Terbit online pada laman web jurnal: **http://journal.isas.or.id/index.php/JACOST**

JOURNAL OF APPLIED COMPUTER SCIENCE AND TECHNOLOGY (JACOST)

Vol. 6 No. 1 (2025) 37 – 45 | ISSN: 2723-1453 (Media Online)

News Classification using Natural Language Processing with TF-IDF and Multinomial Naïve Bayes

Nadira Alifia Ionendri[1], Feri Candra[2], Afdi Rizal[3]
[1,2]Teknik Informatika, Fakultas Teknik, Universitas Riau, Pekanbaru, Indonesia
[3]Badan Pusat Statistik Provinsi Riau, Pekanbaru, Indonesia
[1]nadira.alifia4460@student.unri.ac.id*, [2]feri@eng.unri.ac.id, [3]afdi@bps.go.id

Abstract

Online news contains valuable insights into public phenomena that can support statistical analysis by institutions like BPS Riau. However, current methods of classifying news are manual, time-consuming, and prone to human error. This study proposes an automated news classification system using Natural Language Processing (NLP) techniques with Term Frequency–Inverse Document Frequency (TF-IDF) for feature extraction and the Multinomial Naïve Bayes algorithm for classification. The dataset was collected via web scraping and manually labeled across five statistical categories: poverty, unemployment, democracy, inflation, and economic growth. The system achieved a validation accuracy of 83%, a test accuracy of 90%, with an average precision of 0.85, recall of 0.93, and f1-score of 0.87. These results demonstrate that the proposed system can significantly reduce the manual workload of news classification and be practically implemented by BPS Riau to support accurate and timely statistical reporting.

Keywords: *news, classification, NLP*

1. Introduction

News is a report on the latest facts or ideas that are interesting or important to most audiences through media such as newspapers, radio, television, or the Internet. The facts or events in the field are covered, written, and edited by journalists and then distributed through the mass media [1]. Accessing online news sites is one way that can be used to get the desired news information because it is current and fast in reporting news as well as a form of technological development. This online news site helps facilitate government institutions, such as BPS Riau, in knowing and monitoring things happening in society. The BPS Riau is one of the state institutions responsible for providing data and administering statistics in Riau Province. In carrying out its duties, The BPS Riau routinely publishes data to provide the data needed by the government and the public. This data requires supporting data as proof that the data they publish is actual. The BPS Riau uses the online news site to collect news related to phenomena occurring in the people of Riau Province. The news is on specific topics according to the needs of The BPS Riau, namely poverty, unemployment, democracy, inflation, and economic growth. By searching for and collecting this phenomenon, The BPS Riau has vital supporting data related to their published data.

News data collection carried out by The BPS Riau is still done manually. They open, read, and copy-paste the title, content, images, and screenshots of news one by one from a news site every day. After the officer reads the news, the news is then classified into the appropriate group based on the officer's assessment. This is done in order to obtain news data according to the needs and categories of specific statistical data required by The BPS Riau. The large amount of news routinely uploaded by several existing online news sites requires much staff time to process.

The BPS Riau frequently lacks news as adequate supporting data for the data or information they release because officers put off gathering news which is time-consuming and prone to human errors. Additionally, the large volume of news makes this process inefficient, and they have other tasks to complete. This cannot continue to happen, especially if there are parties who ask for evidence to support the news release. Therefore, a practical automatic approach is needed to classify news into appropriate categories.

Natural Language Processing (NLP) has experienced significant advances in recent years that are changing how human language is processed and understood. NLP studies the relationship between human language and computers using calculative techniques to analyze and re-present the text or speech and achieve human-like language processing for various tasks or applications [2]. NLP has also been getting the attention of several researchers and seems a promising and challenging area to work on [3]. News classification using NLP could be one method that can help The Central Bureau of Statistics Riau overcome news collection problems.

Several previous studies have used Natural Language Processing (NLP) to classify text. Research [4] classified the sentiment of news headlines reporting on COVID-19 in 2021 on DetikHealth media. The results obtained were that after going through the entire text pre-processing process with NLP and implementing the Naïve Bayes algorithm in the sentiment classification process, the accuracy percentage obtained was 72.5%. Furthermore, research from [5] obtained higher accuracy for the Multinomial Naïve Bayes algorithm of 84% than the Support Vector Machine algorithm of 78%. This shows that implementing NLP and the Multinomial Naïve Bayes algorithm better analyzes student feedback sentiment towards institutional facilities.

In this research, NLP is utilized to classify text with the aid of TF-IDF for feature extraction and Multinomial Naïve Bayes for classification. These methods were chosen due to their effectiveness in processing and categorizing textual data efficiently. The choice of the TF-IDF method is based on its ability to measure the importance of a word in a document based on its frequency while minimizing the impact of common words that do not provide significant information [6]. TF-IDF is widely used in various NLP applications due to its effectiveness in extracting relevant text features.

Meanwhile, the Multinomial Naïve Bayes algorithm is chosen for classification due to its efficiency in handling text data composed of words with a multinomial distribution. This algorithm is also well-known for text classification because of its simplicity, speed, and ability to maintain high accuracy in large datasets [7]. This choice is also supported by previous studies showing that Naïve Bayes performs competitively compared to other algorithms for text classification tasks.

