# Predicting News Article Popularity With Multi Layer Perceptron Algorithm

**Arie Rachmad Syulistyo[1] , Dwi Puspitasari [2], Vira Meliana Agustin [3]**
[1,2,3] Politeknik Negeri Malang, Indonesia
Email: [1]arie.rachmad.s@polinema.ac.id, [2]dwi.puspitasari@polinema.ac.id
[3]virameliana18@gmail.com

**Abstract.** In the current era of information and technology, all media seems to have been digitized. One of them is news in print media which has now turned into online news. The increasing use of social media has made people's interest in reading news online to increase. The rapid flow of online news certainly creates new problems for writers to continue to innovate and present news that is always up to date. More and more, online news enthusiasts are booming. News needs to attract readers with headlines or news headlines, to anticipate reader preferences. Various online news media businesses want to know the future demand of readers, as well as whether the released news can reach more readers so that the news becomes popular.

Therefore, with the increasing interest in online news today, the authors are interested in analyzing the performance of Neural Network Algorithm and other techniques in predicting the popularity of news articles that can help the media to know whether their news will become popular. The news article popularity prediction system can increase their revenue if there are advertisements in the news. This system is widely used in various types of applications such as media advertising, traffic management, and economic trend forecasting. Artificial Neural networks are models with a high enough level of accuracy to perform tasks such as classification and prediction. This model is considered as a Multi Layer network of logistic regression units. Predicting News Article Popularity with Multi Layer Perceptron Algorithm can help news media in classifying whether the news is worthy / can be published with an accurate prediction of the popularity. The media can do the classification first before the news is in the hands of the public, so that the media can provide interesting content and headlines to achieve popularity.

**Keywords:** News, Popularity, Multi Layer Perceptron, Random Forest

## INTRODUCTION

In the current era of information and technology, all media seems to have been digitized. One of them is news in print media which has now turned into online news. Reading, writing and sharing information has become a part of life for people's entertainment (Ren & Yang, 2015). The emergence of online news makes people very interested in discussing all information for public consumption. This is supported by the development of social media such as YouTube, Instagram, Twitter, and Facebook so as to make people's interest in reading online news become increasing. It is undeniable, the advancement of social media seems to increase the distribution of online news media. That's why, information in online news flows so fast, so news becomes more dynamic with low cost but relatively short life span (Rezaeenour et al., 2018).

The fast flow of online news certainly creates new problems for writers to continue to innovate and present news that is always up to date. More and more, online news enthusiasts are booming. News needs to attract readers with headlines or news title, to anticipate the preference of readers. In other hand, the reader is able to anticipate the content of a news article before a headline, because a knowledge inside the content of news certainly in accordance with the content of the news (Lamprinidis et al., 2018). Various online news media businesses want to know the future demand of readers, as well as whether the released news can reach more readers. If they can find out if news can reach more readers, of course they will be better prepared to make decisions immediately in implementing news on their online platform (Rathord et al., 2019).

Therefore, the increasing interest in online news today, the authors are interested in analyzing the performance of Neural Network Algorithm and other techniques in a predicting news article popularity can help the media to know whether their news will become popular.

The predicting news article popularity system can increase their income if there are advertisements in the news. This system is widely used in various types of applications such as media advertising, traffic management, and economic trend forecasting. The author uses an online news data set from Tribun News Articles which containing almost 1000 news prediction. This model is considered as a Multi Layer network of logistic regression units. This model also has more layers and a complex structure, so the authors assume that neural networks are stronger for prediction systems than one-layer parametric logistic regression. Artificial Neural Network has widely used as one of predictive modelling. This method has good ability in analyzing data patterns, that's why this algorithm good in prediction. One of Artificial Neural Network that are often used as predictive model is Multi Layer Perceptron.

Previous research conducted by Priyanka Rathord (2019) with the title A Comprehensive Review on Online News Popularity Prediction using Machine Learning Approach conducting research with Comparative analysis of various popularity prediction methods, namely Random Forest, SVM, AdaBost, KNN, Naive Bayes, Linear Regression, Logistic Regression and Genetic Algorithm (Rathord et al., 2019). This research results in the accuracy of each algorithm in predicting the popularity of news, where Random Forest occupies the highest accuracy position. However, this research will still be improved by using the Neural Network algorithm and will be compared with the previous algorithm.

In Jalal Rezaeenour's (2018) research in a journal entitled Developing a New Hybrid Intelligent Approach for Prediction Online News Popularity, he conducted research on Algorithm can help news media in classifying whether the news is worthy / can be published with an accurate prediction of the popularity. The media can do the classification first before the news is in the hands of the public, so that the media can provide interesting content and headlines to achieve popularity.

title, news_articles descriptions, publish time, publish date, number of views from February 01, 2021 to April 08, 2021 to be processed in a model so that it can be classified to predict the popolarity.

Artificial Neural networks are models with a high enough level of accuracy to perform tasks such as classification and popular news prediction by utilizing the ELM (Extreme Learning Machine) Neural Network algorithm (Rezaeenour et al., 2018). This research shows that the most important predictors of popularity are the time for publishing news (higher number of visitors on weekends) and news topics (lifestyle and social media are the most popular topics on the site).

In Feras Namous' (2018) research entitled Online News Popularity Prediction, he conducted a study to determine popular news predictions using data sets from the Mashable News Website and compared algorithms for classification and prediction. The best algorithms with the highest level of accuracy for popular news prediction cases are Random Forest and Multi Layer Perception Neural Network.

In Feras Namous' (2018) research entitled Online News Popularity Prediction, the method with the best accuracy is Multi Layer Perceptron and Random Forest, so the researcher will compare the level of accuracy in the application of Multi Layer Perceptron with Random Forest. The dataset used consists of most popular articles from various publishers enriched with readers engagement (top articles) and "The Tribune" which is popular local English news website (https://www.tribuneindia.com/) of Punjab state and New Delhi.

