News Classification Based On News Headline Using SVC Classifier

Goldius Leonard
Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
goldius.leonard@binus.ac.id

Fukriandy Sisnadi
Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
fukriandy.sisnadi@binus.ac.id

Nicholas Vigo Wardhana
Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
nicholas.wardhana@binus.ac.id

Muhammad Abdul Aziz Al-Ghofari
Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
muhammd.ghofari@binus.ac.id

Abba Suganda Girsang
Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480,
agirsang@binus.edu

2022 16th International Conference on Telecommunication Systems, Services, and Applications (TSSA) | 978-1-6654-5594-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/TSSA56819.2022.10063879

***Abstract*—News gives new insight and information from all over the world. News has many categories, such as politic, economy, science, and other common news categories. Every news will have their own category based on its content. The classification of news is usually done manually by inputting the category during the news posting. Some of the categories may be inputted incorrectly. The news classifier can be the solution for problem, but the news classifications out there are usually based on the news content. The classifier will receive the word vector inputs that are taken from the news content and try to classify it into one of certain categories. Unfortunately, news contents can be longer and harder to be processed rather than processing the news headline. The news headline is shorter and packs a decent information for the classifier to find out what category it is. Besides the news headline usage, the classifier also needs to be chosen correctly. In this paper, the SVC model will be tested using the news headline data to classify the news and compare with several other models, such as Linear Regression, Multinomial Naïve Bayes, Decision Tree, and Random Forest. The common variables to be compared are the accuracy, recall, and precision to evaluate the SVC model.**

Keywords—News Classification, News Category, Text Classification, TF-IDF, SVC

I. INTRODUCTION

News is a closely related thing to human's daily life. Everyone should have read news either daily or quite sometimes. News keeps us up to date with the world and our surroundings. The information gained from news is huge because it varies from every of life aspect.

News can be classified into hard news, soft news, and 'general' news based on Sam N and Michal [1]. The hard news is news that is important to be published as soon as possible and has a great impact to the public like news about people or events. The soft news is a less important and less impact news that talks about gossips, fashion, and consumerism. This news type is usually published in infotainment. The last type, 'general' news, is in between the hard news and soft news. It is an up-to-date news but not necessarily immediately [1]. It is also influencing certain groups only.

Besides the news type above, every news is categorized into a news category or several categories. Every news category has its own word characteristics. The category of a news can be seen from the words used in the content. For example, the technology category will use words like 'computer', 'gadget', and other technology related words. Readers able to recognize the category easier with more words that represents the category. The more knowledge about words that represents the category, the better they can categorize the news. This process can be called as category classification [2].

Aside the news content, every news also has a news headline. The news category can also be categorized by reading the news headline. Readers can conclude the news category based on the news headline, which is quite short and concise. This makes readers able to know what news category faster without reading the whole news content.

Nowadays, internet has spread a lot of benefits to the users. One of it is the faster information delivery worldwide. The information from all around the world are spread through news channel. The news volume is increasing year by year and creating a huge volume of data. The huge volume of news data makes the process of categorizing news manually impossible.

News category classification is one of the Text classification or text categorization examples. Text classification is a method of categorizing and/or sorting the text-based entities into some predefined set of semantic categories or labels.[7] Text classification can classify document either into one category, several categories or even no category. Text classification or text categorization is also one of the methods in Natural Language Processing (NLP) [3].

Some researchers have proposed and done some experiments on news category classifier. Jafreezal et al tried the category classification using some algorithms, Naïve Bayes, Support Vector Machine, and KNN, to classify Indonesian and Malay news documents [4]. Garin et al proposed the Indonesian news classification based on NaBaNA (Naïve Bayes and Nazief-Adriani stemming) [5]. Pooja and Vaibhav conducted experiment on news category classification for multi-category using the Least Square Twin Support Vector Machine (LSTSVM) [2]. Krishnalal et al also proposed the text mining approach based on HMM-SVM (Hidden Markov Model Support Vector Machine) for web news classification [6]. The other researchers from India, Gurmeet and Karan from Chitkara University uses Neural Network for news classification [7]. Another researcher group from Japan, Ali et al also uses Neural

978-1-6654-5594-7/22/$31.00 ©2022 IEEE

Network with the added PCA for web news classification [8].