While many studies have explored text classification in news analysis, most focus on English texts, with limited attention to Bahasa Indonesia. Moreover, although TF-IDF and Naive Bayes have shown effectiveness in various tasks, there's still a lack of research applying them to classify local Indonesian news—especially for government institutions like BPS Riau. This study addresses that gap by using TF-IDF and Multinomial Naive Bayes to classify Indonesian-language news from BPS Riau, aiming to develop lightweight NLP tools for real-world public data workflows.

This research contributes by offering an automated Natural Language Processing (NLP)-based approach for news classification, utilizing Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction and Multinomial Naïve Bayes for classification. By automating this process, the research has the potential to reduce workload and improve the accuracy of news classification used by BPS Riau as supporting data for statistical analysis.

2. Research Methods

This research was conducted using the Natural Language Processing (NLP) method. The NLP method was used to pre-process and classify news texts. Text classification is by far the most popular task in NLP and is used in various tools, one of which is that NLP can also classify news articles into a series of news topics [6]. Text pre-processing in this research utilized NLP in several stages: case folding, data cleaning, removing stop words, stemming, and tokenizing. In classifying text, NLP utilizes Machine Learning. The type of machine learning algorithm that is widely used in text data classification is Multinomial Naïve Bayes [7]. This research applied this algorithm to classify news texts.

2.1 Research Data

The dataset used in this research was obtained from BPS Riau Province and the Riau Pos news site. Data originating from BPS Riau Province is news that has been used as supporting data in many BPS data products and is a list of news that BPS Riau Province officers have labeled. News from the Riau Pos website was taken to increase the data used. The Riau Pos site was chosen because it routinely publishes news stories daily and still provides news from more than five years ago. Data was collected by searching for news with the keywords required by BPS Riau: poverty, unemployment, democracy, economic growth, and inflation. News with these keywords is in large quantities on the Riau Pos site. News data was collected by applying web scraping techniques using Python programming. Web scraping is a technique for converting unstructured web data into structured data that can be saved and analyzed in a central spreadsheet or database [8]. The data taken consisted of title, link, date, and news content and was immediately labeled according to keywords, as shown in Table 1.

Table 1. News Data Scraping Results from the Riau Pos Website

Title	Link	Class	Date	Content
Pemkab Komitmen Turunkan Angka Stunting Balita	https://riaupos.jawapos.com/riau/21/06/2019/201917/pemkab-komitmen-turunkan-angka-stunting-balita.html	1	21/06/2019	RIAUPOS.CO) -- Bupati Kampar Catur Sugeng Susanto yang diwakili Sekretaris Daerah Yusri membuka sosialisasi percepatan penurunan kasus stunting

DOI : https://doi.org/10.52158/jacost.v6i1.1099

2.2 Research Framework

This study consisted of 6 stages: data collection (data scraping), text pre-processing, feature extraction, splitting data into training data and testing data, multinomial naïve Bayes stage, and data evaluation stage, as shown in Figure 1.

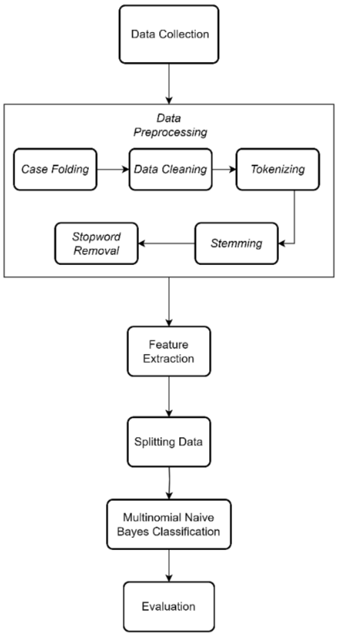


Figure 1. System Flowchart

In the data collecting process, news data with the keywords ‘Kemiskinan’ or poverty, ‘Pengangguran’ or unemployment, ‘Demokrasi’ or democracy, ‘Inflasi’ or inflation, and ‘Pertumbuhan Ekonomi’ or economic growth was collected. Data was collected on the Riau Pos news site using web scraping techniques with a Python script. Data pre-processing is considered the core stage in machine learning and data mining [9]. Data pre-processing in this study aims to clean data from unnecessary elements and produce data ready for the next stage. Data pre-processing is often needed to ensure the reliability of data analysis using various techniques [10]. This stage consisted of case folding, data cleaning, removing stop words, stemming, and tokenizing. Furthermore, the TF-IDF method was used at the feature extraction stage to extract essential features from news text data. Then, the dataset was divided into three parts: training data, validation data, and testing data.

After that, the Multinomial Naïve Bayes algorithm was used as a model algorithm built to classify news data into several specified categories. The following process evaluated the model using a confusion matrix to measure system performance, such as calculating precision, recall, f1-score, and accuracy values.

2.3. *Pre-processing Dataset*

Before the model processes the data, pre-processing must be done first so that the data entered into the model is suitable to the algorithm used. The type of data used in classifying news data was in text form. In this research, text pre-processing utilized NLP in several stages: case folding, data cleaning, tokenizing, stemming, and removing stop words. The case folding process aims to change all the letters in a text document into lowercase letters [11]. Data cleaning is the process of cleaning data from unnecessary components that can interfere with the next stage, in this case, removing punctuation and numbers. Tokenizing is breaking complex data, like paragraphs, into simple units called tokens [5]. Stemming is converting the word into its basic form [12]. Stop words Removal is a process of eliminating words that are not important; this process can reduce the dimensions of space that look heavy [11]. The results of the pre-processing process can be seen in Table 2.

