Fake News Detection Using Machine Learning

Boon Yu Jing[*a], Thagirarani Muniandy[b]

[a]*Nilai University, Malaysia*

Abstract

Social media has become significant in people's daily life and various news can be shared easily among users on multiple platforms. This leads to various versions of information spreading on the internet which makes it hard for users to determine the credibility of the news. Fake news detection systems enable users to verify the credibility for news that posted on social media. It would be helpful for user to avoid being misled by fake news. The objective of this research aims to improve the accuracy with web scrapping, data cleaning, ensemble method and validation techniques. This research will mainly focus on the news posted on Facebook and Instagram which are recognized as the popular social media platforms. The dataset for this research will be collected from multiple fact checking websites. HTML Parsing technique is used to scrape and collect the latest data from the data sources. Several data cleaning techniques will be implemented to improve the dataset, such as normalization, text cleaning, correcting missing value, checking character encodings and fixing inconsistent data entry. There are various machine learning algorithms that can be applied for binary classification, such as Logistic Regression, K Nearest Neighbors, Decision Tree, Support Vector Machine and Naïve Bayes. In this research, ensemble method which consists of Support Vector Machine, Logistic Regression and Decision Tree will be implemented to perform classification due to the error occurred can be compensated by each other. K-fold cross validation technique will be used to determine the accuracy while confusion matrix will be used to evaluate the result. The outcome of this research will be a model to verify the integrity of the messages shared in the social media platform.
Keywords: Fake news, data cleaning, classification, validation

1. INTRODUCTION

Social media has become a more effective news and information sharing source because social media allows the news to be published and accessed in real time. Tankovska (2021) published a statistical report which shows that the number of active global social media users are increasing annually and will continue to increase to 4.12 billion in 2023. With the increasing usage of social media, the fake news has spread more rampantly According to West (2017), the America election 2016 was analysed. The analysis found that more traffic have been generated over the Internet and led to more fake news spread to people.

Fake news refers to news stories that are incorrect without verifiable facts and sources and it can be easily published by people to mislead others. For instance, misinformation and disinformation are the two types of fake news. Misinformation refers to news created by mistake and having some incorrect information but is not intended while disinformation refers to the whole news having incorrect information and is created intentionally to influence public opinion or cover up the truth (Desai et al., 2021). The authors of fake news can be someone who would like to make money of it, influence people's beliefs or poor journalists. The impact of fake news can be severe in different situation. For instance, when people mistrust the fake news that can affect their behaviour in the workplace, an organization's

culture will be affected and people will be unsure who to trust. Besides, fake news can be harmful to a business if an incorrect product review is written. It can mislead consumer when purchasing the product and can cause serious damage to the reputation of the business. There are several steps can be performed to determine the credibility of the news, such as checking the source and the author. News that published from unfamiliar source are usually not verified and maybe a fake news even published from a familiar source. Moreover, sometimes news can also be published wrongly by a professional news source, so reader need to be careful by examining the evidence. The evidence can be either quotes from experts or official statistics that have a proven result.

Hence, a lot of methodologies have been used to address the issue, such as checking the news on fact-checking website or even a predictive model that is built with machine learning. There are various fact checking websites created for the ease of readers, such as Snopes that investigate news with a complex process to prove the news either is true or fake and Sebenarnya.my is another reliable fact checking website that mainly focus on the news in Malaysia. Sebenarnya.my is created by Malaysian Communications and Multimedia Commission (MCMC) which is a government agency to check on the news that shared among social media platforms. However, people sometimes will prefer to simply trust the news instead of spending time to examine the credibility of the news. Besides, fact-checking websites only store the news that have collected by their journalists so readers unable to search for the news that is not existing in the database of the website.

There are some popular social media sites that visited by people daily, such as Facebook, Instagram and Twitter. Based on Tankovska (2021), Facebook owns the greatest number of active users which are 2,449 million. The number of active users for Instagram is 1,000 million, Sina Weibo is 497 million and for Twitter is 340 million. Hence, various news can be shared by users easily on multiple platforms. It also leads to misinformation spreading on the internet. Social media users who read the news is hard to determine the credibility of the news.

Therefore, fake news detection system can be used to verify the news that are posted on social media. The existing system reported in previous studies mostly utilize news that come from certain social media platform such as Twitter (Tankovska, 2021). Although Facebook and Instagram are more popular than Twitter, the data from both social media sites are rarely used in previous studies. Besides, the dataset used for fake news detection is from 2016 to 2018 which is not up-to-date that may cause the system unable to verify the news properly and may provide wrong output. In addition, the accuracy of each algorithm can be improved by data cleaning techniques.

This paper discusses a proposal to develop a fake news detection system based on data from multiple social media sites using machine learning. The proposed techniques to build a fake news detection system are by using data cleaning methods on the collected dataset, ensemble machine learning algorithms to perform classifications on the dataset, k-fold cross validation to select the model with highest accuracy and evaluation metrics to evaluate the performance of the model.

2. STEPS IN A FAKE NEWS DETECTION SYSTEM

Several steps are used in a fake news detection system, including data collection, data cleaning, machine learning classification, validation and evaluation.

2.1 Data Collection

In machine learning, data is important and is required for the model to train and test to make predictions. The types of data including text, numbers, time series and categorical. A dataset is a collection of data which can be classified in structured or unstructured dataset. The difference between two type of dataset is the format. The format of structured datasets is in tabular that a row of the dataset corresponds to a record and column corresponds to a feature while unstructured datasets refer to images, text and audio.

2.2 Dataset

Based on research, "FakeNewsNet" and "LIAR" datasets are used in most published journals. The data of both datasets are collected from "Politifact" and "GossipCop" websites and published on GitHub (Shu, 2019). Both websites are used to do fact checking on news published on newspapers and web. As current year is 2021, both latest dataset that can be obtained from the official repository was published in 2019 which is not up to date. As a lot of news has been published in 2020-2021, hence web scraping is required to collect recent data.

2.3 Web Scraping

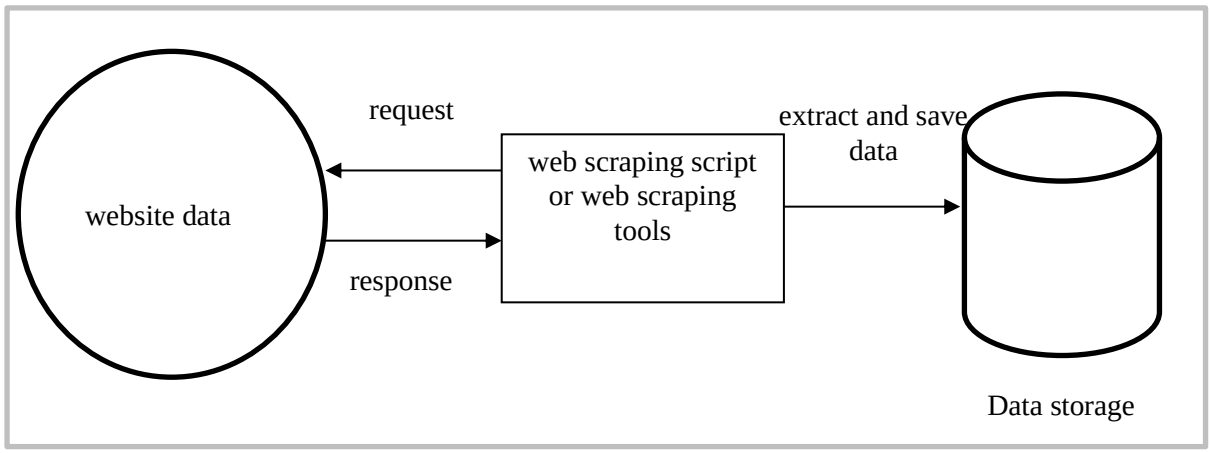

Fig. 1. Architecture of web scraping process (Karthikeyan T. et al., 2019)

Most websites display data that can view with typical web browser software. The data of the website usually can be copied and pasted manually to extract the data but it takes time when large amount of data is required. Hence, web scraping techniques can be implemented to extract data automatically to solve the issue. There are three types of web scraping techniques, which are manual scraping, automated web scraping with programming languages and web scraping tools.

