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

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

Institute of Computer Science, University of the Philippines Los Baños

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.

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.

METHODOLOGY

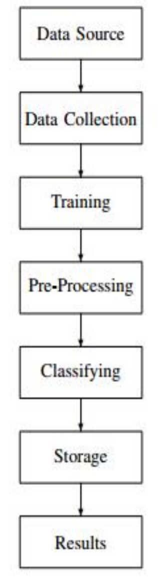
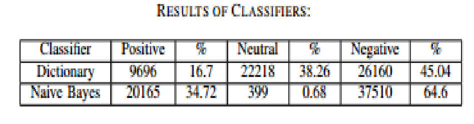


Fig. 1. Theoretical Framework of the Project

1. Data Source –Data were collected from Facebook and Twitter.
2. Data Collection – used tools such as
   1. Rfacebook
   2. twitteR
   3. Apache Flume
   4. Apache Hive
3. Training –
   1. Dictionary Based: Online Lexicon
   2. Naïve Bayes: crowdsourced 11020 tagged as positive or negative data containing tweets and posts .
4. Pre-Processing - The data extracted on each post or tweet were: id, username, and text.
5. Classifying –
   1. Dictionary Based
   2. Naïve Bayes
6. Storage- Hadoop Distributed File System(HDFS) in a Multi-node cluster in Peak Two Cloud.

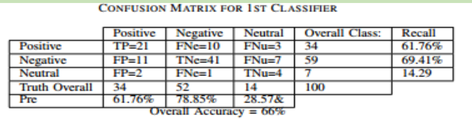
RESULTS



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

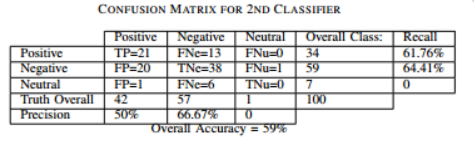
Out of 100 sampled Tweets and Posts:



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%



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%

CONCLUSIONS 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.

ABOUT THE AUTHOR

Rahmat Peter I. Dabalos, is a 4th year undergraduate student of UPLB taking up BS Computer Science. His fields of interests are Artificial Intelligence, Networking, and Software Development. He is currently a member of Systems Research Group of UPLB for the AY. 2015-2016.