Application of C5.0 Algorithm to Flu Prediction Using Twitter Data

[1]LZ Albances, [2]Beatrice Anne Bungar, [3]Jannah Patrize Patio, [4]Rio Jan Marty Sevilla, [5]*Donata Acula
University of Santo Tomas
Manila, Philippines
[1]lz.albances.iics@ust.edu.ph, [2]beatriceanne.bungar.iics@ust.edu.ph, [3]jannahpatrize.patio.iics@ust.edu.ph,
[4]riojan.sevilla@ust.edu.ph, [5]*ddacula@ust.edu.ph

***Abstract*—Since one's health is a factor considered, data coming from Twitter, one of the most popular social media platforms often used by millions of people, is beneficial for predictions of certain diseases. The researchers created a system that will improve the precision rate of the current system conducted by Santos and Matos using C5.0 algorithm instead of Naive Bayes algorithm for classifying tweets with flu or without flu. For the testing part, a total of 1000 tweets which is only limited within the Philippines were gathered to evaluate the system. Moreover, both English and Tagalog tweets are included in the dataset. The researchers found that the proposed system, after examination, has achieved a rate of 62.40% in terms of precision, and 66% in terms of accuracy. It was concluded that the C5.0 algorithm is less precise but more accurate than the Naïve Bayes algorithm.**

Keywords—C5.0 Algorithm; Naive Bayes Classification Algorithm; Flu Prediction

I. INTRODUCTION

To quote Dewing [8], "the term "social media" refers to the wide range of Internet-based and mobile services that allow users to participate in online exchanges, contribute user-created content or join online communities".

At present, social media is considered important not only because it is used as a medium of communication, but it is also used to gain specific and relevant information.

Social media is growing rampant. Millions of people are currently using it to interact, communicate, and express themselves online. These individuals use social media as a part of their daily routine.

People post or "tweet" about how and what they are doing on a daily basis - including their thoughts, opinions, feelings, and current status in life. These tweets are also updated by mobile phones through which the researchers can trace the user's exact location and its weather condition.

According to Grover and Aujla, "Twitter provides free APIs from which a sampled view can be easily obtained" [3]. It helps build a map spread model. In line with this, the researchers came up with an idea of working on a research using these Twitter data, focusing on the so-called "tweets" regarding diseases and health, specifically, the flu or influenza.

Furthermore, previous studies were made regarding this topic. In 2008, there was a website called "Google Flu Trends" that took data from Google, Facebook, and Twitter itself [5]. It was widely used in its release, but it was proven to have inconsistencies with regard to its data [5].

Back in 2011-2013, Google Flu Trends overestimated the number of people who had flu by 50% [4]. One possible cause was its inclusion of data which contained the word "flu" or "influenza", wherein the person who "tweeted" or posted about it did not have the flu at all [4].

There was also a study that made a precision of 0.78 or 78% using the Naïve Bayes Classification algorithm to identify tweets that mention flu or flu-like symptoms [6]. As mentioned earlier, previous researchers had difficulty classifying these tweets as some data gathered got accepted even though the person did not have it.

According to the Department of Health of the Philippines [1], getting flu is possible by contacting items that are contaminated through the discharges of an infected person. In addition, the Department of Health mentioned that a person can get flu by the entry of influenza virus in the respiratory tract of a person when someone coughs or sneezes. Moreover, they said that all people who have a weakened immune system can be at risk of flu.

Furthermore, the World Health Organization stated that being infected with this disease may lead a person to death or may contribute to his or her death [7]. However, an early detection of a disease may prevent or may delay some problems with the disease [2]. With all this information in mind, it is important to improve the current solution for predicting flu in order to limit the spreading of this disease as well as to make people's lives healthier and better.

II. CONCEPT, MODEL, AND/OR METHODOLOGY

As shown in Figure 1, tweets were gathered by tracking words related to flu through Twitter API. These gathered tweets serve as the data of the new system. Afterwards, these tweets undergo the preprocessing process. These were cleaned by removing any Unicode format, hashtags, @USER, links or URLs, and punctuations. These words in the tweets were converted to lowercase. With the use of the Natural Language Toolkit (NLTK) in Python, tweets were tokenized in order for it to be useful in the classification process. Words that are not related to flu prediction like stop words would then be removed to reduce the unneeded data that could be redundant. Stemming was done to make the word in each tweet be in its base form using the NLTK. Feature selection and extraction technique were used to convert the tweets into an m by n matrix using the scikit-learn in Python. Subsequently, the researchers used 10-fold cross-validation to know which tweets would serve as the training data and test data. The training data would then be inserted in the Naive Bayes Algorithm and C5.0 algorithm

978-1-5386-4710-3/18/$31.00 ©2018 IEEE

followed by the test data in the new system. The output of the classifiers was used to determine the number of flu incidence rate in the Philippines which is considered as the output for this system.

