ISSN 2278-3091
Volume 9, No.3, May - June 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse125932020.pdf
https://doi.org/10.30534/ijatcse/2020/125932020

A Semantic Approach of the Naïve Bayes Classification Algorithm

Juvi C. Tesoro[1], Michael Joseph M. Buen[2], Romeo C. Sullera, Jr.[3], Meljohn V. Aborde[4],
[1-4]College of Computing Education, University of Mindanao, Davao City, Davao del Sur, Philippines
juvi_tesoro@umindanao.edu.ph[1], michaeljoseph_buen@umindanao.edu.ph[2],
romeo_sullera@umindanao.edu.ph[3], mjaborde@umindanao.edu.ph[4]

ABSTRACT

This study aims to study the semantic approach of Naïve Bayes Classification Algorithm. From a statistical, probabilistic machine learning model, the classical decision-level classification algorithm which is the Naïve Bayes classifier shows to be efficient on a variety of sentiment classification problems. Naive Bayes is often used in sentiment classification applications and practical experiments because of its simplicity and effectiveness. However, its performance is often degraded because of the reliability of the result. This paper focuses on developing a different approach to the primary sentiment analysis of the NB classifier. The approach leads to the implementation of providing semantic information from lexicon resources together with the semantic calculator. Addressing the problems of the baseline algorithm show that it can be solved by incorporating other algorithm approaches. The comparative results show that the semantic approach is statistically superior and yielded improvements to the baseline classifier. In comparison with the baseline NB algorithm, SNB achieved a relatively favorable classification accuracy with the threshold of 50% to 60%and an average improvement of 11.394% for its accuracy rating while significantly reducing the training time.

Key words: Naïve Bayes, lexicon-based, semantic approach, Semi Naïve Bayes

1. INTRODUCTION

Sentiment analysis is a classification method under data mining that predicts qualitative response from a document and categorizing them into its polarity. It is a task that identifies the class labels for exemplification based on sets of features [1].Sentiment analysis is the most popular task in classification. It involves extracting information as well as hidden meanings from an individual using an electronic source or document. As the words imply, it is to detect the sentiment or emotion of a person and give a proper response to it. This concept was a center of attention because of its array of application especially in business industries [2].

The Naïve Bayes (NB) classifier is a highly popular algorithm used in classification mostly in categorizing text data sets because of its simplicity. The significant advantage of Naïve Bayes is it only requires low memory in processing and less time for execution. This algorithm is the baseline algorithm for researches in decision-level classification problems. The most common application of this algorithm is categorizing data into positive and negative sentiments. It is used as a classifier in various real-world issues like Sentiment Analysis, Email Spam Detection, Email Auto Grouping, Email Sorting by Priority, Document Categorization and Sexually Explicit Content Detection [3].

However, the NB algorithm needs to be improved because of its simplicity and also to adapt to the changing needs of today and in the future. An area of concern in this algorithm is the ambiguity of the context of data when it encounters or reads a large chunk of text. It will lose its accuracy on rating the encountered large texts. While it is good to determine positive or negative sentiments on short statements, it cannot identify if the reports are neutral. The NB relies on the keywords from the bag of words, but it could hardly capture the intensity of the statement from a long text such as "very good," "very bad," and others text, and the negation such as "not good." These terms are not optimal because it will lose its accuracy.

Along with other information management challenges, classification of e-mail messages is becoming more important especially in business. E-mail classification is a way of flagging or tagging messages as being of a particular type. For example, a message might be sorted in priority as "high," "medium," or "low" based on the hidden meaning or sentiment on the words. In more complicated cases, message classifications may be hierarchical or relevant to only some people in the organization. Therefore, the purpose of classifying email is to make the handling of messages more precise and more accessible. With machine-assisted classification, it will limit human involvement. Once configured, the machine does all the heavy lifting. Potentially all inbound, outbound, and internal messages can be processed automatically.

This study enhances the Naïve Bayes algorithm using a semantic analysis approach on classification and is applied in email dataset. This paper explores different method of

improving the accuracy, efficiency, and functionality of the Naïve Bayes Classification Algorithm. A semantic is described as the meaning or the interpretation of the sense of a word, sign, sentence, and others. The approach is based on the definition to interpret the meaning of words or sentence by incorporating lexicon-based methods which define a set of words that have an equivalent score and a Semantic Orientation Calculator that will calculate the intensity score of words and thus determine the result of the sentiment analysis. This study also, intends to apply a Semantic Approach for the Naïve Bayes Classification Algorithm and implement it on emails from personal collection and Enron's email dataset.

