Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

Implementation of Enhanced Confix Stripping Stemming and Chi-Squared Feature Selection on Classification UIN Walisongo Website with Naïve Bayes Classifier

Muhammad Naufal Muhadzib Al-Faruq*[1], Wenty Dwi Yuniarti[2], Maya Rini Handayani[3], Khotibul Umam[4]

[1,2,3,4]Information Technology, Science and Technology, Walisongo State Islamic University, Indonesia

Email: [1]naufalfaruq082@gmail.com

Received : Apr 28, 2025; Revised : Jun 2, 2025; Accepted : Jun 3, 2025; Published : Jun 23, 2025

Abstract

Academic news classification on university websites remains a challenge due to the growing volume of content and lack of efficient categorization systems. At UIN Walisongo Semarang, this problem hinders students, faculty, and the public from easily accessing relevant information. This study aims to develop an automated academic news classification system to address this issue. We applied a Naïve Bayes Classifier model, enhanced with Term Frequency weighting, the Enhanced Confix Stripping Stemmer for Indonesian language preprocessing, and Chi-Squared feature selection to identify the most informative terms. The dataset consisted of 880 academic news articles from UIN Walisongo's website, split into 704 training and 176 testing documents. The system achieved 95% accuracy on the test set. To evaluate generalizability, we used a separate evaluation set of 12 new articles, obtaining 83.3% accuracy. The preprocessing stage played a vital role in reducing morphological complexity, while Chi-Squared scoring improved the relevance of selected features. This research highlights the importance of robust text classification techniques in academic information systems, particularly in Indonesian language contexts where language morphology poses unique challenges. The proposed model demonstrates strong performance, scalability, and potential for integration into academic portals to improve information retrieval. This study contributes significantly to the field of Natural Language Processing and applied machine learning in academic settings, especially for Indonesian-language content. It provides an effective solution for automated academic content management in institutional information systems.

Keywords : *Academic News, Chi-squared, Classification, Enhanced Confix Stripping Stemmer, Naive Bayes Classifier*

1. INTRODUCTION

The digital age has placed a premium on effortless access to information, a critical cornerstone for research and academic development within higher education institutions (HEIs). However, Walisongo State Islamic University Semarang (UIN Walisongo), despite its dedication to Islamic knowledge and science, faces challenges in ensuring efficient and effective information access for its diverse stakeholders, encompassing students, faculty, researchers, and broader community.

This article proposes a methodology for classifying news information on the UIN Walisongo's website. The ever-growing volume and diversified nature of news items necessitate the implementation of an automated news categorization system. Recent research has demonstrated the effectiveness of Naïve Bayes combined with Support Vector Machine in classifying user sentiment on mobile applications [1]. The objective of this study is to categorize news articles into four distinct categories: campus, education, events, and announcements. This classification framework is designed to significantly enhance the accessibility and user experience of news content on the UIN Walisongo's

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

website. To achieve this objective, data mining techniques are employed to analyze and automatically classify news articles.

This study addresses the gap in the availability of robust, domain-specific automated classification systems tailored for Indonesian academic news content. While prior studies have focused on general news classification, few have explored the challenges of classifying academic news articles, which often use formal, domain-specific language and unique structural patterns.

Data mining is a field of computer science to analyze and extract valuable information from previously unknown data [2]. Naïve Bayes classifiers have shown robust performance in hoax news classification in recent studies [3]. In this context, the data will be extracted from UIN Walisongo's website. Classification, another concept crucial to this study, refers to the process of assigning data points to specific categories based on a predefined set of criteria [4]. Naïve Bayes has also been successfully applied for classifying public complaints data, demonstrating its versatility across different text classification tasks [5]. The ultimate goal is to develop a classification model that can automatically categories news articles into designated classes, such as campus news, educational announcements, upcoming events, and general university updates. This model will be built upon the analysis of the extracted dataset [6].

Prior to news classification, a preprocessing stage is implemented to prepare the data for analysis. This stage encompasses several essential steps, including case folding, tokenization/parsing, filtering, and stemming [7]. Stemming is a particularly important step within preprocessing, aiming to reduce inflected words to their base forms. This process is achieved through the Enhanced Confix Stripping Stemmer Algorithm, a development of the Confix Stripping Algorithm that addresses several limitations of its predecessor. The original Confix Stripping Algorithm presents certain limitations in processing Indonesian words, especially when dealing with complex affix combinations. This model often struggles to accurately remove prefixes and suffixes for specific word structures, which can lead to incorrect base forms. A comprehensive explanation of these limitations is provided in the methodology section [8].

This study explores the application of the Naïve Bayes Classifier for the automated categorization of online news articles. The Naïve Bayes Classifier has emerged as a prominent method for data classification across diverse domains, including medical diagnosis, computer network analysis, and text categorization [9]. Its enduring popularity stems from its inherent simplicity, effectiveness, and ability. The Naïve Bayes Classifier demonstrates competitive performance in terms of computational efficiency when compared to more complex classification algorithm [10].

Research on online news classification has been conducted extensively by previous researchers. Pramudita et al. (2018) in their study, a Naïve Bayes classifier with Enhanced Confix Stripping Stemmer was employed to classify sports news articles. Training data consisted of 18 randomly selected sports news articles chosen by the user. User analysis determined that 14 out of 18 articles were classified correctly, resulting in an accuracy 77%. This finding suggest that the proposed method holds promise for classifying sports news into predefined categories such as Football, Basketball, Racket, Formula 1, Motto GP, etc. [11].

Building upon the work of Ariadi and Fithriasari (2022) who compared Naïve Bayes (NB) and Support Vector Machine (SVM) classifier for Indonesian news classification, this study investigated the performance of these methods. Their finding indicated that the SVM classifier achieved superior performance using a Radial Basis Function (RBF) kernel and a linear kernel on a 10,000-word vector, with accuracy, precision, recall, and F-measure scores of 88.1%, 89.1%, 88.1%, and 88.3% respectively. In comparison the NB classifier achieved an accuracy of 82.2%, precision of 83,9%, recall of 82.2%, and F-measure of 82.4%. These results suggest that SVM with RBF and linear kernel may be a more effective approach for classifying Indonesian news articles [12].

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

Hidayat and Rizqi (2020) investigated document classification using a dataset of 600 news articles obtained from the Jawa Pos news portal (https://www.jawapos.com/). The data was categorized into four balanced groups: Sports, Technology, Economy, and Miscellaneous (each containing 150 articles). The researchers employed a preprocessing stage involving case folding, tokenization, filtering, and stemming to prepare the data for classification. Their findings revealed high classification accuracy across all categories, achieving 90% for Sports, and Technology, 100% for Economy, and Miscellaneous. The overall average accuracy reached 95% [13].

Prakoso et al. (2019) explored Naïve Bayes classifier with feature selection and boosting techniques for news articles classification on the detik.com website. Their research involved data modelling and processing. They started with preprocessing, implemented Information Gain for feature selection, and using Bayesian Boosting for the algorithm. Information Gain feature selection fielded an accuracy of 69.5%, while a simpler model using Naïve Bayes Classifier with Bayesian Boosting achieved an accuracy of 73.2%. Furthermore, employing Bayesian Boosting for polynomial labels resulted in a 4.3% improvement in evaluation metrics. This finding suggest that Information Gain may not significantly enhance performance, particularly for dataset with polynomial labels [14].

Unlike previous works, this study introduces improvements to the Enhanced Confix Stripping Stemmer by modifying prefix-handling rules such as for words with the patterns “mem+p”, “peng+k”, and “penge+” and implementing an iterative suffix recovery mechanism. These enhancements address stemming errors commonly found in academic Indonesian texts, which are not sufficiently handled in prior research.

