Mining Of Tweets For Possible Investment In Lipa City Using Natural Language Processing And Naive Bayes Classifier

Francis G. Balazon, Myrna A. Coliat

Batangas State University, College of Informatics and Computing Sciences Graduate School
Lipa City, Batangas, Philippines
fbalazon@yahoo.com

Batangas State University, College of Informatics and Computing Sciences Graduate School
Batangas City, Philippines
myrna_coliat@yahoo.com

Abstract: Investing is one of the most popular businesses to make money grow. With that, one cannot have a mistake of just investing into something. The use of social media nowadays is fast and widespread all over the world. It could help a lot in gathering the information that an investor needed. With the use of a social media site, surveying will be easy and there will be no cost in surveying on what do people like or need these days. The study focused on the development of a website which collects all the tweets about Lipa City and then analyzes the sentiments of the tweets if it is positive or negative. It then categorizes the sentiments and displays the percentage of the tweets from different parameters. The researchers started the website application through data gathering, planning, and designing the functional features of the website. The processes involved in the sentiment analysis of the website are Naive Bayes Classifier to analyze the tweets and Natural Language Processing to properly categorize the parameters of the tweets. To see the errors of the website, the researchers evaluated the system by a series of testing and it was proven to be a functional and helpful website.

Keywords: tweets, sentiment, investing, social media, naïve bayes

1. Introduction

Investing is a trend for Filipinos nowadays. Working and earning is not just enough to have a better living. Now, people have learned to invest into something that could benefit them or earn more money. There are many things that are good for investing nowadays and an investor should be knowledgeable on the things that interest him. Investing is a very serious thing to be done. It needs wise decision-making and broad research on the information about the possible investments. There are many factors to consider in this generation when it comes to investing whether it is for business, insurance, fine-living, or being part of a successful community. Sharing of information is an important factor of everyday living and people depend more on the facts and opinions of other people, articles, and studies that are available on the internet. With that, different websites and social media sites grow and became popular like Facebook, Twitter, and Instagram. With these social media sites, the passing of information is massive and viral. These sites are regularly used in the business industry for fast ad campaigns, business deals, informative views of any matter, and for easy communication purposes. One of the most important things that people always look for in social media sites is the comments or sentiments of the internet users. The sentiments of other people towards a subject are persuading and at the same time informative. Most of the time, people tend to rely on other people's opinions in making a decision. It is good for the people who do not have first-hand experiences or encounters on certain things. Especially in starting businesses or investing into something for a better future and for a better living, it helps the people to know what's in or what's not, what's working out well with the community and what things are disappointing in a community. Today, many people want to start their own businesses or invest in good businesses, invest into finer things in life that could benefit one's income and lifestyle. One way to succeed in deciding what businesses are good for investments is through using social media sites. One of the best social media sites to use is Twitter. Twitter is good for discovering new ideas and perspectives because it is a tool that most people use to express their feelings and thoughts, share things and inform people. With that, searching and reading tweets and comments about the possibilities of investing in a certain area would be easy and fun for the internet users. It will be a big help if there is a website that can collect all the tweets about Lipa City so that people would know the good things or areas to invest in the city and what things are still needed on this place.

2. Design and Methodology

The study was created aiming to develop an interface which is a website, to filter and analyses the tweets about Lipa City to be able to know the possibilities of investing in the area. This project was created by implementing the Sentiment Analysis algorithm using Naive Bayesian Classifier. Sentiment analysis of the tweets about Lipa City is the process of reviewing and exploring the tweets on the internet to determine the overall opinion or feeling about Lipa City. Reviews represent the user-generated content, which greatly attract the attentions of businessmen, sociologists and psychologists and others who might be concerned with opinions, views, public mood and general or personal attitudes. It is treated as a classification task as it classifies the orientation of a text into either positive or negative. Machine learning is one of the widely used approaches towards sentiment classification. Machine learning classifiers such as Naive Bayes was used in sentiment classification to achieve accuracies. [2] Machine learning methods are based on training an algorithm, mostly classification on a set of selected features for a specific mission and then test on

another set whether it is able to detect the right features and give the right classification. [3] Sentiment Analysis has the following processes: [4]

- Data Collection – Identify sources and gather data through tweets from Twitter.
- Connectivity – Pull the data via authentication API to connect through the analytical tools.
- Pre-processing – Cleanse and structure the data.
  - NLP (Tokenization)
- Training Data - This data is the fuel for the classifier; it will be fed to the algorithm for learning purpose.
- Classification by Sentiment Analysis using Naive Bayesian Classifier
- Results.

