Sentiment Analysis on UP-related Communities in Facebook and Twitter Using Hadoop

Rahmat Peter I. Dabalos, Lailanie R. Danila, Joseph Anthony C. Hermocilla

I. Introduction

A. Background of the Study

As of today, the world is seeing the rise of technology. We are fortunate to witness the advent 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 mean 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 is 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.

B. 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 researchs 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.

C. Objectives of the Study

The main objectives of this study are 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.

D. Significance of the Study

The tool that will be developed will be of use to examine, understand, and interpret the remarks of the students of the

Presented to the Faculty of the Institute of Computer Science, University of the Philippines Los Baños in partial fulfillment of the requirements for the Degree of Bachelor of Science in Computer Science

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.

E. Scopes and Limitations

This study is intended to produce 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 and MapReduce will be the only modules that will be used from the Hadoop framework. The results from running this tool will be made 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.

II. 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 out-performed 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 inte ed 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 to 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].

III. Theoretical Framework

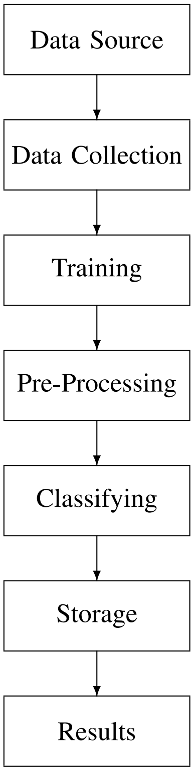

Fig. 1. Theoretical Framework of the Project

A. Data Source

Data that will be collected from:

1) Facebook:

- University of the Philippines Group
- Narinig ko sa UP
- Nakita ko sa UP Campus Group
- The Elbi Files
- The Diliman Files
- UPLB University Student Council
- UP Diliman University Student Council
- UP Manila University Student Council
- Uplb Iskolar Ng Bayan

2) Twitter: For Twitter, the ’#’ will be used in order to search for tweets that are related to the topic of interest.

B. Data Collection

In order to collect the data from the given sources, Graph API will be used in order to implement a program that will be collecting posts from the given pages and Twitter API, for twitter.

C. Training

Before processing the tweets, each word in the dictionary that has subjective meaning will be classified as positive, negative. The dictionary will also be composed of Filipino words.

D. Pre-processing

Because the data that will be gathered will surely be composed of many unnecesary data, at this stage, the dataset will be filtered so that, only the relevant data will be used in the next processes. The data to be extracted on each post or tweet are: date, keyword, and content.

E. Classifying

Because the tool will be run on top of a Hadoop cluster, an implementation of the MapReduce Function will be used.

Algorithm 1 Map Class

```
procedure MAP($Keyword$,$Text$,$Output$)
    line = Text.toString()
    Score=computeProbability(line)
    output.collect(Keyword, Score)
```

Algorithm 2 Reduce Class

```
procedure REDUCE($Keyword$,$values$,$Output$)
    int sum = 0
    while values.hasNext() do
        sum += values.next.get()
    output.collect(keyword,sum)
    output.collect(Keyword, Score)
```

computeProbability(): Naive Bayes with Laplace Smoothing will be used to classify the post or tweet. Based on the dictionary generated in the training stage, the probability of the tweet being a negative or positive will be computed in order to tell if the tweet or post’s polarity is negative or positive.The formula is as follows:

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

where:

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

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

$$P(neg) = \frac{count(neg) + k}{count(neg \cup pos) + 2k}$$

$$P(pos) = \frac{count(pos) + k}{count(neg \cup pos) + 2k}$$

$$P(w|neg) = \frac{(count(w)inNeg) + k}{count(totalNeg) + k \times (dicSize + count(newWords))}$$

$$P(w|pos) = \frac{(count(w)inPos) + k}{count(totalPos) + k \times (dicSize + count(newWords))}$$

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

and k is the smoothing factor, which is 2.

F. Storage

The storage for the data that will be processed is the Hadoop Distributed File System(HDFS). That has already been set up in the Peak Two Cloud.

G. Results

The results will be plotted into graphs and charts in order to visualize the results of the sentiment analysis clearly.

IV. Methodology

A. Development Tools

1) Peak Two Cloud(P2C): Big Data problems requires a platform that can support parallel and distributed computing. The Peak Two cloud can provide the infrastructure needed in order to create a Hadoop cluster. 4 instances will be used from the cloud, 1 will be the Master Node and three others will be slave nodes.

2) Apache Hadoop: Hadoop is an open-source framework that will be used to store the data via the Hadoop Distributed File System, and to run the MapReduce function that I will be implementing on the on the gathered data to be processed.

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.

4) Twitter API: Twitter has an Application Programming Interface used to gather data from Twitter.