2.4 Manual Scraping

Manual scraping refers to copying a part or complete web content from a website and pasting manually. This method works well when a website is protected with anti-scraping techniques. However, this technique is ineffective as it needs much effort, and the process is repetitive.

2.5 Automated Web Scraping with Programming Languages

Automated web scraping with programming languages refers to writing a script in programming language to extract information from the websites. The script turns the unstructured data into a structured form. These data can be stored anywhere in the local computer or remote servers. This process is often efficient, faster, and less prone to errors when automated. In recent times, many researchers are widely using this technique to create their own data sets for journal article which related to information extraction, text mining-related projects and etc. These scrapers can be built using any technology such as PHP, Node JS, and Python etc. In most cases, python will be the choice to develop such programs, which is useful in terms of flexibility. However, the knowledge on the structure of HTML scripts and data formats are required to extract the useful data. There are several techniques can be used when writing a script in programming language, such as Hyper Text Markup Language (HTML) Parsing, Document Object Model (DOM) Parsing and XML Path Language (XPath).

2.6 Hyper Text Markup Language (HTML) Parsing

HTML parsing is a common technique to target HTML pages to retrieve and transform page content. It works faster and identifies HTML tags from webpages. In Python, a popular library called as Beautiful Soup is available for parsing HTML and XML. By searching the tag names or class names, we can get the data easily. When parsing HTML, the website performance can be affected if making too many requests to scrape the data from the website (Singrodia et al., 2019).

The advantages of HTML parsing are the code is easy to be written and maintained as there are readily available HTML parsers in most of the programming languages. The output will be very clean and structured as specific piece of information from the page can be selected using the HTML parser. The disadvantages of HTML parsing are it works based on structure of the page hence the selectors need to be changed if website's structure changes and the chances of getting blocked by the server is more as a regular Internet browser is not being used.

2.7 Document Object Model (DOM) Parsing

DOM defines the style, structure and the content of an XML document. The internal functionality of a website must be transparent to the scrapers for content extraction, and DOM delivers its parsing modules to do it. The nodes organized with DOM parsers and XPath and other related tools help to retrieve the content from the sites. DOM parsers can even work well with dynamic websites (Anand et al., 2018).

The advantage of DOM parsing is it can obtain dynamic content as pages are generated dynamically on client side. Besides, the chances of getting blocked will be less as a real browser is being used and writing selectors will be easier as processing the same data as seen on the browser. However, DOM parsing will be resource intensive as a complete web browser program is running in the background and is slower compared to HTML parsing as the page build happens on the client side.

2.8 XML Path Language (XPath)

XPath is a query language that used when the data is being extracted from the nodes of the Extensible Markup Language (XML) documents which follow a hierarchical tree-like form. XPath provides a simple way to handle exact nodes and to extract data from the nodes. It is used with DOM parsing technique to retrieve data from the webpages.

The advantage of XPath is it can scrape almost all type of information because Unicode is supported. XPath can scrape the data efficiently and consistently because its syntax is strict. However, the syntax of XML is redundant which will cause the file size larger compared to HTML parsing and is hard to debug when there is an error in the code.

2.9 Web Scraping Tools

Web scraping tools refers to tools that are ready to extract information from the website and not require knowledge of programming language. This tool can be configured based on user requirement and the type of the website. With a single click, the data can be scraped, formatted, and stored as a file in the system. The data can be replicated anywhere with the help of the scraper software (Karthikeyan T. et al., 2019).

2.10 Import.io

Import.io is a web-based tool for extracting data from website without writing code. Import.io allows users to scrape the data from the website in a few seconds.

Import.io is able to scrape specific parts of webpages. The tool is very effective and accurate when it comes to scraping data of large URL lists. However, unwanted columns and unexpected data will be extracted. Besides, users require a basic understanding on using the tool and then where to use it. Sometimes, coding required if extracting data from complex pages. The URL of website must be static or Import.io unable to reach to each page if the pages are continuous.

2.11 ParseHub

ParseHub is also a web scraping tool but it works with single-page apps and multi-page apps. ParseHub can handle JavaScript, AJAX, cookies, sessions and redirects. ParseHub also allows users to easily fill in forms, loop through dropdowns, login to websites, click on interactive maps and even deal with infinite scrolling.

Unlike Import.io, ParseHub can handle website with dynamic URL and extract exact columns as well as expected data. However, there is a disadvantage of ParseHub which is XPath selection is required when scraping the data on the web because the technique is advanced that would be difficult for a beginner user to use.

2.12 Summary

By comparing among the three types of web scraping techniques, manual scraping is eliminated at the first stage. Among automated scraping with programming languages, HTML parsing is better than DOM parsing due to the speed and able to provide clean and structured output. HTML is also better than XPath as it is easier to write and maintain the code. Among web scraping tools, both Import.io and ParseHub have some significant limitations as Import.io may extract unwanted column and data while ParseHub requires user to use XPath selection which is an advanced technique. In summary, HTML parsing is the best approach for web scraping in this project because the code is easy to be written and maintained to provide clean and structured output. Besides, by using HTML parsing, only the useful column will be selected hence it saves time in data cleaning that do not need to perform removing unnecessary columns.

2.13 Data Cleaning

Data cleaning refers to eliminate the errors in a dataset that may affect the performance of the machine learning model negatively. The errors including missing values, incorrect character encoding and inconsistent data format (Tandel et al., 2019).

2.14 Handle Missing Values

Missing values refer to some values are missing in columns that will affect the accuracy of the machine learning model. There are several techniques can be done to address the missing values issue, such as dropping the rows, filling in value and flagging the data. When the value is missing randomly and rarely in the dataset, the row that having missing values can be dropped. However, if most of the values in a column are missing, the whole column should be dropped. Besides, when the values are missing, filling in value can also be implemented especially for numerical data. By using statistical values such as mean or median, the data will not become biased. The values can also be copied and referenced from other similar records to fill in at the missing values.