Predicting News Article Popularity with Multi Layer Perceptron

## METHODS

The data used to conduct this research are articles from Tribun News https://www.kaggle.com/waseemakramkhan/the-tribune-news-articles. This dataset contains almost 1000 news titles, news_articles descriptions, publish time, publish date, number of views and popularity from February 01, 2021 to April 08, 2021 collected from the tribune news papers. On Tribun News Article Dataset, there is a column which define the popularity, conducted of is_popularity. This research will apply prediction of popularity by headlines/title of news article. The dataset used is balanced, the two classes have a number that is not much different, which is 432 for popular news data and 589 for non-popular news data.

The dataset is also taken from https://www.kaggle.com/datasets/szymonjanowski/internet-articles-data-with-users-engagement?resource=download. In the dataset there is a top_article attribute which indicates that the article is popular. Of the total 10436 data, only 3853 data were taken for research so that the overall dataset was more balanced for both popularity classes. The dataset with the top article class has unbalanced data, so the researcher only takes some and combines it with the first dataset. The result of the two datasets is a balance.

From total dataset will be split into three sets, consist of train, validation and test. From research by Islam, M. M., Karray, F., Alhajj, R., & Zeng, J. (2021) entitled "A Review on Deep Learning Techniques for the Diagnosis of Novel Coronavirus (COVID-19), the data partitioning step splits the data into training, validation, and testing set for the experiment. In the "Diagnosis Using Computer Tomography (CT) Images" scheme, this study uses a split dataset train, testing, validation scheme of 80:10:10, 60:20:20, 60:25:15, 60:30:10, 50:30:20.

In this research, the dataset will divided into three sets namely train, testing, and validation for the experiment with 3 scheme, consist of 80:10:10, 60:20:20, 60:25:15, 60:30:10 and 50:30:20 experiment pattern.

### System Design

Main design of Predicting News Article Popularity with Multi Layer Perceptron Algorithm research as follows :

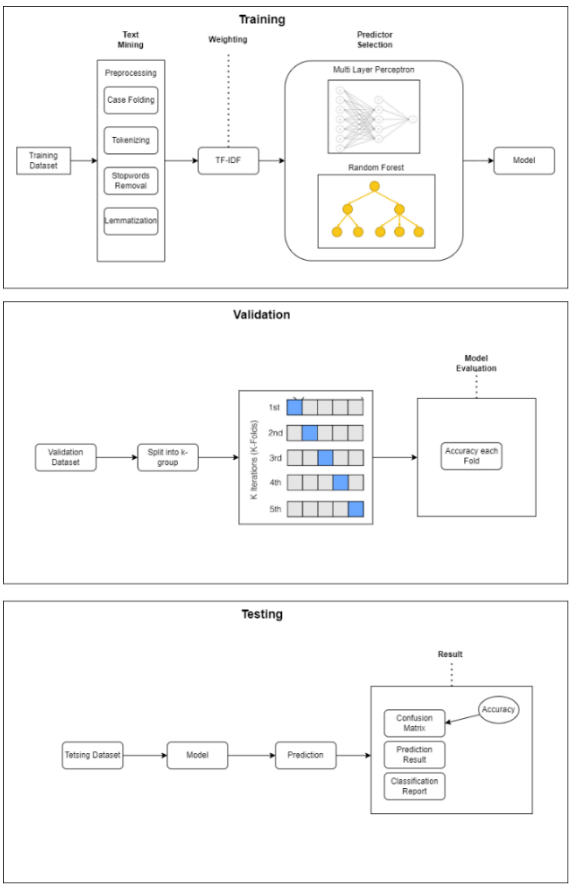


Figure 1. System Design

In this study, researcher will apply the text mining method to preprocess the data which has been divided into training, validation and testing stage. At the preprocessing stage, the data will be processed using the text mining method. In this method, the data will go through a case folding stage to convert letters to lowercase, tokenizing, Stopwords removal, and lemmatization.

The next training data will be a prediction process using the Multi Layer Perceptron method and Random Forest with input in the form of news article titles. First method is MLP. For the example, this method have 3 layers, that is input layers, hidden layers, and output layer. After modelling saved, next step is validation process.

After the training stage for the MLP algorithm, the data is then trained with the Random Forest algorithm which is also stored in the model. The preprocessing stage for this method is the same as previously described, the difference lies in the classification process on news popularity.

Validation will applied by 5-fold cross validation, which is split into a K number (on this stage K = 5) of section or fold where each fold is used as a testing set at some point. This step also gives an output the prediction and accuracy of each k-subset value. Testing step will use testing data and use model to process it. At the testing stage, the system will apply a confusion matrix to find out what percentage of the classification accuracy is.

Both of model will be compare between Random Forest and Multi Layer Perceptron, it means which method has the best accuracy rate for news popularity prediction. In today's prejudice, MLP has a better level of accuracy and performance. However, the results will still be determined based on the accuracy results at the testing stage.

Researcher only input news article title to show the prediction. Algorithm will predict the popularity. The entire process is built in the python programming language using the Scikit-learn, NLTK, Numpy, Pandas to Regular Expression libraries and is integrated on the website using the Flask RESTful API so that users can use the system more easily

**Preprocessing**

This stage is the stage for processing the data set. First begins with inputting data on article news titles which will be processed using Stopwordss, tokenizing and filtering all words.

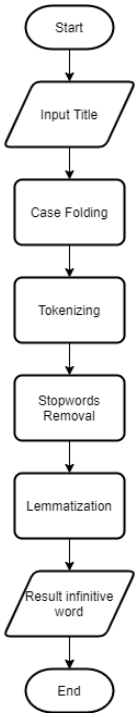


Figure 2. Flowchart Preprocessing

At the preprocessing stage figure 2 will produce clean data. This means that the data is in lowercase format, does not contain meaningless data, and consists of infinitive words with a valid meaning from lemmatization step.