This study employs the Chi-Squared method to identify words with the highest Chi-Squared scores within each news category. To reduce inflected variants to their base form, the Enhanced Confix Stripping Stemmer is then applied to segment words by removing prefixes and suffixes. Following this, a Naïve Bayes Classifier algorithm is implemented to the model to train a news category prediction for classifying news articles from UIN Walisongo’s website.

Additionally, this research highlights the synergy between Enhanced Confix Stripping and Chi-Squared feature selection. While the stemmer reduces morphological noise and ensures consistency of word forms, Chi-Squared scoring filters the most statistically relevant terms. This combination significantly improves both linguistic and statistical quality of features used in classification.

This study investigates the effectiveness of the Naïve Bayes Classifier in accurately categorizing academic news documents on UIN Walisongo’s website. It also examines the extent to which the Enhanced Confix Stripping Stemmer improves the stemming process for Indonesian text and evaluates the role of the Chi-Squared method in enhancing feature selection for the classification model.

2. METHOD

In machine learning, classification refers to the task of constructing a model or function that can effectively categorize data points into predefined classes. The primary objective is to predict the class label of a new, unseen data point. This process typically involves two key stages: first, the model is trained using a set of labeled data points (training data), and then it is evaluated by classifying a separate set of labeled data points (test data) to assess its accuracy [15].

In machine learning, classification refers to the task of constructing a model or function that can effectively categorize data points into predefined classes. The primary objective is to predict the class label of a new, unseen data point. This process typically involves two key stages: first, the model is trained using a set of labeled data points (training data), and then it is evaluated by classifying a separate set of labeled data points (test data) to assess its accuracy.

In this study, the first stage begins with text preprocessing, which includes case folding, tokenization, filtering, and stemming using the Enhanced Confix Stripping Stemmer algorithm. Next,

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

feature selection is performed using the Chi-Squared method, followed by model training with the Naïve Bayes algorithm. The complete research flow is visualized in Figure 1, outlining each stage of the classification process from data preprocessing to evaluation.

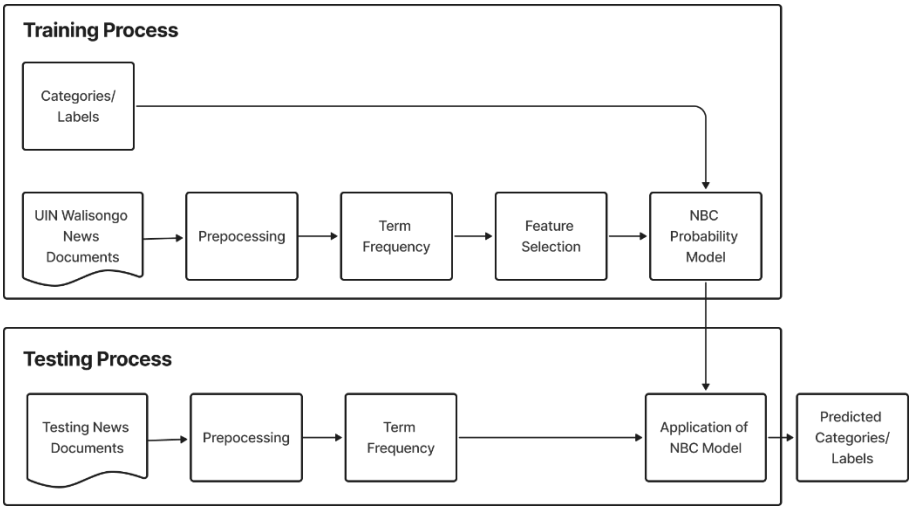

Figure 1. Research Flow of The Academic News Classification System

The second stage involves testing the model with unseen test data, which also undergoes preprocessing (case folding, tokenization, filtering, and stemming). The preprocessed test data is then classified using the trained model. Finally, the model's performance is evaluated by comparing its predictions with the actual labels to calculate the accuracy, ensuring its reliability for real-world applications.

2.1. Text Preprocessing

Text Preprocessing constitutes the foundational step in text processing pipelines. It aims to transform raw, unstructured text data into a format that facilitates efficient and accurate computational analysis. This process involves several stages [7]. Enhanced stemming algorithms like Modified Enhanced Confix Stripping are crucial for handling Indonesian language morphology [16].

a. Case Folding: this initial step involves converting all characters in the text to lowercase. This normalize the data by eliminating inconsistencies arising from capitalization, which can be irrelevant for many text analysis tasks.
b. Tokenizing: the text is then segmented into meaningful units, typically word or terms. This process creates a stream of token that serve as the basic elements for further analysis.
c. Filtering: following tokenization, filtering techniques are employed to select informative words or tokens. This stage often utilizes stopword removal algorithms, which eliminate common and uninformative words (e.g., “the”, “a”, “is” etc.). Alternatively, custom wordlist can be used to retain domain-specific or relevant terms.
d. Stemming: The goal of stemming is to reduce inflected words to their base forms. This process aims to identify the morphological root (stem) of each word, thereby grouping related words together and improving the overall effectiveness of subsequent analysis tasks.

2.2. Chi-Squared

The Chi-Squared test, a well-established statistical method, proves valuable in determining the association between two categorical variables [17]. Within the domain of text classification, this technique is employed to identify the most relevant features that contribute to distinguishing between different categories. Feature selection techniques such as Chi-Square have been proven to improve

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

Naïve Bayes performance significantly [18]. The Chi-Squared score quantifies the discrepancy between the observed (O) and expected (E) frequencies of features within each category. The specific formula for calculating this score is presented in equation (1).

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

In equation (1), X^2 represents the Chi-Squared score, while O_i and E_i denote the observed and expected frequencies of a word in a specific category, respectively. The parameter n refers to the number of categories being analyzed. This equation calculates the sum of the squared differences between observed and expected frequencies $(O_i - E_i)^2$ for each category, normalized by the expected frequency E_i. A higher Chi-Squared score signifies a stronger association between a feature and a particular category. Therefore, features with the highest Chi-Squared scores are selected as the most informative, improving the accuracy of the classification model by emphasizing these significant features.

The expected frequency E_i, which is crucial for calculating the Chi-Squared score, is determined based on the overall distribution of features across categories. It is calculated using the formula presented in equation (2).

$$E_i = \frac{R \; x \; C}{N} \quad (2)$$

In equation (2), R refers to the total occurrences of a feature across all categories, C represents the total number of occurrences in a specific category, and N is the total number of occurrences of all features across all categories. By dividing the product of R and C by N, the formula estimates the expected frequency of a feature within a given category under the assumption of independence. This calculation ensures that the Chi-Squared score effectively captures the degree of association between a feature and its corresponding category, facilitating the selection of the most relevant features for classification tasks.

2.3. Enhanced Confix Stripping Stemmer Algorithm

The Enhanced Confix Stripping Stemmer algorithm builds upon its predecessor, the Confix Stripping Stemmer, aiming to address limitations identified during experimentations and analysis. These limitations manifested in the inability the Confix Stripping Stemmer to effectively stem certain words. The specific issues encountered are as follows [8].

a. The absence of prefix-breaking rules for words with the format “mem+p...”. This results in the inability to correctly stem words like “mempromosikan” and “memproteksi”.
b. Other limitation is the absence of prefix-breaking rules for words following the format “men+s...”. This leads to the incapability of correctly stemming words such as “mensyaratkan” and “mensyukuri”.
c. There are limitations in handling specific verb. For instance, the Stemmer fails to separate the prefix “menge+...” from the verb stem in words like “mengerem”.
d. The Stemmer cannot handle words with the prefix “penge+…” followed by consonant. This leads to incomplete stemming in words like “pengeboman”.
e. Inability to handle prefix words beginning with “peng+k...” followed by a consonant, leading to mis-stemmed words like “pengkajian”.
f. The Stemmer incorrectly identifies certain suffixes “-an”, “-kan”, and “-ku” in words like “pelanggan”, “perpolitikan”, and “pelaku”as prefixes, leading to the removal of these essential morphological elements

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

Due to limitations in the prefix-handling capabilities of the original Confix Stripping Stemmer algorithm, the Enhanced version introduces several rule modifications to improve prefix segmentation accuracy, especially for complex Indonesian word formations. As summarized in Table 1, these enhancements define specific transformations for word structures such as "mem+V", "peng+C", and "penge+V". Each rule aims to correct errors previously unhandled in the standard approach. By making these rule-based refinements explicit, the Enhanced Confix Stripping Stemmer contributes to more reliable root word extraction during preprocessing—an essential component in improving classification accuracy. [19].