Table 2. Example of data after the pre-processing stage

Content	Class
['potensi', 'daerah', 'gali', 'pekanbaru', 'riauposco', 'plt', 'bupati', 'antan', 'singingi', 'drs', 'suhardiman', 'amby', 'ak', 'mm', 'potensi', 'daerah', 'gali', 'maksimal', 'sampai', 'hadir', 'acara', 'halalbihalal', 'ikat', 'keluarga', 'antan', 'singingi', 'ikks', 'pekanbaru', 'hotel', 'grand', 'suka', 'pekanbaru', 'sabtu', 'malam', 'sempat', 'suhardiman', 'amby', 'potensi', 'wisata', 'potensi', 'daerah', 'maksimal', 'bangun', 'infrastruktur', 'jalan', 'jembatan', 'irigasi', 'program', 'tunjang', 'kait', 'sektor', 'pariwisata', 'kuansing', 'indah', 'banding', 'daerah', 'suhardiman', 'amby', 'suhardiman', 'amby', 'sda', 'pada', 'wajib', 'tunjang', 'sdm', 'unggul', 'bangun', 'fungsi', 'jalan', 'bidang', 'ekonomi', 'fokus', 'mudah', 'modal', 'laku', 'umkm', 'meminimalisir', 'miskin', 'masyarakat', 'beber', 'suhardiman', 'amby', 'ketua', 'ikks', 'jakarta', 'wan', 'hasyim', 'maju', 'kuansing', 'kembang', 'bidang', 'infrastruktur', 'layan', 'masyarakat', 'program', 'obat', 'gratis', 'rasa', 'manfaat', 'masyarakat', 'pasien', 'ktp', 'tugas', 'sehat', 'langsung', 'obat', 'wan', 'hasimyas']	1

2.4 Feature Extraction

Feature extraction is extracting relevant information from raw data [13]. In this research, which used text data, feature extraction is done by changing raw data into numerical features without reducing its original information [14]. The feature extraction used was Term Frequency–Inverse Document Frequency (TF-IDF). TF-IDF is a method in text analysis that was used to give weight to words in a document based on two main factors: Term Frequency (TF) and Inverse Document Frequency (IDF). This method aims to measure the importance of a particular word compared to other words in a document and corpus [6].

Term Frequency (TF) is how often a term or word appears in a document. TF of a word t in a document d is defined as formula (1):

$$\mathrm{TF}\ (t, d) = \frac{number\ of\ words\ t\ in\ the\ document\ d}{total\ number\ of\ words\ in\ the\ document\ d} \quad (1)$$

DOI : https://doi.org/10.52158/jacost.v6i1.1099

Inverse Document Frequency (IDF) is how unique or essential a word is in the corpus. The more often a term appears in many documents, the smaller the IDF value will be [15]. The calculated IDF of a word t is shown in the formula (2):

$$\text{IDF}(t,D)=log\frac{Total\ number\ of\ documents\ in\ the\ corpus\ D}{Number\ of\ documents\ with\ term\ t\ in\ them} \quad (2)$$

The TF-IDF weight for a word t in a document d can be calculated by multiplying the TF by the IDF:

$$\text{TF-IDF}\ (t,d,D)=\text{TF}\ (t,d)\times \text{IDF}\ (t,D) \quad (3)$$

This method minimizes the weight of common words often appearing and increases the weight of unique or specific words in a news text. The results of the feature extraction process can be seen in Table 3.

Table 3. Feature Extraction

Words			
'karet'	'korban'	'perintah'	'isu'
'zakat'	'partai'	'ras'	'komitmen'
'parkir'	'tol'	'sosial'	'bupati'
'lansia'	'harap'	'manfaat'	'kantor'
'ekonomi'	'indonesia'	'wabah'	'negeri'
..	..	..	..
'kerja'	'tumbuh'	'dampak'	'calon'

2.5 Multinomial Naïve Bayes Classification

In classifying text, NLP utilized Machine Learning. The part of machine learning used in this research was supervised learning, which was applied to the Naive Bayes algorithm. Supervised learning is used to describe prediction tasks because the goal is to forecast/classify a specific outcome of interest [16]. In text classification or NLP, Naïve Bayes is a probabilistic classifier that uses Bayes' theorem to classify text based on evidence seen in the training data [6]. The type of algorithm that is widely used in text data classification is Multinomial Naïve Bayes [7]. Multinomial Naïve Bayes takes into account the number of occurrences of words in a document so that it assumes independence of the presence of words in the document by not estimating word order or information context. [17]. The calculation of the probability that a word t appears in a class c in Multinomial Naïve Bayes is expressed as shown in equation 4. Where Wct is the number of word t occurrences in class c documents, α is the Laplace smoothing parameter to avoid zero values, and B is the number of W unique words (IDF value) in the entire document.

$$P\ (t|c)=\frac{Wct+\ \alpha}{(\sum W\prime\in V\ Wct')+B\prime} \quad (4)$$

Multinomial Naïve Bayes was chosen for its strong performance in text classification with feature independence and efficiency, making it ideal for real-time systems. It will predict a text's category based on the probability of word occurrences in the relevant class.