B. Functional Requirements

1) Gathering of Data: This is a functional requirement that will be implemented according to the Graph API from Facebook and Twitter API from Twitter. Its goal is to collect all the relevant data from the two social media sites.

2) Training of Data: This is where the dictionary will be initialized based on the polarity of each word. Polarity is the term used to define if a word is either negative or positive.

3) Sentiment Classification: This is the main requirement of the tool. The sentiment classification that will be used for this project is Naive Bayes with laplace smoothing that will be implemented with the MapReduce Function. Map Reduce will split the data gathered into independent chunks then it will processed in parallel manner, As seen in the implementation above, the Map function will take in the raw data in the form of posts or tweets, then it will compute the score of each post according to the sentiments in the post, it will produce a key value pair which is the keyword and the score. Then it will be used by the reduce function. The reduce function will only be run after all of the map functions in the node has finished execution. Reduce will aggregate the scores according to the keyword. Afterwards all of the data will be saved in the Hadoop File system. Apache Hadoop, takes care of the scheduling and monitoring of the nodes and in case of failed tasks, the framework can execute it [12].

4) Hadoop Distributed File System: This requirement is to distribute the data gathered from the tweets in the Hadoop cluster.In a Hadoop Cluster there is a Name Node which is also the Master Node and there are Data Nodes or the slave nodes. The role of the NameNode is to manage the whole data in the cluster. In order to prevent single point of failure in the cluster, it first partitions the data into blocks of 64MB. Then, each block is distributed in the data nodes according to the value of the Data Replication. Although, this project will use 3 slave nodes, the data replication is still 3, this means that each block of size 64MB data will stored in all of the data nodes. To set-up the Hadoop Cluster, follow the steps from: http://chaalpritam.blogspot.com/2015/01/hadoop-260-multi-node-cluster-setup-on.html

5) Visualization: This module is to present the output of the tool in graphs or charts so that the result will be understood clearly.

C. Non-Functional Requirements

1) Performance Requirements: The status of the system will always be checked, system crash, hang, or error will always be detected. And the performance according to the efficiency and integrity of the system.

2) Safety Requirements: This is in terms of the data, the storage will always be backed up in case of failures in the hadoop cluster.

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.

V. Expected Output

In Table 1, that will be the expected inputs that will be coming from facebook and twitter. The data will be used as input to the map function. The data will be distributed into each of the data nodes in the hadoop cluster, and then each data will be processed by the map function as explained above. The output of the map function on the sample data is on Table II. After the Map function, the output will be shuffled and sorted and will be used by the reduce function to aggregate the scores. The output of the Reduce Function is on Table III. It can be seen that based on the sample data that was gathered, 2/2 tweets related to APEC show positive sentiment and 2/2 for both UPLBWalkout and ParisAttacks give out negative sentiment.

TABLE I
SAMPLE INPUT DATA

Date	Keyword	Content
11/19/15	APEC2015	Trudeau: I've adored my stay in the PH. It was great to immerse in a culture that I got to know very well from Filipinos in Canada #APEC2015
11/17/15	APEC2015	Obama: The good news is that more and more companies are realizing that climate change presents good business opportunities. #APEC2015
11/17/15	ParisAttacks	Its not a setback, its a catastrophe! - @GovernorPataki disputing #Obama 's description of the #parisattacks
11/17/15	ParisAttacks	We're all scared - but eroding civil liberties & marching into war isn't the answer. @CarolineLucas on #ParisAttacks
11/13/15	UPLBWalkout	Hundreds of students walk out from their classes to protest against the commercialization of UP especially the recent failure and mismanagement of the UP Administration in the implementation of the flawed and dubious SAIS under President Pascual's 'eUP.' SERBISYO SA UP, WAG GAWING NEGOSYO! ISANG PAMANTASAN, ISANG PANAWAGAN! WALKOUT! WALKOUT! WALKOUT! #UPLBWalkout
11/12/15	UPLBWalkout	CAS-SC: Bukod sa komersyalisasyon ng edukasyon, nakararanas ng mga represibong polisiya, katulad ng org recognition. #UPLBWalkout

TABLE II
OUTPUT FROM MAP FUNCTIONS

Keyword	Values
APEC2015	1
UPLBWalkout	-1
ParisAttacks	-1
APEC2015	1
ParisAttacks	-1
UPLBWalkout	-1

TABLE III
OUTPUT FROM REDUCE FUNCTIONS

Keyword	Values
APEC2015	2
ParisAttacks	-2
UPLBWalkout	-2

REFERENCES

[1] L. Mearian, "Scientists calculate total data stored to date: 295+ exabytes," Feb. 2011. [Online]. Available: http://www.computerworld.com/article/2513110/data-center/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-software-products/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 e-learning environments," in *2012 International Symposium on Computers in Education (SIIE)*, Oct. 2012, pp. 1–6.

[12] "Mapreduce tutorial." [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html