Comparison of Machine Learning Algorithms for Sentiment Classification on Fake News Detection

Yuzi Mahmud, Noor Sakinah Shaeeali, Sofianita Mutalib
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA
40450, Shah Alam, Selangor, Malaysia

***Abstract*—With the wide usage of World Wide Web (WWW) and social media platforms, fake news could become rampant among the users. They tend to create and share the news without knowing the authenticity of it. This would become the most critical issues among the societies due to the dissemination of false information. In that regard, fake news needs to be detected as early as possible to avoid negative influences on people who may rely on such information while making important decisions. The aim of this paper is to develop an automation of sentiment classifier model that could help individuals, or readers to understand the sentiment of the fake news immediately. The Cross-Industry Standard Process for Data Mining (CRISP-DM) process model has been applied for the research methodology. The dataset on fake news detection were collected from Kaggle website. The dataset was trained, tested, and validated with cross-validation and sampling methods. Then, comparison model performance using four machine learning algorithms which are Naïve Bayes, Logistic Regression, Support Vector Machine and Random Forest was constructed to investigate which algorithms has the most efficiency towards sentiment text classification performance. A comparison between 1000 and 2500 instances from the fake news dataset was analyzed using 200 and 500 tokens. The result showed that Random Forest (RF) achieved the highest accuracy compared to other machine learning algorithms.**

***Keywords*—Data mining; fake news; sentiment classification; supervised machine learning; text mining**

I. Introduction

Fake news is now viewed as one of the greatest threats to democracy, journalism, and freedom of expression. Fake news is typically produced by people, the so-called "fakesters", who generate an article with fake content often injected to an original real and trusted news content [1]. Although fake or satirical news can be less deceptive, intentional readers may still be deceived. Satirical news may deliberately establish a false expectation in the minds of readers through traditional meanings of dissatisfaction, taken as a face value. The untruthfulness is badly dissimulated and demand to be known [2]. According to Parikh and Atrey [3], researchers around the world have been very involved in the issue of fake news detection. Their studies have been carried out on the impact of fake news and how people react to it by viewing the title of the story, and cover image of the story. These factors might convince the readers about the content in the story or in news is realistic. Thus, the headline and image should be given more attention and take a step back and analyze the story or news after reading it so that readers might not believe the news fast enough. The fake news issues have become more popular after the Presidential election of U.S. which makes many researchers trying to find out better solutions for machine learning classification [4].

Sentiment analysis study has taken a long time. Sentiment analysis in science and development has been the main problem of today's world. As the number of users on social networking websites increases daily, enormous quantities of data are produced in text, audio, video, and images. Sentiment analysis as messages or posts must be carried out to decide if the sentiment is positive, negative, or neutral. Many automated classifiers are introduced to identify the text in the basic phrases, but new informal terms are applied to the current environment in the minimal spheres, which implies everything in the social realm [5].

This research focuses on filling the research gap between the machine learning algorithm and fake news challenges and assessments. Therefore, the research aims to perform research for automated prediction on fake news detection and investigate the performance of the machine learning technique to predict the fake news using text classification of the data. Manual analysis of the textual review can be frustrated and tedious. Some of data contains a lot of textual unrelated and unimportant message and this would be some challenges to define the best text representation for the textual classification.

In this research work, the textual classification and prediction can help an organization, a group of teams, or the other people to understand and expose more to the efficiently and effectively of fake news detection. Despite that, the automated natural language processing concept will be proposed and implemented that can be adaptive by the business or some organization to handle the hugely massive fake news textual data that show the genuinely comes from the truth sources. The remainder of this work is structured as follows: Section 2 presents the literature review. Section 3 explains the methodology of the research. Section 4 describes the result and discussions, and Section 5 explains the conclusion and future work that can be made to improve the research.

II. Literature Review

Sentiment analysis or opinion mining, as it is often called, is indeed one of the computational studies that discuss the analysis of opinion-oriented natural languages [6, 7]. These opinion-oriented work comprises, along with other aspects, gender disparities, emotion, and attitude detection, ranks, evaluations of significance, textual perspective, description of source documents, and descriptive opinion [8]. The sentiment

analysis puts together several fields of research, such as natural language processing, data mining and text mining. The purpose is to use artificial intelligence tools in the activities, and to simplify and develop their goods and services which are becoming extremely essential for the enterprises. The goal is to discover views of people articulated in the written language (text) in sentiment analysis or opinion mining [9].

Machine learning techniques are particularly effective for classifying sentiments in positive, negative, or neutral types for classified document [10]. Training and testing datasets are needed in machine learning techniques. A testing data collection is used to study the documents and to verify the accuracy of the evaluation dataset. To classify and evaluate the performance of fake news data, some machine learning techniques were utilized and modelled. Based on the literature findings, there are four common classification machine learning models that have been used in many research to build the model which are Support Vector Machine, Naïve Bayes, Logistic Regression and Random Forest classifier [e.g: 11, 12, 13].