Table 1. Additions and Modifications to Word Prefix Rules in Enhance Confix Stripping Stemmer

Rules	Word Structure	Prefix Segmentation
14	men {e|c|d|j|z}...	men-{e|c|d|j|z}...
17	mengV	eng-V… | meng- kV… | (mengV-... If V='e')
19	mempA…	mem-pA…where A!='e'
28	pengC…	peng-C…
29	pengV…	peng-V… | peng- kV… | (pengV-... If V='e')

As shown is the table above, the Enhanced Confix Stripping Stemmer algorithm incorporates several enhancements. These include the modification of certain rules utilized in the decapitation of words and the introduction of a novel rule, designated as "to overcome suffix decaptitating errors". The specific details of loop "Pengembalian Akhiran" (Ending Return) algorithm is defined as follows [11].

a. Following prefix segmentation, the algorithm attempts to reconstruct the original word by restoring the previously removed prefix. This reconstructed word, referred to as the "word model", is then subjected to a dictionary lookup process. If the dictionary lookup successfully identifies the word model, the process terminates. Conversely, if the dictionary lookup fails to locate the word model, the algorithm proceeds to the next step.
b. The algorithm employs an iterative process to reinstate previously removed suffixes. This restoration prioritizes suffixes in the following order: Derivational Suffixes ("-i", "-kan", "-an"), Possessive Pronoun ("-my", "-mu", "-nya"), Inflectional Particle Affixes ("-lah", "-kah", "-tah", "-pun"). For each suffix type, the algorithm attempts to reconstruct the word by appending the suffix to the current stem and checks if the resulting from exists in a base word dictionary. If a match is found in the dictionary, the process terminates. Otherwise, the algorithm proceeds to the prefix beheading rules for further segmentation. Additionally, if the base word dictionary lookup remains unsuccessful after appending all the aforementioned suffixes, the previously removed prefixes are restored again, followed by iteration of the suffix restoration process.

2.4. Term Frequency

Term Frequency is a fundamental technique used in information retrieval and text mining to assess the significance of individual words within a document. This method operates under the assumption that a words importance is directly proportional to its frequency of occurrence within the text. In simpler term, words that appear more frequently are considered to hold greater weight compared to those that appear less often. This weighting scheme is based on the intuition that words that are repeated more frequently are likely to be more relevant or informative (higher weight) in conveying the document's central theme [7]. Chi-Square feature selection has been effectively applied for optimizing text classification models [20].

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

2.5. Naive Bayes Classifier

Naive Bayes Classifier, also known as Bayesian Classification, represent a family of machine learning algorithm that utilize Bayes' theorem for classification purposes. Bayes' theorem is a fundamental principle in probability theory that allows for the calculation of the conditional probability of an event (class membership) occurring given another event (specific features). In essence, the algorithms leverage the theorem to predict the likelihood of a data point belonging to a particular class based on its observed features. The mathematical formulation of Bayes' theorem is commonly expressed as equation (3).

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (3)$$

The conditional probability of A given B, represented by P(A|B), and the conditional probability of B given A (P(B|A). P(A) is the probability of event A, and P(B) is the probability of event B.

This study employs the Naive Bayes Classification method for text classification due to its appealing combination of simplicity and efficiency. The algorithm's core principle lies in its probabilistic approach, where each document is represented as a feature vector. This vector denoted as "a_1, a_2, a_3, ..., a_n" captures the presence or absence of individual words (a_1) to (a_n) within the document.

During classification, the Naïve Bayes algorithm seeks to identify the category with the highest posterior probability. Optimization of text classification models using Chi-Square and TF-IDF has been explored to increase classification accuracy [21]. This probability is calculated using V_{MAP}, which is detailed in equation (4). The V_{MAP} equation is derived from Bayes' theorem and assumes conditional independence, enabling efficient computation of posterior probabilities for each category.

$$V_{MAP} = argmax\, v_j \in v\ P(v_j) \prod_i P(a_i|v_j) \quad (4)$$

V_{MAP} presents the category with the maximum posterior probability. The term $P(v_j)$ denotes the prior probability of category v, while $P(a_i|v_j)$ is the conditional probability of feature a_i given that the data belongs to category v_jThe product $\prod_i P(a_i|v_j)$ combines the probabilities of all features assuming independence.

$$P(v_j) = \frac{|doc\ j|}{|training|} \quad (5)$$

Where |doc j| represents the number of documents (news articles) belonging to category j, divided by the total number of documents in the training set. Additionally, equation (6) governs the calculation of the probability of word a_i occurring within each category $P(a_i|v_j)$ during the training process.

$$P(a_i|v_j) = \frac{|n_i+1|}{|n+kosakata|} \quad (6)$$

The n_i represents the number of times word a_i appears in documents belonging to category v_j, while n is the total number of words in documents belonging to category v_j, and vocabulary is the total number of words in the examples in the training sample. To measure classification accuracy, equation (7) can be used.

$$Accuracy = \frac{Number\ of\ Correct\ Test\ Documents}{Total\ Test\ Documents}\ x\ 100\% \quad (7)$$

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

3. RESULT

The classification structure employed for Walisongo State Islamic University Semarang (UIN Walisongo) news documents involves two distinct stages: training and testing. Data collection was conducted through the ppid.walisongo.ac.id news portal, yielding a corpus of 880 news documents. Subsequently, the documents were partitioned into two subsets with a ratio of 80% designated for training and 20% reserved for testing purposes. During the training stage, a probability model was constructed utilizing 704 training documents. The remaining 176 documents were used as test documents to evaluate the efficacy of the constructed probability model.

Prior to the commencement of the training process, the training documents outlined in Table 2 undergo a series of preprocessing steps, including term frequency analysis and feature selection. Test documents are subjected to a similar preparatory process before classification based on probability model constructed during the learning phase. The preprocessing stage aims to transform unstructured text data into a structured format suitable for machine learning algorithms. This stage encompasses several techniques, such as case folding, tokenization, filtering, and stemming.

As detailed in Table 3, the first step involves case folding, where all characters are converted to lowercase to ensure consistency. Subsequently, the tokenization stage (Table 4) parses the text content into individual words. Following tokenization (Table 5), a filtering process is employed to retain only relevant words and eliminate stopwords, which are common words that hold minimal semantic value.

The final preprocessing step involves stemming, which reduces words to their root forms. This research employs the Enhanced Confix Stripping Stemmer algorithm for this purpose. Table 6 exemplifies the outcome of this process on a document.