2.6 Confusion Matrix

Confusion Matrix is a table in the form of a matrix used to describe the performance of classification models [17]. The confusion matrix is widely used in machine learning for supervised classification or determination of the behavior of classification models [18]. This matrix consists of four cells, as shown in table 4.

The confusion matrix provides a more detailed picture of model performance than just looking at accuracy alone. This allows researchers to understand how the model can address a particular class and identify whether the classification has biases or specific problems.

The confusion matrix contains information that compares the classification results carried out by the system with the default results of the classification [19].

Table 4. Confusion Matrix

		Predicted Value	
		Negative	Positive
Actual Value	Negative	TN	FP
	Positive	FN	TP

In measuring the performance of the confusion matrix, there are four terms to re-present the classification results. TP (True Positive), namely the amount of positive data that is classified as true by the system; TN (True Negative), namely the amount of negative data that is classified as true by the system; FP (False Positive), is the amount of positive data that is classified as wrong by the system, and FN (False Negative)) is the amount of negative data that is classified incorrectly by the system [14].

There is some information related to measuring the performance of the classification model, which can be calculated based on the confusion matrix: accuracy, precision, recall, and f-1 score, see formula 5-8.

$$Accuracy\ =\frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Precision\ =\frac{TP}{TP+FP} \quad (6)$$

$$Recall\ =\frac{TP}{TP+FN} \quad (7)$$

$$\frac{1}{f1}=\frac{1}{2}\left(\frac{1}{precision}+\frac{1}{recall}\right) \quad (8)$$

Accuracy is a value that measures the extent to which the model can predict correctly from the total predictions made. The precision value can be obtained using equation (5). Precision is a value that calculates how many samples predicted as positive are positive samples. Recall is a value that calculates how many positive samples are captured by positive predictions. The f1-score is the harmonic average between the precision and recall values[14]. The f1-score value can be obtained from equation (8).

3. Result and Discussions

News data obtained from BPS Riau Province was news labeled by BPS Riau Province officers, totaling 81 news stories. The news consisted of 16 'kemiskinan' news, 4 'pengangguran' news, 16 'demokrasi' news, 20 'inflasi' news and 23 'pertumbuhan ekonomi' news. Besides that, 30 news data were also taken from the BPS data product “Bahan Resmi Statistik” for the 'inflasi' class. The Bahan Resmi Statistik provides only supporting news on the topic 'inflasi' from 2022, which, if collected, amounts to 30 news stories. The news data collected using web scraping techniques with Python scripts can be seen in Table 1. The data scraping process produced 1220 news stories for five news data classes with a composition of 250 news stories per class, except for the 'inflation' class. As previously mentioned, the 'inflation' class data amounted to 220 news stories, which were added with 30 news items taken from the Bahan Resmi Statistik Riau Province data product. Web scraping news data was taken from May 2019 to January 2024. The data consisted of titles, links, dates, and news content, and the contents were immediately labeled according to keywords. The data were then cleaned through the pre-processing stage, which consists of case folding, data cleaning, removing stop words, stemming, and tokenizing.

Data labeling is carried out in parallel with the data scraping process. The labels consist of “Kemiskinan” or poverty, “Pengangguran” or unemployment, “Demokrasi” or democracy, “Inflasi” or inflation, and “Pertumbuhan Ekonomi” or economic growth. Kemiskinan is labeled 1, Pengangguran is labeled 2, Demokrasi is labeled 3, Pertumbuhan Ekonomi is labeled 4, and Inflasi is labeled 5, as shown in Table 2.

The amount of news data for each class is balanced, as shown in Figure 2. As shown in Figure 2, each class's news data composition is balanced. A word cloud was also created to make it easy to get acquainted with the content so that only the most significant terms are retained and presented. The most significant and frequently appearing words are visualized in different colors and larger font sizes to catch attention immediately [20]. Figures 3, 4, 5, 6, and 7 show the word cloud for each data class.

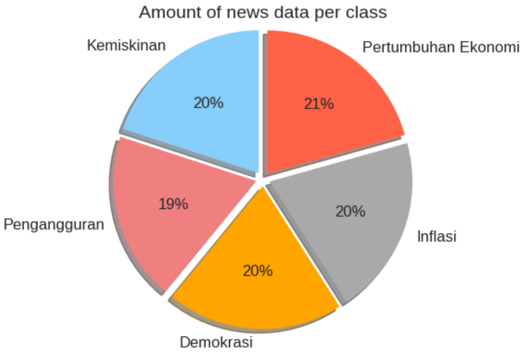


Figure 2. Composition of news data

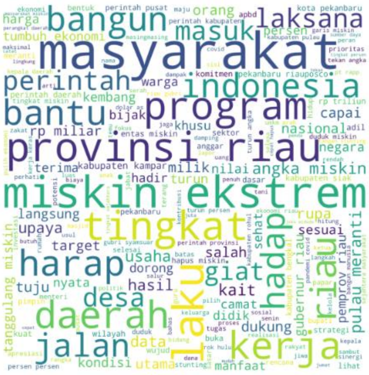


Figure 3. Word Cloud Data ‘Kemiskinan’