Few studies on comparing the machine learning classification algorithms on fake news have been conducted. For example, Hasan, et al. [11] has performed lexicon-based sentiment analysis (W-WSD, SentiWordNet and TextBlob) with two machine learning algorithms, Naïve Bayes and Support Vector Machine. The finding shows that W-WSD has a better result when analyzing the Tweets. Another research was by Aphiwongsophon and Chongstitvatana [12] who have conducted the experiments using Naïve Bayes, Neural Network and Support Vector Machine classification algorithms to detect fake news. The result shows that Naïve Bayes has the accuracy of 96.08%, and Neural Network and SVM provide the accuracy of 99.09%. Next, Hiramath and Deshpande [13] proposed fake news detection system based on classification using Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, and Deep Neural Network for detecting fake news. The result shows that deep neural network is more crucial in detecting the fake news.

III. Research Methodology

This research is conducted using CRISP-DM methodology. CRISP-DM is a modelling process which provides a data mining framework that could be used in technology and industry sectors to improve cost-effectiveness, reliability, repeatability, and speed for large data mining projects [14]. Fig. 1 shows the six phases of CRISP-DM methodology namely business understanding, data understanding, data preparation, modelling, evaluation, and deployment. The explanation of each phase will be explained in the next subsections.

A. Phase 1: Business Understanding

Business understanding is the first place of this research area. In this phase, the main area that will be examined is issues that are related with fake or real news and the representation of text and the lexicon-based method [15] to preprocess the textual data that are consider the point and significant phase in this text mining project. The initial business understanding phases emphasis on understanding the business objective from business point of view, then changing over this knowledge into a research question, and after that create a research plan to accomplish the research objectives. This phase involves two activities. The first activity focuses on delivering research title, problem statement, research objectives, and research significance. While the second activity focuses on delivering the literature review to understand how previous scholars conduct the research in this area, for example what techniques have been used, what are the research findings, and what are the limitation of their research.

B. Phase 2: Data Understanding

This phase requires the researchers to obtain the required data and transformed it into a format that could be mined using data mining tools. Two activities are involved in this phase. The first activity is conducting data gathering. In this study, we used dataset from Kaggle website. However, this dataset might occur data incompleteness and data redundancy. Hence, several alternatives need to be applied to solve the problems, such as replace with alternative data source, gather new data, or narrow down the research scope. For this research, the dataset of Fake News Prediction consists of 12999 instances news with 20 attributes included the news title, authors, and others in year 2016 when the US President Election was happened. The second activity is to verify the quality of data. Up to this phase, data has been examined and studied, hence, it is crucial to confirm whether the data is good enough to support the objective of this research. Any missing value or error need to be identified and come out with the lists of action that can be taken to overcome this issue.

C. Phase 3: Data Preparation

Data preparation refers to preprocess the dataset for the modelling phase. This activity needs to perform multiple times to ensure the quality of the data. Fig. 2 illustrates the five stages of data preparation process which consists of data selection, data cleaning, data construction, data integration and data formatting. However, for this text mining research, the process is slightly different from the data mining process where some of the stages will happen in the middle of the modelling phases and not before. This is the flexibility of adapting CRISP-DM framework for textual mining research.

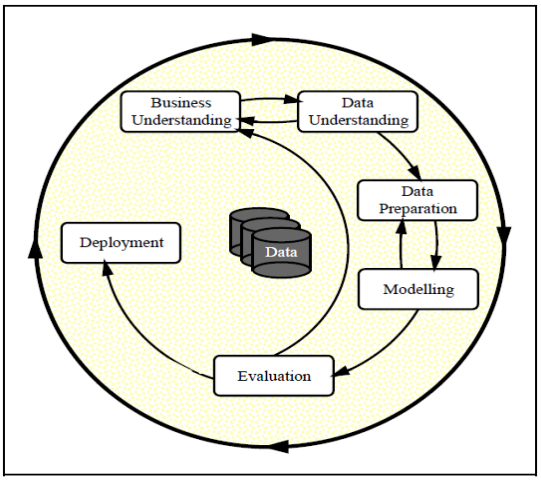


Fig. 1. Life Cycle of CRISP-DM [14].

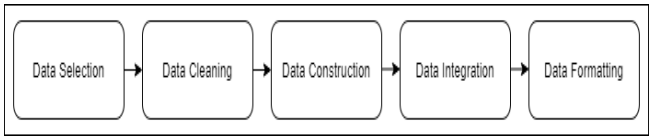


Fig. 2. Data Preparation Process.

D. Phase 4: Textual Data Labelling

Before the sentiment text classification performance can be compared, the model of the sentiment text classification must be constructed first. Fig. 3 shows the work flows of textual representation process.

The process starts with preprocessing the raw text data from the Fake News Dataset. Then, the data is transformed into a tokens and tags formation in Data Selection, which has been filtered and cleaned. When the data is ready, the preprocessed text will be sent to the sentiment analysis to label the sentiment from the review text data. Hu and Liu [16] sentiment analysis widget is based on the lexicon has been chosen for these activities. Unlike Vader method [17], Liu Hu method is simpler, which generate a single output of sentiment integer. However, the sentiment integer label is not represented as sentiment classification. The comparison result between the use of Liu Hu and Vader methods will be shown later in the results and discussions section. Therefore, the unsupervised label integer sentiment needs to classify using the hierarchical clustering technique. The purpose of this technique is to cluster integer sentiment that has close relations to classify into three groups of sentiment classification. Thus, the negative integer sentiment label will cluster under the negative values, the positive sentiment label will cluster under the positive values and the neutral sentiment label will cluster between -1 and 1 values.

