Fake News Detection using Machine Learning Algorithms

Akash Patra[1], Tamojit Roy[2]
[1]Vellore Institute of Technology, Chennai, India
[2]Vellore Institute of Technology, Chennai, India
[1]E-mail: akash.patra2019@vitstudent.ac.in
[2]E-mail: tamojit.roy2019@vitstudent.ac.in

Introduction

In today's Internet world, people rely on different online services / platforms for news. As the use of social media platforms such as Facebook, Twitter, and news websites increases, news spreads rapidly to millions of people in a short period of time. News websites publish news and provide authentication sources. The problem is how to authenticate messages and articles distributed on social media such as WhatsApp groups, Facebook pages, Twitter, other microblogging and social networking sites. Believing in rumors of pretending to be news is bad for society. The spread of fake news has widespread consequences, including the generation of biased opinions that affect the outcome of elections in favor of a particular candidate. In addition, spammers use compelling headlines to monetize their ads through Clickbait. Especially in developing countries like India, it takes time to put an end to rumors and focus on the right certified news articles. This white paper presents models and methods for detecting fake news.

Social media platforms such as Facebook, Instagram, and Twitter provide a cheaper way to deliver news online much faster by spreading the network more easily. It's better than traditional news media, but it's a lot of fake news. News articles contain intentionally false information created online for a variety of purposes for economic and political gain. Therefore, there is an urgent need for a fake news detection system that not only distinguishes between fake news and real news, but also displays the relevant real news articles that are closest to the original.

Since dealing with fake news requires precision and with current advances in machine learning algorithms to detect such fake news, it would be foolish to depend on a single algorithm/method to perform such complex classification. Therefore, it makes much more sense to divide the work into phases to completely separate the work by data mining operations such as data collection, data pre-processing, feature extraction, feature selection and implementation of machine learning models to perform predictions to classify the news as True or False and also predict the probability that the news belongs to the predicted label.

Several machine learning models will be used to classify news as true or false. Each will be evaluated and compared against each other based on metrics like accuracy, f1 score, precision and recall. After the following machine learning models - SVM, Logistic Regression, Naïve Bayes and Random Forest will be trained and adjusted. A voting classifier will be implemented that will combine all the models mentioned above and form a composite classifier that uses all these classifiers to predict class labels and probabilities, and use soft voting to make the final prediction.

A suitable dataset that has been divided into training set, validation set and test set will be used and will be preprocessed to apply feature extraction techniques in the next steps. Models like Random Forest Classifier, Naïve Bayes, Logistic Regression and SVM Classifier will be used to train the dataset. Finally, the different scores for each category will be calculated and compared. Below is a detailed flowchart that summarizes the entire process.

Figure 1. Flow Chart

Literature Survey

This paper presents a model as well as a methodology for detecting fake news. Machine learning and natural language processing were used to aggregate messages, and later attempts were made to use support vector machines to determine if the message was genuine or fake. Results of the proposed model were compared to the existing models. The proposed model was found to be working well and had defined the correctness of results upto 93.6% of accuracy [1].

This work helped to judge the accuracy of the false news by using various different classification techniques. This work presents the solution for the Fake news problem by implementing fake news detection models by using different classification techniques. When resources are taken into consideration, Fake News Detection becomes complicated. Datasets are limited, which are the primary resources. In this model, classification techniques like Support Vector Machine (SVM), Naïve Bayes, and Passive Aggressive Classifier are used. Output of the model uses feature extraction techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and Support Vector Machine (SVM) as classifier and has accuracy of 95.05% [2].

This paper deals with binary classification of various news articles that are present on the internet with the help of concepts that are related to Artificial Intelligence, Machine Learning and Natural Language Processing. This paper supplies the ability of news classification as false or true and also verifies its source [3].

The importance of machine learning to detect fake news is proved in this research paper. The various ways in which fake news creeps into authentic sources are also mentioned in this research paper. The authors have also discussed how machine learning can be used for detecting false news. Machine learning algorithms that are used to detect false news have been discussed in the research paper [4].

This paper deals with the application of Natural Language Processing techniques to identify when a news source may be producing fake news. Collection of labeled real and fake news articles to build a classifier that can make decisions about information based on the content from the collection has been implemented. A text classification approach, using four different classification models to analyze the results has been implemented. LSTM implementation was the best performing model. The model

focused on identifying fake news sources, based on multiple articles originating from a source. High predictions were made when the source was flagged as a producer of fake news that future articles from this source will also be fake news. As multiple data points come from each source, it broadens the misclassification tolerance of the article [5].