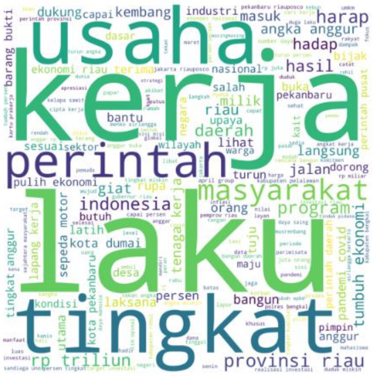


Figure 4. Word Cloud Data ‘Pengangguran’

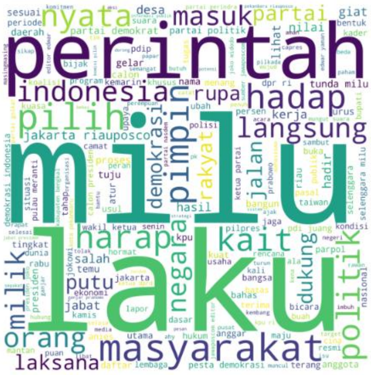


Figure 5. Word Cloud Data ‘Demokrasi’

DOI : https://doi.org/10.52158/jacost.v6i1.1099

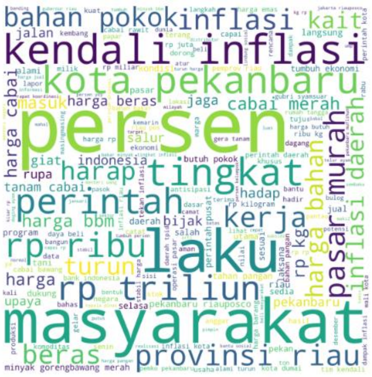


Figure 6. Word Cloud Data 'Inflasi'

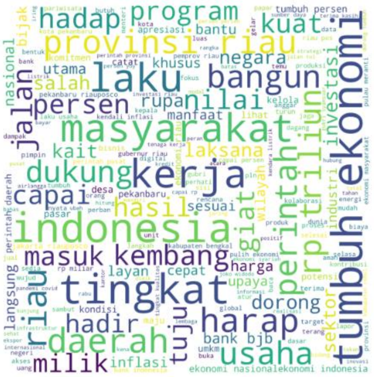


Figure 7. Word Cloud Data 'Pertumbuhan Ekonomi'

The next step was feature extraction using the TF-IDF (Term Frequency-Inverse Document Frequency) method to see how often the frequency of a word appears in the document. TF-IDF converts a document into a matrix per word to calculate its weight. The more often a word appears, the greater the weight of the word, as shown in Table 5.

Table 5. TF-IDF Weight

Words	Weight
'karet'	0.832838
'zakat'	0.806143
'parkir'	0.779234
'calon'	0.067982
'ekonomi'	0.004013
..	..
'kerja'	0.004087

The division of the data into training and validation data was carried out in a ratio of 80:20. Meanwhile, test data used news labeled by 2 BPS Riau officers. Table 6 shows the division of data into training and testing data.

Tabel 6. Split Datasets

Division of the Dataset	
Training Data	1000
Validation Data	250
Testing Data	81

After splitting the data, the Multinomial Naïve Bayes classification process continued. Multinomial Naïve Bayes is one of the algorithms of the Naïve Bayes classifier often used for text data classification. By implementing this algorithm, machines can learn and test the data sets that have been provided [4].

The Multinomial Naïve Bayes algorithm considers the average value of every feature for every class in its workflow. By comparing a data point with the statistics for every class, a prediction is made for the class that best fits the data point. The application of the Multinomial Naïve Bayes algorithm in this research was predicting the category or class of news text based on the probability of the appearance of words in the news text in the appropriate class.

The final process in this research was to evaluate using a confusion matrix to measure system performance. The confusion matrix compared the actual data with the resulting classification data. The confusion matrix produced in this research is shown in Figure 8. Performance evaluation is measured using the values of accuracy, precision, recall, and f1-score with the formula, which are shown in equations (5), (6), (7), and (8).

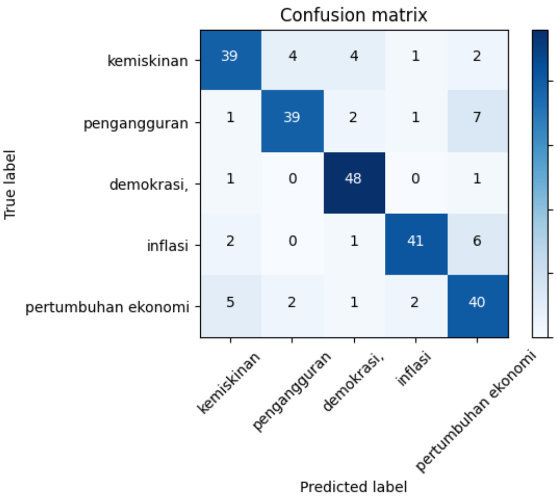


Figure 8. Confusion Matrix of Validation Accuracy

Figure 8 shows the confusion matrix of validation accuracy for news classification. It is known that 'kemiskinan' news in the data is predicted as 39 data as 'kemiskinan' news in the system, 4 data are predicted as 'pengangguran', 4 data are predicted as 'demokrasi', 1 data is predicted as 'inflasi', and 2 data are predicted as 'pertumbuhan ekonomi' in the system.