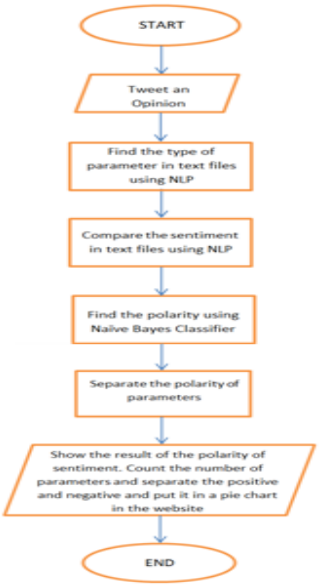

Figure 1. *Sentiment Analysis Process of InveStation Website*

Figure 1 shows how the polarity of a tweet can be categorized by using Sentiment Analysis and NLP. When the user tweets a sentiment about Lipa City with the hashtag "#WhatsInLipa", the tweet will be synched with the website. The website contains a knowledge base which is a collection of possible tweets comments, and a collection of the keywords for each parameter. The collections of the possible tweets are the positive and the negative tweets. The parameters are location, population, climate, resources, economy, education, government, and culture. In the process, the computer will search the type of parameter from the text files by comparing the text from a tweet with the keywords in each parameter. This method is called natural language processing. After the system finds matched the keyword with the text, the computer will categorize what parameter the tweet is about. Then, the computer will compare the words from tweets to the text files of the possible positive or negative tweets. Once the words were matched by phrase or by sentence, the system will then prove the polarity of the sentiment from the tweet. The process on this part is Naive Bayes Classifier. Once the polarity was proved and classified if it is negative or positive, the computer shows the result of the tweet in the pie chart if it is positive or negative and what parameter it will count. There are three pie charts on the website. The first pie chart is the positive pie chart of the parameters, the second is the negative pie chart of the parameters, and the third one is the summary of the tweets that were collected from Twitter. The collected tweets can be viewed on the tweets page and it contains the sentiment analysis result of each tweets. The next step is the NLP or Natural Language Processing. There are five steps of NLP. These are Lexical Analysis, Syntactic Analysis (Parsing), Semantic Analysis, Disclosure Integration and Pragmatic Analysis. Lexical Analysis focuses on identifying the sentiment if it is a word, a sentence or a paragraph. This process divides the sentence into words. Example is below:

The air during summer is not good.

|The| |air| |during| |summer| |is| |not| |good|

The next step is Syntactic Analysis. Syntactic Analysis analyses the chunk words. The chunk words will analyze the type of part of speech of it. Syntactic Analysis has two simple methods to get the sentiment. These are Context Free Grammar and Top down Parser. The first is Context Free Grammar. It is the grammar that consists rules with a single symbol on the left-hand side of the rewrite rules. Create grammar to parse a sentence − "The air during summer is not good."

NP(Noun Phrase) – Article + Nouns
Article – The
Nouns – air, summer
Preposition - during
VP(Verb Phrase) – Verb + Adjective
Verb – Is
Adjective – not good

Break this sentence in a parse tree. These rules say that a certain symbol may be expanded in the tree by a sequence of other symbols.