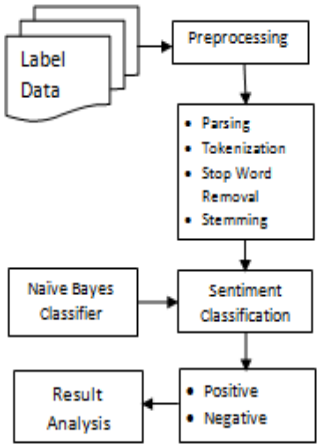

Figure 1: The baseline Naïve Bayes Classification process flow

Figure 1 gives an architectural overview of the preprocessing and the sentiment classification flow of the NB algorithm. This section presents the baseline algorithm of the Naïve Bayes classifier. In sentiment classification, the one related component is the pre-processing. To avoid incorrect and ambiguous result, data must be processed before training and analysis. A typical dataset contains the different variation of symbols, abbreviations, and non-standard language.

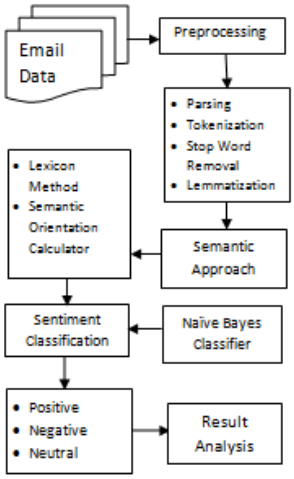

Figure 2: The proposed Semantic Approach to the Naïve Bayes Classifier process flow

The proposed Semantic Approach of NB is expected to enhance the original NB. A Python web application is used to test its implementation. The Enron's dataset undergoes preprocessing, transforming the data into the format as per requirement followed by the feature extraction with a semantic approach process to extract desired features. The improved algorithm determines the sentiment of the dataset and classify its polarity if it is negative, neutral and positive. The program displays the percentage result of the original NB and is compared as against the improvedNB with semantic approach. See Figure 2 for the proposed Semantic Approach to the Naïve Bayes Classifier process flow.

2. RELATED LITERATURE

2.1 A Comparative Study of Classification Algorithms used in Sentiment Analysis

The field of data mining has grown exponentially in the past and today. Sentiment analysis is a process of identifying polarity in a given text using different text processing and classification. The paper presented a comparative study of different classifying algorithms that commonly used in sentiment analysis. It discussed the pros and cons of each algorithm. The study shows the weakness of the Naïve Bayes classifier in comparison with other algorithms such as SVM[1].

2.2 Improving the Naive Bayes Classifier Using Conditional Probabilities

The Naïve Bayes classification algorithm is considered a primary classifier and has proven to be efficient on data classification tasks. However, as the study suggests, the Naïve Bayes features a strong independent assumption among the features given the category which is often violated considering the different applications in text classification. The study aimed to develop an improvement to Naïve Bayes which alleviates the assumption of independent features which utilizes conditional probabilities as it has been a concern for the improvement of Naïve Bayes[2].

2.3 A Naïve-Bayes Strategy for Sentiment Analysis on English Tweets

Established as a strategy based on the Naïve Bayes classifier to detect the polarity of English Tweets. As Twitter could be seen as an abundant source of short texts, the task of making positive and negative sentiment analysis from tweets is a hard challenge for this study, much more for a tweet without polarity or neutral sentiment. The study performs an experimental approach to achieve the best performance using a binary classifier between two polarity categories: the negative and positive. Identification of the fundamental polarity of context which is the positive and negative is well suited to be performed by the classifier. In detecting neutral tweets or tweets without a specific polarity, this study used a polarity condition for the tweet that has no identified overall polarity and is classified under the Neutral category[4].

2.4 A Comparison of Lexicon-based Approaches for Sentiment Analysis of Microblog Posts

A fine-grained approach to lexicon-based approaches for sentiment classification in which a given tweet is split into several micro-phrases using specific cues that occur in the message which aims to break down each part for faster analysis of sentiment. The study used specific prompts for extraction such as punctuations, conjunctions, etc. Each of the terms in the message will also implement a sentiment score as its representation of polarity which is part of the concept of weighted words. The collective polarity from each extracted phrase will then represent the overall sentiment of a single post [5].