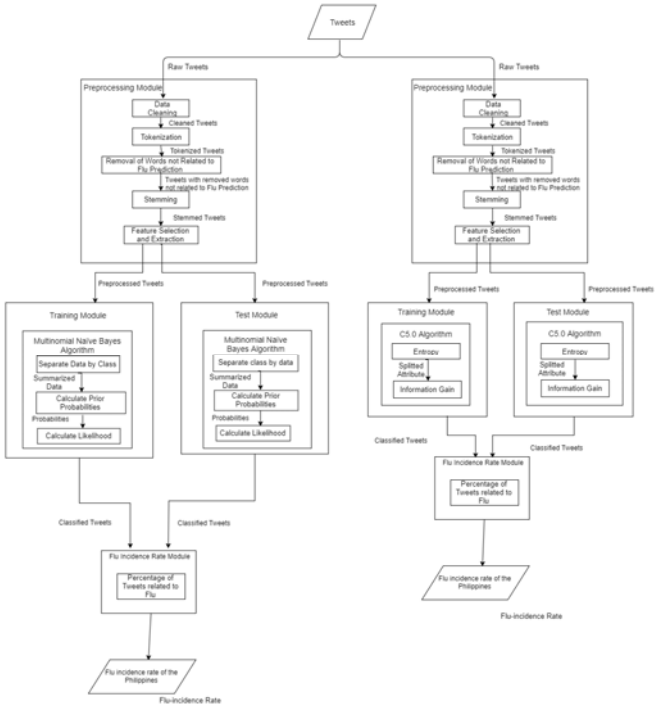

Fig. 1. System Architecture of the study

III. Experiment, Simulation, and/or Analysis

TABLE I. Summary of Results

Algorithm	Mean Accuracy of Folds (10-Fold)	Big-O (Time and Space Complexity)	Precision
Naïve Bayes	54.9%	O(n)	100%
C5.0	66.0%	$O(mn^2)$	62.40%

Listed in Table 1 is the summary of accuracy, big-o notation, and precision of the two algorithms. In terms of precision and efficiency, the Naive Bayes algorithm is better than the C5.0 algorithm. However, the C5.0 is better in terms of accuracy yielding 66%.

In line with this, the proponent's system is more accurate but less precise and efficient. In the Naive Bayes algorithm, the researchers have noticed that by calculating the probabilities of both classes, class 0 or "without flu" has a higher probability as compared to class 1 or "with flu" due to a certain fact that as they further analyzed the dataset, there were several tweets containing the keyword "sakit" and "ulo" which affected the probabilities of the classes. With this, the probability of getting a class 0 as a classification for the other tweets is high as well that led for these to be classified as without flu.

Meanwhile, for the C5.0 Algorithm, since the information gain is computed per attribute, there are instances wherein the decision is made based on the root or the first nodes without checking the rest of the attribute. This is the reason why C5.0 algorithm predicted wrongly in some of the tweets but somehow outperformed the naive in terms of accuracy.

TABLE II. Difference Accuracies Calculations

C5.0 Algorithm	Naïve Bayes Algorithm	Difference Between Accuracies
0.69	0.55	0.14
0.70	0.48	0.22
0.64	0.54	0.10
0.62	0.54	0.08
0.66	0.52	0.14
0.61	0.59	0.02
0.63	0.56	0.07
0.75	0.50	0.25
0.68	0.55	0.13
0.66	0.66	0.00
$Mean_{C5.0} = 0.664$	$Mean_{NB} = 0.549$	$Mean_{Diff} = 0.115$
Standard $Deviation_{C5.0} = 0.042479$	Standard $Deviation_{NB} = 0.049766$	Standard $Deviation_{Diff} = 0.079197$
		$S_x = 0.025044$