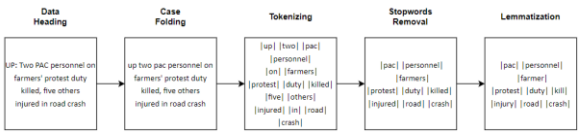


Figure 3. Preprocessing Process Example

Figure 3 above, the heading data will be processed through case folding. All letters will be returned to lowercase and remove punctuation marks. From this sentence, it goes to the tokenizing stage, which is dividing it into several word tokens. And each of these words will go through a stopword removal process that filters the words listed in the stopword list. Furthermore, the clean words enter the lemmatization stage, which is changed to basic words in English

**TF-IDF Process**

In the TF-IDF weighting the process for each word has gone through the preprocessing stage. In the stage of giving weights to words, it is necessary to use the TF-IDF method. This weighting aims to assign a value to a word that will be used as input in the

implementation of the model.

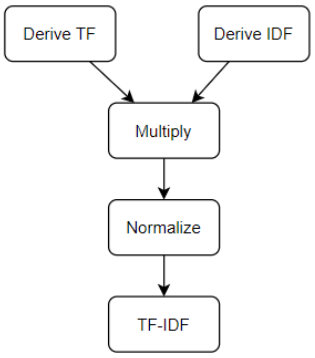


Figure 4. TF-IDF Process

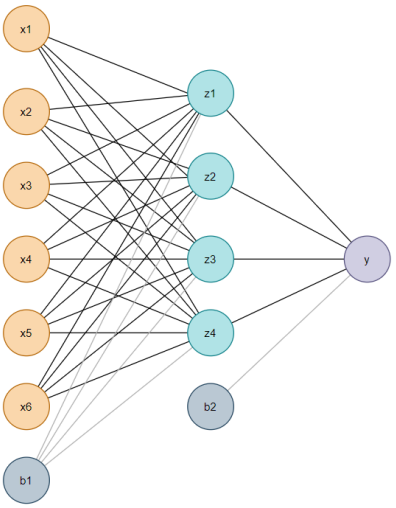


Figure 5. MLP Structure

**Data Processing**

At this stage the data processing will be carried out. To predict the popularity of news articles, it is necessary to apply several techniques. After collecting data and through all preprocessing step, the processed data will be classified using the Multi Layer Perceptron Algorithm method, which will classify predictions of popularity from news articles

**Prediction using Multi Layer Perceptron**

From the results of the words that have been processed by TF-IDF, then the prediction process will be carried out using the Multi Layer Perceptron method. The steps that need to be done in the Multi Layer Perceptron method are as follows:

a. Determine the number of input inputs, hidden layers, and outputs as training targets.
b. Randomly assigns initial values to all weights between the input-hidden and hidden-output layers.
c. Doing Feedforward.
d. Processing backpropagation.

In figure 5, Example Multi Layer Perceptron's structure consist of 3 layers, namely 1 input layer, 1 hidden layer, and 1 output layer. For input layer, consist of 6 inputs neuron and 1 bias neuron. The hidden layer consist of 4 hidden neurons and 1 bias neuron. And the last one is the output layer consist of 1 output neuron.

In the table 1 below, is a sample calculation of the Multi Layer Perceptron using 2 input layer neurons, 2 hidden layer neurons and 1 output layer.

Table 1. Table Sample of input value

| Prediction | Y | X1 | X2 |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 0 |
| 1 | 1 | 0 | 1 |
| 0 | 0 | 1 | 1 |

From the sample above, X1 and X2 is value of input layer and Y is actual result or popularity of news article. The prediction result need to be same with Y value. So first thing to do is define Network Parameters :

- Epoch = 150
- Bias = 1

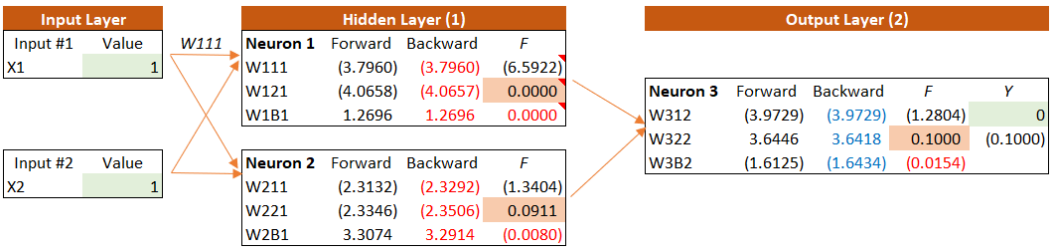


Figure 6. MLP Process Sample

The input layer neurons are forwarded to the Hidden Layer in Neurons 1 and 2. The first step is to calculate the Weight. W111 means Weight of Hidden Neuron Layer 1, Input Layer 1, first Weight while B means Bias. The value at the input layer will be calculated using the formula described in the previous chapter.

The higher the epoch value, the more accurate the results will be. Giving an epoch

value that is too high also does not have a good effect on training performance, so it is necessary to determine the right epoch value.

After calculate weight for forward, next is update weight for backward. From updated weight, next is define induced field and neuron output, which is 0.0 and 0.9 in both neuron of hidden layer.

**System Testing**

Testing will be carried out after the implementation phase is complete. Testing is very helpful for research to find out whether the system is running properly and appropriately. Testing of the Predicting News Article Popularity system with Multi Layer Perceptron Algorithm can be done by:

a. Perform User Acceptance Testing to run website-based applications that have implemented algorithms.
b. Testing the accuracy of all implemented methods and comparing the accuracy results between Multi Layer Perceptron and Random Forest

Calculation of accuracy can be done with the Confusion Matrix table. From the table, the calculation of accuracy, recall and precision can be displayed.