Then, it was discovered that the news 'pengangguran' in the data was actually predicted as 39 data as 'pengangguran' news in the system, 1 data is predicted as 'kemiskinan', 2 data are predicted as 'demokrasi', 1 data is predicted as 'inflasi', and 7 data are predicted as 'pertumbuhan ekonomi' news in the system.

After that, 48 news stories were predicted correctly as 'demokrasi' news, 1 data is predicted as 'kemiskinan', 1 data is predicted as 'pertumbuhan ekonomi' and no data

DOI : https://doi.org/10.52158/jacost.v6i1.1099

predicted as ‘pengangguran’, and 'inflasi' news in the system.

Furthermore, 41 news stories were predicted correctly as 'inflasi' news by the system, 2 data were predicted as 'kemiskinan', and 0 data predicted as ‘pengangguran’, 1 data was predicted as 'demokrasi', and 6 data were predicted as 'pertumbuhan ekonomi' news in the system.

And, it is known that the news 'pertumbuhan ekonomi' in the data was actually predicted as 40 data as 'pertumbuhan ekonomi' news by the system, 5 data predicted 'kemiskinan', 2 data predicted 'pengangguran', 1 data predicted 'demokrasi', and 2 data predicted 'inflasi' in the system.

Figure 9 shows the confusion matrix of test accuracy for news classification. It is known that ‘kemiskinan’ news in the data was predicted as 15 data as ‘kemiskinan’ news in the system, 0 data are predicted as 'pengangguran', 0 data were predicted as 'demokrasi', 0 data are predicted as 'inflasi', and 1 data are predicted as 'pertumbuhan ekonomi' in the system.

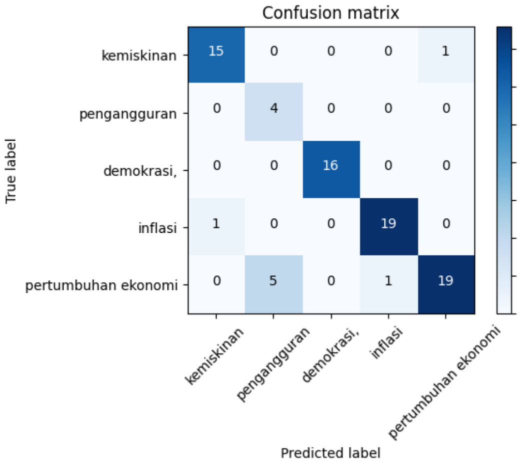


Figure 9. Confusion Matrix of Test Accuracy

After that, 4 news stories were predicted correctly as 'pengangguran' news, and no data predicted as ‘kemiskinan’, 'demokrasi', ‘inflasi’ and 'pertumbuhan ekonomi' news in the system.

Then, it was discovered that the news 'demokrasi' in the data was predicted as 16 data 'demokrasi' news in the system and no data predicted as ‘kemiskinan’, 'pengangguran', ‘inflasi’ and 'pertumbuhan ekonomi' news in the system.

Furthermore, 19 news stories were predicted correctly as 'inflasi' news by the system, 1 data were predicted as 'kemiskinan', and no data predicted as ‘pengangguran’, 'demokrasi', and 'pertumbuhan ekonomi', news in the system.

Moreover, finally, it is known that the news 'pertumbuhan ekonomi' in the data was actually predicted as 19 data as 'pertumbuhan ekonomi' news by the system, 0 data predicted 'kemiskinan', 5 data predicted 'pengangguran', 0 data predicted 'demokrasi', and 1 data predicted 'inflasi' in the system.

Table 7. Value of Confusion Matrix of News Classification

News Category	Precision	Recall	F1-Score
kemiskinan	0.94	0.94	0.94
pengangguran	0.44	1.00	0.62
demokrasi	1.00	1.00	1.00
inflasi	0.95	0.95	0.95
pertumbuhan ekonomi	0.95	0.76	0.84
Validation accuracy			0.83
Test Accuracy			0.90

Information comparing the expected and actual classification outcomes of the system's classification is generated by the confusion matrix. The value of the confusion matrix in this research is shown in Table 7. The precision value is obtained by comparing the amount of relevant information obtained by the system with the total amount of information retrieved. Furthermore, the recall value is obtained by comparing the amount of relevant information obtained by the system with the total amount of relevant information contained in the information, whether retrieved or not retrieved by the system. Then, the f1-score value is obtained from the average harmonic result between the precision and recall values. Meanwhile, the accuracy value indicates the effectiveness of the test based on the effectiveness between the predicted value and the actual value [14].

In Table 7, the model evaluation results indicate that the ‘demokrasi’ category has perfect precision, recall, and f1-score values (1.00), demonstrating that the model classifies news in this category very well. However, the ‘pengangguran’ category has low precision (0.44) despite its high recall (1.00). This indicates that although the model captures almost all news in the ‘pengangguran’ category, many news items from ‘pertumbuhan ekonomi’ category are also classified as ‘pengangguran,’ leading to an increased number of false positives.