E. Phase 5: Modelling

Orange data mining toolkit has been used in this phase to show the relationships between data in an understandable figure. The modelling method in this study is divided into three tasks, which consist of textual representation using a sentiment rating model, a content comparative model, and a predictive model. The process flow of architecture design in this research is shown in Fig. 4. The cleaned dataset derived from this phase was fed into the Designing Test task. There is one process occurred here, which is training process. To train the models, four machine learning algorithms are selected namely Naïve Bayes, Logistic Regression, Support Vector Machine and Random Forest. The selection of a best classification model is based on the highest coefficient accuracy result. The prediction model will use the same dataset during the model building phase. This is to produce the best classification model in predicting the fake news dataset.

A predictive model will be developed during this process. The Orange data mining prediction model is transparent and simple. Fig. 5 shows the prediction model process workflow. The prediction analysis uses the same data sets for classification model as the data set. The model is therefore nearly similar to the one for the performance analysis, in which the beginning component before the word bag is the same. The data sampler module is then used to separate the datasets into training and test results. The method of sampling is the one-on-10-fold testing of the selected fold. One-fold will select 2000 instances from 6525 instances of datasets of input. The data sample serves as the test data and connects to the prediction module. In the meantime, the rest of the data acts as training data and connects to the learning algorithms.

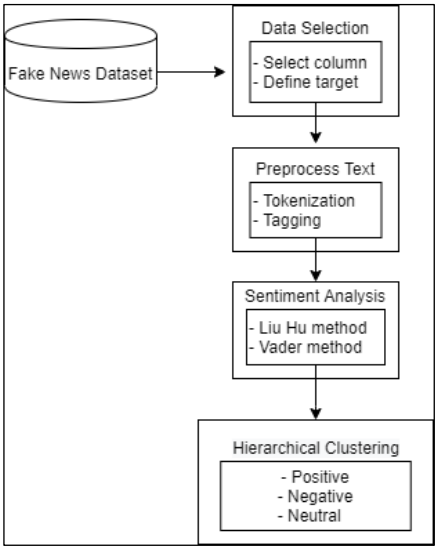


Fig. 3. Textual Representation Process Workflow.

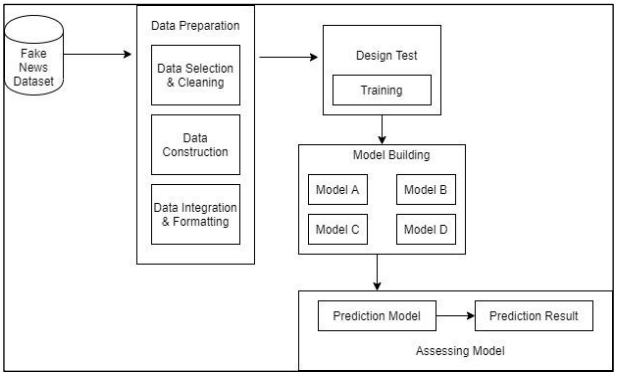


Fig. 4. The Process Flow of Architecture Design .

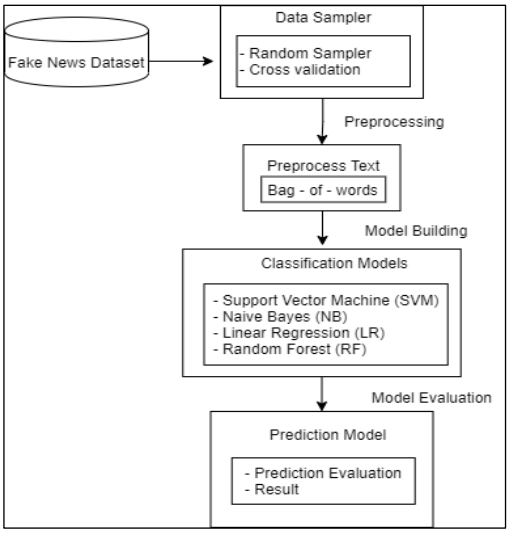


Fig. 5. Prediction Model Process Workflow.

Then, performance of each model is compared for sentiment classification. The process begins with the Data Sampler where the random sampler and ten-fold cross-validation will be done in this process using the same dataset. Then, the preprocess text activity and bag-of-words activity is combined to produce the clean data. For classification models, four accessible machine learning algorithms in the literature [18] has been determining to assess the performance of machine learning on sentiment classification. The selected machine learning algorithms are Support Vector Machine, Naive Bayes, Logistic Regression, and Random Forest. Each of classification model was developed through the training and testing process, following the supervised method. For that training and testing process, both the cross-validation approach was applied with 10 folds and hold out method through random data sampler, with 66% of dataset as training set and 34% of dataset as testing set.

F. Phase 5: Evaluation

In the evaluation phase, the performance of the predictive model is evaluated. The model is assessed in terms of precision, accuracy, recall and value of F1 classification. Error rates are used by supervised classification tasks to assess the consistency of data mining process. The dataset is also measured by the difference in the value of fixed data and the most common tokens for determining whether the data set size may affect the machine performance. If the process is failed, it is necessary to identify any possible reasons why the model did not satisfy the requirement. The data mining process also need to check thoroughly if there is existing additional process of iterations that can be made. The evaluation on the results from the comparison of the classification algorithms is performed on the dataset. In comparison, the assessment of datasets involves 10 cross validation directories and 10 random samples containing 60 per cent training results. This is an important experiment to determine how well the model can predict based on the data sets of training. There are two categories of performance measure that comprises of accuracy measures, as in (1) and error measures, as in (2). These performance measures will be discussed in the results and findings section.