Table 2. Sample Training Document

Title	Content	Category
Walisongo Halal Center Terbitkan 275 Sertifikat Juru Sembelih Halal Se-eks Karesidenan Pekalongan	Walisongo Halal Center (WHC) bekerjasama dengan BI Tegal, Komite Nasional Ekonomi dan Keuangan Syariah ..... tegasnya	Event
FDK UIN Walisongo-PAMHU gelar Sertifikasi Pembimbing Manasik Haji Profesional Angkatan XV	Fakultas Dakwah dan Komunikasi (FDK) UIN Walisongo Semarang menyelenggarakan sertifikasi pembimbing manasik haji profesional ..... PAHMU	Education

Table 2 presents an illustrative example of an academic news document from UIN Walisongo Semarang. This documents comprises the title and content of the news article and will be utilized as a training document. The subsequent step in the preprocessing pipeline involves case folding, as detailed in Table 3.

Table 3. Example Document Result Of Case Folding Process

Title	Content	Output Title	Output Content
Walisongo Halal Center Terbi-tkan 275 Sertifi-kat Juru Semb-elih Halal Se-eks Kar-esidenan Pekalongan	Walisongo Halal Center (WHC) bekerjasama dengan BI Tegal, Komite Nasional Ekonomi dan Keuangan Syariah ..... tegasnya	walisongo halal ce-nter terbi-tkan 275 sertifi-kat juru semb-elih halal se-eks kar-esidenan pekalongan	walisongo halal center (whc) beker-jasama deng-an bi tegal, komite nasi-onal ekonomi dan keuangan syariah ..... tegasnya

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

FDK UIN Walisongo-PAMHU gelar Sertifikasi Pembimbing Manasik Haji Profesional Angkatan XV	Fakultas Dakwah dan Komunikasi (FDK) UIN Walisongo Semarang menyelenggarakan sertifikasi pembimbing manasik haji profesional ..... PAHMU	fdk uin walisongo-pamhu gelar sertifikasi pembimbing manasik haji profes-ional angk-atan xv	fakultas dakwah dan komunikasi (fdk) uin walisongo semarang menyelenggarakan sertifik-asi pembimb-ing manasik hajiprofesio-nal ..... pahmu

Following the case folding process, both the title and news content are converted entirely to lowercase letters. This step serves to eliminate inconsistencies arising from capitalization, thereby mitigating the potential for redundant data representations in Table 4.

Table 4. Example Of Tokenizing Result Document

Title	Content	Output Title	Output Content
walisongo halal ce-nter terbi-tkan 275 sertifi-kat juru semb-elih halal se-eks kar-esidenan pekalongan	walisongo halal center (whc) beker-jasama deng-an bi tegal, komite nasi-onal ekonomi dan keuangan syariah ..... tegasnya	walisongo halal center terbitkan 275 sertifikat juru sembelih halaleks', karesidenan pekalongan	walisongo halal center whc beker-jasama deng-an bi tegal, komite nasi-onal ekonomi dan keuangan syariah ..... tegasnya
fdk uin walisongo-pamhu gelar sert-ifikasi pe-mbimbing manasik haji profes-ional angk-atan xv	fakultas dakwah dan komunikasi (fdk) uin walisongo semarang menyelenggarakan sert-ifik-asi pem-bimbing ma-nasik haji profesional ..... pahmu	fdk uin walisongo-pamhu gelar sertifikasi pembimbing manasik haji profes-ional angk-atan	fakultas dakwah dan komunikasi fdk uin walisongo semarang menyelenggarakan sertifik-asi pembimb-ing manasik hajiprofesio-nal ..... pahmu

Subsequent to case folding, the next step is tokenization. This process involves segmenting the text into smaller units known as tokens. Tokenization simplifies the text by dividing it into individual words or sub-words, facilitating subsequent natural language processing task.

Following tokenization, where the text is segmented into individual words, a filtering process is implemented. This stage aims to select informative terms from the tokenized corpus. Conversely, unimportant words, of the referred to as stopwords, are eliminated using a stopwords removal algorithm.

Table 5. Example Of Tokenizing Result Document

Title	Content	Output Title	Output Content
walisongo halal ce-nter terbi-tkan 275 sertifi-kat juru semb-elih halal se-eks kar-esidenan pekalongan	walisongo halal center whc beker-jasama deng-an bi tegal, komite nasi-onal ekonomi dan keuangan syariah ..... tegasnya	'walisongo', 'halal', 'center', 'terbitkan', '275', 'sertifikat', 'juru', 'sem-belih', 'hal-al', 'eks', 'ka-residenan', 'pekalongan'	'walisongo', 'halal', 'center', 'whc', 'bekerjasama', 'tegal', 'komite', 'nasional', 'ekonomi', 'keuangan', 'syariah', ... 'tegasya'
fdk uin walisongo-pamhu gelar sert-ifikasi pe-mbimbing	fakultas dakwah dan komunikasi fdk uin walisongo semarang menyelenggarakan sert-	'fdk', 'uin', 'walisongo', 'pamhu', 'gelar', 'sert-ifikasi', 'pe-	'fakultas', 'dakwah', 'komunikasi', 'fdk', 'uin', 'walisongo', 'semarang', 'menyelenggarakan',

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

manasik haji profes-ional angk-atan	ifik-asi pem-bimbing ma-nasik haji profesional ..... pahmu	mbimbing', 'manasik', 'haji', 'profe-sional', 'angkatan'	'sertifikasi', 'pembimbing', 'manasik', 'haji', 'profesional', ... 'pahmu'

Figure 2 outlines the stages of the Enhanced Confix Stripping Stemmer Algorithm. This algorithm integrates three distinct components: Enhanced, Confix Stripping, and Stemmer, each addressing specific aspects of word normalization.

The Enhanced component represents improvements over the original Confix Stripping Algorithm. This stage begins with a precedence rule check, introducing new rules to handle complex affix combinations such as "mem-", "meng-", and "penge-". These enhancements address limitations in the original algorithm, allowing it to process a broader range of Indonesian word structures accurately.

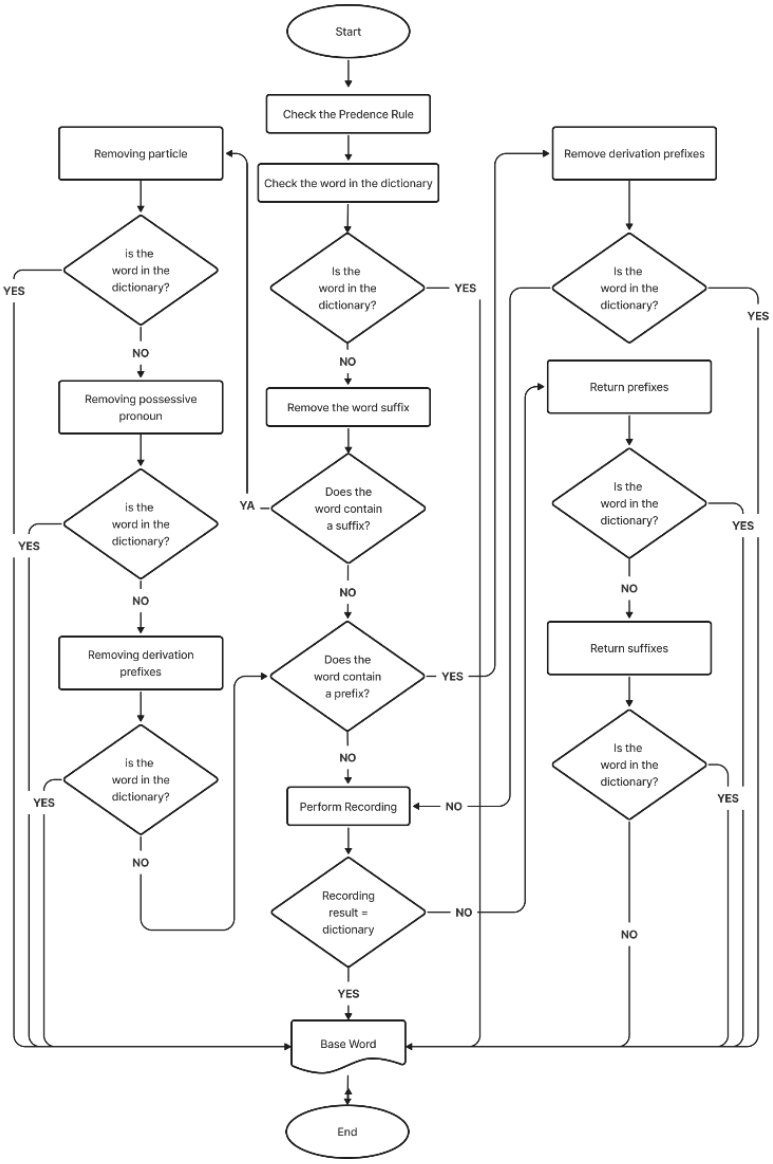

Figure 2. Flowchart Enhanced Confix Stripping Stemmer

The Confix Stripping process focuses on systematically removing affixes from words. It is divided into two main stages:

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

a. Case Folding: this initial step involves converting all characters in the text to lowercase. This Suffix Removal: This stage eliminates inflectional particle affixes (e.g., "-lah", "-kah"), possessive pronouns (e.g., "-ku", "-mu", "-nya"), and derivational suffixes (e.g., "-kan", "-an").
b. Prefix Removal: After suffix removal, prefixes such as "meng-", "pen-", and "ber-" are removed using derivation prefix elimination rules. This ensures that the resulting word aligns more closely with its base form.

Prefix Removal: After suffix removal, prefixes such as "meng-", "pen-", and "ber-" are removed using derivation prefix elimination rules. This ensures that the resulting word aligns more closely with its base form.

The Stemmer serves as the overarching process that integrates the Enhanced and Confix Stripping components to identify the base word. If a word cannot be matched to a base word after prefix and suffix removal, the algorithm employs the word recording process. This process reconstructs the word by restoring previously removed prefixes or suffixes and checking the reconstructed word against a base word database. If no match is found, the algorithm iterates through the Ending Return Alghoritm stage, reapplying suffix rules to refine the normalization process.

By clearly delineating these components, the Enhanced Confix Stripping Stemmer Algorithm effectively addresses the challenges of processing Indonesian text, ensuring robust and accurate stemming results.

Table 6. Document Example Of Enhanced Confix Stripping Stemmer Implementation Result

Title	Content	Output Title	Output Content
'walisongo', 'halal', 'center', 'terbitkan', '275', 'sertifikat', 'juru', 'sem-belih', 'hal-al', 'eks', 'ka-residenan', 'pekalongan'	'walisongo', 'halal', 'center', 'whc', 'bekerjasama', 'tegal', 'komite', 'nasional', 'ekonomi', 'keuangan', 'syariah', ... 'tegasya'	'walisongo', 'halal', 'center', 'terbit', '275', 'sertifikat', 'juru', 'sembelih', 'halal', 'eks', 'karesidenan', 'kalong'	'walisongo', 'halal', 'center', 'whc', 'bekerjasama', 'tegal', 'komite', 'nasional', 'ekonomi', 'uang', 'syariah', ... 'tegas'
'fdk', 'uin', 'walisongo', 'pamhu', 'gelar', 'sert-ifikasi', 'pe-mbimbing', 'manasik', 'haji', 'profe-sional', 'angkatan'	'fakultas', 'dakwah', 'komunikasi', 'fdk', 'uin', 'walisongo', 'semarang', 'menyelenggarakan', 'sertifikasi', 'pembimbing', 'manasik', 'haji', 'profesional', ... 'pahmu'	'fdk', 'uin', 'walisongo', 'pamhu', 'gelar', 'sertifikasi', 'pembimbing', 'manasik', 'haji', 'profesional', 'angkatan'	'fakultas', 'dakwah', 'komunikasi', 'fdk', 'uin', 'walisongo', 'semarang', 'menyelenggarakan', 'sertifikasi', 'pembimbing', 'manasik', 'haji', 'profesional'..... 'pahmu'

Table 6 presents the results of applying the Enhanced Confix Stripping Stemmer to real academic news documents. The output columns demonstrate how complex affixed words are normalized into their root forms. This process significantly improves feature consistency by reducing morphological variation in Indonesian text, which is essential for achieving higher accuracy in text classification. The normalized terms allow the Naïve Bayes model to better generalize and distinguish between categories such as education, event, campus, and announcements.

Following preprocessing, the training data underwent feature selection using the Chi-Square test. This method aimed to identify the most discriminative features for news categorization, thereby enhancing the performance of the Naïve Bayes classifier. The process, visualized in Figure 3, involved calculating the total word count per category, determining observed and expected frequencies for each

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

term within each category, computing the Chi-Square score, and selecting features based on their associated Chi-Square values.

Observed frequency represents the actual occurrences of a term within a specific category, while expected frequency is the theoretically predicted occurrence under the assumption of independence. The Chi-squared statistic measures the deviation between these frequencies, indicating the term's relevance to the category. Higher Chi-Square values correspond to stronger associations between terms and categories. The features with the highest Chi-Square scores were subsequently employed as input for the Naïve Bayes classifier. Table 7 presents the top three terms with the highest Chi-Square values for each category. A Chi-Square score approaching 1.0 signifies a more pronounced association between term and its corresponding category.

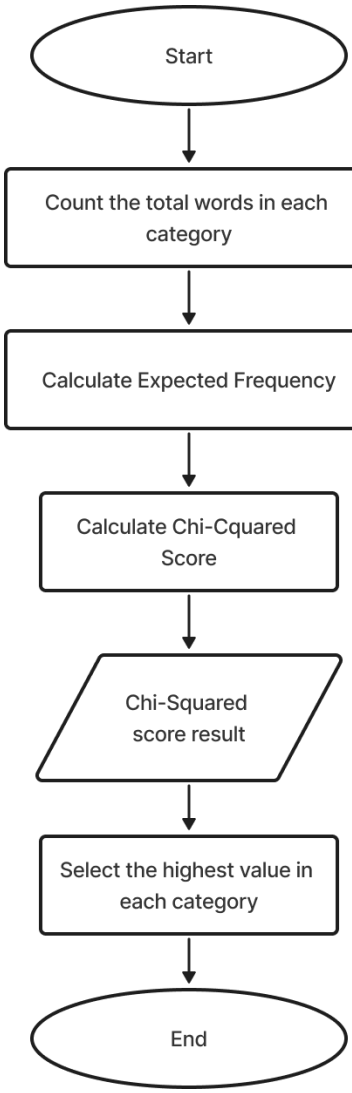

Figure 3. Flowchart Feature Selection Chi-Squared

Table 7. Sample Document of Chi-Squared Implementation Result

Word	Chi-Squared Score	Category
mahasiswa	1.0	education
dosen	0.98	education
kuliah	0.98	education
seminar	1.0	event

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

upacara	0.99	event
diskusi	0.98	event
ukt	1.0	campus
uin	0.99	campus
ptkin	0.99	campus
seleksi	1.0	announcement
Pendaftaran	0.94	announcement
informasi	0.94	announcement

The final stage involved training the Naïve Bayes classifier using the preprocessed training documents. This process encompassed several steps. Firstly, test documents, as outlined in Table 8, were inputted and subjected to preprocessing. Subsequently, term frequency calculations were performed. The system then calculated *Vj = P (vj) n P(ai|vj)* for each category, utilizing the probability model established during the training phase. Upon completion, the system automatically assigned categories to the test documents. Evaluation was conducted using test documents with unknown categories, as listed in Table 8.