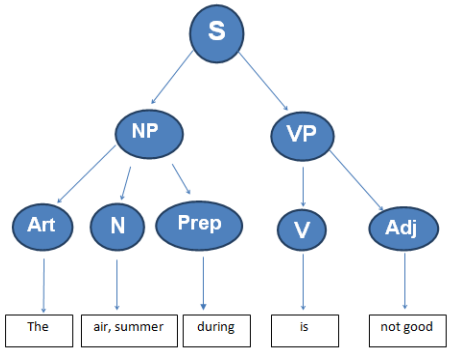

Figure 2. *Syntactic Analysis*

Figure 2 shows the Syntactic Analysis. According to the first order logic rule, if there are two strings Noun Phrase (NP) and Verb Phrase (VP), then the string combined by NP followed by VP is a sentence. The rewritten rules for the sentence are as follows:

S → NP VP
NP → Article | Noun | Preposition
VP → Verb | Adjective

Now consider the above rewrite rules. Wrong subject-verb agreements were also permitted. The next is Semantic Analysis. Semantic Analysis compares the evaluated sentence to the knowledge base of the website. If the sentence matched the process is correct. The computer will choose the compared sentence or phrase from the text files that are the same with the tweeted sentiment. The last is Pragmatic Analysis. During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of language which require real world knowledge. It means that the output will be chosen and it means that the computer accomplishes the task that need.

Integration of Sentiment Analysis Using Naive Bayesian Algorithm

To be able to get the right categorization of tweets, the developers used Naïve Bayesian Classifier in the Sentiment Analysis of the website for the computation of the words. Bayesian classifiers are based around the Bayes Rule, a way of looking at conditional probabilities that allows you to flip the condition around in a convenient way. A conditional probability is a probability that event X will occur, given the evidence Y. That is normally written P(X | Y). The Bayes Rule allows us to determine this probability when all we have is the probability of the opposite result and of the two components individually: In computing for the probability, here is a table that contains the data tweets of the Twitter users. It has a positive and a negative class. It will match the tweets of the users to the database tables that have positive and negative keywords.

Table 1. *Review table of the sentiments about the weather in Lipa City*

DOC	SENTENCE	CLASS
1	The advisory of the weather in Lipa City is not always announce early.	NEGATIVE
2	Sometimes, there's no weather advisory in Lipa City and its important for the people to be updated for the weather forecast.	NEGATIVE
3	There's no weather advisories in Lipa City and its important for the people to be updated for the weather forecast.	NEGATIVE
4	The atmosphere in Lipa is good for anyone who wants a fine weather.	POSITIVE
5	The weather here in Lipa City is breezy.	POSITIVE
6	Its balmy during summer in Lipa but it is a nice balmy weather.	POSITIVE
7	The weather here in Lipa City is cold.	POSITIVE
8	The weather here in Lipa City is cool.	POSITIVE
9	I love the cool weather in Lipa City.	?

Table 1 shows the review table of the collected tweets about the weather in Lipa City. The set of sentiments with the corresponding class is called a train set. The train set has its known value. The boxed sentence is the test set. And its value needs to be classified.

Train Set. It has a training attribute that works with the corresponding class.

Test Set. It is a set of data where the class is unknown.

After creating a Train Set and Test Set, the next step is the Text tokenization which is the process in which the training texts, if they exist, and the texts that are going to be processed are split into the units that will be considered for the classification. [5] The tokenization has different procedures. The first procedure is to use the basic filtering to remove the other symbols in the sentence. The second procedure is tokenization. In tokenization, the sentence will be chunked into words. The third procedure is multi-words grouping. In multi-words grouping, the words that are separated will be grouped. The fourth procedure is removing the stopwords. The final procedure is showing the result. Figure 3 shows the text tokenization and multi-words. The first step is about removing the symbols. The second is tokenization. The sentence is chunk into words. The third is multi-words grouping. It will group the group words that always go together. The next is removing the stop words. And after that, the result can be seen. Repeat the text tokenization and multi-words to get the proper result in the other sentences.

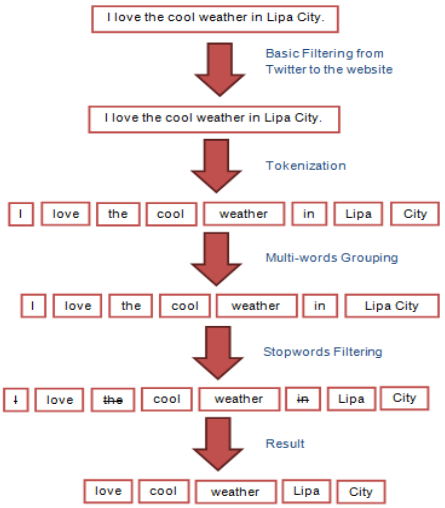

Figure 3. *Tokenization and Multi-words*

In this tokenization of the train set, as can be seen, the researchers remove all the symbols already and remove out the stop words. The red words that are lined are the stopwords.