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{error rate} = \frac{FP+FN}{TP+TN+FP+FN} \quad (2)$$

G. Phase 6: Deployment

Deployment is the final phase of the CRISP-DM approach. This phase requires all the process involves in this research are documented properly. Every part of the experimental results and comparative textual analysis from findings of this research are discusses and presents.

IV. Result and Discussion

This section discusses the results and findings of the applied classification methods that had been chosen. It also provides the analysis on the trained and testing results based on ten folds cross-validation and sampling methods of 66% training set and 34% testing set. The first experiment is performed to determine the label of each record in the Kaggle fake new dataset. This is a crucial process for textual classification using supervised method. The second experiment is focused on evaluating the performance of machine learning algorithms, which are Naïve Bayes, Logistic Regression, Support Vector Machine and Random Forest, for textual classification problems in the fake news data set. The third experiment is to compare the four machine learning algorithms with top frequent tokens. To compare the quality of the machine learning algorithms, the researchers used different performance measures such as accuracy score, precision, recall, and F-Score.

A. Experiment 1: Labelling of Textual Data using Lexicon Scores

Textual representation model experiment is more likely to a sentiment classification model where it has been conducted to prove the effective way to automated sentiment analysis through textual representation. Initially, the textual data and type of data is selected using select columns and corpus. Then, the raw data will be sent to preprocess text module which need to go through several stages to prepare the data for the sentiment analysis process. The text preprocessed activities generate 673862 based on the tokenization and uni-gram with bi-grams technique. Among all the token generated, 45183 types were identified as unique tokens. All the generated tokens will be feed into sentiment analysis widget to predict and label the sentiment on each news text. The sentiment classification labelling is conducted using lexicon-based dictionary approach by Liu Hu and Vader [16, 17] which produces the sentiment value. The sample of two lexicon-based sentiment is shown in Table I. Based on the sample of bias type, the first news text sentiment is bias with a negative value of sentiment. The second news text sentiment is also bias but with a positive value of sentiment.

The next process is to group the sentiment values to represent textual labels. Hierarchical clustering has used with distances to produce ten clusters contains positive sentiment values, negative sentiment values and lastly neutral sentiment values. A result of hierarchical clustering with Top-N = 10 selections for the three textual representation models for (a) Liu Hu [16] and (b) Vader [17] methods using ‘ward’ as the linkage between the attributes and 10 levels of pruning were compared. The 10 Top-N was chosen for this experiment because of the sentiment values show the random and mixture values of sentiments, so to make the next process easier, and an “Edit Domain” widget is used to group the 10-clusters into the three categories which is positive, negative, and neutral groups. Fig. 6 shows how the 10 Top-N cluster was categorized and labelled into three clusters group of positive, negative, and neutral sentiment according to sentiment values produced in hierarchical clustering chart in Fig. 7 (Liu Hu method) and Fig. 8 (Vader method).

TABLE I. SAMPLE OF LEXICON-BASED SENTIMENT

News Text	Type	Sentiment
Print They should pay all the back all the money plus interest. The entire family and everyone who came in with them need to be deported asap. Why did it take two years to bust them? Here we go again …another group stealing from the government and taxpayers!	Bias	-1.961
Share on Facebook You've got to hand it to this guy for such an ingenious, yet simple design. The how-to example in the video below is made from approximately 12 feet of copper tubing plus a few fittings (the stainless steel tube option is shown too). Follow the instructions in the video below to learn how to build it yourself. If a torch isn't something you have in your tool kit you can find "push on" fittings from a hardware store that you won't need to solder.	Bias	1.786
Today Dr. Duke and Dr. Slattery talked about Hillary's clear acts of treason against the United States by providing massive shipments of weapons to Saudi Arabia at a time that she knew they were providing support to ISIS. Dr. Duke, if elected to the Senate, would be in a position to expose Hillary and push for her impeachment should she win (steal) the election. BLOOD ON THE TRAITOR'S HANDS!	Hate	0.784
Today Dr. Duke discussed the state of his campaign, including television commercials that he was preparing. He will be in a televised debate with the other leading candidates, which should be critical in putting him in the run off. Pastor Mark Dankof took over the show at the break. He took calls from listeners. One call asked about Jesus's warning about the Synagogue of Satin. Pastor Dankof ended the show with a passionate warning about the risk of World War III should Hillary be elections. This is another great show that you won't want to miss.	Hate	-1.370
COLUMBUS, OH (AP) — History was made today in Columbus, Ohio when more than 3 million Amish poured into the city to see the American Amish Brotherhood (AAB), an organization which acts as an informal governing body for the Amish community, endorse Donald Trump for president. That number represents a significant portion of the total Amish population, which the United States Census Bureau says numbers more than 20 million men and women nationwide all pledging their vote to Trump for President.	Fake	0.796
64 SHARE President Obama has signed an Executive Order declaring an investigation into the election results and plans for a revote on December 19th. (AP Photo / Dennis System)	Fake	0.385