2.15 Check Character Encoding

Character encoding is a process to convert the text data into binary numbers. If the character encoding is wrongly configured, error may occur when loading dataset which in excel file. The common character encodings are American Standard Code for Information Interchange (ASCII) encoding and Unicode Transformation Format (UTF) encoding. ASCII encodings consists of 128 characters, including English characters, numbers and some symbols. The width of ASCII encoding is fixed that uses one byte per character so the size of the data is smaller and can be read faster by the model. The width encoding of UTF is based on the number of bits. For example, UTF-8 represent the width of UTF encoding is 8 bits so that all the existing characters are supported. At the same time, UTF-8 also supports ASCII encoding which means that UTF-8 is the superset of ASCII.

2.16 Check Data Format and Normalization

In each column, the format of data needs to be consistent to make the data become readable by the model and avoid confusion. For example, the format of date should be consistent, such as 14 March 2021 should be converted into 2021-03-14 which is yyyy-mm-dd international format. Besides, in the dataset, the result of fact checking which represents in text can be normalized to numerical value for the model to read. For example, convert fake result to 1

and true result to 0. This technique can be called as label encoding which refers to assign a unique integer to each type of result.

2.17 Text Cleaning

Text cleaning is a subset of data cleaning that only focus on correcting errors in text data without removing any important content. The advantages of text cleaning are smaller the vocabulary and lower the complexity of word (Pradana and Hayaty, 2019).
Firstly, syntax errors should be removed in the dataset. Syntax errors may occur when the content of the data having extra white spaces. Hence, extra white spaces should be removed to decrease the size of the dataset and increase the system speed when loading the dataset. The second step is to normalize the character case by converting all English characters into lowercase. This normalization will make the vocabulary become simple and reduce the size of vocabulary. After that, common typos and misspellings can be corrected to improve the performance of the model when reading the dataset. Besides, punctuations and stop words can be filtered and remove because both punctuations and stop words do not have meaning in the sentence. Punctuations include comma, full stop, semi colon etc and stop words include "a", "an", "the" and "is". Then, stemming words can be implemented to return the words to its base, such as converting "saying" to "say".

2.18 Classification

Based on the discussion above, the result of dataset is represented in numerical value which fake result is 1 and true result is 0. After completing data collection, a classification predictive model is required to be implemented because classification is to classify the result of an input. To complete a classification predictive model, machine learning algorithms are required to be used to learn from the dataset examples to classify the result from an input. The type of classification needs to be examined to decide which machine learning algorithm is suitable for the dataset. There are three main types of classification, which are binary classification, multi-class classification and multi-label classification (Mendez et al., 2019).

Binary classification refers to classify the input that have a single output and the output can be in two possible result only. For example, the result of phishing email detection is either true or fake. The common machine learning algorithms that can be used in the binary classification are Logistic Regression (LR), K-nearest Neighbors (KNN), Decision Trees (DT), Support Vector Machine (SVM) and Naïve Bayes (NB). All these algorithms are in supervised machine learning where there is an input variable and an output variable during training.

Multi-class classification refers to classify the input that have a single output but the result of output is more than two. For example, classification of cat breeds is one of the multi-class classifications. KNN, DT, NB are the common machine learning algorithms that can be used in both binary classification and multi-class classification. Moreover, in this multi-class classification, Random Forest (RF) and Gradient Boosting (GB) can also be used.

Multi-label classification refers to classify the input that may have more than two outputs, such as object detection in a photo which may detect multiple objects. Unlike other classification, multi-label classification requires specific machine learning algorithms, such as Multi-label DT, Multi-label RF, Multi-label GB.

In summary, binary classification is more suitable to be used for the dataset because the possible result of the dataset is either real or fake. Therefore, machine learning algorithms such as LR, KNN, DT, SVM and NB can be used to train the dataset to output a predictive model.

2.19 Logistic Regression

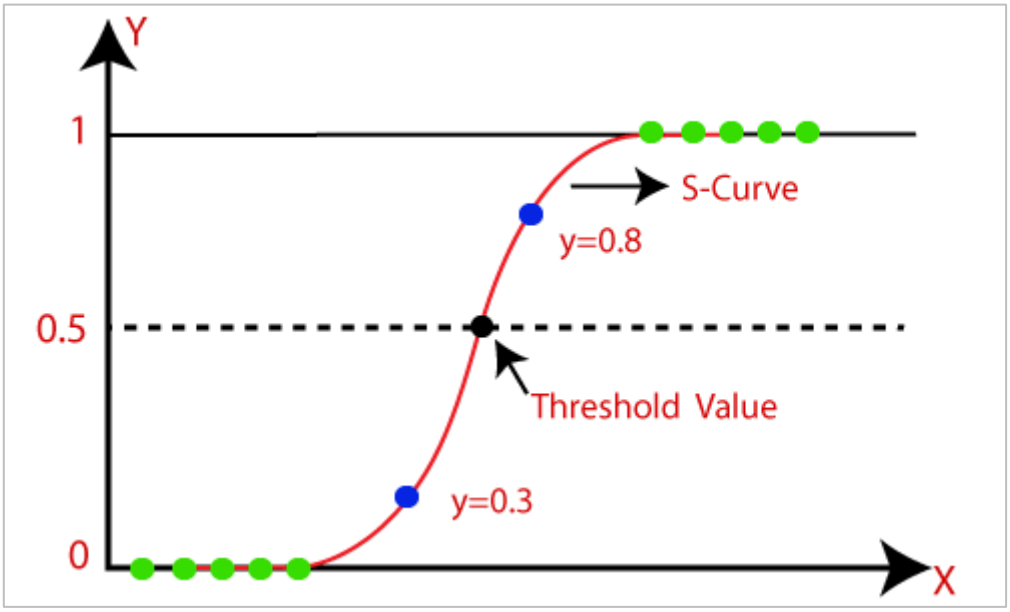

Fig. 2. Logistic Regression (Supervised Machine Learning, 2021)

Logistic Regression is a machine learning algorithm to perform classification by providing the probabilities of the data. LR uses a "S" curve to predict the two maximum values that either in 0 or 1. There will be a threshold value in the curve to determine the relationship of the data. When the input value below the threshold value, it is classified in category 0 while the input value above threshold value, it is classified in category 1 (Ahmad et al., 2020).

The advantage of Logistic Regression is the algorithm is easy to be implemented and is fast when making classification after a model is completed. The algorithm can obtain good accuracy when the dataset is simple and the curve is linear. However, simple dataset is rare and LR only can solve linear problems. Non-linear problems need to be solved by adding more features to make it become linear. When the number of features is increasing, it may lead to overfitting issue which the algorithm able to get high accuracy on training dataset but unable to make prediction with good accuracy on testing dataset (Ying, 2019).

