International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

Analysis of Content Consistency in Scientific Journal Based on Natural Language Processing and Machine Learning

Sitti Mawaddah Umar[1], Ingrid Nurtanio[1*], Zahir Zainuddin[1]

[1] Departement of Informatics, Hasanuddin University, Makassar, Indonesia

E-mail address: umarsm21d@student.unhas.ac.id, ingrid@unhas.ac.id, zahir@unhas.ac.id

Received ## Mon. 20##, Revised ## Mon. 20##, Accepted ## Mon. 20##, Published ## Mon. 20##

Abstract: This groundbreaking research proposes implementing a state-of-the-art monitoring system designed to evaluate the structural cohesion of scientific journal manuscripts and generate accurate sentence interpretations. By utilizing Natural Language Processing (NLP) and the power of Machine Learning (ML), this research aims to answer how authors can know their consistency in writing papers by self-detecting. This system provides accurate comparative analysis results and sentence interpretation ratios to measure author consistency in writing journal manuscripts or papers. By using natural language processing (NLP) as a pre-processing stage and applying the Term Frequency-Inverse Document Frequency (TF-IDF) as a determinant for vectorization by dividing two vectors, this study uses Support Vector Machine (SVM) for predictive classification in machine learning. It uses Cosine Similarity (CS) to distinguish the similarity of sentences. The results were staggering: the study achieved an 83.94% accuracy rate for relevance consistency in content comparison analysis, supported by the activation of 2485 journal datasets, with a yield of 0.740402 obtained from convolution optimization. This remarkable achievement has the potential to revolutionize academia by improving the efficiency and quality of writing, providing easy-to-understand information for novice researchers trying to write scientific papers. By implementing this monitoring system, researchers can ensure that their manuscripts are structurally cohesive and consistent and can enjoy the benefits of a more efficient and smooth writing process, resulting in better quality research and more impactful publications.

Keywords: Journal Consistency, Natural Language Processing (NLP), Support vector Machine (SVM), TF-IDF

1. INTRODUCTION

One important aspect of academic requirements is writing scientific journals [1]. It requires the ability to write and independent analytical skills[2]. Peer review is carried out to assess the consistency of authors, which is meticulously analyzed and compared based on available journal content. A system is built to generate comparative analysis with percentage values, which is presented with sentence meaning to provide accurate information. Machine Learning (ML) is used to perform classification training to determine the features to be classified based on the dataset[3].

Keywords are identified as a structural framework for locating texts with resemblances[4]. The parameters used in determining this consistency analysis are the content in scientific journal manuscripts: Title, Abstract, Introduction, and Conclusion in manuscripts for submission to the publication service provider. This built system can also be used for journals or proceedings if readers need information about sentence meaning and the consistency of authors in writing scientific journal manuscripts. The documents used include scientific journal documents in English using the Institute of Electrical and Electronics Engineers (IEEE) and Association for Computing Machinery (ACM) writing formats.

In the field of Natural Language Processing (NLP), various studies have been conducted to assess the consistency of newspapers and analyze sentence coherence similarity in Japanese text with the help of techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and semantic similarity measures. These methods have achieved an accuracy of 68.02% and 74.82% respectively. Other related writings propose a new approach to evaluating text similarity in a short text mining framework designed explicitly to evaluate the maintenance of electric inventory lists [6]. The main goal

E-mail:author's email

is to develop a text processing framework that can optimize accuracy to 77.23%.

The use of machine learning methods for analyzing thematic consistency in articles has been conducted by researchers in the past [7]. They focused on a system model that could identify topic continuity from 50 documents and successfully achieved a 54.98% continuity of different topics and fields. Exploring the application of machine learning techniques in classifying scientific articles' structure and writing style has also been done [8]. This was to differentiate writing patterns with varied styles. Additionally, researchers have used algorithms to optimize text classification by adding Support Vector Machine (SVM) to classify the unique language of articles on a large scale [9]. Based on this, it is proven that consistency is essential in creating a system, just as consistency in writing is necessary to represent and convey information effectively in a scientific journal manuscript [10]. Natural Language Processing (NLP) can also generate abstractive summaries based on research in 2020 with an accuracy rate of 80.46%.

The attempt to construct a system employing natural language processing to analyze the consistency of authors in academic writing. We propose an innovative approach to evaluate sentence similarity and textual meaning between the contents of journal manuscripts. By using NLP to process natural language that readers more easily understand through preprocessing[11], followed by the use of vectorization with two vectors using TF-IDF for weighting the text, the SVM algorithm is employed for machine learning classification using parameter context from an integral part of the model, influencing the model's performance [12].