2.5 Sentiment Classification by Semantic Orientation Using SentiWordNet Lexicon from Online Customer Reviews

A method which utilized semantic approaches in sentence-level sentiment analysis for evaluation of customer reviews from online sources. The rule-based method which is also considered to be domain-dependent identifies and decides the polarity and subjectivity of a single sentence from analysis of lexical resources. The study utilizes the SentiWordNet lexicon in which it will assign a specific weight to a single word that will define the polarity of the review by calculating the weights from each sentence[6].

2.6 A Deep Analysis on Aspect based Sentiment Text Classification Approaches

Feedback plays an important role with the evolution of a new version of a product or service. In business that invested majority of its operation to customers, analyzing feedbacks could be overwhelming. Likewise, identifying feedbacks from employees performance is a challenging tasks. The study utilized multiple techniques in sentiment analysis and utilized different datasets. The study showed that comparing the results would be unjustifiable since it comes from different domains. But the study also gave emphasis on understanding the contribution of every approach in doing sentiment analysis [8].

2.7 A Naïve Bayes Sentiment Analysis for Fintech Mobile Application User Review in Indonesia

Classifying user reviews and knowing user opinion in Realtime can help Fintech industries in facing their competitors in the market. In this study, the Naïve Bayes algorithm was used to do sentiment analysis in Bahasa Indonesia and in English. Two datasets were used in this study, both tested with its accuracy in doing sentiment analysis. Both datasets also implemented data cleansing. Results showed that a 3% margin of accuracy with Bahasa Indonesia and English compared with Bahasa Only [9].

3. METHODOLOGY

3.1 Data Gathering

First, the collection of the Enron corpus in the format of a CSV file retrieved from the Carnegie Mellon School of Computer Science website[7]will serve as the data set for this study and then will undergo preprocessing. The study utilized the Enron email **c**orpus which consists of over 500,000 emails from 150 employees of the Enron's Corporation. From the dataset, this study developed a module to extract certain parts of the corpus into the CSV file format to be fed into the system which is then analyzed and tested. The information from the email contains the sender and receiver email addresses, the subject of the email, the date received and the message content to be exclusively used in the study. Next, in applying the classification phase, which is the baseline Naïve Bayes algorithm, this study integrated another preprocess procedure that consists of the lexicon dictionary and calculator methods. Lastly, the semantic approach is implemented to obtain information about new approaches and techniques on refinement of the baseline algorithm to further improve the results.

3.2 Data Analysis

The process of data source gathering is then succeeded by the feature extraction and selection of the dataset which the study will undertake in data analysis. To achieve the best representation of the text dataset as a feature vector, this study employed various techniques before the fitting model of the machine learning for data analysis.

3.2.1Features

In natural language processing, the commonly used model is called the ***bag-of-words*** model. The idea behind this model is the creation of vocabulary. The collection of different words that occurs in the training dataset and each word is associated with a count of when it occurs. This bag of words can be understood as an unordered set word of non-redundant items where the order doesn't matter. The only concern of this model is whether the known words occur in the training set and is not in the testing dataset. In addition to data analysis, some unrecognizable stop words which would appear to be irrelevant are excluded from the vocabulary entirely but in most cases, stop words add semantic values to the context of the word and provide very poor context on their own.

3.2.1.1 Feature Vector Formation

Let *T1* and *T2* be two the text data samples in the training dataset:

- *T1* = "I think Fletch good CPA I still"
- *T2* = "Chris What latest PG E We good discussions regarding EOL Call Phillip"

Based on these two samples, the vocabulary could be written as:

V = {I:2, think:1, Fletch:1, good:2, CPA:1, still:1, Chris:1, What:1, latest:1, PG:1, E:1, We :1, discussions:1, regarding:1, EOL:1, Call:1, Philip:1}

Feature vectors for the individual sample sentence is shown as where the dimensionality of different words is equal to words in the vocabulary. Each vectors form represents the count of each word. In this case, each sample sentence will generate 17 elements of vectors.