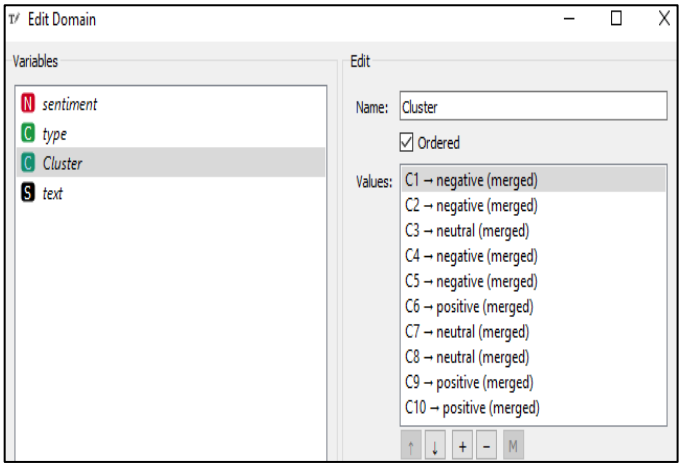
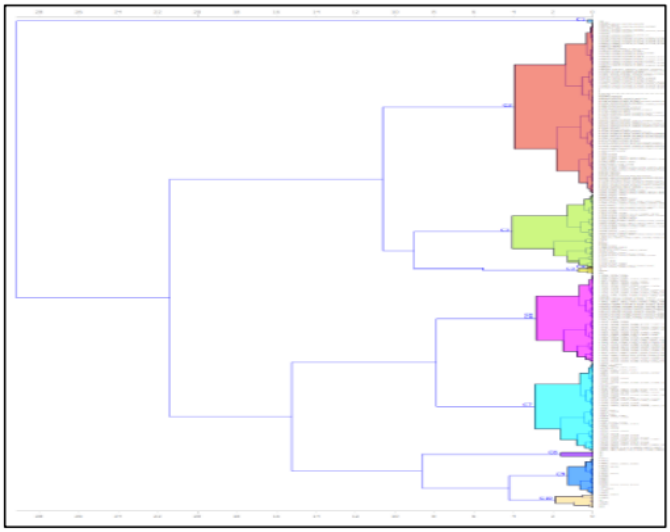


Fig. 6. Edit Domain Widget Process.

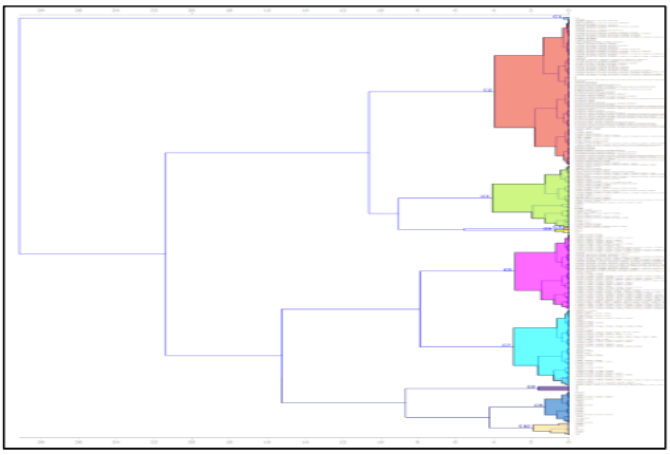
The comparison results of hierarchical clustering between Liu Hu [16] and Vader [17] methods does not show a lot of differences, but from other perception, Liu Hu method shows the most useful information that needed in this experiment as it uses lexicon-based sentiment analysis to classify each data in the dataset. Liu Hu method is easier because it shows the result of the sentiment directly and the values of sentiment for each cluster using the 'edit domain' widget. While for Vader method, the result from the hierarchical clustering shows a 'pos', 'neg', 'neu' and 'compound' values. The result from Vader method makes this experiment confusing because there is too much value of sentiment with the compound values that we define it was not useful for this experiment. So, for the next experiment Liu Hu's method will be chosen.

By referring to the text reviews in Table I for the bias texts, the negative and positive value sentiments are due to hierarchical clusters of Ward linkage combine with Euclidean on distances widget. The probability of correcting the precise opinion would evaluate the distance between two points in the line and then measure the number of squared differences within each of the clusters. This means that most of the result is close to the average neutral sentiment. The findings of the output sentiment can be acknowledged therefore by analyzing the dispersal plot between the output sentiment and the text types ranking scale.



title	type	text	Cluster	sentiment
1	bias	Print They shou...	positive	-1.961
2	bias	Why Did Attorn...	positive	-1.488
3	bias	Red State : Fox ...	positive	-1.460
4	bias	Email Kayla Mu...	negative	0.000
5	bias	Email HEALTH...	positive	-0.804
6	bias	Print Hillary go...	neutral	-5.490
7	bias	BREAKING! NY...	positive	-1.446
8	bias	BREAKING! NY...	neutral	-3.319
9	bias	Limbaugh said ...	positive	-1.224
10	bias	Email These pe...	positive	-1.316

Fig. 7. Hierarchical Clustering and Sample of Data Sentiment Values using Liu Hu Method.



title	type	text	Cluster	pos	neg	neu	compound
1	bias	Print They shou...	negative	0.113	0.09	0.797	0.3129
2	bias	Why Did Attorn...	negative	0.038	0.064	0.898	-0.7562
3	bias	Red State : Fox ...	negative	0.085	0.034	0.882	0.8955
4	bias	Email Kayla Mu...	negative	0.267	0.237	0.496	0.2375
5	bias	Email HEALTH...	negative	0.13	0.077	0.793	0.9664
6	bias	Print Hillary go...	positive	0.049	0.232	0.719	-0.9887
7	bias	BREAKING! NY...	negative	0.071	0.06	0.87	0.8119
8	bias	BREAKING! NY...	positive	0.041	0.09	0.869	-0.9305
9	bias	Limbaugh said ...	neutral	0.055	0.012	0.933	0.8151

Fig. 8. Hierarchical Clustering and Sample of Data Sentiment Values using Vader Method.