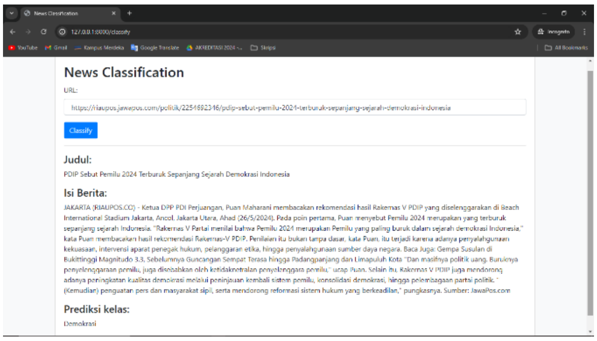
The misclassification of ‘pengangguran’ with categories like ‘pertumbuhan ekonomi’ suggests a semantic overlap, likely due to the similarity in frequently used words across both categories, as illustrated in Figures 4 and 7. Since the model uses TF-IDF as a feature representation, it relies heavily on word frequency without deeper contextual understanding. This means that shared keywords—like “kerja,” “laku,” “tingkat,” “usaha,” or “masyarakat”—can confuse the Multinomial Naive Bayes classifier, which assumes word independence. So, when vocab overlaps, it kinda just vibes and throws the label based on probability, even if the actual context doesn’t match.

Other categories, such as ‘kemiskinan,’ ‘inflasi,’ and ‘pertumbuhan ekonomi,’ have relatively high f1-scores, indicating that the model classifies news in these

DOI : https://doi.org/10.52158/jacost.v6i1.1099

categories effectively. The overall model accuracy reaches 90% on the test data, demonstrating that the selected method is effective. This result outperforms previous findings in [4], which achieved 72.5% accuracy using a similar method, indicating that domain-specific training improves classification accuracy.

The practical system was implemented using an interface containing a classification program that the author previously created. The interface was created using the Python Django framework to create simple websites. At this implementation stage, a link from a news item originating from a local Riau website, such as Riau Pos, Tribun Pekanbaru, Go Riau, or Haluan Riau, is entered into the column provided in the program. Next, the program ran the text pre-processing function and continued with the news classifier function. The program's output is in the form of category names predicted by the model for the news entered. Figure 10 shows the prediction results from the model in this research, which correctly predicted that the news was news in the 'kemiskinan' category



News Classification

URL:

https://riaupos.jawapos.com/politik/2254692346/pdip-sebut-pemilu-2024-terburuk-sepanjang-sejarah-demokrasi-indonesia

Judul:

PDIP Sebut Pemilu 2024 Terburuk Sepanjang Sejarah Demokrasi Indonesia

Isi Berita:

JAKARTA (RIAUPOS.CO) - Ketua DPP PDI Perjuangan, Puan Maharani membacakan rekomendasi hasil Rakernas V PDIP yang diselenggarakan di Beach International Stadium Jakarta, Ancol, Jakarta Utara, Ahad (26/5/2024). Pada poin pertama, Puan menyebut Pemilu 2024 merupakan yang terburuk sepanjang sejarah Indonesia. "Rakernas V Partai menilai bahwa Pemilu 2024 merupakan Pemilu yang paling buruk dalam sejarah demokrasi Indonesia," kata Puan membacakan hasil rekomendasi Rakernas-V PDIP. Penilaian itu bukan tanpa dasar, kata Puan, itu terjadi karena adanya penyalahgunaan kekuasaan, intervensi aparat penegak hukum, pelanggaran etika, hingga penyalahgunaan sumber daya negara. Baca Juga: Gempa Susulan di Bukittinggi Magnitudo 3.3, Sebelumnya Guncangan Sempat Terasa hingga Padangpanjang dan Limapuluh Kota "Dan masifnya politik uang. Buruknya penyelenggaraan pemilu, juga disebabkan oleh ketidaknetralan penyelenggara pemilu," ucap Puan. Selain itu, Rakernas V PDIP juga mendorong adanya peningkatan kualitas demokrasi melalui peninjauan kembali sistem pemilu, konsolidasi demokrasi, hingga pelembagaan partai politik. "(Kemudian) penguatan pers dan masyarakat sipil, serta mendorong reformasi sistem hukum yang berkeadilan," pungkasnya. Sumber: JawaPos.com

Prediksi kelas:

Demokrasi

Figure 10. News Classification sample

4. Conclusion

Based on the results and discussion, it can be concluded that news classification using Natural Language Processing with TF-IDF and Multinomial Naïve Bayes method was successfully carried out with an accuracy validation value of 83% and an accuracy test value of 90%. This result proves that this system is 90% accurate and can be used by BPS Riau Province in classifying news. The division of news data into train and validation data was 8:2, and 81 test data obtained an average precision result of 0.85, recall 0.93, and f1-score of 0.87.

For future research, the model can be further developed by utilizing alternative feature extraction methods such as Word2Vec or BERT to enhance text representation, exploring different classification algorithms, such as Support Vector Machines or Random Forest, to compare results and improve accuracy, and also expanding the dataset to include more diverse news variations and broader categories.

With further development, this news classification system could evolve into a reliable tool integrated into a pipeline to help BPS Riau or similar institutions automatically and efficiently categorize news, significantly reducing human workload and improving data validation and decision-making processes.