2.20 K Nearest Neighbors (KNN)

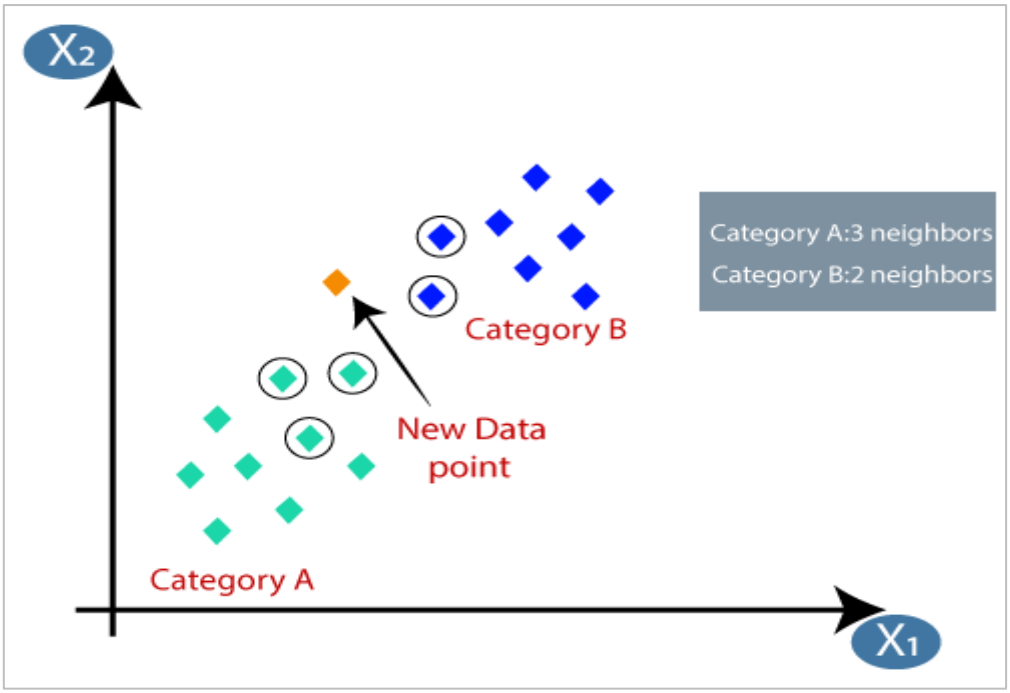

Fig. 3. K Nearest Neighbors (Supervised Machine Learning, 2021)

K Nearest Neighbors (KNN) is a simple algorithm which perform classification based on the similarity of the data (Elshrif and Abdelouahed, 2018).

Factor K refers to the number of neighbors of the new input. In Figure 4, a new data point is input to get the result. Factor K is set to 5 to find the 5 nearest neighbors of the new data point. The algorithm will start to calculate between the neighbors and the new data point. Then, the algorithm found three neighbors in category A and two neighbors in category B having the nearest distance to the new data point. Due to the number of neighbors in category A are more than in category B, the new data point is classified in category A.

The advantage of K Nearest Neighbors is the algorithm is simple to be implemented and is robust when there is error in the training data. The algorithm also works effectively when the size of the dataset is large. The disadvantage of KNN is the training time is longer as the value of factor K takes time to determine and configure. The complexity is also higher due to the distance calculation between the data.

2.21 Decision Tree

Decision Tree is an algorithm which structure is like a tree. The algorithm starts with a root node and expands based on the decision to have two output. A final output that cannot expand further is called as a leaf node. A decision that expands to two leaf nodes represents the features of the dataset and is called as a decision node. The first decision node during classification is called as root node (Ozbay and Alatas, 2020).

Decision Tree starts performing classification from the root node and dividing the dataset into subsets based on the features. The best feature will be selected by using Attribute Selection Measure (ASM) and contained in a decision node. The examples of popular ASM are Information Gain or Gini Index. Then, the algorithm continues to select new feature and split into new subsets until the decision node cannot be classified.

Decision Tree is easy to understand because the tree structure is divided clearly with decisions and it is similar to a human when making decision. The algorithm requires less data cleaning compared to other algorithms and is able to provide all the possible output. When there is a lot of subsets, Decision Tree will be more complex. Overfitting issue may occur when there is a huge difference between the accuracy of training dataset and testing dataset because the algorithm fits too perfectly in the training dataset.

2.22 Support Vector Machine (SVM)

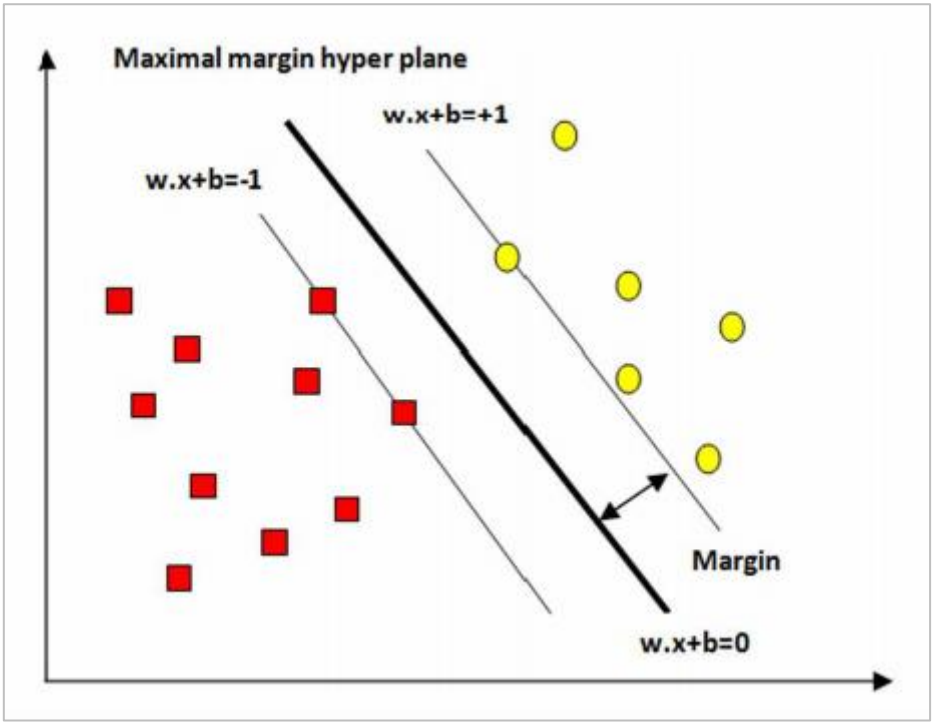

Fig. 4. Support Vector Machine (Rajan et al., 2019)

Support Vector Machine is an algorithm which creates a boundary to clarify the relationship between the nearest variables. The boundary is called as a hyperplane and the variables are called as support vectors. The distance between the hyperplane and support vectors is called as margin. Then, SVM will find the hyperplane with the maximum margin to get the best performance (Rajan et al., 2019).

For example, in fake news detection, SVM starts performing classification by examining the difference between the features of real news and fake news in the dataset. A hyperplane will be created between the nearest data. A multi-dimensional hyperplane can be created when there are multiple columns of features in the dataset. Then, when there is an input that mix with the features of both real news and fake news, the output will be classified as fake news by SVM based on the hyperplane. Hence, SVM is also suitable in image classification that have a lot of features.

The advantage of SVM is it works well with unstructured data like text or image. SVM will archive good performance when the margin is clear between support vectors and works effectively when the dimension is high. SVM also uses less memory compared to other algorithms. However, SVM will not get a good result when the dataset having missing values or other errors. Besides, there is no clear explanation for the hyperplane that created by SVM during classification.