B. Experiment 2: Classification Model Construction

In the evaluation of prediction models, the same dataset was used to compare the output sentiment result of the automated sentiment classification based on the lexicon-based approach in Experiment 1, and the result of the prediction model using LR supervised machine learning in Experiment 3. The only different in dataset for Experiment 2 is the dataset will be split into two subsets of data using 'data sampler' widget for the training dataset and the testing dataset. Based on the confusion matrix for prediction model in Fig. 9, negative prediction on sentiment text classification is 100%, which shows that the accuracy is perfect. However, the percentage results for this dataset are likely to change after inserting more data. The highest value of misclassified is through neutral sentiment, which is 5.1%. However, the misclassified sentiment cannot assume as wrongly predicted. For example, in Table II, it shows the differences value of misclassified for Naïve Bayes is quite high but for correctly prediction, Naïve Bayes provides the highest accuracy among the other classifiers. Based on the situation, the researchers conclude that by using machine learning algorithms, the automated sentiment text classification can be improved.

Actual	negative	positive	neutral	Σ
negative	100.0 %	0.0 %	1.4 %	366
positive	0.0 %	100.0 %	3.7 %	83
neutral	0.0 %	0.0 %	94.9 %	801
Σ	354	52	844	1250

Fig. 9. Sample of Confusion Matrix of Prediction Model.

TABLE II. DATA AND PREDICTION USING MACHINE LEARNING CLASSIFIERS

Classifier	Support Vector Machine (SVM)	Random Forest (RF)	Naïve Bayes (NB)	Logistic Regression (LR)
Correctly Classified	99.8%	99.9%	84.0%	96.7%
Correctly prediction	99.6%	99.3%	99.7%	92.8%
Misclassified	12.0%	8.4%	56.4%	7.4%

C. Experiment 3: Comparison Performance Model with Top Frequent Tokens

The third experiment was conducted to compare four machine learning strategies for sentiment text classification with selected terms. The models are developed using a bag of words and four supervised machine learning technology, which includes Support Vector Machine, Naïve Bayes, Logistic Regression and Random Forest. The news text will be pre-processed by text and sent to bag-of-words to count the number of words occurring in the news text. Then two setups for experiments using four machine learning techniques of supervision are carried out. To assess the efficiency of this machine learning, four main efficiency indicators (KPI) will be evaluated which are classification accuracy, F-1 Score, Precision and Recall.

Table III shows the machine learning classification accuracy based on fixed proportion data and the number of tokens that are most common, as the first set up. The effect based on the number of tokens can be seen by comparing Set A for 1000 instances with 250 most frequent token, Set B for 1000 instances with 500 most frequent token, Set C for 2500 instances with 250 most frequent token and Set D for 2500 instances with 500 most frequent token.

TABLE III. CLASSIFICATION ACCURACY BASED ON SIZE DATA AND THE AMOUNT OF TOKENS

Label	Classifier	SVM	NB	LR	RF
Set A	50% proportion of data (1000 instances) with 250 most frequent tokens	96.2%	87.6%	94.9%	**98.7%**
Set B	50% proportion of data (1000 instances) with 500 most frequent tokens	**99.4%**	86.3%	93.4%	97.8%
Set C	50% proportion of data (2500 instances) with 250 most frequent tokens	55.6%	48.9%	45.1%	**58.4%**
Set D	50% proportion of data (2500 instances) with 500 most frequent tokens	49.2%	35.8%	54.6%	**60.2%**

The results show that the amount of the most frequent tokens (Set B and Set D) does not affect the classification accuracy, while the lowest number of tokens show a higher classification accuracy (Set A and Set C). By contrast, the effect based on the size of proportion data can be seen by comparing Set A with Set C and Set B with Set D. The

different of classification accuracy for comparing these set is quite high. Thus, this can conclude that the size of proportion data is affecting the classification accuracy but not heavy. One unanticipated finding was that all the classification accuracy of A is higher when the size proportion of data is bigger for Set A with Random Forest accuracy 98.7% except Set B for Naive Bayes, Set C for Logistic Regression and Set D for Naive Bayes. Surprisingly, Support Vector Machine has the highest achievable classification accuracy of 99.4% with a variable of high proportion data and high frequent token in Set B. However, with 2500 instances for 250 and 500 most frequent tokens in Set C and Set D show that the classification accuracy for all four methods was unsatisfied, whereby the classification accuracy for Set C shows that Random Forest with highest accuracy of 58.4% and the highest accuracy for Set D is 60.2% for Random Forest. In conclusion, variable Set B shows the best option for creating classification accuracy with the highest result and can be considered to adopt for the next experiment.