Acknowledgement

The author declares no conflict of interest. The author would like to thank the University of Riau and The Central Bureau of Statistics (BPS) Riau for facilitating the authors in carrying out this research.

References

[1] Fakultas Hukum, Universitas Muhammadiyah Sumatera Utara, T. H. Lubis, I. Koto, and Fakultas Hukum, Universitas Muhammadiyah Sumatera Utara, "Diskursus Kebenaran Berita Berdasarkan Undang-Undang Nomor 40 Tahun 1999 Tentang Pers Dan Kode Etik Jurnalistik," *LEGA LATA J. Ilmu Huk.*, vol. 5, no. 2, pp. 231–250, Jul. 2020, doi: 10.30596/dll.v5i2.4169.

[2] M. Agarwal, "An Overview of Natural Language Processing," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 7, no. 5, pp. 2811–2813, May 2019, doi: 10.22214/ijraset.2019.5462.

[3] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed. Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: 10.1007/s11042-022-13428-4.

[4] F. Delfariyadi, A. Helen, and S. Yuliawati, "Klasifikasi Sentimen Judul Berita Pemberitaan COVID-19 Tahun 2021 pada Media DetikHealth," *J. Inf. Eng. Educ. Technol.*, vol. 6, no. 2, pp. 50–57, Dec. 2022, doi: 10.26740/jieet.v6n2.p50-57.

[5] F. K. Khaiser, A. Saad, and C. Mason, "Sentiment Analysis Of Students' Feedback On Institutional Facilities Using Text-Based Classification And Natural Language Processing (NLP)," *J. Lang. Commun.*, vol. 10, no. 1, pp. 101–111, Mar. 2023, doi: 10.47836/jlc.10.01.06.

[6] Sowmya V. B., B. Majumder, A. Gupta, and H. Surana, *Practical natural language processing: a comprehensive guide to building real-world NLP systems*, First edition. Sebastopol, CA: O'Reilly Media, 2020.

[7] A. C. Müller and S. Guido, *Introduction to machine learning with Python: a guide for data scientists*, First edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly, 2016.

[8] M. Khder, "Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application," *Int. J. Adv. Soft Comput. Its Appl.*, vol. 13, no. 3, pp. 145–168, Dec. 2021, doi: 10.15849/IJASCA.211128.11.

[9] H. S. Obaid, S. A. Dheyab, and S. S. Sabry, "The Impact of Data Pre-Processing Techniques and Dimensionality Reduction on the Accuracy of Machine Learning," in *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*, Jaipur, India: IEEE, Mar. 2019, pp. 279–283. doi: 10.1109/IEMECONX.2019.8877011.

[10] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Front. Energy Res.*, vol. 9, p. 652801, Mar. 2021, doi: 10.3389/fenrg.2021.652801.

[11] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing For Student Complaint Document Classification Using Sastrawi," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, p. 012017, Jun. 2020, doi: 10.1088/1757-899X/874/1/012017.

[12] Muhammad Ikram Kaer Sinapoy, Yuliant Sibaroni, and Sri Suryani Prasetyowati, "Comparison of LSTM and IndoBERT Method in Identifying Hoax on Twitter," *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 7, no. 3, pp. 657–662, Jun. 2023, doi: 10.29207/resti.v7i3.4830.

[13] A. O. Salau and S. Jain, "Feature Extraction: A Survey of the Types, Techniques, Applications," in *2019 International Conference on Signal Processing and Communication (ICSC)*,

NOIDA, India: IEEE, Mar. 2019, pp. 158–164. doi: 10.1109/ICSC45622.2019.8938371.

[14] Nanda Ihwani Saputri, Yuliant Sibaroni, and Sri Suryani Prasetiyowati, “Covid-19 Fake News Detection on Twitter Based on Author Credibility Using Information Gain and KNN MethodsCovid-19 Fake News Detection on Twitter Based on Author Credibility Using Information Gain and KNN Methods,” *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 7, no. 1, pp. 185–192, Feb. 2023, doi: 10.29207/resti.v7i1.4871.

[15] B. P. Zen, I. Susanto, and D. Finaliamartha, “TF-IDF Method and Vector Space Model Regarding the Covid-19 Vaccine on Online News,” *SinkrOn*, vol. 6, no. 1, pp. 69–79, Oct. 2021, doi: 10.33395/sinkron.v6i1.11179.

[16] T. Jiang, J. L. Gradus, and A. J. Rosellini, “Supervised Machine Learning: A Brief Primer,” *Behav. Ther.*, vol. 51, no. 5, pp. 675–687, Sep. 2020, doi: 10.1016/j.beth.2020.05.002.

[17] Angga Aditya Permana *et al.*, *Machine Learning*. in I. PT Global Eksekutif Teknologi, 2023.

[18] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, “Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking,” *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.

[19] F. Rahmad, Y. Suryanto, and K. Ramli, “Performance Comparison of Anti-Spam Technology Using Confusion Matrix Classification,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 879, no. 1, p. 012076, Jul. 2020, doi: 10.1088/1757-899X/879/1/012076.

[20] Y. Kalmukov, “Using Word Clouds For Fast Identification Of Papers’ Subject Domain And Reviewers’ Competences15,” vol. 60, 2021.