2.23 Naive Bayes

Naïve Bayes is a machine learning algorithm that calculate the output based on the probability of the features' occurrence. This algorithm is a combination of two words which Naïve refers to the assumption of a feature's occurrence is independent to others' while Bayes refers to use prior knowledge to determine the likelihood of hypotheses (Prannay et al., 2019).

$$P(A|B) = \frac{\mathrm{P(B|A)\ P(A)}}{\mathrm{P(B)}}$$

Equation 1: Formula of Naïve Bayes

Equation 1 shows the formula of Naïve Bayes. P(A|B) refers to the probability of A occurs on the feature B while P(B|A) refers to the probability of feature B to make the occurrence of A. P(A) refers to the probability of the total occurrence of A while P(B) refers to the probability of the total occurrence of feature B.

Naïve Bayes can perform classification within a shorter time compared to other algorithms due to its formula. The algorithm can work well with binary classification and multi-class classification. However, the algorithm is unable to learn the relationship between the features because it makes an assumption that all the features are not related to each other. Hence, NB is limited to certain applications as it is hard to have dataset which features do no need to relate to each other.

2.24 Ensemble Method

Ensemble method uses multiple machine learning algorithms instead of a single algorithm to do classification. Ensemble method can be formed by any two or more type algorithms. There will be some errors occurred when using a single algorithm only. By using ensemble method, a single algorithm can compensate the errors of another algorithm. Besides, each algorithm will find different hypothesis and the risk of choosing an incorrect hypothesis for classification can be reduced to avoid overfitting (Sagi and Rokach, 2018). Therefore, ensemble method can be used to achieve higher performance than single algorithm in terms of accuracy when doing classification.

2.25 Summary

In conclusion, ensemble method will be used in this research because using multiple algorithms can help to achieve higher accuracy than using a single algorithm. Support Vector Machine, Logistic Regression and Decision Tree are selected to be used in this ensemble method. SVM is selected because the algorithm is able to handle unstructured data like text and also uses less memory. Besides, there should be no missing value in the dataset hence SVM is able to get a good accuracy. LR is selected because it is easy to be implemented and the training time is short while Decision Tree able to provide all possible output without missing any important features. K Nearest Neighbors is not selected because it takes longer time on training. Besides, Naïve Bayes is also not suitable because it unable to learn the relationship between the features although it has a shorter training time.

2.26 Validation

Validation is a technique to determine the effectiveness of the machine learning model before performing evaluation with evaluation metrics. It will test on the unseen data to determine whether the model is stable to avoid overfitting issue. However, if the unseen data having some important information that left during training on the training dataset, the performance of the model may not be accurate. Therefore, cross validation is a technique that can be used to perform validation with both training data and unseen data to ensure no important information is not being trained. Cross validation is suitable for small dataset or it takes long time to compute. There are two well-known cross validation techniques in machine learning, which are Leave one out cross validation (LOOCV) and K-fold cross validation.

2.27 Leave One Out Cross Validation (LOOCV)

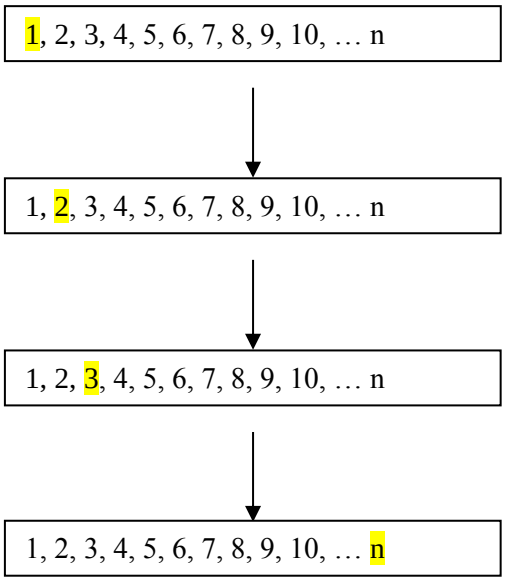

Fig. 5. Leave one out cross validation

Leave one out cross validation (LOOCV) is a technique to make use of all the data in the dataset. LOOCV perform training on the whole dataset but left one row of data for validation. The technique will keep running until the last row of data is validated (Ghojogh and Crowley, 2019).

The advantage of LOOCV is it reduce the bias as all the data is used in the validation. However, it will lead to higher variation because only one row of data is validated. Besides, LOOCV requires long execution time because the technique will run based on the number of data row.

2.28 K-fold Cross Validation

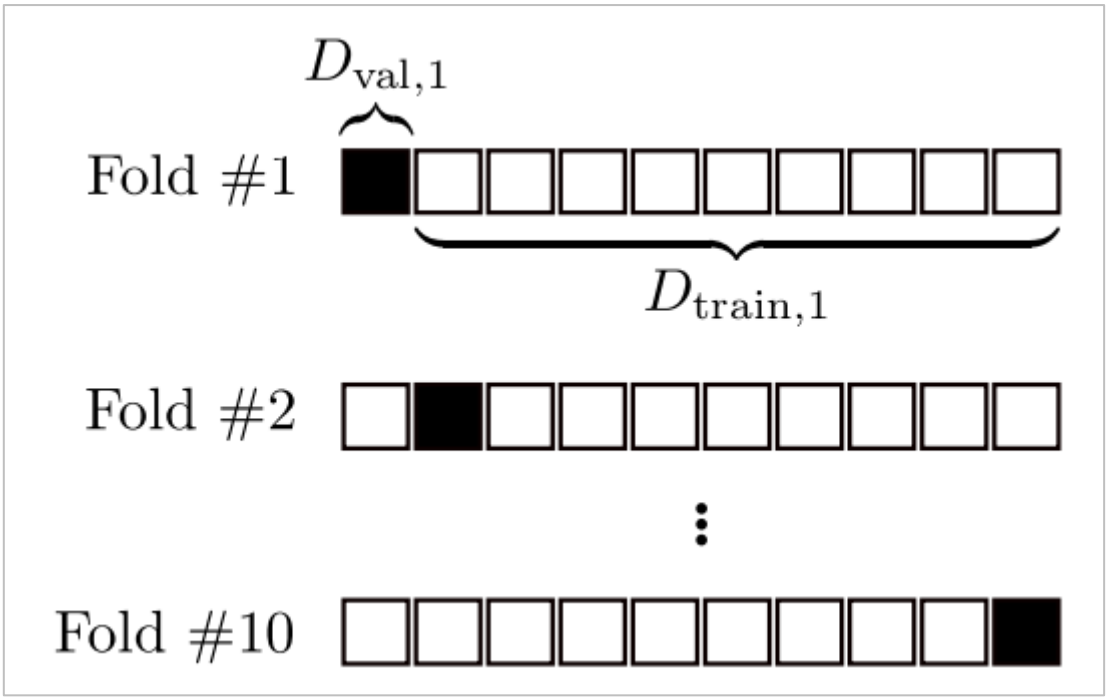

Fig. 6. 10-fold cross validation (Berrar, 2019)

K-fold cross validation split the dataset into "k" number of subsets. The subsets can also be called as folds. Figure 7 shows the dataset split into 10 subsets. It will perform 10 times of validation. For the first time, it will train on the first 9 training folds and validate using the last test fold. For the second to tenth time, it will train on the other 9 training folds respectively and validate with the remaining test fold (Berrar, 2019).