Advancements in science and technology have had a significant impact on various aspects of human life, including academic writing. Writing scientific journal manuscripts and other academic documents requires independent writing and analytical skills, with consistency being a crucial assessment observed in the peer review process. To address this challenge, researchers have developed a system that utilizes Natural Language Processing (NLP) and Machine Learning (ML) techniques to automatically analyze and evaluate the consistency of scientific journal manuscript writing. The Cosine Similarity algorithm is one such technique that can determine the similarity in sentences composed of text to produce weighted sentence meaning. These techniques have been implemented on a comprehensive academic text dataset, and the findings suggest that integrating these algorithmic methods dramatically improves the analysis of author consistency and enhances the comprehension of sentences in academic texts.

The analysis is conducted by building a system that utilizes Machine Learning (ML) classification techniques and Cosine Similarity algorithms to identify similarity within sentences found in journal manuscript, such as title, abstract, introduction, and conclusion. This method provides a concrete solution to support authors and journal editors in ensuring the consistency and quality of submitted manuscripts. Additionally, this system can be widely applied to various types of journals and proceedings, thereby enhancing the overall quality of scientific publications.

This research indicates that integrating these algorithmic methods dramatically improves the analysis of writing consistency and enriches the understanding of sentences in academic texts. The contribution of this research is significant in understanding and enhancing the consistency of scientific journal manuscript writing and can provide valuable guidance for researchers and practitioners in improving the quality of academic writing.

2. RELATED WORK

They conducted a study using TF-IDF and TextRank algorithms to extract keywords from English news texts. The research evaluated the algorithm's performance using the Sina News Corpus and reported average macro values. The study showed that the combination of TF-IDF and TextRank algorithms performed comparably to conventional algorithms regarding performance parameters and extraction impact.

It provided a detailed explanation of the algorithms and application in keyword extraction from news articles. The proposed algorithm integrated TextRank and TF-IDF while enhancing the weight on titles, thus improving the effectiveness of keyword extraction, which is limited to a maximum of 800 words[13].

Clustered Indonesian articles from KNS&I 2017 using word frequency and cosine similarity, achieving an accuracy of 69.8%. The system successfully grouped articles based on computer science and information system knowledge, with the most significant research publication in the "information management" cluster[14].

Highlighted that while natural language processing can aid in identifying ambiguities and incompleteness in software requirements, manual inspection remains necessary for ensuring accuracy and completeness. Combining manual inspection and natural language processing tools is the most effective approach to identifying nearly synonymous terms that can cause terminological ambiguity[15].

Note that the dataset used could have represented the overall population adequately, and measurement errors in data collection or measured variables could have affected the accuracy of the results. Subjectivity in evaluating keyword extraction method performance could lead to different interpretations[16]. They highlighted the simplicity and efficiency of cosine similarity in calculating similarity, supporting its application in considerable data classification efforts for economic journals and providing significant benefits in managing and organizing academic literature. Combining web scraping and text processing with cosine similarity demonstrates a more robust and accurate system for identifying and presenting relevant images. The study proved that cosine similarity can be compelling in text and image search relevance[17].

It is suggested that combining two standard text representation methods, Count Vectorizer and TF-IDF, provides a more comprehensive understanding of the text[18]. they discussed the measurement of sentence similarity for Bengali text summarization, exploring various methods and proposing a method using data collection, pre-processing, summarization maker, and applying similarity measurement methods like cosine similarity, achieving an accurate text similarity rate of up to 71% for detecting the meaning of Bengali language similarity [19].

It introduced semantic search, involving understanding the meaning behind user queries and not just matching keywords literally. NLP allows the model to understand sentence structure, entities, or topics in a text document using vector representation, word embedding, or other techniques to understand better the true meaning of words or phrases [20].

They utilized text mining to automate the identification of relevant information from claimed texts. Text mining demonstrated the ability to understand natural language through entity recognition, text classification, and information extraction, achieving a significant accuracy of 67.98% using NLP [21].

Indicated that NLP, used for understanding and analyzing essays, provides deeper semantic analysis and understanding using vector representation techniques. The model evaluating specific essay aspects accurately achieved 74.68% of the score. They discussed the challenges in understanding text messages in informal language or abbreviations, adding uncertainty and difficulty for systems due to ambiguous meanings, language variations, or handling vocabulary variations[22].