## RESULT AND DISCUSSION

**Results**

After implementing and testing, the results and discussion of the research that has been carried out are obtained. This section describes the results of the accuracy tests on the Multilayer Perceptron and Random Forest methods. The output of testing the accuracy of the system as a whole is as follows:

- The highest percentage accuracy of the Multilayer Perceptron method is 76% of the 80:10:10 (train:validation:test) split dataset from total dataset.
- The highest percentage accuracy of the Random Forest method is 70% of the 80:10:10 (train:validation:test) split dataset from total dataset.
- Validation testing using KFold Cross Validation get the result is quite much different. Each validation result in each ratio will be averaged.
- From the results of the 2 tests, it can be concluded that the Multi Layer Perceptron gets the highest accuracy in the "Predicting News Article Popularity with Multi Layer Perceptron Algorithm" which is 76% using MLPClassifier Library with 80:10:10 ratio split dataset.
- Accuracy increases when training data is added with a ratio of 80:10:10. The comparison between the training data ratio of 50% to 80% is 68% to 76%, meaning that the more and varied the data processed for the model, the higher the accuracy results. The highest result is 76%, can be influenced by inconsistent data (there are the same words in 2 classes). The use of news datasets from various online news sources influences the accuracy of the results.
- The results of the User Acceptance Testing concluded that the application that was developed based on the website was quite satisfactory for the head of the news article writer using English. The website is stated to make it easier for writers to manage news and article data to be done by other writers.
- The difference in accuracy between the Multi Layer Perceptron and Random Forest for this research is 6% different.

**Discussion**

After implementing the algorithm in the program code, as well as doing the testing, it is time to discuss the results of the research. Each stage of the research provides outputs that need to be discussed in detail. The following is a discussion of the results of the research "Predicting News Popularity with Multi Layer Perceptron":

**Split Dataset**

Before starting, the first stage that must be prepared is the dataset. The total dataset is 4877 data. The entire data will be split into 3 parts, namely train, validation and test. This data comparison produces an output in the form of a CSV file with a total split of data:

Table 2. Table Ratio Split Dataset

| Ratio | Train | Validation | Testing |
|---|---|---|---|
| 80:10:10 | 3901 | 488 | 488 |
| 60:20:20 | 2926 | 975 | 976 |
| 60:25:15 | 2926 | 1219 | 732 |
| 60:30:10 | 2926 | 1463 | 488 |
| 50:30:20 | 2438 | 1463 | 976 |

From table 2, researcher already have 5 ratio split dataset. All split data will return to CSV file. Each ratio will be tested on testing phase and select the best accuracy from 5 ratio.

**Preprocessing**

Entering the training stage, the training data CSV file is used as input to run the

preprocessing function that produces clean data. This clean data will be input for weighting in the next stage. The following is a representation of the preprocessing results:

Table 3. Table Preprocessing Stage Result

| Row | Clean Data |
|---|---|
| 0 | david,duffy,led,cybg,share,plunge,ppi |
| 1 | another,gop,congressman,texas,say |
| 2 | tory,lose,working,majority,ahead,crunch |
| 3 | never,mind,politics,get,brexit,deal,done |

The table 3 provides a representation of the results of preprocessing on the training data of 3901 data. Each description develop different sizes of words.

**TF-IDF Vectorizer**

The results of the preprocessing become input in the word weighting stage using the TF-IDF Vectorizer. The results of the TF-IDF Vectorizer are represented in the following table.

Table 4. Table TF-IDF Stage Result

| Term | TF-IDF |
|---|---|
| ability | 0.434973 |
| abroad | 0.418352 |
| absolut | 0.625648 |
| accept | 0.25768 |
| accident | 1.886203 |
| account | 4.427424 |
| achievement | 0.383056 |
| ... | ... |
| advertising | 0.34767 |
| advice | 0.364052 |
| affair | 0.759717 |
| affect | 1.420336 |
| aftermath | 1.358144 |
| afternoon | 0.249418 |

Table 4 is representation of TF-IDF Vectorizer result. All word in the table above saved in get_features_name function. The whole word is propagated to be the TF-IDF value for each word. If there is no such word in a sentence, then the stored value is 0.

**Multi Layer Perceptron**

The vectorizer results stored in the pickle can be extracted to perform the training, validation, and testing processes. In the training process using Multi Layer Perceptron, there are many parameters that we need to fill below:

Table 5. MLP Train Parameters

| | |
|---|---|
| Activation Function | Relu |
| Hidden Layer Sizes | 256 Neurons (3 Layers) |
| Solver | Adam |
| Input Layer | 8537 Neurons |
| Output Layer | 2 Neurons (2 classes) |
| Alpha | 0.0001 (default) |
| Batch Size | 256 |
| Learning Rate | Constant learning rate |
| Learning Rate Init | 0.001 |
| Max Iteration | 200 |
| Shuffle samples | True |
| Random State | None |
| Tol | 0.001 |
| Verbose | True |

From table 5, there is 1 input layer with a total of 8537 neurons. All neurons will be associated with each neuron in the hidden layer, where there are 3 hidden layers with 256 neurons. Each hidden layer has Relu Activation Functions. In the third hidden layer, it is connected to 1 output layer with 2 neurons. When it reaches the output layer, it is called forward. Furthermore, back propagation will be carried out from the output - hidden layer by updating the weight of each neuron. Learning is carried out for a maximum of 200 iterations and 256 batch sizes.