In the second set up of experiment, the machine learning performance were evaluated based on the sampling approach, which are ten folds cross-validation and ten repeat train or test random sampling with 60% training set size. The same data for the four KPIs will be analyzed in these experiments to studies the performance of a machine learning technique for textual sentiment classification. Later, the confusion matrix will be applied to observe the proportion between the actual and predicted class. With the confusion matrix, a misclassified instance can use to review the textual data in detail to discover the reason behind it. Table IV and Table V show the KPI of machine learning algorithms for textual sentiment classification using 1000 instances and 2500 instances. The results for show that the highest classification accuracy is 98.5% for 1000 instances and 99.9% for 2500 instances using Random Forest algorithm. By comparing between sampling type, the different of KPI for all algorithms are not much except for Naïve Bayes algorithm. All algorithms have slightly higher KPI result for using ten folds cross-validation sampling except for Naïve Bayes, which KPI result is better on ten folds cross-validation. As the conclusion from this experiment, both sampling approaches give almost the same value of classification accuracy results and do not contribute to classification performance.

For further analysis, the confusion matrix in Fig. 10 and Fig. 11 shows the comparison of the proportion of the predicted sentiment on the text datasets with 1000 instances and 2500 instances. The results show that most predicted correctly is positive sentiment of fake news for 1000 instances is 82.1%, while for 2500 instances is 98.8% of positive sentiment for fake news. However, neutral sentiment for fake news in Fig. 10 archived 100% accuracy, while Fig. 11 shows the highest accuracy of neutral sentiment is 99.7%. In contrast, negative sentiment for fake news in both result shows lower accuracy compared to the rest. In conclusion, the bigger proportion of data to be tested, the bigger accuracy can be classified and evaluated, compare to the small proportion of data that predicted as positive, negative, and neutral sentiment.

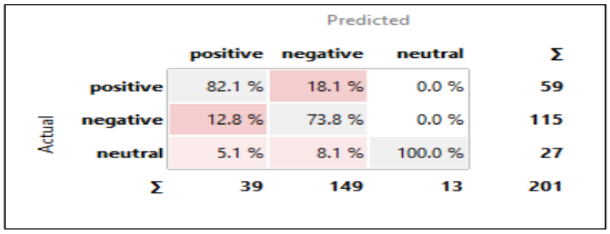


Fig. 10. Confusion Matrix of Predicted Sentiment using 1000 Instances.

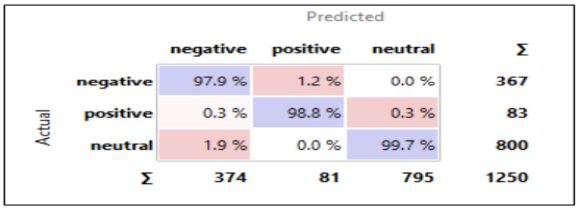


Fig. 11. Confusion Matrix of Predicted Sentiment using 2500 Instances.

TABLE IV. KPI OF MACHINE LEARNING ALGORITHMS FOR TEXTUAL SENTIMENT CLASSIFICATION USING 1000 INSTANCES

Machine Learning Algorithms	10-folds cross validation				10 repeat train/test random sampling with 60% training set size			
	CA	Precision	Recall	F1	CA	Precision	Recall	F1
Logistic Regression (LR)	**94.5%**	**94.9%**	**95.4%**	**94.3%**	**93.1%**	**93.4%**	**93.1%**	**92.8%**
Support Vector Machine (SVM)	**96.0%**	96.2%	96.0%	96.0%	97.5%	99.7%	97.5%	97.5%
Naïve Bayes (NB)	**84.5%**	87.6%	84.5%	84.9%	84.1%	86.3%	84.1%	84.7%
Random Forest (RF)	**98.5%**	98.5%	98.5%	98.5%	**97.8%**	97.8%	97.8%	97.8%

TABLE V. KPI OF MACHINE LEARNING ALGORITHMS FOR TEXTUAL SENTIMENT CLASSIFICATION USING 2500 INSTANCES

ML Algorithms	10-folds cross validation				10 repeat train/test random sampling with 60% training set size			
	CA	Precision	Recall	F1	CA	Precision	Recall	F1
Logistic Regression (LR)	**96.6%**	**96.7%**	**96.6%**	**96.3%**	**95.7%**	**96.0%**	**95.7%**	**95.3%**
Support Vector Machine (SVM)	**99.1%**	99.1%	99.1%	99.1%	99.4%	99.8%	99.8%	99.8%
Naïve Bayes (NB)	**89.0%**	84.0%	89.0%	86.0%	88.9%	83.9%	88.9%	84.7%
Random Forest (RF)	**99.9%**	99.9%	99.9%	99.9%	**99.8%**	99.8%	99.8%	99.8%

V. Conclusion and Recommendation

The study includes the comparison of four classification algorithms to evaluate the performance of classification accuracy in sentiment text classification. The evaluation process includes several aspects such as the size of data, amount of token, and test sampling approach. A 1000 instances dataset and 2500 instances dataset with different values of 250 and 500 most frequent tokens were applied in this experiment. The four classification algorithms used in this experiment are Support Vector Machine, Naïve Bayes, Random Forest and Logistic Regression. By that, a model was evaluated to measure the accurateness and the exactness of the model to make a prediction on a new dataset. Prior to the model evaluation, the model was able to predict all the fake news correctly, which makes the model reliable and trustworthy to be used to predict the fake news detection status.