Table 1: The Bag-of-words representation

V	I	think	Fletch	good	CPA	still	Chris	E	What
T1	2	1	1	1	1	1	0	0	0
T2	0	0	0	1	0	0	1	1	1
∑	2	1	1	2	1	1	1	1	1

latest	PG	We	discussion	regarding	EOL	Call	Philip
0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1

Given an example on Table 1, each element represents the number of counts in occurrence for each word in the sample dataset. The feature vectors in the bag-of-words are count 1s or 2s if the word occurs in a particular sample and 0 otherwise.

Bag-of-words is one of the implementations that this study used to evaluate text documents by forming a vector that outlines its features and content.

3.3 Data Pre-processing

In preparation for text categorization and classification, the data preprocessing technique is the process of cleaning and filtering of data. The text from the email dataset usually contains huge noise and uninformative parts commonly found in spams and parts of less subjectivity in the whole message. Also, orientations of many texts or words do not have an impact in general. Keeping those words may affect the quality of the data. The dimensionality of the problem becomes high, and the classification will suffer less precision since each word in the text is equal to or represented as one dimension. Therefore, for improvement of the quality of data classification, the raw dataset will undergo a preprocessing procedure. Data preprocessing deals with the preparation that reduces and removes unnecessary words and punctuations and improves the efficiency of the dataset.

Table 2: Data Pre-processing of the Sample Email Data

Original Email	To: lynn.blair@enron.com,steve.january@enron.com,dan.fancler@enron.com From: raetta.zadow@enron.com Subject: Top imbalances for November 2001 Attached is a worksheet that shows the top imbalances for November 2001. If you have any questions, please let me know. Thanks, Raetta
Parsing	Attached is a worksheet that shows the top imbalances for November 2001. If you have any questions, please let me know. Thanks, Raetta
Tokenization	Attached is a worksheet that shows the top imbalances for November 2001 If you have any questions please let me know Thanks Raetta
Stop Words Removal	Attached worksheet shows top imbalances November 2001 If questions please let know Thanks Raetta
Lemmatization	Attach worksheet showtop imbalance November 2001 If question please let know Thank Raetta

The contents of Table 2 display the pre-processing techniques. From the original email dataset sample retrieved from the Enron's corpus, this study provided the filtering process technique. By utilization of these methods, it can produce an appropriate result in which it can now implement the proposed semantic approach.

3.4 Implementing the Semantic Approach

Implementing the semantic approach would be the main part of this study since it is focused on an experimental study on improving the efficiency and accuracy of sentiment classification through the semantic approach application on Naïve Bayes. This study applied two main approaches to sentiment analysis that involves calculating the sentiment from the semantic orientation of the word or the context of the phrases that occur in an email text data. Applying Lexicon Method is one of the two main approaches to sentiment classification. With this approach a dictionary of words that has a polarity is required, each word having been assigned a positive or negative sentiment value. Generally speaking, a piece of email text is represented as a bag-of-word in this lexicon-based approach. Aside from a sentiment value or polarity of a text, the context of a word is also usually taken into consideration that includes negation and intensification. From evidence based on perspective, it is known that some occurring words take a neutrally oriented meaning. By revising the polarity of that text that stands next to the negated word may also change the context of a word. To mitigate this issue, this study integrated another approach which utilizes the Semantic Orientation Calculator collection of functions and is the second main approach in this study. This approach formulates and takes into consideration negation and intensification that calculates the sentiment value of a negated word. The advantage in this approach is leaning towards a result that is more accurate in the manner of assigning the polarity value to the negated words. Once the polarity of words is identified and the context of the word is verified, a combining process is performed that includes the baseline algorithm which the Naïve Bayes classifier applied to produce the final prediction of the sentiment classification. The output of the process is represented as a positive, negative and neutral class label.

Integrating the proposed Semantic Approach: the lexicon-based method and the semantic orientation calculator collection of functions, this study accomplished the main part of this study which is implementation of the semantic approach to the Naïve Bayes classification algorithm.

3.5 Testing

To test and verify the efficiency of the proposed semantic approach to the baseline algorithm, this study performed an experimental classification with a number of preprocessed email text data from the dataset gathered.