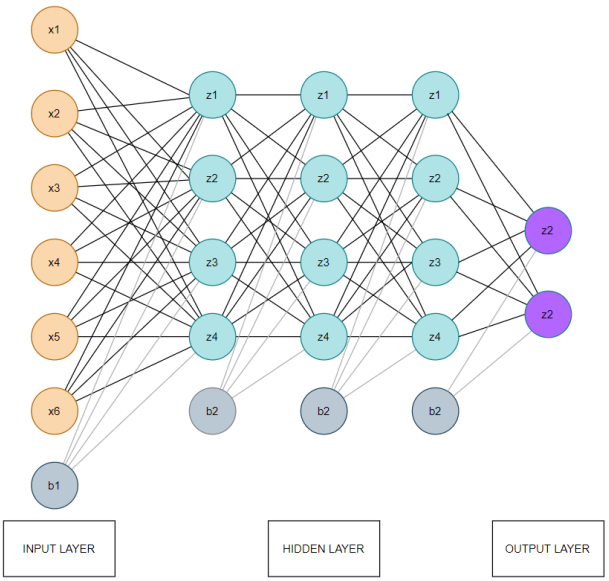


Figure 7. MLP Structure Representation

Figure 7 is a representation of the Multi Layer Perceptron network. The number of input layers is adjusted to the number of terms from the training results which consist of 8537 neurons through 3 hidden layers, each layer consisting of 256 neurons which ends up in the output layer as 1 feed forward. The number of output layers is adjusted to the number of classes being trained, namely 2 classes.

**Random Forest**

In the Random Forest algorithm, the input still uses the vectorizer results from clean data. The parameters used in the Random Forest algorithm are 100 n_estimators, which means that there are 100 trees in the algorithm that is run in training.

Table 6. Table Random Forest Training Parameters

| | |
|---|---|
| N Estimators | 100 |
| Criterion | gini |
| Max Depth | None |
| Min Samples Split | 2 |
| Min Samples Leaf | 2 |
| Min Weight Franction Leaf | 0.0 |
| Max Features | Sqrt |
| Max Leaf Nodes | None |
| Min Impurity Decrease | 0.0 |
| Random_state | None |

From table 6, the algorithm will created a trees as much as 100 trees. The maximum depth of tree will be stopped until all leaves contain less than min_samples_split, which is 2 for default.


Figure 8. Random Forest Representation

From figure 8, there example of representation of random forest with decision tree (until 100). All the result of each decision tree will be calculate by majority vote to get final result.

**System Testing Result**

After building the system, it is necessary to do black box testing according to the scenario prepared in the previous chapter. The results of black box testing are represented in the following table:

Table 7. Table Black Box Testing

| | Scenario | Hoped Result | Result | Status |
|---|---|---|---|---|
| 1 | Submit form news popularity prediction on home page | Prediction result will show in the bottom of form with keyword | Pass | Succeed |
| 2 | Register new account for user | User success create new account | Pass | Succeed |
| 3 | Submit form login account | If the form validations is true, user redirect to dashboard page. If not, user still on login page | Pass | Succeed |
| 4 | Submit form news popularity prediction on dashboard page | Prediction result will save on database and redirect to history prediction | Pass | Succeed |
| 5 | Export data prediction | Export success with PDF, CSV, Word format | Pass | Succeed |
| 6 | Edit prediction data in history page and resubmit | The system will generate new prediction result and edit the last one | Pass | Succeed |
| 7 | Delete data | The system | Pass | Succeed |

| | | | | |
|---|---|---|---|---|
| | prediction | will delete the selected data | | |
| 8 | Display dashboard of classification and accuracy report | Display the dashboard with classification report | Pass | Succeed |

From all testing scenarios, the results are appropriate which can be concluded as successful. All features have been tested with expected result.

**User Acceptance Testing Result**

Testing needs to be done to find out the Design and Build of a News Popularity Prediction Website as needed and has been running correctly. Tests are carried out using the User Acceptance Testing (UAT). The following are the results of the questionnaire testing the UAT method which is implemented in News Popularity Prediction system. This stage aims to obtain information whether the system that has been built is in accordance with user needs. Testing is intended to test the extent to which the application can function and be useful according to needs.

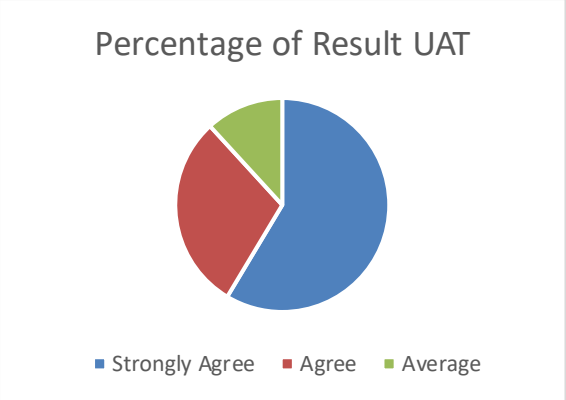


Figure 9. UAT Result

From each percentage of respondents as many as 15 users taken on August 2, 2022, then the highest and lowest score can be calculated as follows:

Table 8. Table Highest and Lowest Score

| | |
|---|---|
| Highest Score | 15 x 8 x 5 = 600 (if all respondent answer strongly agree) |
| Lowest Score | 15 x 8 x 1 = 120 (if all respondent answer strongly disagree) |

From the calculation which states the highest value is 4800 so that the results of the percentage of UAT tests can be found as follows:

Percentage = 486 x 600 x 100% = 81%

From the results of the percentage above, it can be concluded that the level of usability of the system is strong, which is 81% from 100%.

**Accuracy Testing Result**

Accuracy testing aims to determine the level of success of the system in predicting the popularity of news by using several testing samples that have been split in the system.

**Multilayer Perceptron Testing (Library)**

Before entering the testing stage, researchers need to calculate the validation value using KFold Cross Validation. From the model parameters used for training and data validation, there are 5 ratio split dataset model to validate.