The advantage of K-fold cross validation is it will achieve higher accuracy and run faster than LOOCV. The technique also reduce bias in the model. However, the technique requires a lot of computer resource as it runs "k" times to perform validation.

2.29 Summary

In conclusion, both cross validation techniques are able to reduce the bias in the model because all the data is used during validation. K-fold cross validation will be used for the machine learning model because it is able to achieve higher accuracy and run faster than LOOCV.

2.30 Evaluation Metric

Evaluation metrics refers to the performance measurement of the machine learning model. Multiple evaluation metrics should be used to evaluate the model because the performance may vary in different evaluation metric. It is important to use multiple evaluation metrics to ensure the performance of the machine learning model is optimal. There are several evaluation metrics can be used to evaluate the model, such as classification accuracy, Confusion Matrix and F1 Score.

2.31 Classification Accuracy

Classification accuracy is the simplest metrics to perform evaluation of the machine learning model performance. The formula of classification accuracy is the number of correct

predictions divided by the total number of predictions. For instance, when there are 1000 predictions in the dataset and the number of correct predictions is 900, the result of the classification accuracy is 0.9 which is 90%.

2.32 Confusion Matrix

		Predicted	
		Fake news	Real news
Actual	Fake news	True positive (TP)	False negative (FN)
	Real news	False positive (FP)	True negative (TN)

Fig. 7. Confusion matrix

A confusion matrix is usually a 2 x 2 matrix which able to display the evaluation of the performance completely. There are several components in a confusion matrix, which are True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) and accuracy.

TP refers to the prediction is positive and the actual result is positive while FP refers to the prediction is positive but the actual result is negative. TN refers to the prediction is negative and the actual result is negative while FN refers to the prediction is negative but the actual result is positive. TP and FP can be used to calculate Precision to determine the percentage of correct positive prediction. TP and FN can be used to calculate True positive rate which also called as Recall to determine the percentage of correct actual positive prediction. False positive rate can be calculated by using FP divided by addition of FP and TN.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Equation 2: Formula of Precision

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Equation 3: Formula of Recall

Then, the accuracy of the confusion matrix is calculated as similar as classification accuracy by using addition of TP and TN divided by the number of total predictions made.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

Equation 4: Formula of classification accuracy

2.33 F1 Score

In previous section, Precision and Recall has been discussed to evaluate the model. F1 Score is a metric to get the average of Precision and Recall at the same time.

$$Accuracy = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

Equation 5: Formula of F1 Score

2.34 Summary

In conclusion, all the three evaluation metrics having different formulas from easy to complex. All the evaluation metrics that are discussed will be used because using multiple evaluation metrics can ensure the performance of the machine learning model is optimal.

2.35 Summary

In conclusion, a lot of techniques are discussed for each step that need to be implemented in a fake news detection system. For data collection, HTML parsing is selected because the code is easy to be written and maintained. The execution time of HTML parsing is short and it is able to provide clean and structured output. Besides, by using HTML parsing, only the useful column will be selected hence it saves time in data cleaning that do not need to perform removing unnecessary columns.

For data cleaning, there are some techniques will be implemented, such as normalization to convert the fact checking result from text value to numerical value, check unexpected missing values, check character encoding and check inconsistent data entry. Several text cleaning techniques will also be used, such as remove syntax errors, case normalization to convert all words from upper case to lower case, fix typos, stem words, remove punctuation as well as stop words.

For classification, several machine learning algorithms are discussed. Among all the algorithms, Support Vector Machine is selected as it is able to handle unstructured data and uses less memory. Besides, there should be no missing value in the dataset hence SVM is able to achieve a good result in terms of accuracy and execution time. LR is selected because it is easy to be implemented and the training time is short while Decision Tree able to provide all possible output without missing any important features. Ensemble method will be used in this research by using these three algorithms to achieve higher accuracy than using a single algorithm.

For validation, although both cross validation techniques are able to reduce the bias in the model, K-fold cross validation will be used for the machine learning model because it is able to achieve higher accuracy and have shorter execution time compared to Leave one out cross validation.

For evaluation metrics, all the three evaluation metrics, such as classification accuracy, Confusion Metrix and F1 Score will be used. The reason is using multiple evaluation metrics can ensure the performance of the machine learning model is optimal.

3. DEVELOPMENT OF A PROPOSED FAKE NEWS DETECTION SYSTEM

3.1 Overview of the proposed system

The system will run few steps to create a classification predictive model. The steps are loading dataset, data cleaning, splitting dataset, classification, validation, evaluation for individual algorithm. After that, top 3 algorithms with highest accuracy will be selected and used together as ensemble method in the system to analyse the input and provide output.

HTML parsing will be used for web scraping to create a dataset. After loading dataset, several data cleaning techniques are used to improve the dataset for classification to improve the accuracy. The machine learning algorithms that suitable in binary classification are Logistic Regression, K-nearest Neighbors, Decision Tree, Support Vector Machine and Naïve Bayes. K-fold cross validation will be used to calculate the accuracy. Based on the accuracy, the best three predictive model will be decided to be used in ensemble method to do classification again. After the model is created, user input the news and the model will do prediction to detect the legitimacy of the news. The system will generate a result in numerical value