Based on the studies done before, the most used algorithm for news classification is the Support Vector Machine either native or improved SVM. SVM able to give impressive result around 83.13% to 96.88% accuracy and 5s to 52s based on the experiment conducted by Jafreezal et al on various number of categories and number of words taken from each document [4]. The other result from improved SVM, HMM-SVM, gives even better result with 90.76% to 96.34% of accuracy on news data with three categories [6]. The HMM-SVM has an added feature extraction process using the Hidden Markov Model (HMM) before the data being classified using the SVM.

In this paper, a news classification based on the news headline using Support Vector Classifier (SVC) will be proposed and evaluated based on its performance compared to the other machine learning models. The result aim is a fast et accurate news classifier with lower processing power requirement.

II. TEORITICAL EXPLANATION

A. Machine Learning

Machine Learning is a branch of Artificial Intelligence which focuses on the use of data and algorithms to imitate the way humans learn. Machine learning has been developed since 1950s. It is developed over several decades until it has emerged computer vision, speech recognition, natural language processing, robot control, and other applications. The impact of machine learning has been felt broadly across a range of industries concerned with data-intensive issues [9].

A diverse machine learning algorithms has been developed through these several decades. Machine learning can solve classification, regression, clustering, association, and dimensionality reduction problems. Classification and regression problems are the problems that uses supervised learning from machine learning. In the other part, clustering, association, and dimensionality reduction problems use the unsupervised learning instead.

The famous machine learning algorithms are Support Vector Machine (SVM), Linear Regression, Logistic Regression, Decision Tree, Naïve Bayes, K Nearest Neighbor (KNN), K-Means, and Random Forest.

B. Supervised Learning

Supervised learning is a type of machine learning where the model is trained using labeled datasets to classify data or predict outcomes accurately. The training dataset includes inputs and correct outputs. It adjusts the weight until the model fits the data appropriately in the cross-validation process.

The supervised learning is used to train a model that gives a desired output which can be categorical or numerical. The categorical outputs are used in classification problems, such as news category and cat vs dog. It classifies the inputs to several categories based on the trained datasets. The common classifiers used in classification are linear classifiers, support vector machines, decision trees, k-nearest neighbor, and random forest.

The numerical or continuous outputs can also be predicted using the supervised learning. The numerical or continuous output prediction can solve regression problems, such as revenue prediction, salary prediction, weight prediction, and other prediction tasks. Most of the problems are projections for the future. The most popular regression models are linear regression, logistical regression, and polynomial regression.

C. Support Vector Machine

Support Vector Machine (SVM) is one of the supervised learning models that is used for classification (Support Vector Classifier) and regression (Support Vector Regression). The SVM is mostly used by a lot of researchers as it has a robust performance, fast computation, and easy training. It also solves a convex quadratic problem and converges to a global solution in a definite time [2].

Support Vector Classifier has similar principal with the linear classifier, but it also has added kernel to work with non-linear problems in high dimension workspace. It will find the hyperplane to maximize margin between each data class. The hyperplane notation for linear SVM can be shown in Eq.(1), Eq(2) and Eq(3).

$$f(x) = \boldsymbol{w}^T\boldsymbol{x} + b \quad (1)$$

$$[(\boldsymbol{w}^T.\boldsymbol{x}_i) + b] \geq 1 \text{ for } y_i = +1 \quad (2)$$

$$[(\boldsymbol{w}^T.\boldsymbol{x}_i) + b] \leq 1 \text{ for } y_i = -1 \quad (3)$$

$\boldsymbol{x}_i$ = training data
i = 1, 2, 3, .. , n
$\boldsymbol{y}_i$ = class label for $\boldsymbol{x}_i$

D. Natural Language Processing

Natural Language Processing or usually called as NLP is the branch of Artificial Intelligence concerning on the computer's ability to understand text and spoken words such the way as human being. NLP was originally distinct from text information retrieval (IR), which employs highly scalable statistics-based techniques to index and search large volumes of text efficiently [9]. It combines linguistic computational with machine learning or deep learning models to enable the computers to process natural human language either in text or speech complete with its meaning. It has many usages in daily life, such as speech recognition, text translation, etc.