Table 9. Table MLP K-Fold Cross Validation

| Ratio | Fold | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 80:10:10 | 0.56 | 0.67 | 0.61 | 0.49 | 0.57 |
| 60:20:20 | 0.57 | 0.6 | 0.61 | 0.55 | 0.58 |
| 60:25:15 | 0.65 | 0.57 | 0.6 | 0.64 | 0.63 |
| 60:30:10 | 0.64 | 0.65 | 0.6 | 0.61 | 0.61 |
| 50:30:20 | 0.59 | 0.63 | 0.65 | 0.6 | 0.61 |

From the validation test, the highest average value lies in ratio 50:30:20 which is equal to 61%. Values from k-fold 1 to 5 have insignificant differences.

Testing the accuracy of the Multi Layer Perceptron method using the MLPClassifier library with 5 dataset split data represent by confusion matrix on table as follows:

Table 10 MLP Confusion Matrix for Ratio 80:10:10

| Actual/Prediction | Not Popular' | Popular' |
|---|---|---|
| Not Popular | 181 | 59 |
| Popular | 58 | 190 |

From the table 10 above, there are 190 popular classes and 181 not popular class data which are predicted to be correct. Meanwhile, the other 118 data were predicted to be incorrect. From the table above, the results of the classification report calculation are as follows:

Table 11. Table MLP Classification Report

| Ratio | Precision | Recall | F1 | Acc |
|---|---|---|---|---|
| 80:10:10 | 1 = 0.76<br>0 = 0.76 | 1 = 0.77<br>0 = 0.75 | 1 = 0.76<br>0 = 0.76 | **0.76** |
| 60:20:20 | 1 = 0.76<br>0 = 0.68 | 1 = 0.63<br>0 = 0.79 | 1 = 0.69<br>0 = 0.73 | 0.71 |
| 60:25:15 | 1 = 0.73<br>0 = 0.67 | 1 = 0.63<br>0 = 0.77 | 1 = 0.67<br>0 = 0.71 | 0.69 |
| 60:30:10 | 1 = 0.74<br>0 = 0.69 | 1 = 0.63<br>0 = 0.79 | 1 = 0.68<br>0 = 0.73 | 0.71 |
| 50:30:20 | 1 = 0.69<br>0 = 0.68 | 1 = 0.66<br>0 = 0.72 | 1 = 0.67<br>0 = 0.70 | 0.68 |

The table 11 is a classification report. From the report, it is stated that the MLPClassifier model gives a highest accuracy at ratio 80:10:10 split dataset which is 76% with the precision value for not popular predictions is 76%, recall is 77%, and f1-score is 76%. Meanwhile, for popular predictions, precision is 76%, recall is 75% and f1-score is 76%.

**Random Forest Testing**

In this study, the Random Forest Algorithm was used as a comparison method with the Multilayer Perceptron. This method uses the same split dataset with Multi Layer Perceptron. The stage before testing is to run the validation function using KFold Cross Validation with a total of 5 Folds.

Table 12. Table Random Forest K-Fold Cross Validation

| Ratio | Fold | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 80:10:10 | 0.55 | 0.54 | 0.52 | 0.50 | 0.59 |
| 60:20:20 | 0.62 | 0.62 | 0.59 | 0.56 | 0.62 |
| 60:25:15 | 0.63 | 0.58 | 0.64 | 0.62 | 0.62 |
| 60:30:10 | 0.61 | 0.61 | 0.62 | 0.57 | 0.57 |
| 50:30:20 | 0.63 | 0.64 | 0.67 | 0.61 | 0.62 |

From the validation test, the highest average value lies in ratio 50:30:10 which is equal to 63%. Values from k-fold 1 to 5 have insignificant differences. Testing the accuracy of the Random Forest method using the RandomForestClassifier library with 5 dataset split data represent by confusion matrix on table as follows:

Table 13. Table Random Forest Confusion Matrix Ratio 80:10:10

| Actual/Prediction | Not Popular' | Popular' |
|---|---|---|
| Not Popular | 193 | 47 |
| Popular | 98 | 150 |

From the confusion matrix table, there are 193 not popular data and 150 popular data which are predicted to be correct, while the other 155 data are predicted to be wrong. Table 6.16 below is a classification report is also given that describes the value of precision, recall, and f1-score in each class.

Table 14. Table Random Forest Classification Report

| Ratio | Precision | Recall | F1 | Acc |
|---|---|---|---|---|
| 80:10:10 | 1 = 0.76<br>0 = 0.66 | 1 = 0.60<br>0 = 0.80 | 1 = 0.67<br>0 = 0.73 | **0.70** |
| 60:20:20 | 1 = 0.72<br>0 = 0.61 | 1 = 0.48<br>0 = 0.81 | 1 = 0.58<br>0 = 0.69 | 0.64 |
| 60:25:15 | 1 = 0.73<br>0 = 0.60 | 1 = 0.46<br>0 = 0.83 | 1 = 0.57<br>0 = 0.70 | 0.64 |
| 60:30:10 | 1 = 0.73<br>0 = 0.62 | 1 = 0.48<br>0 = 0.83 | 1 = 0.58<br>0 = 0.71 | 0.65 |
| 50:30:20 | 1 = 0.71<br>0 = 0.62 | 1 = 0.50<br>0 = 0.80 | 1 = 0.58<br>0 = 0.70 | 0.65 |

From the table 14, the highest accuracy is at ratio 80:10:10 split data with 70% accurate. The classification report for the not popular class has a precision of 76%, recall of 60% and an f1-score of 67%. Meanwhile, the popular class has a precision of 66%, a recall of 80% and an f1-score of 73%.