3.2 Description of Methodology

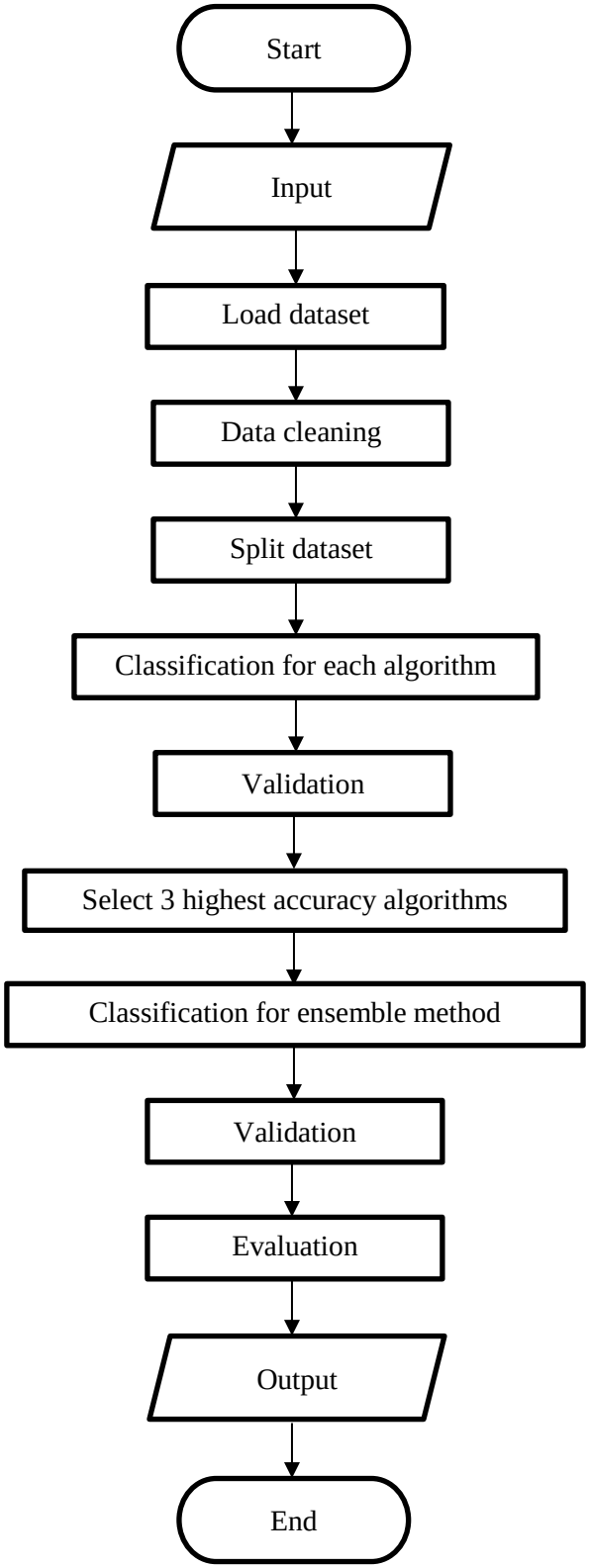

Fig. 8. Flowchart of proposed system

Input

The system starts when user input the news for checking.

Load dataset
The system will start to load the dataset that scraped from web. The dataset contains of real news and fake news that are verified. The number of news that contained in the dataset may increase because there is more news uploaded on the website every day. For web scraping, HTML parsing will be used because the code is easy to be written and maintained to provide clean and structured output. Besides, by using HTML parsing, only the useful column will be selected hence it saves time in data cleaning that do not need to perform removing unnecessary columns.

Data cleaning
After scraping the data, data cleaning techniques will be implemented to improve the performance of machine learning algorithm when doing classification. Normalization is one of the technique by converting the fact checking result to numerical value for the machine learning model to read, fake to 1 and true to 0. Then, the system will handle missing value and check character encodings to avoid error when loading CSV files if necessary. Next, inconsistent data entry will be checked to fix data format. The system will perform text-cleaning on the dataset, such as remove syntax errors, normalizing case, fix typos, remove punctuation, remove stop words and stem words.

Split dataset
After loading the dataset, the system will split the dataset into train dataset and test dataset with a ratio 70%:30%. This means that the train dataset will have 70% amount of news and test dataset will have 30% amount of news from the whole dataset. Train dataset is used to do classification and selection on the best machine learning algorithms while test dataset is used to check the accuracy of the selected model.

Classification for each algorithm
All the five machine learning algorithms that discussed in section 3.4 will be used separately to do binary classification on the training dataset.

Validation
The system will do K-fold cross validation on the algorithms and will calculate the accuracy.

Select 3 highest accuracy algorithms
The three algorithms with highest accuracy will be selected and combined to be used in the ensemble method based on the result of validation. Therefore, an initial predictive model is created.

Classification for ensemble method
The model of the ensemble method will be used to do binary classification again on the training dataset.

Validation
K-fold cross validation will be implemented on the ensemble method and will calculate the accuracy.

Evaluation
Predictions can be made by the initial model on the testing dataset for evaluation. Evaluation metrics such as classification accuracy, confusion matrix and F1 score can be used to

evaluate the predictions by comparing them to the expected results in the testing dataset. A final accuracy of the complete predictive model will come out after evaluation.

Output

After getting the final accuracy, the system will start to analyse the input and do classification to output the result. The result of the output will be represented in numerical value, either 0 or 1. 0 represents the news is real while 1 represents the news is fake.

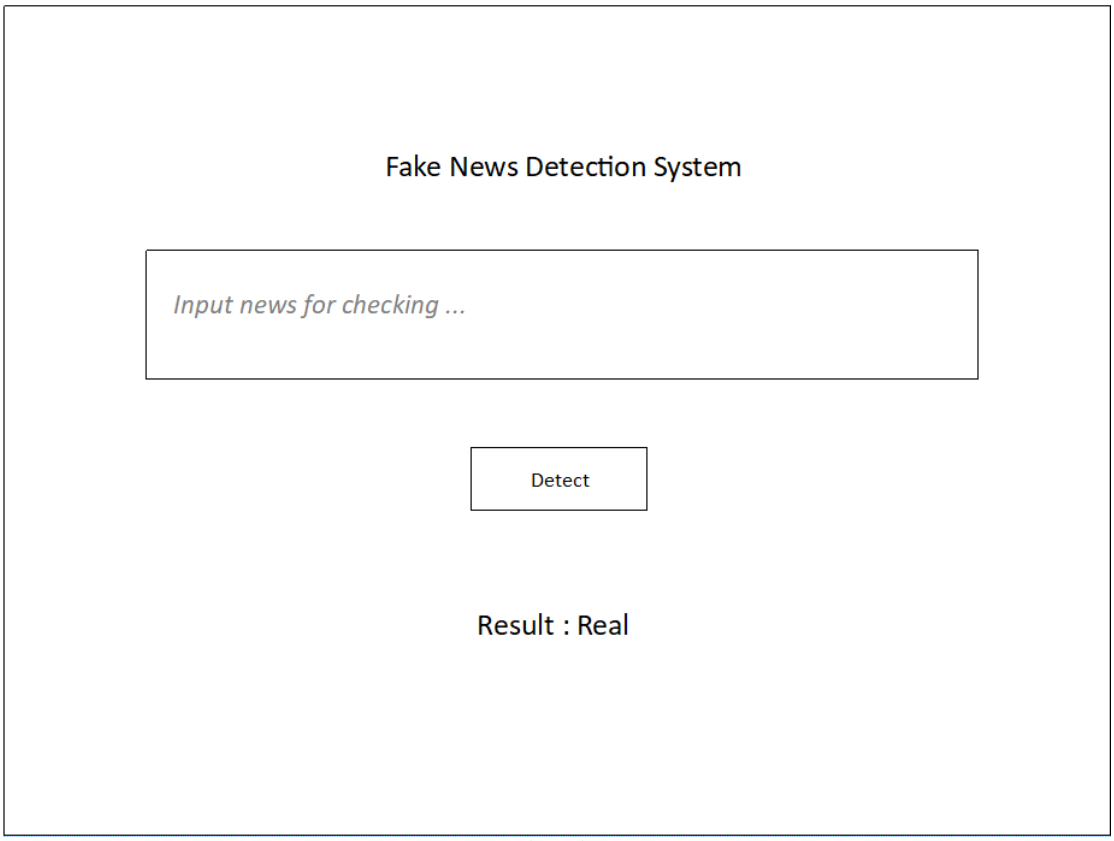

Fig. 9. User interface of proposed system

Figure 9 shows the interface of the proposed system. In the interface, there is an input field which has a placeholder to ask user to input the news for checking. Once input the news and click the "Detect" button, the result will show as real or fake based on the trained classification model.

3.3 Expected Results

Figure 10 shows the accuracy of machine learning algorithms implemented by previous studies (Guo et al., 2019). Support Vector Machine, Decision Trees and Logistic Regression had the highest accuracy among all the five machine learning algorithms. Therefore, these three algorithms will form an ensemble method in the proposed system.