In this paper, authors looked at how Natural Language Processing techniques may be used to detect when a news source is spreading bogus news. To develop a classifier that can make judgments about information based on the content in the corpus, the authors employed a corpus of labeled actual and false new articles. A text classification approach was applied and assessed the outcomes using four distinct classification models. The LSTM implementation produced the best results. Based on many articles originating from a source, the algorithm focuses on identifying fake news sources. It is anticipated with high confidence that any future publications from a source classified as a producer of fake news will also be fake news. Focusing on sources broadens the scope of our article misclassification tolerance, because we then have multiple data points coming from each source [6].

The goal of this paper is to investigate four well-known machine learning algorithms, namely the random forest, the Naïve Bayes, the neural network, and the decision trees, to see how effective they are at detecting fake news. On a widely used public dataset, LIAR, an experiment was performed, and the findings demonstrate that the Naïve Bayes classifier outperforms the other algorithms significantly [7, 12].

This paper deals with the collection of dataset from Kaggle. The information is divided into two categories: true and false news, and then integrated into a single dataset. The machine learning model is trained using this dataset. The authors attempted to develop a machine learning model utilizing four different classifiers and the Tf-idf vectorizer in this project. The goal is to foresee news that would mislead the user and cause pandemonium [8].

The authors extracted and analyzed real-time data from multiple domains across Twitter. The dataset was preprocessed and the user_verified column played an important role. Next, some machine algorithms were run on the features extracted from the preprocessed dataset. Logistic regression and support vector machines each yielded promising results with an accuracy of over 92%. Naive Bayes and long / short term memory did not achieve the desired accuracy. The model can also be applied to images and videos for better detection of fake news [9].

To distinguish bogus news from genuine news, a deep learning-based technique was applied in this study. The proposed model was created using an LSTM neural network. A gloVe word embedding was employed for vector representation of textual words in addition to the neural network. Tokenization has also been used for feature extraction and vectorization. The concept of N-grams is applied to improve the suggested model. The results of a comparison of several false news detection systems were examined. The suggested model's outcomes were assessed using accuracy measures. The model outperformed with a precision of 99.88 percent [10].

Machine learning is offered as a method for detecting bogus news. Using vectorization of the news title and our dataset to analyze the tokens of words. The dataset that the authors' have worked with is a pre-curated collection of news items that have the property of being false or not. Their goal was to create a model that can determine whether an article is real or false [11].

This research looks into advanced and cutting-edge false news detection systems in depth. The authors started with the negative repercussions of bogus news. Then they talked about the dataset that was used in earlier research and the NLP approaches that were used. To categorize representative methods into several categories, a complete overview of deep learning-based techniques has been presented. The most often used evaluation measures in the detection of false news are also reviewed. Nonetheless, in

future research paths, they proposed additional recommendations to improve fake news detection techniques [13].

Problem Statement

There are many ways to detect fake news and articles using machine learning algorithms. Given today's advances in machine learning and the level of knowledge reached, it is not appropriate to blindly rely on one or two such algorithms to classify fake news. Therefore, instead of collecting more data and extensive training to improve accuracy and comparing accuracy scores for better understanding, a holistic approach for detecting fake news is the need of this crucial moment. Therefore, we decided to take a step-by-step systematic approach to classify fake news using data mining technology. The implementation of data mining operations such as data collection, data preprocessing, feature extraction, feature selection, and machine learning models can be used to make predictions to classify messages as true or false and predict the probability of the news belonging to the predicted level. A set of machine learning models will be implemented to compare the performance of machine learning models based on metrics such as accuracy, f1 value, fit rate, and recall. The primary determinant for assessing model performance can be selected as the f1 score, which takes into account the trade-off between fit and recall. After the following machine learning models (naive Bayes, SVM, logistic regression, random forest) have been trained and tuned, all of the above models are combined to form an ensemble classifier that predicts using all of these classifiers. A voting classifier will be implemented. Label and class probabilities that use the soft voting method to make the final prediction. There are two approaches for machine learning models to gain insights from headings: CountVectorizer and TfidfVectorizer. These models are trained using features extracted from both CountVectorizer and TfidfVectorizer. All models are then hyperparameter adjusted in GridSearchCV with 5 holdout cross-validation sets for all different possible parameters. This hyperparameter adjustment is intended to improve the model's f1 score. After the models have been tuned, they are tested in the test set and the model's evaluation metrics are calculated. The trained and voted models are combined into a voting classifier that uses them all as the basic estimator. When new test data is passed to the voting classifier, all underlying models are built to predict sample specifications. After receiving the labels from all models, the final label of the test sample is predicted using the soft voting mechanism.