Table 8. Sample Test Document

Title	Content	Category
Moderasi Beragama, Hidup Damai di Tengah Masyarakat Plural	Gagasan moderasi beragama ala Walisongo harus disebarkan kepada seluruh masyara-kat Indonesia, karena sikap Walisongo dalam mengapresiasi local wisdom ..... pelatihan	Unknow
Sekretaris RMB UIN Walisongo mengajar Political Theology Moderasi Beragama di Universitas Vienna, Austria	Luthfi Rahman, diundang untuk mengisi mata kul-iah Approaches of Islamic Theology to Dialogue di sa-lah satu kelas ma-hasiswa jurusan ..... beragam	Unknow

Table 8 presents an example of an uncategorized academic news document from UIN Walisongo Semarang. Comprising a headline and news content. This document will serve as a test case for evaluating the previously constructed Naïve Bayes probability model, which was trained using a separate dataset. Prior to classification, the test document undergoes preprocessing, including case folding, tokenizing, filtering, and stemming.

Table 9. Example Of Categorized Test Document

Title	Content	Category
Moderasi Beragama, Hidup Damai di Tengah Masyarakat Plural	Gagasan moderasi beragama ala Walisongo harus disebarkan kepada seluruh masyara-kat Indonesia, karena sikap Walisongo dalam mengapresiasi local wisdom ..... pelatihan	Education
Sekretaris RMB UIN Walisongo mengajar Political Theology Moderasi Beragama di Universitas Vienna, Austria	Luthfi Rahman, diundang untuk mengisi mata kul-iah Approaches of Islamic Theology to Dialogue di sa-lah satu kelas ma-hasiswa jurusan ..... beragam	Event

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

Following the preprocessing stage, the machine automatically categorized the test documents using the constructed Naïve Bayes probability model. A total of 176 test documents were evaluated, achieving the accuracy rate of 95%. As detailed in Table 9, 167 documents were classified correctly.

In the context of classifying academic news documents from UIN Walisongo Semarang, three documents categories exist: training, testing, and new documents. The initial data collection retrieved 880 news documents from ppid.walisongo.ac.id. these documents were subsequently divided into two subsets:80% designated as training documents and 20% reserved for testing purposes. During the training phase, the probability model was built utilizing 704 training documents. The remaining 176 documents were employed as test documents to evaluate the efficacy of the model. To further assess the system's categorization accuracy, an additional set of 12 latest news was created and compared against user-assigned categories.

Table 10. Testing Results

No	News Title	News Content	Category (System)	Category (User)
1	Pengumuman Beasiswa KIP-K untuk Mahasiswa Baru Tahun 2024	UIN Walisongo memberikan pengumuman terkait program beasiswa KIP-K ..... kampus	Announcement	Announcement
2	Fakultas Sains dan Teknologi Adakah Pelatihan Kewirausahaan Bagi Mahasiswa	'Fakultas Sains dan Teknologi akan menyelenggarakan pelatihan kewirausahaan ..... mereka	Event	Event
3	Civitas Akademik Melakukan Kerja Bakti Untuk Menjaga Kebersihan Lingkungan Kampus	Civitas akademik mengadakan kegiatan membersihkan lingkungan kampus. kegiatan ini ..... nyaman.	Event	Campus
4	Fakultas Dakwah dan Komunikasi adakan Pelatihan Desain Grafis untuk Mahasiswa	Kompetisi ini akan memberikan kesempatan bagi mahasiswa untuk menunjukkan ..... kemam-puan desain grafis mereka.	Education	Education
5	UIN Walisongo Gelar Diskusi Etika dalam Kehidupan Sehari-hari	UIN Walisongo akan mengadakan diskusi tentang etika dan moralitas ..... kehidupan sehari-hari.	Campus	Acara
6	UIN Walisongo Gelar Webinar Internasional Tentang Pendidikan Islam di Era Digital	Universitas Islam Negeri (UIN) Walisongo Semarang menggelar webinar internasional ..... mahasiswa	Event	Event
7	Kantin FST UIN Walisongo Longsor Akibat Hujan Deras	Akibat hujan deras selama sehari penuh kantin FST longsor ..... perbaikan	Campus	Kampus
8	PTIPD Lakukan Workshop Pelatihan Jurnalistik	Dalam meningkatkan kualitas kepenulisan mahasiswa ..... jurnalistik	Education	Education
9	Mahasiswa Teknologi Informasi UIN Walisongo Raih Prestasi di Tingkat Internasional	Tiga mahasiswa Teknologi Informasi UIN Walisongo Semarang raih penghargaan ..... Malaysia	Campus	Campus

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

10	Perubahan Tanggal Pelaksanaan Wisuda Periode Juni 2024	Tanggal Pelaksaan Wisuda Periode Juni 20024 dirubah ..... mahasiswa	Announcement	Announcement
11	UIN Walisongo Buka Jalur Prestasi untuk 34 Jurusan, Ini Jadwal dan Syaratnya	Jalur Prestasi UIN Walisongo kembali dibuka untuk 34 jurusan ..... syarat kenentuan	Announcement	Announcement
12	Lantik Pejabat Struktural, Ini Pesan Rektor UIN Walisongo	Dalam rangka pelantikan pejabar struktural periode baru rektor berpesan ..... kualitas	Event	Event

The system was evaluated using twelve novel news documents. Results indicate that ten news stories were accurately categorized according to user assessments, as detailed in Table 10. The overall accuracy rate was 83.33%, with an error rate of 16.67%.

$$Accuracy = \frac{10}{12} 100\% = 83{,}33\%$$

Two misclassifications were identified in the test dataset. In the third news item, the system predicted “Event” while the user labeled it as “Campus”. This discrepancy likely occurred because the article included action-oriented phrases such as “mengadakan kegiatan” (holding an activity), which semantically align with event contexts despite referring to campus maintenance.

The fifth news item was also misclassified. The system labeled it as “Campus” whereas the user labeled it as “Event”. This error may have been caused by the system focusing on institutional terms like “diskusi tentang etika dan moralitas” (discussion on ethics and morality), which may have been interpreted as campus-related content rather than a scheduled event.

These findings indicate that the classification model could benefit from further refinement in distinguishing nuanced language, particularly when semantic overlap exists between categories such as campus activities and organized events.

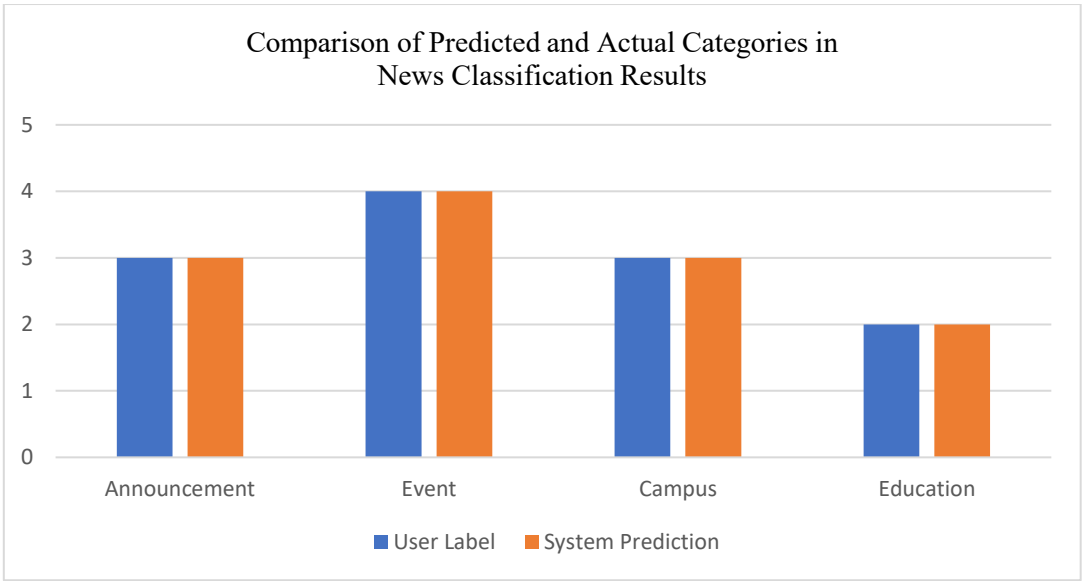

Figure 4. Comparison of Predicted and Actual Categories in News Classification Results

Figure 4 presents a bar chart comparing the predicted categories by the system and the actual categories labeled by users for the 12 test news documents. The distribution shows a strong alignment

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

between system predictions and user labels across all four categories. No significant overclassification or underclassification was observed, indicating balanced category recognition by the model.