**Comparison Method Between MLP and Random Forest**

After getting the accuracy results from the three implementations of the method and 2 algorithms, the next step is to provide an accuracy comparison. The three methods use the same amount of data and balance in each split dataset. The comparison of the Multi Layer Perceptron and the Random Forest is represented in the following table:

Table 15. Table Comparison Accuracy

| | MLP | Random Forest |
|---|---|---|
| Accuracy | 76% | 70% |

From the results of the comparison of the method models, it can be seen that the highest accuracy is obtained by the Multi Layer Perceptron algorithm with MLPClassifier library with an accuracy value of 76%. The difference in accuracy with the Random Forest algorithm is quite large, namely 6% because this algorithm reaches 70% accuracy for news headline data.

**Experimental Result**

In the system, the user can make predictions on the landing page. The following are the results if the input gives an not popular prediction result.

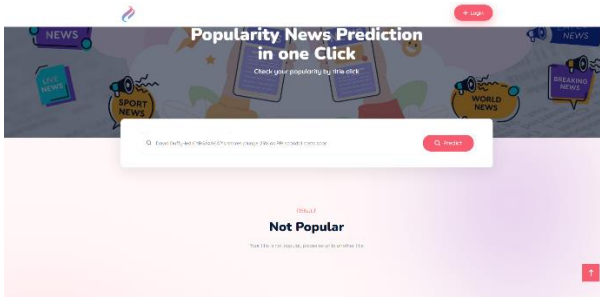


Figure 10. Not Popular Landing Page

Users can correct the title until the check gives popular prediction results. Here are figure 6.5 for popular prediction results.

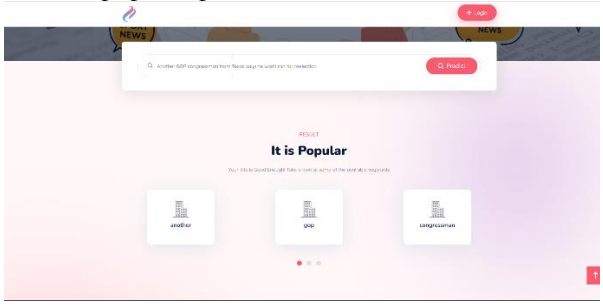


Figure 11. Popular Landing Page

If the prediction results are popular, the system will provide keyword suggestions that can help users choose keywords according to the title. The result of this keyword is the result of preprocessing, so meaningless words have been removed.

If the user registers an account and logs into the dashboard, the user can save the prediction results into the database. This feature helps users in managing headline data before the article is published.

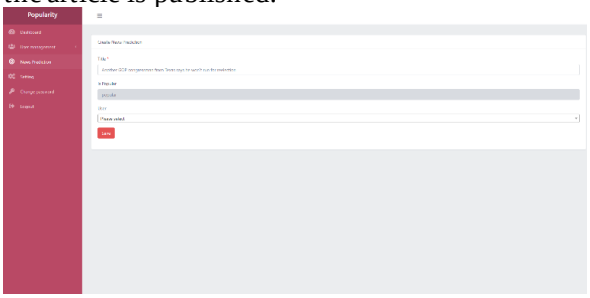

Figure 12. Prediction Page

The dashboard page also displays the results of the comparison accuracy of the three methods implemented in the system.

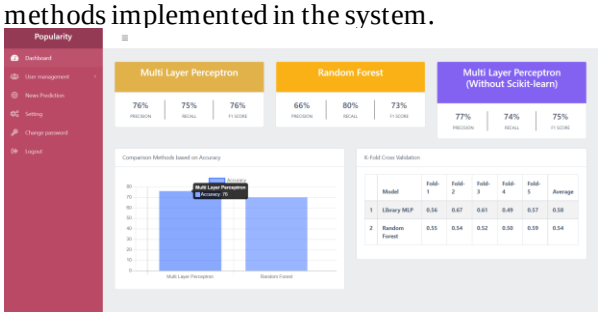


Figure 13. Dashboard Page

On the page above, there are comparison results of the classification report consisting of recall, precision and f1-score of the 2 methods. After that, there is a comparison of accuracy in the form of a bar chart. Besides that, there are validation results using k-fold cross validation. highest accuracy value in bold in each method.

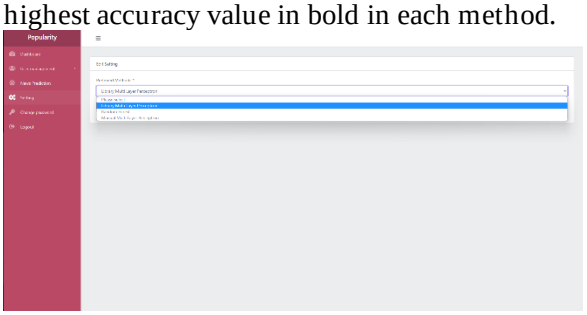

Figure 14. Setting Page

Figure 6.8 above is the settings page for changing the preferred method. If you change the method, the prediction results will be adjusted to the selected model. So it could be with the same title, the results obtained may be the same and may not. Because the accuracy results of the Multi Layer Perceptron and Random Forest methods are too 6% the difference.

## CONCLUSION AND SUGGESTION

Based on the research that has been done by the author, it can be concluded as follows:

1. From the test results that have been described in detail in chapter V, the accuracy of the Multi Layer Perceptron algorithm using the library is 76%. Comparison of accuracy is done with Random Forest which gives an accuracy of 70%. The difference in accuracy is 6%, so it can be concluded that for the Predicting News Popularity with Multi Layer Perceptron research, the algorithm with the best classification and accuracy results is achieved by Multi Layer Perceptron.
2. The News Popularity Prediction system is built with a website-based front-end so that it can be accessed flexibly and can assist writers in managing the right headlines for news content development.
3. Predicting the popularity of news in the system is given the Setting feature, which is a feature that can choose the preferred method for prediction in the system. Prediction can be processed using Multi Layer Perceptron Method or Random Forest.

From the results of the research that has been done, there are several suggestions as follows:

1. The prediction process still uses 1 input in the system. In the future, it can be developed so that it can predict more than 1 input in 1 prediction form.
2. Added other features such as news

management for internal company.
3. Can be developed with mobile-based applications.