NLP requires a large volume of text for training and outputs a great model. The text will be transformed using word embedding into vectors for model training as the model cannot receive text inputs. NLP able to produce several types of output, such as category, summary, and part of speech tagging. The examples of NLP usage in daily life are email spam classifier, search results, predictive text, and smart assistant.

E. TF-IDF Vector

TF-IDF Vector is the most common vectorizer used for word embeddings. The concept of this vectorizer is calculating the uniqueness of every word from every document / data. The TF-IDF consist of Term Frequency (TF) and inverse Document Frequency (IDF). Term Frequency describes how many the term exists in a

document and Inverse Document Frequency tells the term uniqueness from all documents. The TF-IDF is calculated using Eq.(4), Eq.(5) and Eq.(6)

$$TF(w,d) = \frac{occurences\ of\ w\ in\ document\ d}{total\ number\ of\ words\ in\ document\ d} \quad (4)$$

$$IDF(w,D) = \ln\left(\frac{Total\ number\ of\ documents\ (N)\ in\ corpus\ D}{number\ of\ documents\ containing\ w}\right) \quad (5)$$

$$TFIDF\ (w,d,D) = TF(w,d) * IDF(w,D) \quad (6)$$

III. METHODOLOGY

The proposed news category classifier based on the news headline is designed and developed to classify news into its category faster and still accurate. The main classifier used in this experiment is the Support Vector Classifier with several comparison models, such as Linear Regression, Multinomial Naïve Bayes, Decision Tree, and Random Forest.

The programming language used in this experiment is the famous text mining programming language which is Python as it is easy to use and fast. Some of the libraries that will be used are Pandas, NLTK, Matplotlib, Seaborn, OS, RE, and Scikit-Learn.

The data used in this experiment is taken from www.kaggle.com . It is a dataset consists of news taken from the inshorts news webapp. The dataset contains 12,120 data classified into seven categories, which are technology, sports, politics, entertainment, world, automobile, and science as shown Table I. Each data consists of two features (news_headline and news_article) and one target (news category).

TABLE I. NUMBER OF DATA FROM EACH DATASET CATEGORY

Category	Number of data
automobile	1293
science	1437
politics	1596
technology	1791
sports	1900
entertainment	2036
world	2067

The raw data is not balanced between each category. Data rebalancing is needed by reducing the data for each category to the minimum number of data from a category. The balanced data will help the model to fit every category.

Dataset link: https://www.kaggle.com/kishanyadav/inshort-news

There are several steps to conduct this experiment. The steps for the experiments are:

A. Data preprocessing

The raw data taken from Kaggle dataset needs to be preprocessed. The preprocessing includes feature selection, data balancing and splitting (training & testing dataset), label encoding (for target), and corpus creation using TF-IDF Vector.

The feature selected for this experiment is the news_headline. The data will be rebalanced into 1000 data per category for training and 200 data per category for testing. Besides the feature, the target category also need to be preprocessed to numeric using the label encoding. The last one is the corpus creation using the TF-IDF vectorizer based on the formula.

B. Model Creation & Training

The first step is creating the SVC model and other comparison models (Logistic Regression, Multinomial Naïve Bayes, Decision Tree, and Random Forest) to accommodate the classification task. After all models are created, all models are trained until it reaches acceptable accuracy. The model will also need to be check whether it experience overfitting or underfitting during the training through checking the testing and training accuracy and loss difference. If the difference is quite far, there is an indication of overfitting. Overfitting is the condition where model only fits the train data.