Table 11. Evaluation Metrics

Category	Precision	Recall	F1-Score	Support
Event	0.75	0.75	0.75	4
Campus	0.67	0.67	0.67	4
Education	1.0	1.0	1.0	3
Announcement	1.0	1.0	1.0	2

As shown in Table 11, the evaluation metrics indicate that the system achieved perfect precision, recall, and F1-score for the "Announcement" and "Education" categories, reflecting that no misclassifications occurred in those groups. In contrast, the "Event" and "Campus" categories recorded lower F1-scores of 0.75 and 0.67 respectively, which is consistent with the two classification errors identified earlier. These results demonstrate the model's effectiveness in recognizing most academic news categories, while highlighting the need for further refinement in handling semantically overlapping categories such as campus-related activities and organized events.

4. DISCUSSIONS

The findings of this study demonstrate that the integration of the Enhanced Confix Stripping Stemmer, Chi-Squared feature selection, and Naïve Bayes Classifier achieves a robust level of performance in categorizing academic news documents from UIN Walisongo's website. The proposed model attained an accuracy of 83.33% on newly collected data, indicating the model's generalizability and adaptability in real-world academic contexts. Aspect-based sentiment analysis using SVM has been successfully implemented on e-commerce review datasets [22].

Compared to prior studies, the performance of this model is promising. For instance, Pramudita et al. [8] applied a similar approach for sports news classification and achieved an accuracy of 77%. The higher accuracy observed in this research suggests that improvements made to the stemming process and feature selection significantly enhance classification performance.

Ariadi and Fithriasari [9] compared Naïve Bayes and Support Vector Machine (SVM) for Indonesian news classification and reported that SVM outperformed Naïve Bayes, with SVM achieving 88.1% accuracy while Naïve Bayes achieved 82.2%. Despite this, Naïve Bayes remains a favorable choice due to its computational simplicity and interpretability, which align with the objectives of this study focusing on scalable academic categorization.

In addition, the study by Hidayat and Rizqi [10], which utilized the Enhanced Confix Stripping Stemmer and Naïve Bayes, obtained a 95% accuracy in a general news classification task. Although this is higher than the result in the current study, it is important to note that their dataset consisted of standard news content, whereas this study focused on academic news with specific linguistic patterns and domain terms.

Overall, the classification accuracy obtained confirms that the selected approach is well-suited for academic environments. The Enhanced Confix Stripping Stemmer effectively handled morphological complexities in the Indonesian language, while Chi-Squared feature selection allowed for the identification of discriminative terms across categories. Recent comparative studies highlight the advantage of combining FastText and Chi-Square for improving sentiment analysis models [23]. The Naïve Bayes Classifier, despite its simplicity, proved capable of maintaining competitive accuracy while offering efficient computational performance, making it an appropriate choice for institutional applications. Social media data poses significant challenges for text classification, especially during

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

events like the COVID-19 pandemic [24]. Hoax news detection leveraging Naïve Bayes and SVM continues to be an important research trend [25].

Despite the promising results, this study has several limitations. The classification model was trained on a relatively small dataset, which may limit its ability to generalize across a broader and more diverse range of academic news. In addition, although the probabilistic classification method used in this study is computationally efficient and easy to implement, it lacks the capability to understand contextual meaning within the text, especially when compared to more advanced deep learning-based classification models.

Future research could focus on expanding the dataset and evaluating the performance of the classification model in more dynamic environments. Another important direction would be to implement this system in real-time academic content platforms to enable automatic classification of news upon publication. Furthermore, integrating this system with more advanced language models that are capable of understanding word context and sentence structure could significantly enhance classification accuracy, particularly in academic texts that are longer and more complex.

5. CONCLUSION

To provide a clearer understanding of the outcomes of this research, the key conclusions have been structured into the following points. These conclusions are derived from the analysis and evaluation of the classification model applied to academic news documents from UIN Walisongo's website, highlighting the effectiveness of the techniques used and the contributions made to the field of academic text classification.

- The Naïve Bayes Classifier, integrated with the Enhanced Confix Stripping Stemmer and Chi-Squared feature selection, successfully categorized academic news documents on UIN Walisongo's website, achieving an accuracy of 83.33% and an error rate of 16.67%.
- The Enhanced Confix Stripping Stemmer improved the stemming process by addressing challenges in processing complex Indonesian word structures, contributing to the overall performance of the model.
- The Chi-Squared method effectively selected informative features, enhancing the model's ability to differentiate between news categories.
- These findings support the advancement of academic text classification, especially for Indonesian-language educational content, by offering a lightweight and interpretable model suitable for institutional use.
- Future research is encouraged to compare this model with other classification algorithms such as Support Vector Machine and neural networks, to explore improvements in contextual understanding and accuracy.
- Overall, this study contributes meaningfully to the field of Informatics, particularly in the development of intelligent information systems for academic environments.

REFERENCES

[1] M. D. Rizkiyanto, M. D. Purbolaksono, and W. Astuti, "Sentiment Analysis Classification on PLN Mobile Application Reviews Using Random Forest Method and TF-IDF Feature Extraction," *INTEK : Jurnal Penelitian*, vol. 11, no. 1, pp. 37–43, Apr. 2024, doi: https://doi.org/10.31963/intek.v11i1.4774.

[2] W. Purba, W. Siawin, M. N. K. Nababan, N. P. Dharshinni, and S. Aisyah, "Implementasi Data Mining untuk Pengelompokkan dan Prediksi Karyawan yang Berpotensi PHK dengan Algoritma K-Means Clustering," *JUSIKOM PRIMA : Jurnal Sistem Informasi Dan Ilmu Komputer Prima*, vol. 2, no. 2, pp. 85–90, 2019, doi: https://doi.org/10.34012/jusikom.v2i2.429.

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

[3] R. Qubra and R. A. Saputra, "Classification of Hoax News Using the Naïve Bayes Method," *IJSECS : International Journal Software Engineering and Computer Science*, vol. 4, no. 1, pp. 40–48, Apr. 2024, doi: https://doi.org/10.35870/ijsecs.v4i1.2068.

[4] N. Buslim, L. K. Oh, M. H. Athallah Hardy, and Y. Wijaya, "Comparative Analysis of KNN, Naïve Bayes and SVM Algorithms for Movie Genres Classification Based on Synopsis," *JTI : Jurnal Teknik Informatika*, vol. 15, no. 2, pp. 169–177, Dec. 2022, doi: https://doi.org/10.15408/jti.v15i2.29302.

[5] K. Wabang, Oky Dwi Nurhayati, and Farikhin, "Application of the Naïve Bayes Classifier Algorithm to Classify Community Complaints," *Jurnal RESTI : Rekayasa Sistem dan Teknologi Informasi*, vol. 6, no. 5, pp. 872–876, Nov. 2022, doi: https://doi.org/10.29207/resti.v6i5.4498.