~~The~~ advisory ~~of the~~ weather ~~in~~ Lipa City ~~is~~ ~~not~~ always announced early
Sometimes, there's no weather advisory in Lipa City and its important for the people to be updated for the weather forecast
~~There's~~ ~~no~~ weather advisories ~~in~~ Lipa City ~~and its~~ important ~~for the~~ people ~~to~~ ~~be~~ updated ~~for the~~ weather forecast
~~The~~ atmosphere ~~in~~ Lipa ~~is~~ good ~~for~~ anyone ~~who~~ wants ~~a~~ fine weather
~~The~~ weather ~~here~~ ~~in~~ Lipa City ~~is~~ breezy
~~Its~~ balmy ~~during~~ summer ~~in~~ Lipa ~~but~~ ~~it~~ ~~is~~ ~~a~~ nice balmy weather
~~The~~ weather ~~here~~ ~~in~~ Lipa City ~~is~~ cold
~~The~~ weather ~~here~~ ~~in~~ Lipa City ~~is~~ cool

Stopwords are those words that do not provide any useful information to decide in which category a text should be

classified. This may be either because they do not have any meaning (prepositions, conjunctions, etc.) or because they are too frequent in the classification context. To know the polarity of the test set, a frequency table is needed. It is based on the number of positive and negative tweets that are categorized in the table. This frequency table counts the words in each class. If a word is repeated in a class, the count of same word is one. If the word has the same word in the other class, the count is one as well. Table 2 shows the words in the positive sentence that are counted in the frequency table. One point for each word from the training set. If the word is both positive and negative, it can be counted as one for positive and one for negative. The words in the negative sentence will be counted as well. This is done to know how many times a word appears in each class. It is based on the number of positive and negative comments that are categorized in the table. The third step is to compute the prior. To compute the prior use this formula:

$P(C) = N_c / N$

Where:

P(C) is the probability of the class

N_c is the total count of a particular class in a training set (count of positive class and negative class)

N is the total count of class in the training set (total number of the items)

So the prior will be:

Positive → 5/8 = 0.625

Negative → 3/8 = 0.375

The next step is to compute the Conditional Probability/Likelihood. To compute this, this formula is needed:

$$P(w/c) = \frac{count(w,c)+1}{count(c)+|v|}$$

Table 2. *Frequency Table of the Data Set*

WORDS	POSITIVE	NEGATIVE
advisory	0	2
weather	5	5
Lipa	3	3
city	5	3
always	0	1
early	0	1
sometimes	0	1
important	0	2
people	0	2
updated	0	2
forecast	0	2
announced	0	1
atmosphere	1	0
good	1	0
anyone	1	0
wants	1	0
fine	1	0
balmy	2	0
summer	1	0
nice	1	0
cold	1	0
advisories	0	1
cool	1	0

Where:

P(w,c) is the conditional probability/likely-hood; where "w" is the word attribute (comment itself) and "c" is the class (positive or negative).

count(w,c) is the total count of the word attribute that occurs in a "c".

+1 is the Laplace smoothing, this is needed for the computation when the count(w,c) is zero, the whole computation will not be equal to zero.

count(c) is the total count of the word attribute in a particular class that occurs in the training set (count of class).

|v| is the vocabulary; it is the total count of the word attribute in the training set.