3.5.1Training Data

The main idea of learning the classification experiment is through the instance of using a selection of training dataset. Given a volume amount of the dataset of over 500,000 emails from the Enron's Corpus, the efficiency of selecting the suitable dataset for training is tedious work for this study. To construct the best possible model for the training data, this study used some variation of entities from Enron's email corpus. However, submitting a large volume of training data for the development becomes even more difficult for this study and

resulted to the slowing down of the process of building the classifier. While the training data is added to the system, the study found out that the dataset suffers from what is considered as an imbalance with a proportion of more or less 10:1, which translates to 90% positive/neutral and the rest is 10% negative. The dataset is imbalanced due to unequal distribution between classes. Dataset imbalance became a bit challenging for this study in training the data. To address this problem, this study applied a resampling technique which is using the concept of the SMOTE method or the Synthetic Minority Over-sampling Technique. It can be done through over-sampling the minority labels and under-sampling the majorities which normalize the sensitivity of the classifier. Balancing distribution is the appropriate technique for cross-validation and for training and testing the data.

3.5.2Enron Corpus Email Testing

In testing the semantic approach to the algorithm, this study utilized the same dataset source which are the Enron Corpus emails and are separated from the training set accordingly. This study also tried to use other sources for testing such as movie reviews and from society opinion collection but has concluded that for appropriation and categorized as a best fit in this study, the Enron email dataset is ultimately used.

As observed, the Enron dataset contains a high number of emails which is estimated to be more than 500,000 from its users. This study allowed the decision to include only ten example data as displaying the whole dataset is deemed above the capabilities of the study.

Table 3: Enron Corpus Emails Classification Test Results Sample

Email	Sentiment
Lynn, In reviewing the Duke contract 27291, I did find that the rate was not correct for four months in the past year. I also reviewed Duke contract 27349 since it also had the same type of alternate to California deal. Two months on the second contract were not billed correctly. If Marketing can input the correct rates into the system, I can regenerate the invoices and bill Duke immediately. Please see the attached file for details. Thanks, Bert	negative
As you recall this list is to be used with the Bushton Processing shut down = plan to determine what receipts may have to be curtailed in the event of a = shut down of the Bushton processing plant. Eric Faucheaux is working with = ETS Measurement Technology to set up a process to regularly update this lis= t. In the future these updates will not be emailed, but will be available = on Optimization website, under Operating Guidelines. The Optimization team website is at: http://www.ets.enron.com/optimization/ A corp id and login are required to access the site.	negative
Chris, after reviewing TMS data, I can confirm that NNG's storage was allocated during the ID2 process on Nov. 14, 15, 16, 17, 18, 21, 22, 23, 24, & 25, only.	positive
Steve, Steve, and Lynn- Attached is a draft of our discussions from this morning for distribution to the parties. Please provide your input & comments. Mike	positive
I thought you both might want to work together for the Customer Service Teams. I hope this helps.	negative
Lynn attached find the notes I took at our meeting with Alliant this week. I would like to conduct a conference call with you to go over a few of the items which will require Gas Logistics' input. I'll call you at 2:30 today. Chris Sebesta Northern Natural Gas Office: 402-398-7064 Fax: 402-398-7413 chris.sebesta@enron.com	positive

Table 4: Significant words for semantic approach to sentiment classification

Result	Word	Modifier
positive	confirm, attached, built	-
negative	correct, shutdown	not

Table 3 shows the emails that represent the two features intended for categorization in this study. For the first ten emails from the dataset that undergone the initial training process yields the following classification results. The positively classified statements could be attributed to the context which contains the words, “confirm”, “attached” and “built”. The otherwise negatively classified data from identification of the context of the message contains words such as, “correct” and “shutdown” with an addition of a modifying word, “not” as exhibited on Table 4. The email dataset has already been preprocessed to obtain the lemmatized form of the following words. The semantic scoring of the significant words under consideration of the context which contributed to the classification to the following categories were to be performed.

