? Repeat.DormBitch, CA, 201*-*****	**Negative**	**Negative**
Ngayong 2014,Magmomove on nako sa boyfriend kong 4 years kong nakasama sa buhay. di ko pa alam kung paano pero gagawin ko :) kelangan ehh.. exBF . mag LoL ka nalang :D-happyMe 20** CAS	**Neutral**	**Negative**
dont quote me. i will deny it. magaling ako dyan! hahaha - sir ****** (theater)	**Neutral**	**Positive**
Math 11 Simplifying fractions. example to ni sir, habang sinasagutan namin yung example, natawa nalang kami nung matapos namin.	**Positive**	**Negative**
Pembebasan WNI Sandera Abu Sayyaf Wewenang Filipina: Tak ada komando untuk TNI bergerak. https://t.co/F1hqoiJV4O	**Neutral**	**Negative**
Wala ng ginawang mabuti yang mga Abu Sayyaf nakakatulog pa kaya yang mga yan?	**Negative**	**Negative**
[Inquirer.ph] AFP vows to bring Abu Sayyaf to justice after beheading of Canadian hostage https://t.co/X5xa7Zaq94	**Negative**	**Negative**
Duterte camp calls on supporters to exercise civility and decency https://t.co/Oh4akgDg4M via @gmanews	**Positive**	**Positive**
Ang dami kaseng pwede i-argue sa Current ish ngayon na Duterte and UPLB student	**Neutral**	**Negative**
On March 30 & April 5 senators will be visiting the campus. HUH APRIL 6 @OhFlamingoMusic! HAHAHAHA @zyrine mae @slowniB #DuterteVisitsUPLB	**Neutral**	**Negative**
Nasusunog na Faculty Center ng UP-Diliman. — via Reynante Ponte https://t.co/VR2orCzKay	**Negative**	**Negative**
Ang sakit wala na ang Faculty Center.	**Negative**	**Negative**
Stupidest person of today, tomorrow and forever. #MiriamUtakAtPuso	**Negative**	**Negative**
I wont even state the dorm pero tangina punung puno nako (hindi to pun kase hindi kami nasa forestry sigh). Sobrang ingay na nga sa umaga, sobrang ingay pa din sa hapon at gabi. Aba ate, hindi lang ikaw nakatira sa dorm natin, sobrang dami tayo putangina. Umagang umaga sobrang lakas ng pagkanta or pakikipagdaldalan. Puta. Di ko naman sinasabi na pangit yung boses mo, gusto ko lang sabihin na pakihinaan naman please,	**Negative**	**Negative**

may natutulog/nag-aaral po. Tapos kung makapangtahimik samin pag nag-ingay kami ng konti, kala mo naman kung sinong tahimik. Ulol, what a hypocrite.P.S. Sana mabasa mo to at matauhan ka na. Yun lang, thanks.- *****, 201*, BS**, CAS		

END