Proposed Work

The entire work revolves around the implementation of invaluable concepts of data mining operations like data collection, data preprocessing, feature extraction, feature selection and implementation of machine learning models for making the predictions of classifying the news into categories of true or false. Special emphasis will also be laid upon predicting the probability of the news belonging to the predicted label. As mentioned earlier, the entire work will be segregated into specific stages or modules based on the field of data mining. A set of machine learning models will be implemented to compare the performance of machine learning models based on a variety of metrics namely the notable ones such as accuracy, f1 value, fit rate, and recall. The primary factor for determination for assessing model performance will be selected as the f1 score, which will be held responsible for the trade-off between fit and recall. The machine learning models namely Naïve Bayes, Logistic Regression, SVM and Random Forest will be used for training. Based on the outcome of the training and tuning, a Voting Classifier will be implemented that will combine all the above machine learning models and form an ensemble classifier that will predict the class probability and label. It will also use the soft voting method to make the final prediction.

The proposed system will consist of a lineage of progressive steps, which encompasses all the nooks and crannies of accessing the data upon which we are going to operate upon, the extraction of the features, selection of the features, data modeling and also its evaluation. The steps for doing so are discussed in detail below:

1. Data Collection: We implemented and tested the proposed system using a new benchmark dataset for fake news detection. The dataset made use of was called William Yang Wang's Liar, Fire Liar Pants. The dataset in this repository is already divided into training sets, validation sets and test sets. The dataset contains 12,836 short statements labeled by truth, subject, context / location, speaker, state, party, and history. First, the dataset categorized the message into six fine-grained labels to assess its authenticity. Pants fire, fake, barely true, semi-true, almost true, true.

2. Data Preprocessing: In the proposed system, the 6-label classification problem is transformed into a binary classification problem with labels such as True and False. We converted the label using the following mapping.

   - pants-fire - False
   - false - False
   - barely-true - False
   - half-true - True
   - mostly-true - True
   - true - True

   Also, only the news headline was used as the classification input. Therefore, in the preprocessing phase, the labels are first mapped using the mapping defined above, then only the labels and news statement columns are extracted from the dataset, and the extracted dataset is for future use. The extracted datasets are saved in csv format for future use.

   After preprocessing, we got three clean files:

   - train.csv
   - valid.csv
   - test.csv

3. Feature Extraction: To make the machine learning models gain insights from the news headlines, two approaches of feature extraction were used:

   I. Count Vectorizer: Using scikit-learn's CountVectorizer, the English stopword was first removed from all news headlines and then tokenized with spaces and punctuation as delimiters. After all headlines were tokenized, a sparse matrix was returned with all news headlines as rows and tokens as columns. Also, in addition to morphological usage, a series of n-grams of tokens was returned to represent the context in which the token was used.

   II. Tfidf Vectorizer: Using scikit-learn's TfidfVectorizer, the English stopword was first removed from all news headlines and then tokenized with spaces and punctuation as delimiters. After all headlines were tokenized, a sparse matrix was returned with all news headlines as rows and tokens as columns. The similarity between different headings was calculated using the Tf Idf similarity measure. Also, in addition to morphological usage, a series of n-grams of tokens was returned to represent the context in which the token was used.

4. Modeling: In this step, we will be using some of the machine learning models for training the model using the feature extracted data. The machine learning models such as Naïve Bayes, Logistic Regression, SVM and Random Forest will be made use of for this purpose.

5. Methodology: The above model was trained using features extracted from both CountVectorizer and TfidfVectorizer. All models were then hyperparameter adjusted in GridSearchCV with 5 holdout cross-validation sets for all possible parameters. This hyperparameter adjustment was done to improve the model's f1 score. After the models were tuned, they were tested in the test set and the scoring metrics for the model were calculated.