Table 5: Initial training process time comparison

Algorithm	Process time
Naïve Bayes	3.0243s
Semantic NB	3.013s

Figure 1: Accuracy Score after initial testing of the Baseline and Semantic Approach on Naïve Bayes

The results from the initial classification of the emails from Table 3 into the positive and negative polarity using the Semantic Approach of Naïve Bayes yields more or less an 80% accuracy score for the first 10 trained data which reflects on how the training data set and the preprocessing as well as the over-sampling techniques could affect the results. The simple bag-of-words approach as part for concept of Naïve Bayes together with the application of lexicon-based methods and semantic scoring allows the context of the sentence to be considered in classification. This, in turn, allows the yielding of a much-improved categorization process. The process time for comparison displayed in Table 5 shows a slight difference which exhibits the Semantic NB taking the minor edge in terms of performance. The accuracy score comparison of the baseline against the Semantic Approach to Naïve Bayes shown on Figure 3 were retrieved from the web application.

4 RESULTS AND DISCUSSION

This study evaluated the results of the testing done with the proposed approach to the Naïve Bayes classifier opposed to its baseline algorithm within system development in terms of its classification accuracy and efficiency in the training of the Enron email corpus dataset. The proposed approach will be referred to as Semantic NB.

4.1 The Experimental Evaluation

The initial findings from testing done with two sets of data with both polarity sentiments and the proposed approach to the algorithm reflect the preliminary outcome of the accuracy ratings.

Table 6: Initial accuracy result comparison of the proposed approach from the baseline algorithm

Email Set No.	Naïve Bayes	Semantic NB
First set	60%	80%
Second set	60%	80%

The result of identifying the classification accuracy of the fitted model on the Enron email dataset with the implementation of the baseline Naïve Bayes and the proposed approach to the algorithm upon preliminary testing and training is shown on Table 6. It indicates that the proposed Semantic NB approach exceeded the traditional NB classifier for sentiment classification for the initial ten training dataset applied to two example testing sets which composed of 10 emails each. The classification for the two-testing set differed in terms of resulting polarity based on different exhibited context by each message on the datasets. By demonstrating the effectiveness of the proposed algorithm approach, it is observed that the traditional classifier obtained an accuracy rating of 60%.It can be noticed that by applying the semantic approach to baseline NB algorithm, the accuracy rating increased to 80% with a consistent result between two example testing sets.

Also, this study evaluated how the proposed approach to the algorithm performed with a larger scale training dataset and how the accuracy rating will respond relative to the volume of the dataset and the minority over-sampling technique used. After testing the algorithm approach with small-scale training data, the results were still not up to par. To gain an acceptable comparative result, this study performed an expansion to the size of the training dataset and investigate whether a consistent or better accuracy and performance can be achieved while performing the email sentiment classification with the proposed algorithm approach against the baseline NB algorithm. Further evaluation is required in this section to provide more accurate findings and evaluation overall.

4.2 Extended Experimental Testing

Due to the undesirable evaluation of the above testing process, empirical testing is also performed using an appropriate validation for dataset selection. Apart from the classification accuracy ratings, the training time and the amount of dataset are also considered for comparison in this section of experimental processing. This comparison is to assess whether the classification accuracies of the two classification models differ on a single dataset. This study obtained the data through random selection from the Enron email dataset and divided this data into eight category class distributions according to the number of emails contained in a single training dataset ranging from 50 to 2000 emails. The categorized dataset is used to test the classification performance of the proposed approach to the algorithm and the traditional classifier. The eight categorized training datasets were also evaluated upon two sample test email set for comparison of polarity classification. The evaluation results are presented in the following section to provide more insight for each categorized dataset.

Table 7: Devices Specifications for Training and Testing Process

Components	Device 1 (D1)	Device 2 (D2)
CPU	AMD A8-7600 APU ~3.1GHz	Intel Core i5-5200U ~2.20GHz x 4
GPU	Integrated Radeon R7	Intel HD Graphics 5500
RAM	8GB (2 x 4GB) 1866MHz	8GB (2 x 4GB) 1600MHz
OS	Windows 10 Home	Ubuntu 18.04.1 LTS

This study conducted the evaluations as part of the extending testing across two machines. The corresponding specifications and operating systems for the training and testing phases that may factor in the comparison are shown in Table 7.

The accuracy rating of each of the testing phases was obtained through the implemented Python library of NLTK which is based on the results of the classification with strong consideration of context and the most informative features of the testing dataset. The fundamental confusion matrix is also implemented based on true values derived from actual and predicted results of empirical testing and the classifiers which could show a variation from the utilization of the accuracy method from the NLTK library.