## REFERENCES


Anggara, B. T. (2019). *Sistem Prediksi Tingkat Inflasi Provinsi Jawa Timur Menggunakan Metode Multilayer Perceptron*. 1–8.

B, P. C., & Oliveira, L. (2019). *Confusion Matrix-Based Building* (Vol. 1). Springer International Publishing. https://doi.org/10.1007/978-3-030-13469-3

Boumans, J., Trilling, D., Vliegenthart, R., & Boomgaarden, H. (2018). The Agency Makes the (Online) News World Go Round: The Impact of News Agency Content on Print and Online News. *International Journal of Communication*, *12*, 1768–1789.

Deolika, A., Kusrini, K., & Luthfi, E. T. (2019). Analisis Pembobotan Kata Pada Klasifikasi Text Mining. *Jurnal Teknologi Informasi*, *3*(2), 179. https://doi.org/10.36294/jurti.v3i2.1077

Gerlach, M., Shi, H., & Amaral, L. A. N. (2019). A universal information theoretic approach to the identification of stopwords. *Nature Machine Intelligence*, *1*(12), 606–612. https://doi.org/10.1038/s42256-019-0112-6

Halimah Khoirunisa, T. (2018). *Implementasi web service untuk handwriting recognition dengan rest api*. 7–16. http://repository.ittelkom-pwt.ac.id/5425/

Henri. (2018). 済無No Title No Title No Title. *Angewandte Chemie International Edition*, *6(11)*, *951–952.*, 1–14.

Khoirunissa, H. A., Widyaningrum, A. R., & Maharani, A. P. A. (2021). Comparison of Random Forest, Logistic Regression, and MultilayerPerceptron Methods on Classification of Bank Customer Account Closure. *Indonesian Journal of Applied Statistics*, *4*(1), 14. https://doi.org/10.13057/ijas.v4i1.41461

Lamprinidis, S., Hardt, D., & Hovy, D. (2018). Predicting News Headline Popularity with Syntactic and Semantic Knowledge Using Multi-Task Learning. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 659–664. https://aclanthology.info/papers/D18-1068/d18-1068

Meesad, P. (2021). Thai Fake News Detection Based on Information Retrieval, Natural Language Processing and Machine Learning. *SN Computer Science*, *2*(6), 1–17. https://doi.org/10.1007/s42979-021-00775-6

Namous, F., Rodan, A., & Javed, Y. (2018). Online News Popularity Prediction. *2018 Fifth HCT Information Technology Trends (ITT)*, *November 2018*, 180–184. https://doi.org/10.1109/CTIT.2018.8649529

Ngantung, R. K., & Pakereng, M. A. I. (2021). Model Pengembangan Sistem Informasi Akademik Berbasis User Centered Design Menerapkan Framework Flask Python. *Jurnal Media Informatika Budidarma*, *5*(3), 1052. https://doi.org/10.30865/mib.v5i3.3054

Rathord, P., Jain, D. A., & Agrawal, C. (2019). A Comprehensive Review on Online News Popularity Prediction using Machine Learning Approach. *Smart Moves Journal Ijoscience*, *5*(1), 7. https://doi.org/10.24113/ijoscience.v5i1.181

Ren, H., & Yang, Q. (2015). Predicting and Evaluating the Popularity of Online News. *Conference Proceedings*. https://pdfs.semanticscholar.org/9e91/6a3469e9e2fc5f0c8f927d7d1d05f5575729.pdf%0Ahttp://cs229.stanford.edu/proj2015/328_report.pdf

Rezaeenour, J., Eili, M. Y., Hadavandi, E., & Roozbahani, M. H. (2018). Developing a new hybrid intelligent approach for prediction online news popularity. *International Journal of Information Science and Management*, *16*(1), 71–87.

Rosid, M. A., Fitrani, A. S., Astutik, I. R. I., Mulloh, N. I., & Gozali, H. A. (2020). Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi. *IOP Conference Series: Materials Science and Engineering*, *874*(1). https://doi.org/10.1088/1757-899X/874/1/012017

Rostianingsih, S., Sugianto, S. A., & Pustaka, S. (2012). *Aplikasi Predictive Text Berbahasa Indonesia Dengan Metode N-Gram*. 1–6.

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2018). Using text mining techniques for extracting information from research articles. *Studies in Computational Intelligence*, *740*, 373–397. https://doi.org/10.1007/978-3-319-67056-0_18

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *PLoS ONE*, *16*(8 August). https://doi.org/10.1371/journal.pone.0254937

Sholih 'afif, M., Muzakir, M., Al, M. I., & Al Awalaien, G. (2021). Text Mining Untuk Mengklasifikasi Judul Berita Online Studi

Kasus Radar Banjarmasin Menggunakan Metode Naïve Bayes. *Kumpulan JurnaL Ilmu Komputer (KLIK)*, *08*(2), 199–208.

Sriyano, C. S., & Setiawan, E. B. (2021). Pendeteksian Berita Hoax Menggunakan Naive Bayes Multinomial Pada Twitter dengan Fitur Pembobotan TF-IDF. *E-Proceeding of Engineering : Vol.8, No.2*, *8*(2), 3396–3405.

Wahid, F., Ghazali, R., Shah, A. S., & Fayaz, M. (2017). Prediction of Energy Consumption in the Buildings Using Multi-Layer Perceptron and Random Forest. *International Journal of Advanced Science and Technology*, *101*, 13–22. https://doi.org/10.14257/ijast.2017.101.02

Islam, M. M., Karray, F., Alhajj, R., & Zeng, J. (2021). A Review on Deep Learning Techniques for the Diagnosis of Novel Coronavirus (COVID-19). *IEEE Access*, *9*, 30551–30572. https://doi.org/10.1109/ACCESS.2021.3058537