Voting Classifier: The trained and matched models were combined into a voting classifier that used them all as basic estimators. When the new test data is passed to the voting classifier, all underlying models predict sample specifications. After receiving the labels from all models, the soft voting mechanism is used to predict the final label of the test sample.
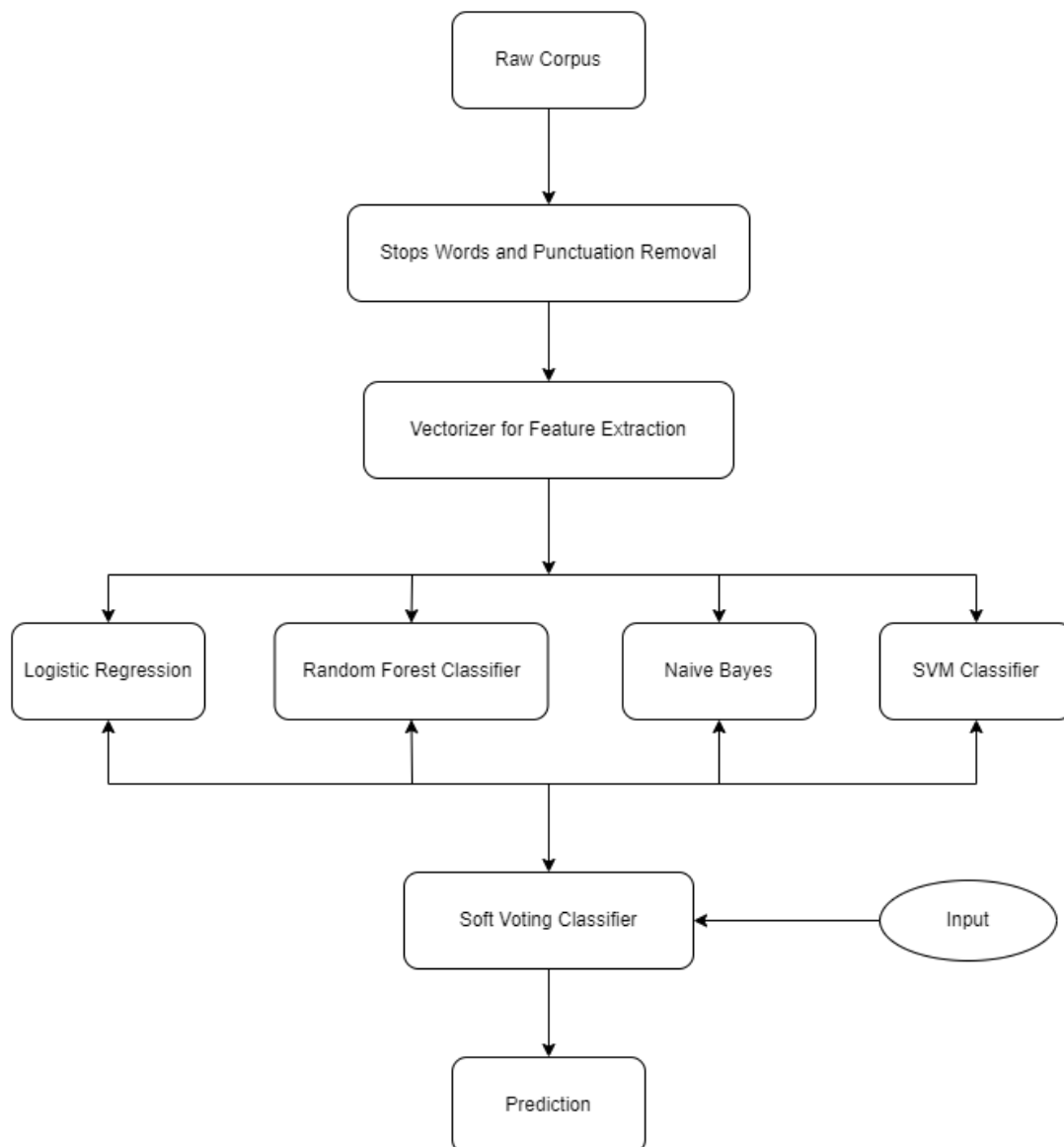
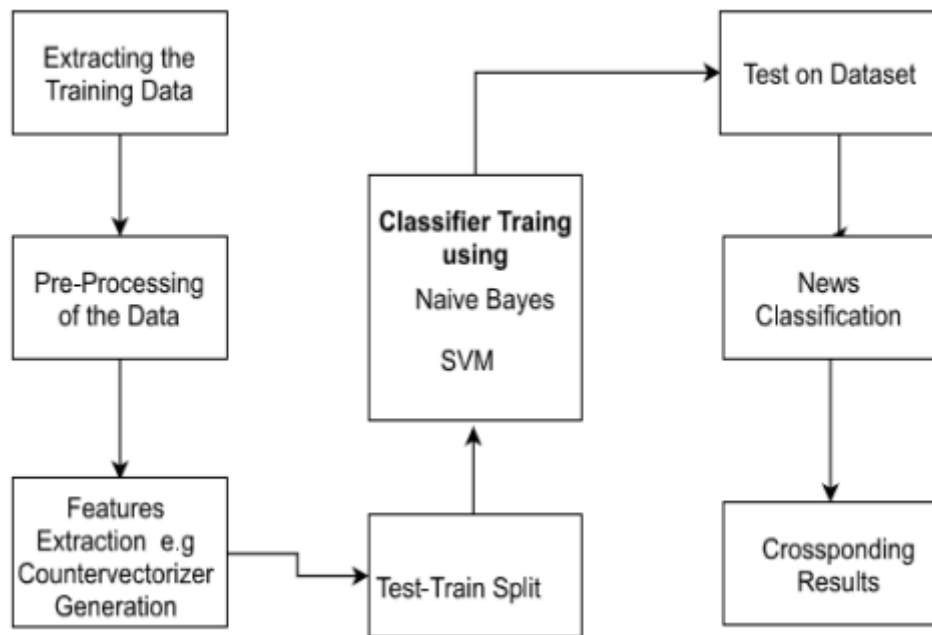## Architecture