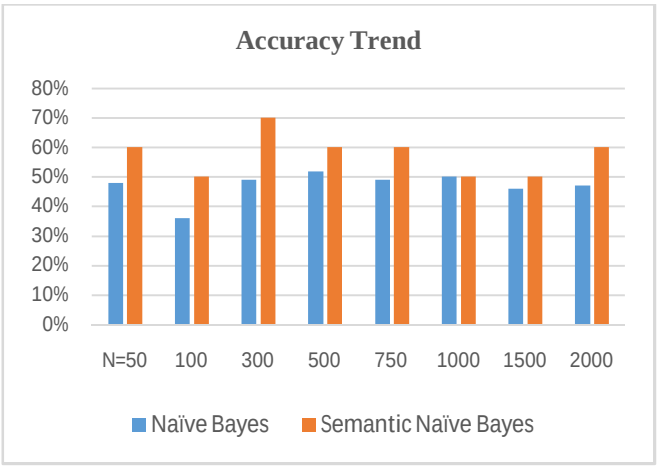

Figure 4: Accuracy Trend of Tested Algorithms

The accuracy percentage trend is depicted in Figure 4 which reflects on how the variance of accuracy between the baseline Naïve Bayes and the Semantic NB approach is changing as the number of data from the dataset is continuously trained and increased. An optimal number of data to be trained could also be retrieved in the trend chart which in this case is 1000 trained data which depicts that both approaches to the NB algorithm could be at par with each other and tells how the baseline Naïve Bayes classifier could be the most sensitive to the dataset without observance and consideration of the context of data and applying the simple bag-of-words approach in conducting the classification process.

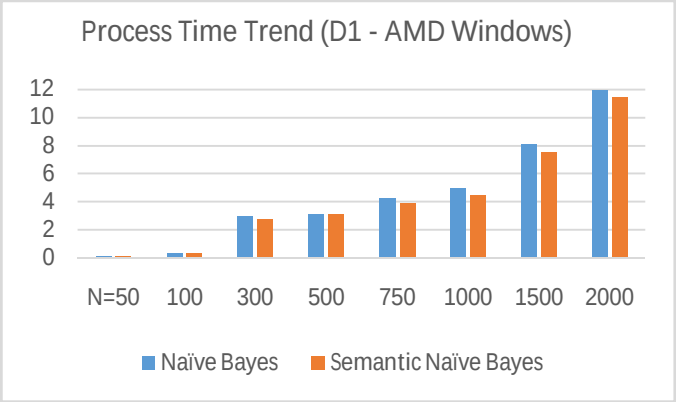

Figure 5: Process Time Trend of Primary Testing Device

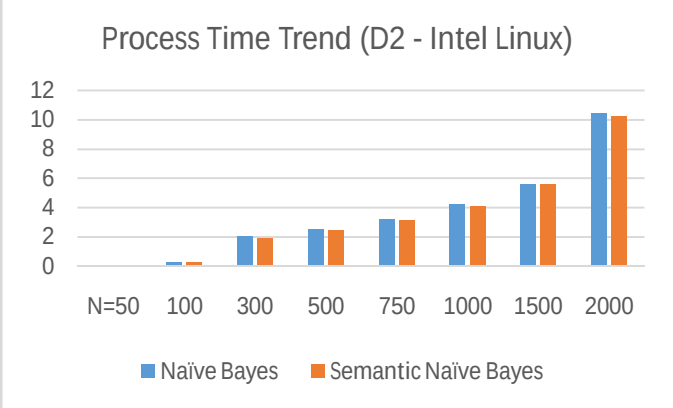

Figure 6: Process Time Trend of Secondary Testing Device

Figure 5 and 6 show the process time trend for both devices used for training and testing the Naïve Bayes and the Semantic NB approach which is measured in seconds represented in thousands. It reflected a consistent output which shows that the Semantic NB approach was able to report a lesser training time compared to the baseline NB algorithm. An optimal average percentage of increase for both devices was also evident and the second device in the Ubuntu OS was able to report a lesser process time overall in comparison to the first device in the Windows OS. The apparent efficiency of the second training device was attributed mostly to the specifications and also the operating systems used in training the data. The whole training process was able to take more time in less decent specifications for a device and less load on the processor.