However, there are some limitations in this research that need to be highlighted. Firstly, Orange toolkits have some technological weakness and limitation when managing massive databases. This work will therefore process randomly 1,000 fake news and 2500 fake news from the initial data sets out 12999 total of fake news during US Presidential Election in 2016. Secondly, this research focuses on the Fake News Detection dataset, which focused on a single objective. The research can be applied to another dataset such as from twitter or social media to mine the knowledge from the text deeper to understand its sentiment. Lastly, in the case of model validation using a cross-validation and/or an individual validation method, Orange Toolkits provide the facilities but cannot save the model and need to rebuild the model each time for the next data set.

For some future work, there is another feature and word representation approach that can be used for text mining project. Nevertheless, in this research there is only focuses on bag-of-words feature approach. There is a possibility that another feature approach can increase classification accuracy performance. Based on literary research, the sentiment analysis may be carried out with the modification of lexicon-based dictionary using a specific language. For more study, an automatic sentiment analysis multi-classification is also can be done for future work.

Acknowledgment

The authors would like to thank the Faculty of Computer and Mathematical Sciences and Research Management Center Universiti Teknologi MARA, Shah Alam, Selangor for supporting this research.

References

[1] R. K. Nielsen and L. Graves. ""News you don't believe": Audience perspectives on fake news." Reuters Institute (accessed 30 September, 2021).

[2] N. M. N. Mathivanan, N. A. M. Ghani, R. M. Janor, and Ieee, "Analysis of K-Means Clustering Algorithm: A Case Study Using Large Scale E-Commerce Products," presented at the 2019 IEEE Conference on Big Data and Analytics (ICBDA), 2019.

[3] S. B. Parikh and P. K. Atrey, "Media-Rich Fake News Detection: A Survey," in 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 10-12 April 2018 2018, pp. 436-441, doi: 10.1109/MIPR.2018.00093.

[4] T. Abdullah All, E. M. Mahir, S. Akhter, and M. R. Huq, "Detecting Fake News using Machine Learning and Deep Learning Algorithms," in 2019 7th International Conference on Smart Computing & Communications (ICSCC), 28-30 June 2019 2019, pp. 1-5, doi: 10.1109/ICSCC.2019.8843612.

[5] A. Shelar and C.-Y. Huang, "Sentiment analysis of twitter data," in 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2018: IEEE, doi: 10.1109/CSCI.2018.00251.

[6] M. Puteh, N. Isa, S. Puteh, N. A. Redzuan, and A. M. Korsunsky, "Sentiment Mining of Malay Newspaper (SAMNews) Using Artificial Immune System," presented at the World Congress on Engineering - WCE 2013, Vol III, 2013.

[7] N. A. S. Abdullah and N. I. A. Rusli, "Multilingual Sentiment Analysis: A Systematic Literature Review," PERTANIKA JOURNAL OF SCIENCE AND TECHNOLOGY, vol. 29, no. 1, pp. 445-470, JAN 2021, doi: 10.47836/pjst.29.1.25.

[8] A. Kumar and T. M. Sebastian, "Sentiment Analysis: A Perspective on its Past, Present and Future," International Journal of Intelligent Systems and Applications, vol. 4, no. 10, pp. 1-14, 2012, doi: 10.5815/ijisa.2012.10.01.

[9] M. Farhadloo and E. Rolland, "Fundamentals of Sentiment Analysis and Its Applications," in Sentiment Analysis and Ontology Engineering, vol. 639: Springer, Cham, 2016.

[10] N. N. Yusof, A. Mohamed, and S. Abdul-Rahman, "Reviewing Classification Approaches in Sentiment Analysis," in 1st International Conference on Soft Computing in Data Science (SCDS), Putrajaya, MALAYSIA, Sep 02-03 2015, vol. 545, in Communications in Computer and Information Science, 2015, pp. 43-53, doi: 10.1007/978-981-287-936-3_5. [Online].

[11] A. Hasan, S. Moin, A. Karim, and S. Shamshirband, "Machine Learning-Based Sentiment Analysis for Twitter Accounts," Mathematical and Computational Application, vol. 23, no. 1, 2018, doi: https://doi.org/10.3390/mca23010011.

[12] S. Aphiwongsophon and P. Chongstitvatana, "Detecting Fake News with Machine Learning Method," in 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 18-21 July 2018 2018, pp. 528-531, doi: 10.1109/ECTICon.2018.8620051.

[13] C. K. Hiramath and G. C. Deshpande, "Fake News Detection Using Deep Learning Techniques," in 2019 1st International Conference on Advances in Information Technology (ICAIT), 25-27 July 2019 2019, pp. 411-415, doi: 10.1109/ICAIT47043.2019.8987258.

[14] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 01/01 2000.

[15] S. B. bin Rodzman et al., "Experiment with Lexicon Based Techniques on Domain-Specific Malay Document Sentiment Analysis," presented at the 2019 IEEE 9TH Symposium on Computer Applications & Industrial Electronics (ISCAIE), 2019.

[16] M. Hu and B. Liu, "Mining opinion features in customer reviews," in AAAI'04: Proceedings of the 19th national conference on Artifical intelligence, San Jose California 25 - 29 July 2004: AAAI Press.

[17] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in Proceedings of the International AAAI Conference on Web and Social Media, 2014, vol. 8, no. 1, pp. 216-225.

[18] W. Wang and K. Siau, "Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda," Journal of Database Management, vol. 30, no. 1, pp. 61-79, 2019, doi: 10.4018/JDM.2019010104.