Table 3. *Computing the Conditional Probability/ Likelihood of Positive Class*

love	= 0 + 1 / 50 + 104 = 1 / 154 = 0.00649
cool	= 1 + 1 /50 + 104 = 2 / 154 = 0.01298
weather	= 5 + 1 / 50 + 104 = 6 / 154 = 0.03896
Lipa	= 5 + 1 / 50 + 104 = 6 / 154 = 0.03896
City	= 3 + 1 / 50 + 104 = 4 / 154 = 0.02597

Table 4. *Computing the Conditional Probability/ Likelihood of Negative Class*

love	= 0 + 1 / 54 + 104 = 1 / 158 = 0.00632
cool	= 0 + 1 /54 + 104 = 1 / 158 = 0.00632
weather	= 5 + 1 / 54 + 104 = 6 / 158 = 0.03797
Lipa	= 3 + 1 / 54 + 107 = 4 / 158 = 0.02531
City	= 3 + 1 / 54 + 107 = 4 / 158 = 0.02531

In the Tables 3 and 4, the numerators need to be computed. Add the count of words that appears in a class with the Laplace smoothing. After that, compute the value of the denominator. Add the count(c) to |v|. After that, divide the numerator with the denominator to get the probability. The next process is to get the posterior probability. To maximize the posterior probability, use this formula:

$$C_{map} = \text{argmax } P(X_1, X_2, ...X_n)\ P(C)$$

Where:

argmax $P(X_1, X_2, ...X_n)$ is the answer in the conditional probability/likelihood. All answers must be multiplied to the conditional probability/likelihood.

P(C) is the prior answer. Prior is the probability of a class.

Posterior Probability of Positive Class:

C_{map} = [(0.00649) x (0.01298) x (0.03896) x (0.03896)] x [(0.02597)] x [(0.625)]

= **$2.075435556 \times 10^{-9}$**

Posterior Probability of Negative Class:

C_{map} = [(0.00632) x (0.00632) x (0.03797) x (0.02531)] x [(0.02531)] x [(0.375)]

= **$3.643261226 \times 10^{-10}$**

The next is to determine the class of the test set. The highest posterior probability of the class will be the basis of the status of the uploaded tweet. This will be added to the count to the positive tweets. Same goes with the negative tweet if the result is negative. Therefore, the highest posterior probability of the class is $2.075435556 \times 10^{-9}$. So the test set was proven to be a positive sentiment.

Used Programming Language

The proponents used PHP programming language for the back-end development and HTML, CSS and Bootstrap for the front-end development. PHP executes on the server, while a comparable alternative, Angular JS, executes on the client. PHP is an alternative to Microsoft's Active Server Page (ASP) technology. AngularJS extends HTML with new attributes. AngularJS is perfect for Single Page Applications (SPAs).