In a partial conclusion after the final testing, the semantic approach yields a more accurate result consistently compared to the baseline algorithm, yet the appropriate number of data and clear context must be approximated. The best-case scenario in terms of the results is having an optimum number of the dataset used for training which inclines into incorporating a larger quantity which affects and influences the context for the semantic classification. The worst-case scenario is concluded to having an imbalanced dataset or training the data with less subjectivity for classification of its polarity and is much more evident in having a miniscule quantity of data. The processing time of the algorithm may also change which will reflect from the increasing and decreasing of the training dataset and the specifications of training devices.

5 CONCLUSION

The results after the experimental testing and comparison shows that both models, the baseline Naïve Bayes and the semantic approach generated a consistent accuracy percentage and time efficiency from its application in both the sample email sets. However, the Semantic NB yielded a better result in terms of consistency with the threshold of 50% to 60%and an average improvement of 11.394% for its accuracy rating and exhibited a significant change with a total of approximately 34 minutes of a difference in the processing time. Also, between the two devices tested posted an approximate difference of 1 hour and 30 minutes which addresses the objective of applying the semantic approach to Naïve Bayes for improvement of the algorithm particularly in terms of accuracy and efficiency. In terms of concern with the functionality, the semantic approach can also be fully utilized when using the approximate number of data for the training dataset which contains the appropriate context and structure fit for overall testing. Further down into the research, this study has also concluded that the Enron email dataset was an average fit in applying the algorithm which contains a substantial amount of data with context yet may contain unwanted noise which the algorithm may fail to analyze. In most cases, it could be concluded that the semantic approach to NB performed better than the baseline NB. This study can also conclude that the lexical resource and semantic scoring applied within creating the Python web application as an objective are factors that can mitigate the independence assumption of the traditional Naïve Bayes Classifier which addresses the objective for providing solution on the limitation of the baseline Naïve Bayes algorithm.

REFERENCES

[1] A. Gupte, S. Joshi, P. Gadgul, A. Kadam, "Comparative Study of Classification Algorithms used in Sentiment Analysis", *(IJCSIT) International Journal of Computer Science and Information Technologies*, Vol. 5 (5) , pp. 6261-6264, 2014

[2] P. Bogawar, K. Bhoyar,"Soft Computing Approaches to Classification of Emails for Sentiment Analysis", *ICIA-16: Proceedings of the International Conference on Informatics and Analytics,* Article No. 22,Pages 1-7, 2016
https://doi.org/10.1145/2980258.2980304

[3] S. Taheri, M. Mammadov, A. Bagirov, "Improving Naive Bayes Classifier Using Conditional Probabilities", *Retrieved from https://www.researchgate.net/profile/Musa_Mammadov2/publication/262394061_Improving_Naive_Bayes_classifier_using_conditional_probabilities/links/00b7d5380424ad3ce2000000.pdf*

[4] P. Ganallo, M. Garcia, "Citius: A Naïve-Bayes Strategy for Sentiment Analysis on English Tweets", *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 171-175, 2014
https://doi.org/10.3115/v1/S14-2026

[5] C. Musto, G. Semararo, M. Polignano, "A Comparison of Lexicon-based Approaches for Sentiment Analysis of Microblog Posts", Retrieved from *ceur-ws.org/Vol-1314/paper-06.pdf*, 2014

[6] M. Ghosh, G. Sanyal, "Sentiment Classification by Semantic Orientation Using SentiWordNet Lexicon from Online Customer Reviews",*International Journal of Computer Science Trends and Technology (IJCST)*, Volume 3 Issue 1, pages 206-213, 2015

[7] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, "Lexicon-Based Methods for Sentiment Analysis", Retrieved from *https://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00049*

[8] M. Syamala, N.J. Nalini, "A Deep Analysis on Aspect based Sentiment Text Classification Approaches", *International Journal of Advanced Trends in Computer Science and Engineering*, Volume 8, No.5, pages 1795-1801, 2019 https://doi.org/10.30534/ijatcse/2019/01852019

[9] R.R. Putra, M.E. Johan, E.R. Kaburuan, "A Naïve Bayes Sentiment Analysis for Fintech Mobile Application User Review in Indonesia", *International Journal of Advanced Trends in Computer Science and Engineering*, Volume 8, No.5, pages 1857-1860, 2019 https://doi.org/10.30534/ijatcse/2019/07852019