The first and second columns of Table 2 show the accuracy per fold of C5.0 Algorithm and Naïve Bayes Classification Algorithm respectively. Meanwhile, the third column shows the difference between these accuracies. In addition, means of the accuracy of the folds in C5.0 Algorithm and Naïve Bayes Algorithm are $Mean_{C5.0}$ and $Mean_{NB}$, as shown above. Whereas, $Mean_{Diff}$ is the difference between $Mean_{C5.0}$ and $Mean_{NB}$. Moreover, the standard deviation of the accuracy of the folds in C5.0 Algorithm and Naïve Bayes Algorithm are Standard $Deviation_{C5.0}$ and Standard $Deviation_{NB}$ respectively. On the other hand, Standard $Deviation_{Diff}$ is the sample standard deviation of the differences of the accuracies. Further, S_x is the estimated standard error of the mean.

The researchers tested the hypothesis that *there is a significant increase in the proposed solution in predicting flu as compared to Naïve Bayes Classification algorithm*.

Using the results shown in table 2, paired t-test is employed using a significance level of $\alpha = 0.05$. A p-value 0.00065 is obtained, which leads to the rejection of null hypothesis.

IV. Conclusion

The goal of the researchers for this study is to outperform the Naive Bayes Classifier in flu prediction using their proposed solution which uses the C5.0 Algorithm.

Santos and Matos predicted influenza with the use of Twitter data that has a precision of 0.78 using Naive Bayes Classification algorithm and this implementation is not high enough.

Specifically, it aims to answer the following questions:

1. Will the proposed solution accurately predict flu with the use of Twitter data?
2. How efficient is the C5.0 algorithm in predicting flu?
3. Will there be a significant increase in the accuracy of the proposed solution in predicting flu as compared to Naïve Bayes Classification algorithm?

Based on the analysis, the following conclusions were deduced:

1. The proposed solution, C5.0 Algorithm, has a better accuracy as compared to the Naive Bayes classifier.
2. The C5.0 Algorithm is less efficient in terms of the Big O notation since Naive Bayes' complexity is linear as compared to the C5.0 Algorithm.
3. There is a significant increase of accuracy in the proposed solution in predicting flu as compared to the current solution.

ACKNOWLEDGMENT

First, the researchers would like to express their deepest gratitude to Almighty God for giving them the strength, courage, knowledge and all the opportunities that they needed, also for being their inspiration to finish writing this thesis.

Second, the researchers would like to express their deepest gratitude to their thesis adviser, Asst. Prof. Donata Acula, who continuously motivated and supported the group during hardships throughout the completion of the study.

Third, the researchers would like to thank their thesis coordinator, Asst. Prof. Cecil Jose Delfinado, for being patient and considerate and also for guiding them throughout the study.

Besides their adviser and coordinator, the researchers would like to thank their panel members: Ms. Cherry Estabillo, Ms. Charmaine Ponay, and Ms. Ria Sagum, for sharing their insights and knowledge for the betterment of this study.

Lastly, the researchers would like to thank their families and friends, particularly Cinderella M. Alvarez, for their moral support and their inspiring messages all throughout the study.

REFERENCES

[1] "DOH: How to prevent the spread of influenza | GOVPH," 28 May 2014. [Online]. Available: http://www.officialgazette.gov.ph/2014/05/28/doh-how-to-prevent-the-spread-of-influenza/. [Accessed 8 April 2017].

[2] N. M. Ferguson, D. A. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn and D. S. Burke, "Strategies for containing an emerging influenza pandemic in Southeast Asia," *Nature,* vol. 437, no. 7056, pp. 209-214, 2005.

[3] S. Grover and G. S. Aujla, "Prediction model for influenza epidemic based on Twitter data," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 3, no. 7, pp. 7541-7545, 2014.

[4] P. C. Hung, Ed., Big Data Applications and Use Cases, Oshawa, Ontario: Spring International Publishing AG Switzerland, 2016.

[5] D. Lazer and R. Kennedy, "What can we learn from the epic failure of Google Flu Trends," 1 October 2015. [Online]. Available: https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/. [Accessed 3 February 2017].

[6] J. Santos and S. Matos, "Analysing Twitter and web queries for flu trend prediction," *Theoretical Biology and Medical Modelling,* vol. 11, no. Suppl 1, p. S6, 2014.

[7] "Influenza (seasonal)," [Online]. Available: http://www.who.int/mediacentre/factsheets/fs211/en/.. [Accessed 8 April 2017].

[8] M. Dewing, Social Media: An Introduction, Ottawa: Library of Parliament, 2010.