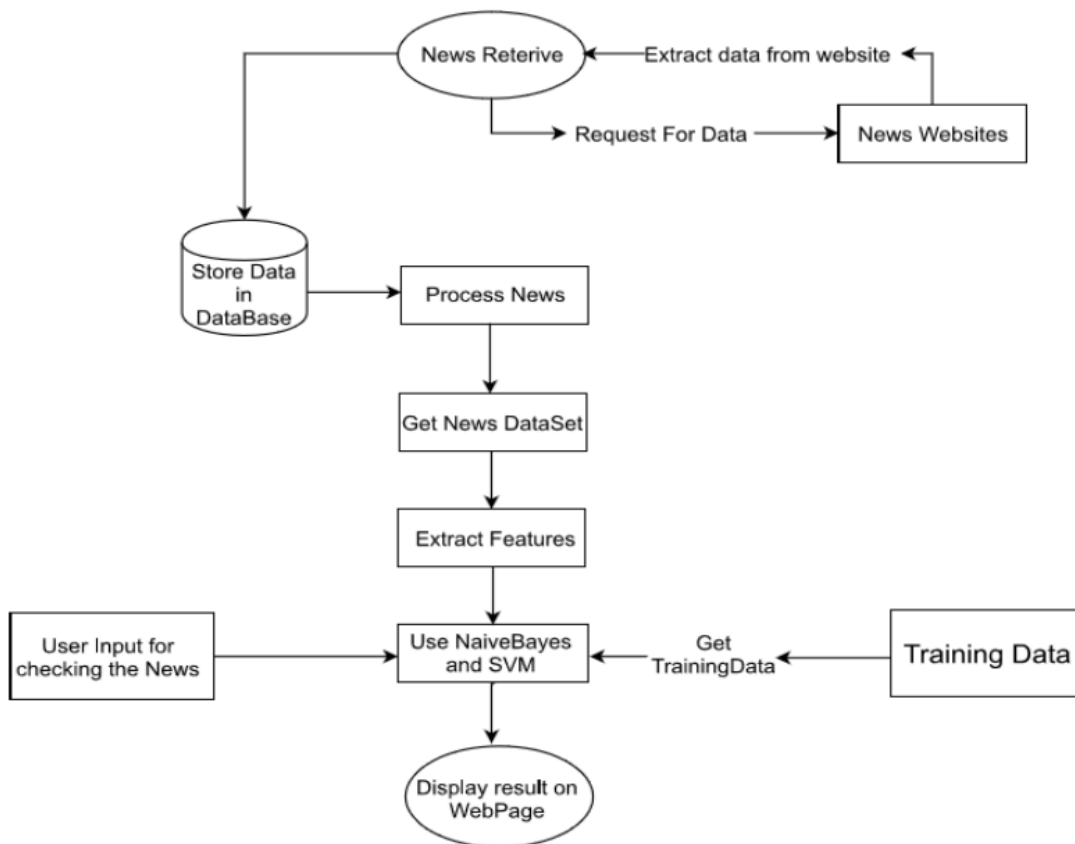Figure 2. Basic Architecture

Figure 3. Classifier Training

Figure 4. Proposed Model

Algorithms

Logistic Regression

This type of statistical analysis (also known as a logit model) is commonly used for predictive analytics and modeling and extends to machine learning applications. In this analytical approach, dependent variables are finite or categorical. Either A or B (binary regression) or a set of finite options A, B, C, or D (polynomial regression). Used in statistical software to understand the relationship between a dependent variable and one or more independent variables by estimating probabilities using logistic regression equations.

Predictive models built using this approach can make a positive difference to our business and organization. These models help you understand relationships and predict outcomes, so you can take action to improve your decisions. For example, a manufacturer's analysis team may use logistic regression analysis as part of a statistical software package to determine the probability of a machine's component failure and how long those components have been in inventory. Using the information obtained from this analysis, the team can decide to adjust the delivery schedule or installation time to avoid future outages.

$f(x) = L / 1 + e\textasciicircum-k(x - x0)$

where,

- $f(x)$ is the output of the function
- L is the curve's maximum value
- e is the base of the natural logarithms
- k is the steepness of the curve
- x is the real number
- x0 is the x values of the sigmoid midpoint

Random Forest Classifier

Random Forest is a widely used machine learning algorithm trademarked by LeoBreiman and AdeleCutler that combines the outputs of multiple decision trees to produce a single result. Its ease of use and flexibility allow it to handle both classification and regression problems, which is increasing its adoption.

This algorithm is an extension of the bagging method because it uses both bagging and feature randomness to create a forest of uncorrelated decision trees. Feature randomness is also known as the "random subspace method" or feature bagging. This produces a random subset of features that guarantees low correlation between all the determinants. This is an important difference between decision trees and random forests. The decision tree considers all possible functional decompositions, but Random Forest selects only a subset of those features.