3. Results and Discussions

The aim of this study is to analyze the collected tweets from Twitter if they are positive or negative. It then categorizes the tweets on what parameter it tackles if it is about population, culture, education, resources, economy, government, location, or climate. These parameters are the major factors that investors take into consideration about an area. This project was intended for searching and analyzing the possibilities for a potential good investment in the area of Lipa City. Sentiment Analysis which is also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space. There are also many names and slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. However, they are now all under the umbrella of sentiment analysis or opinion mining. While in industry, the term sentiment analysis is more commonly used, but in academia both sentiment analysis and opinion mining are frequently employed. They basically represent the same field of study. The term sentiment analysis perhaps first appeared in (Nasukawa and Yi, 2003), and the term opinion mining first appeared in (Dave, Lawrence and Pennock, 2003). However, the research on sentiments and opinions appeared earlier (Das and Chen, 2001; Morinaga et al., 2002; Pang, Lee and Vaithyanathan, 2002; Tong, 2001; Turney, 2002; Wiebe, 2000). Sentiment analysis and opinion mining mainly focuses on opinions which express or imply positive or negative sentiments. [6] The Naive Bayes Classifier is so named because it assumes that each word in the document has nothing to do with the next word. That is a naive assumption. But it turns out that, while naive, it is actually a great simplifying assumption; studying words separately like this actually yields very good results. One could also study bigrams or trigrams (sets of two or three words at a time), at which point the classifier is no longer "naive" but it will require a much larger amount of training data and storage space. [7] The reason the NB classifier works well for document classification is that it de-correlates the number of times a word is seen in a given language from its statistical importance. The word "a" is found in many languages. Perhaps it even appears in 100% of your English training set. But that does not mean that documents that have it are English. Bayes is used to convert the "probability that 'a' appears in an English document" (which is 100%) to the "probability that this document is English because it has 'a' in it" (maybe 50%). [7] Therefore, the common stuff that is found everywhere is given a very weak significance and the stuff that is found more uniquely across a category is given a much stronger weight. The end result is a very smart, simple algorithm that has low error rates ("low" being a relative term). It is not magic, there is no neural network, there is no "intelligence", and it is just math and probability. [7] Machine learning is a subfield of Artificial Intelligence dealing with algorithms that allow computers to learn. This usually means that an algorithm is given a set of data and subsequently infers information about the properties of the data; that information allows it to make predictions about other data that it might come across in the future (Segaran 2007: 3). The ability to make predictions about unseen data is possible because almost all non-random data contains patterns that allow machines to generalize (Segaran 2007: 3). In order to generalize, the computer trains a model with what it determines are the important aspects of the data (Segaran 2007: 3). [8] Machine learning does have its weaknesses; the algorithms vary in their ability to generalize over large sets of patterns, and a pattern that is unlike any seen by the algorithm before is quite likely to be misinterpreted (Segaran 2007: 4). In language, frequently occurring patterns are rare, and rarely occurring patterns are predominant; this makes that machine learning methods can only (limitedly) generalize based on the information that they have already seen (Segaran 2007: 4), while humans have a large world knowledge base that supplies them with countless training data and possibilities for feature construction. [8] Natural Language Processing is involved in the process of Sentiment Analysis. NLP refers to AI method of communicating with intelligent systems using a natural language such as English. The field of NLP involves making computers to perform useful tasks with the natural languages that humans use. The input and output of an NLP system can be speech and written text. [9] There are two components of NLP as given: the Natural Language Understanding (NLU) which maps the given input in natural language into useful representations and analyzing the different aspects of the language. The second is Natural Language Generation (NLG). [9] It is the process of producing meaningful phrases and sentences in the form of natural language from some internal representation. It involves Text Planning, Sentence Planning and Text Realization. Text planning includes retrieving the relevant content from the knowledge base. Sentence planning includes picking up the required words, forming meaningful phrases, and setting the tone of the sentence. Text Realization maps the sentence plan into sentence structure. [9] The researchers used the text planning. The website of this study contains text files that serve as a storage of the collection of words or knowledge base. The proponents conducted various studies about the algorithm and how to develop the project. Testing activities were also conducted to finalize and make sure that the whole system of the website works efficiently.

The processes involved in the Sentiment Analysis are as follows:

1. The process starts when a twitter user tweets a sentiment about Lipa City with the official hashtag of InvestLipa which is "#LipaCity". Then the InvestLipa website will filter and get the tweet that has the official hashtag.
2. The filtered tweet will be analyzed if it is positive or negative by the algorithm implemented on the website which is the Sentiment Analysis.
3. Through Natural Language Processing, the words will be analyzed, and then will be compared to the polarity text files which are a collection of possible tweets or comments and will be computed by the use of Naive Bayesian Classifier.
4. The words will also be compared to the text files of each parameter's keywords. If it matched a word in any of the parameters, the tweet will be categorized to the count of the matched parameter.

After the unit and integration testing were done, the researchers were able to achieve the following:

1. The project achieved the working process of Sentiment Analysis using Naive Bayesian Classifier and Natural Language Processing.
2. The researchers implemented the Sentiment Analysis in the project to be able to categorize the sentiments from Twitter. This is the most suitable algorithm to know the polarity of tweets, coupled with Naïve Bayesian, Natural Language Processing, and Machine Learning along the process of Sentiment Analysis.
3. The website is functioning based on the proposed idea and it was tested, and it passed the evaluation of the testers. The website filters the tweets with the official hashtag through the use of Twitter API so the tweets collected will be all about Lipa City. The website is effective because it analyzes the polarity of tweets and categorizes the parameter of the tweets. The percentage of the total of positive tweets, total of negative tweets, and the total number of tweets will be computed, and its summary can be viewed in the form of pie graphs on the website.