[6] S. Widaningsih, "Perbandingan Metode Data Mining untuk Prediksi Nilai dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika dengan Algoritma C4.5, Naïve Bayes, KNN dan SVM," *Jurnal Tekno Insentif*, vol. 13, no. 1, pp. 16–25, 2019, doi: https://doi.org/10.36787/jti.v13i1.78.

[7] N. Nurdin, M. Suhendri, Y. Afrilia, and R. Rizal, "Klasifikasi Karya Ilmiah (Tugas Akhir) Mahasiswa Menggunakan Metode Naïve Bayes Classifier (NBC)," *SISTEMASI : Jurnal Sistem Informasi*, vol. 10, no. 2, pp. 268–279, 2021, doi: https://doi.org/10.32520/stmsi.v10i2.1193.

[8] A. M. Billah, D. A. R. Wulandari, and Y. A. Auliya, "Rancang Bangun Chatbot Pengaduan Kekerasan Perempuan Anak dengan Metode Fuzzy String Matching dan Enhanced Confix Stripping Stemmer," *INFROMAL : Informatics Journal*, vol. 8, no. 2, pp. 101–109, 2023, doi: https://doi.org/10.19184/isj.v8i2.42310.

[9] I. Sholekha, A. Faqih, and A. Bahtiar, "Sentiment Analysis of Public Opinion Covid-19 Vaccine Using Naïve Bayes and Random Forest Methods," *JTI : Jurnal Teknik Informatika*, vol. 15, no. 1, pp. 34–43, Jun. 2022, doi: https://doi.org/10.15408/jti.v15i1.24847.

[10] N. Agustina, E. Sutinah, and M. Martini, "Kolaborasi Metode Naïve Bayes dan MPE dalam Pengambilan Keputusan Pemilihan Supplier Ban Motor," *Jurnal Media Informatika Budidarma*, vol. 8, no. 2, pp. 1097–1108, 2024, doi: https://doi.org/10.30865/mib.v8i2.7538.

[11] Y. D. Pramudita, S. S. Putro, and N. Makhmud, "Klasifikasi Berita Olahraga Menggunakan Metode Naïve Bayes dengan Enhanced Confix Stripping Stemmer," *JTIK : Jurnal Teknologi Informasi Dan Ilmu Komputer*, vol. 5, no. 3, pp. 269–276, 2018, doi: https://doi.org/10.25126/jtiik.201853810.

[12] R. R. Sani, Y. A. Pratiwi, S. Winarno, E. D. Udayanti, and F. Alzami, "Analisis Perbandingan Algoritma Naïve Bayes Classifier dan Support Vector Machine untuk Klasifikasi Berita Hoax pada Berita Online Indonesia," *JMASIF : Jurnal Masyarakat Informatika*, vol. 13, no. 2, pp. 85–98, 2022, doi: https://doi.org/10.14710/jmasif.13.2.47983.

[13] E. Y. Hidayat and M. A. Rizqi, "Klasifikasi Dokumen Berita Menggunakan Algoritma Enhanced Confix Stripping Stemmer dan Naïve Bayes Classifier," *Jurnal TEKNOSI : Nasional Teknologi dan Sistem Informasi*, vol. 6, no. 2, pp. 90–99, 2020, doi: https://doi.org/10.25077/TEKNOSI.v6i2.2020.90-99.

[14] B. S. Prakoso, D. Rosiyadi, H. S. Utama, and D. Aridarma, "Klasifikasi Berita Menggunakan Algoritma Naïve Bayes Classifier dengan Seleksi Fitur dan Boosting," *Jurnal RESTI : Rekayasa Sistem dan Teknologi Informasi*, vol. 3, no. 2, pp. 227–232, 2019, doi: https://doi.org/10.29207/resti.v3i2.1042.

[15] N. Asmiati, "Penerapan Algoritma Naïve Bayes untuk Mengklasifikasi Pengaruh Negatif Game Online bagi Remaja Milenial," *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, vol. 2, no. 3, pp. 141–149, 2020, doi: http://dx.doi.org/10.35746/jtim.v2i3.102.

[16] E. Lindrawati, E. Utami, and A. Yaqin, "Comparison of Modified Nazief & Adriani and Modified Enhanced Confix Stripping Algorithms for Madurese Language Stemming," *INTENSIF : Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 7, no. 2, pp. 276–289, Aug. 2023, doi: https://doi.org/10.29407/intensif.v7i2.20103.

[17] J. M. Luna-Romera, M. Martínez-Ballesteros, J. García-Gutiérrez, and J. C. Riquelme, "External Clustering Validity Index Based on Chi-Squared Statistical Test," *Information Sciences*, vol. 487, pp. 1–17, 2019, doi: https://doi.org/10.1016/j.ins.2019.02.046.

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670

[18] D. Chen Sami, A. Sugiharto, and F. Jie, "Chi-Square Feature Selection for Improving Sentiment Analysis of News Data Privacy Threats," *JATIT : Journal of Theoretical and Applied Information Technology*, vol. 102, no. 18, 2024, [Online]. Available: www.jatit.org

[19] N. W. Wardani and P. G. S. C. Nugraha, "Stemming Teks Bahasa Bali dengan Algoritma Enhanced Confix Stripping," *IJNSE : International Journal of Natural Science and Engineering*, vol. 4, no. 3, pp. 103–113, Dec. 2020, doi: https://doi.org/10.23887/ijnse.v4i3.30309.

[20] N. Yusliani, S. A. Q. Aruda, M. D. Marieska, D. M. Saputra, and A. Abdiansah, "The Effect of Chi-Square Feature Selection on Question Classification Using Multinomial Naïve Bayes," *Sinkron : Jurnal dan Penelitian Teknik Informatika*, vol. 7, no. 4, pp. 2430–2436, Oct. 2022, doi: https://doi.org/10.33395/sinkron.v7i4.11788.

[21] A. Falasari and M. A. Muslim, "Optimize Naïve Bayes Classifier Using Chi-Square and Term Frequency Inverse Document Frequency for Amazon Review Sentiment Analysis," *JOSCEX : Journal of Soft Computing Exploration*, vol. 3, no. 1, pp. 31–36, Mar. 2022, doi: https://doi.org/10.52465/joscex.v3i1.68.

[22] D. A. Wulandari, F. A. Bachtiar, and I. Indriati, "Aspect-Based Sentiment Analysis on Shopee Application Reviews Using Support Vector Machine," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 15, no. 02, p. 99, Jan. 2025, doi: https://doi.org/10.24843/LKJITI.2024.v15.i02.p03.

[23] R. Fajriah and D. Kurniawan, "Optimalisasi Model Klasifikasi Naïve Bayes dan Support Vector Machine dengan FastText dan Chi-Square," *Faktor Exacta*, vol. 17, no. 4, pp. 1979–276, 2024, doi: 10.30998/faktorexacta.v17i4.24751.

[24] F. J. Damanik and D. B. Setyohadi, "Analysis of Public Sentiment About COVID-19 in Indonesia on Twitter Using Multinomial Naïve Bayes and Support Vector Machine," in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing Ltd, Apr. 2021. doi: 10.1088/1755-1315/704/1/012027.

[25] N. E. Febriyanty, M. A. Hariyadi, and C. Crysdian, "Hoax Detection News Using Naïve Bayes and Support Vector Machine Algorithm," *IJADIS : International Journal of Advances in Data and Information Systems*, vol. 4, no. 2, pp. 191–200, Oct. 2023, doi: https://doi.org/10.25008/ijadis.v4i2.1306.

Vol. 6, No. 3, Juni 2025, Page. 1279-1298
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4670