Naive Bayes

This classification algorithm is a stochastic classifier. It is based on a probabilistic model that involves powerful assumptions of independence. Independence assumptions often have nothing to do with reality. Therefore, they are considered naive. We can use Bayes' theorem (due to Thomas Bayes) to derive a probabilistic model. Depending on the type of probabilistic model, we can train the naive Bayes algorithm in a supervised learning environment.

This is a group or collection of classification algorithms based on Bayes' Theorem. This algorithm is not considered as a single algorithm. This is a family of algorithms that all share the same principle. Each pair of classified features is independent of each other.

$$P (X \mid C_i) = \prod_{k=1}^{n} P (X_k \mid C_i) = P (X_1 \mid C_i) \text{ x } P (X_2 \mid C_i) \text{ x } \ldots \text{ x } P (X_n \mid C_i)$$

SVM Classifier

SVM is a famous supervised learning model that can be used for classification or regression. This approach works well in high dimensional space (many features in feature vectors) and can be used effectively on small datasets. Once the algorithm is trained on the dataset, new observations can be easily and efficiently classified. To do this, one or more hyperplanes are built to separate the dataset between the two classes.

Hyperplane: There can be multiple line/decision boundaries to separate classes in n-dimensional space, but we need to search for the best decision boundaries to help classify the datapoints. This best boundary is called the SVM hyperplane. The dimensions of the hyperplane depend highly on the features present in the dataset. That is, if there are two features (as shown in the image), the hyperplane will be a straight line. Also, if there are 3 features, the hyperplane is a two-dimensional plane.
Always create a hyperplane with the maximum margin, the maximum distance between data points.