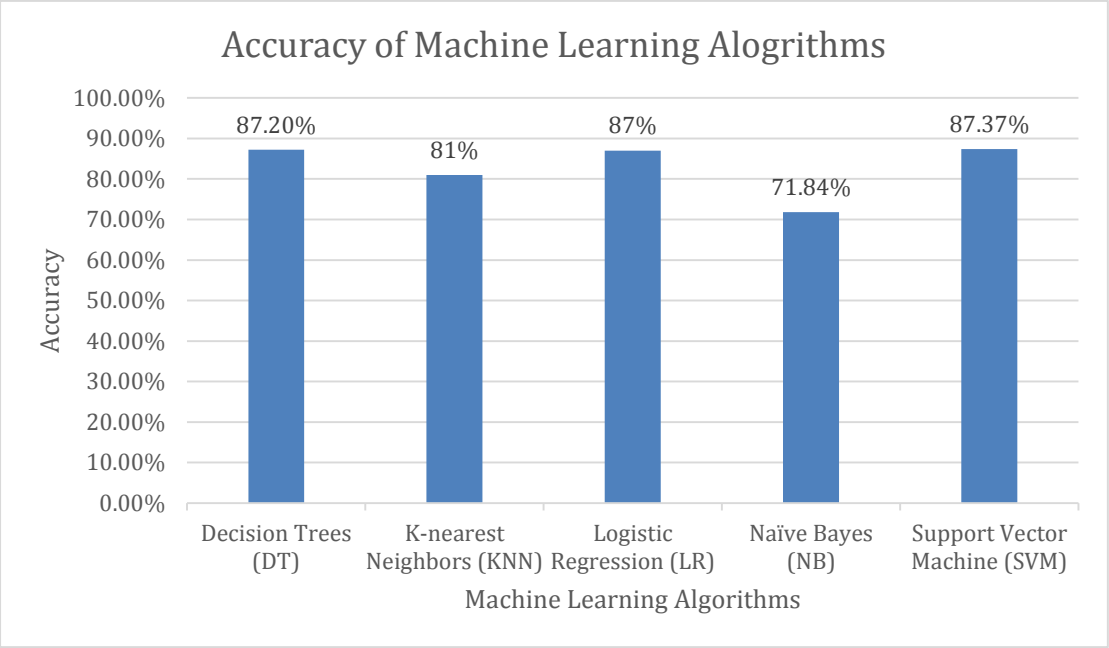

Fig. 10. Accuracy of machine learning algorithms in previous studies (Guo et al., 2019)

The highest accuracy was 87.37% which performed by SVM individually. Therefore, the expected accuracy should be higher than 87.37% by using improved dataset, data cleaning technique, ensemble method and K-fold cross validation.

4. CONCLUSION

In conclusion, the proposed techniques for Fake News Detection include HTML Parsing, data cleaning, ensemble method classification, K-fold cross validation and confusion matrix evaluation which can improve the performance of the system. The dataset that used in previous research was not up to date hence HTML Parsing is used to perform web scraping to get the latest data from verified source and data cleaning techniques will be used to improve the dataset. Besides, various machine learning algorithms were utilized to perform classification on the dataset that was not up to date and the highest accuracy was 87.37% which performed by Support Vector Machine. Therefore, Support Vector Machine, Decision Tree, Logistic Regression will be the initial algorithms to train on the dataset as the algorithms can compensate on each other's errors, but also had the highest accuracy in previous research.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Ms Thagirarani Muniandy who took keen interest and guided me to accomplish this project on the topic Fake News Detection Using Machine Learning I am extremely thankful to my Head of Department, Ts

Harlina Harun who gave me the opportunity and advice to do this project I am also grateful to my parents and friends who encourage and support me during the project

REFERENCES

Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. (2020). Fake News Detection Using Machine Learning Ensemble Methods. *Complexity, 2020*, 1-11. doi: 10.1155/2020/8885861

Anand, V., Kedar G. and Shweta A. (2018). An Overview On Web Scraping Techniques And Tools. *International Journal on Future Revolution in Computer Science & Communication Enginee*, 4(4), pp.363-367.

Desai, S., Mooney, H. and Oehrli, J. (2021). *"Fake News," Lies and Propaganda: How to Sort Fact from Fiction*., Research Guides, Retrieved 1 February 2021, from https://guides.lib.umich.edu/fakenews

Ghojogh, B., & Crowley, M. (2019). The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. *arXiv preprint arXiv:1905.12787*.

Guo, C., Cao, J., Zhang, X., Shu, K. and Yu, M. (2019). Exploiting Emotions for Fake News Detection on Social Media.

Karthikeyan T., Sekaran, K., Ranjith D., Vinoth Kumar V., & Balajee J M. (2019). Personalized Content Extraction and Text Classification Using Effective Web Scraping Techniques. International Journal of Web Portals, 11(2), 41-52. doi: 10.4018/ijwp.2019070103

Mendez, K., Reinke, S., & Broadhurst, D. (2019). A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics*, *15*(12). doi: 10.1007/s11306-019-1612-4

Nagi, K. (2018). New social media and impact of fake news on society. *ICSSM Proceedings, July*, 77-96.

Ozbay, F., & Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. *Physica A: Statistical Mechanics And Its Applications*, *540*, 123174. doi: 10.1016/j.physa.2019.123174

Pradana, A., & Hayaty, M. (2019). The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, And Control*, 375-380. doi: 10.22219/kinetik.v4i4.912

Prannay, R., Diana, E., Manoj, P., Keerthana, M. and Poonam, T. (2019). A Study on Fake News Detection Using Naïve Bayes, SVM, Neural Networks and LSTM. *Jour of Adv Research in Dynamical & Control Systems*, 11(6).

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wires Data Mining And Knowledge Discovery*, *8*(4). doi: 10.1002/widm.1249

Shu, K. (2019). KaiDMML/FakeNewsNet. Retrieved 1 February 2021, from https://github.com/KaiDMML/FakeNewsNet

Singrodia, V., Mitra, A., & Paul, S. (2019). A Review on Web Scrapping and its Applications. *2019 International Conference On Computer Communication And Informatics (ICCCI)*. doi: 10.1109/iccci.2019.8821809

Supervised Machine Learning. (2021). *Supervised Machine Learning*, JavaTpoint, Retrieved 1 April 2021, from https://www.javatpoint.com/supervised-machine-learning

Tandel, S., Jamadar, A., & Dudugu, S. (2019). A Survey on Text Mining Techniques. *2019 5Th International Conference On Advanced Computing & Communication Systems (ICACCS)*. doi: 10.1109/icaccs.2019.8728547

Tankovska, H. (2021). Topic: Social media. Retrieved 20 March 2021, from https://www.statista.com/topics/1164/social-networks/

- Javatpoint. (2021). Retrieved 1 April 2021, from https://www.javatpoint.com/

West, D. (2021). *How to combat fake news and disinformation*., Brookings, Retrieved 20 January 2021, from https://www.brookings.edu/research/how-to-combat-fake-news-and-disinformation/

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal Of Physics: Conference Series*, *1168*, 022022. doi: 10.1088/1742-6596/1168/2/022022