C. Testing & Evaluation

The trained models are tested using the test data. The test results should be evaluated to figure out each model performance. The evaluation variables may be varied according to the input and output. In this experiment, the evaluation variable is accuracy, precision, and recall as shown Eq(7), Eq.(8), Eq(9). The higher accuracy, precision, and recall indicates better model performance.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (7)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (8)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (9)$$

TP = True Positive (positive data that are correctly classified)

TN = True Negative (negative data that are correctly classified)

FP = False Positive (positive data that are wrongly classified)

FN = False Negative (negative data that are wrongly classified)

IV. RESULT

The test data that has been preprocessed are fed into the SVC, Logistic Regression, Multinomial Naïve Bayes, Decision Tree, and Random Forest. The data fed into them are all the same and calculated their accuracy, precision, and

recall. The result from the testing is represented in Figure 1 and Table II.

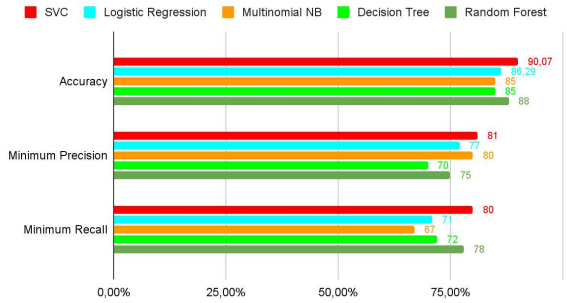

Fig. 1. Performance Comparison Graph

TABLE II PERFORMANCE COMPARISON

	Accuracy	Minimum Precision	Minimum Recall
SVC	90.07%	81.00%	80.00%
LR	86.29%	77.00%	71.00%
Multinomial NB	85.00%	80.00%	67.00%
Decision Tree	85.00%	70.00%	72.00%
Random Forest	88.00%	75.00%	78.00%

V. CONCLUSION

Based on the results that we've got, the accuracy of SVC model is the highest among the other models, with higher minimum precision and minimum recall. It indicates the SVC model is much more accurate and stable compared to the other models.

The conclusion from this experiment is that the SVC model is suitable for this news category classification with news headline. The model has got an acceptable accuracy, precision, and recall to correctly classify the test data after doing some training with the train data.

REFERENCES

[1] Lehman-Wilzig, S. N., & Seletzky, M. "Hard news, soft news, "general" news: The necessity and utility of an intermediate classification", Journalism, 11(1), 2010, 37–56. https://doi.org/10.1177/1464884909350642

[2] Bracewell, D. B., Yan, J., Ren, F., and Kuroiwa, S., "Category Classification and Topic Discovery of Japanese and English News Articles", Electronic Notes in Theoretical Computer Science, 225(C), 2009, 51–65. https://doi.org/10.1016/J.ENTCS.2008.12.066

[3] Saigal, P., & Khanna, V. "Multi - category news classification using Support Vector Machine based classifiers. SN Applied Sciences, 2(3), 2020, 1–12. https://doi.org/10.1007/s42452-020-2266-6

[4] Jaafar, J., Indra, Z., and Zamin, N., "A category classification algorithm for Indonesian and Malay news documents", Jurnal Teknologi, 78(8–2), 2016, 121–132. https://doi.org/10.11113/jt.v78.9549

[5] Septian, G., Susanto, A., and Shidik, G. F. "Indonesian news classification based on NaBaNA". Proceedings - 2017 International Seminar on Application for Technology of Information and Communication: Empowering Technology for a Better Human Life, ISemantic 2017, 175–180. https://doi.org/10.1109/ISEMANTIC.2017.8251865

[6] Krishnalal, G., Rengarajan, S. B., and Srinivasagan, K. G. "A New Text Mining Approach Based on HMM-SVM for Web News Classification'. International Journal of Computer Applications, 1(19), 2010, 103–109. https://doi.org/10.5120/395-589

[7] Kaur, G., and Bajaj, K. "News Classification using Neural Networks. Communications on Applied Electronics", 5(1), 2016, 42–45. https://doi.org/10.5120/cae2016652224

[8] Selamat, A., Yanagimoto, H., and Omatu, S. "Web news classification using neural networks based on PCA", 2003 ,2389–2394. https://doi.org/10.1109/sice.2002.1195784

[9] Nadkarni, P. M., Ohno-machado, L., and Chapman, W. W., "Natural language processing : an introduction", 2011, https://doi.org/10.1136/amiajnl-2011-000464