Support Vector: The data point or vector closest to the hyperplane that affects the position of the hyperplane is called the support vector. These vectors are called support vectors because they support hyperplanes.
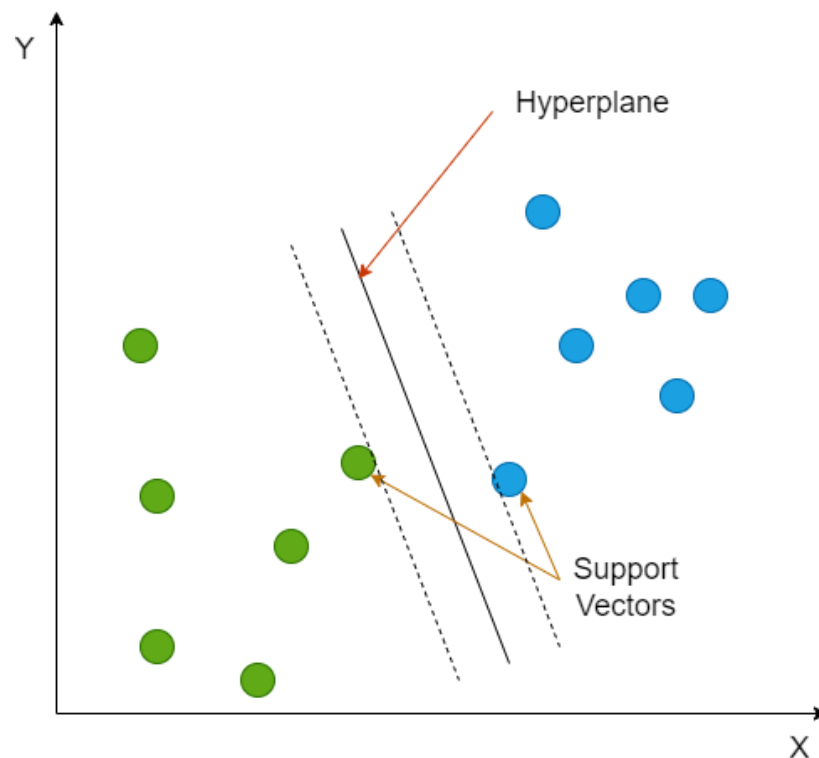

Figure 5. Hyperplane and Support Vectors in SVM

References

[1] Anjali Jain, Harsh Khatter and AvinashShakya, "A Smart System for Fake News Detection Using Machine Learning", International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), September 2019.

[2] Jasmine Shaikh, and Rupali Patil, "Fake News Detection using Machine Learning", 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), 2020.

[3] Uma Sharma, Sidarth Saran, and Shankar M. Patil, "Fake News Detection using Machine Learning Algorithms", International Journal of Engineering Research & Technology (IJERT), NTASU - 2020 Conference Proceedings.

[4] Alim Al Ayub Ahmed, Ayman Aljarbouh, Praveen Kumar Donepudi, and Myung Suh Choi, "Detecting Fake News using Machine Learning: A Systematic Literature Review", arXiv:2102.04458, Computers and Society: Machine Learning.

[5] K.ArunKumar, G.Preethi, and K.Vasanth, "A Study of Fake News Detection using Machine Learning Algorithms", International Journal of Technology and Engineering System (IJTES), Vol.11 No.1 – 2020 Pp.1-7.

[6] Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad, "Fake News Detection Using Machine Learning Ensemble Methods", Hindawi Complexity, Volume 2020, Article ID 8885861, October 2020.

[7] Abdulaziz Albahr, and Marwan Albahar, "An Empirical Comparison of Fake News Detection using different Machine Learning Algorithms", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 11, No. 9, 2020.

[8] Mhualika Dhar, Reet Rai, Shamanth J P, Soumita Das, and Shilpa M, "Detection of Fake News using Machine Learning Algorithms", International Journal of Technology and Engineering System (IJTES), IJARIIE-ISSN(O)-2395-4396, Vol-7 Issue-4 2021.

[9] Harika Kudarvalli, and Dr. Jinan Fiaidhi, "Detecting Fake News using Machine Learning Algorithms", TechRxiv - Powered by IEEE.

[10] Tavishee Chauhan, and Hemant Palivela, "Optimization and improvement of fake news detection using deep learning approaches for societal benefit", International Journal of Information Management Data Insights 1 (2021) 100051, Elsevier Publication.

[11] Nikhil Sharma, "Fake News Detection using Machine Learning", International Journal of Trend in Scientific Research and Development (ijtsrd), Volume-4, Issue-4, pp.1317-1320, June 2020.

[12] Anusha Acharya, Ashwitha Jathan, Deepa Anchan, Krithi Dinesh Kottary, and Mr. Sunil B.N "Fake News Detection using Machine Learning", Sahyadri International Journal of Research, Vol 5, Issue 1, 2019.

[13] M. F. Mridha, Ashfia Jannat Keya, Md. Abdul Hamid, Muhammad Mostafa Monowar, and Md. Saifur Rahman "A Comprehensive Review on Fake News Detection With Deep Learning", IEEE Access, Digital Object Identifier 10.1109/ACCESS.2021.3129329, November 2021.