4. Conclusions

"Mining of Tweets for Possible Investment in Lipa City Using Natural Language Processing and Naive Bayes Classifier" was made to collect tweets regarding the City of Lipa and analyze these sentiments if they are in a positive or a negative perspective. It would then be categorized to a certain parameter that investors consider in investing. In line with this, proponents came up with the following conclusions: Mining of Tweets for Possible Investment in Lipa City Using Natural Language Processing and Naive Bayes Classifier was implemented in the process of analyzing the sentiments. It starts after the website collected the tweets. It then categorizes the tweets on what parameter it tackles. Then, it will categorize the polarity of the tweet. The higher class that was computed will be the result of the tweet. To determine the possibility of investing in Lipa City, the results of the sentiment analysis of the tweets have a display of the pie charts of the percentage of the positive and negative tweets of each parameter. For the testing and evaluation of the website, the developers used a web testing tool that is free and online.

5. Recommendations

The following recommendations were represented to maximize the efficiency of the system. There were some limitations that call for the necessary improvements of this project. The future researchers can further improve this project in the future by expanding its range of use of the social media sites to collect the sentiment. This project can only analyze words sentiment analysis. The future researchers can improve this by expanding the scope of the analysis. They can add it to the system to recognize and analyze emoticons and punctuation marks in the sentiment analysis. This project is limited to filter tweets that are in English. The future researchers can improve the system to be able to collect tweets in different languages that can be classified by sentiment analysis.

References

[1]. About Public and Protected Tweets, 2017. Available from Twitter Help Center: https://support.twitter.com/articles/14016; accessed June 7, 2017.

[2]. Haddi, E., Liu, X., Shi, Y. The Role of Text Pre-Processing in Sentiment Analysis. Available from Procedia Computer Science 17 (2013) 26 – 32: https://ac.els-cdn.com/S1877050913001385/1-s2.0-S187705091300 1385-main.pdf?_tid=8a9824be-c997-11e7-9f96-00000aab0f6c&acdnat=1510704016_b44a-f4d238db021be498a2cdcfea8192; accessed November 2, 2017.

[3]. Machine Learning General Concepts, 2016. Available from http://emma.memect.com/t/4734751241f61c53ff2194e4d3c59bea3412894005afd064799a4de96e13659c/Machine%20Learning%20General%20Concepts.pdf; accessed Nov. 2, 2017.

[4]. Ishwar, R.R. 10 Steps to Begin Sentiment Analysis to Glean Customer Insight, 2016. Available from https://www.infogix.com/blog/10-steps-to-begin-sentiment-analysis-to-glean-customer-insight/; accessed November 2, 2017.

[5]. Text Tokenization and Multiwords, 2017. Available from Meaning Cloud: https://www.meaningcloud.com/developer/resources/doc/models/models/text-tokenization-multiwords; accessed November 13, 2017.

[6]. Liu, B. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012. Available from https://www.cs.uic.edu/~liub/FBS/Sentiment-Analysis-and-OpinionMining.pdf; accessed June 7, 2017.

[7]. Kanber, B. Machine Learning: Naïve Bayes Document Classification Algorithm in Javascript, 2012. Available from https://burakkanber.com/blog/machine-learning-naive-bayes-1/; accessed June 7, 2017.

[8]. Machine Learning General Concepts, 2016. Available from http://emma.memect.com/t/4734751241f61c53ff2194e4d3c59bea3412894005afd064799a4de96e13659c/Machine%20Learning%20General%20Concepts.pdf; accessed June 7, 2017.

[9]. AI-Natural Language Processing, 2017. Available from tutorialspoint: https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm; accessed November 2, 2017.

Author Profile

Francis G. Balazon received his bachelor's degree in Industrial Technology major in Computer Engineering Technology at Batangas State University Rosario Campus in 2004. He completed his Master of Science in Information Technology (MSIT) at Batangas State University Alangilan Campus in 2012. He earned his Doctor of Information Technology (DIT) degree at AMA University Quezon City. His research interests lie in the area of data mining, natural language processing, image processing and artificial intelligence. He has collaborated actively with researchers in several other disciplines of engineering and information technology.

Myrna A. Coliat is an Associate Professor at Batangas State University, where she is the Director for Institutional Planning and Development. She obtained her Master's degree in Computer Science from De La Salle University (DLSU) and is presently pursuing her Doctor of Technology at Batangas State University.