3. METHODOLOGY

The research commenced with compiling a specialized dataset contains 2.485 scientific journal documents within the informatics domain, which were indexed in Scopus and utilized in English. The dataset was divided into 2.237 training data and 249 testing data. This specific dataset was gathered from various journal publishing institutions and aimed at assessing the relevance of the content provided based on title, abstract, introduction, and conclusion. Furthermore, the system to be developed in this research aims to predict the research scope from the input journal topics. Additionally, the system is equipped with sentence meaning interpretation, providing simple conclusions from the journals under study.

Subsequently, the initial stage of the process involved pre-processing, which encompassed several stages, such as text cleaning, which is essential for eliminating non-letter, non-numeric, and non-space characters from the data. This was followed by case folding, aligning diverse letter cases to lowercase. Tokenization was then performed, wherein the data from the journal documents, still in sentence form, were segmented into individual words to facilitate easier filtering and conversion of each word into its root form, along with vectorization to enable the processing of words in machine learning algorithms.

The subsequent stage involved the implementation of class classification utilizing the support vector machine (SVM) method, integrated with Natural Language Processing (NLP). This machine learning model divided into two classes by determining the optimal hyperplane that separates the classes so that the margin (distance) between the hyperplane and the nearest points from each class (referred to as support vectors) is maximized. SVM is also employed for binary classification (two classes) but can be extended to multi-class classification. In the context of the constructed NLP, this classification task can classify documents into relevant categories or topics, facilitating content management systems or information retrieval while providing consistent similarity results and simple sentence interpretation.

Constructed a framework to analyze the consistency of scientific journal manuscripts by classifying the title, abstract, introduction, and conclusion of the paper content. This text mining system aims to uncover author consistency to create a writing style, grammar, and writing structure that aligns with IEEE and ACM standards, using a large-scale corpus and assistance from the Arxiv dataset to make the dataset comprehensive. The corpus is beneficial for processing data from the database and journal manuscript documents. This process is used as a stage in building this research system. The study also focuses on improving accuracy and consistency analysis and how to accurately and efficiently understand sentence meaning.

Calculations can be applied to obtain weights that correspond to text with substantial information skillfully, and these percentage results will also yield easily understandable sentence meaning. It starts with a dataset from scientific journals in related fields. NLP assists with the preprocessing stage in processing text with several phases: preprocessing, TF-IDF feature extraction, Corpus, SVM, and Cosine similarity. In this research, the text mining will involve a classification process to determine author consistency and sentence meaning with better accuracy. The steps in this experiment are shown in Figure. 1

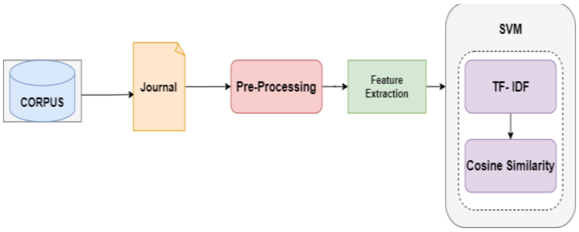


Figure 1. Blok diagram from the NLP methodology.

In this research, the prediction system and analysis of consistency similarity in the content of journal writing are combined, from which a simple conclusion is drawn from the journal documents used as objects. The scenario includes the Title, Abstract, Introduction, and Conclusion of the journal documents, which have been collected (downloaded) from several international journal publishing institutions indexed by Scopus. This approach aims to evaluate the system's ability to provide information with representative and responsible results easily understood by readers, enabling early and independent identification of the paper being prepared by the authors. By conducting this scenario, the effectiveness of the system in processing and providing accurate information in predicting papers can be assessed.

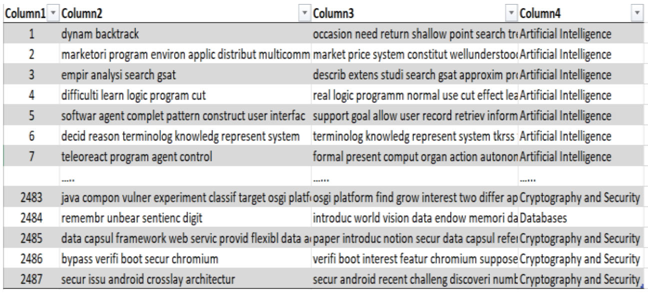
This research scenario will involve the integration of several main methods: Natural Language Processing (NLP), used to process natural language, followed by the Term Frequency-Inverse Document Frequency (TF-IDF), which aims to assess the importance of a word in a document relative to a larger collection of documents. Support Vector Machine (SVM) is used for class classification, as described previously. Lastly, Cosine Similarity (CS) measures two vectors' similarity in multidimensional space. It is also used to compare the similarity between text documents, sentences, or words based on vector representations.

Thus, the result will be in the form of percentage values from the analysis of the consistency of scientific journal content. A prediction of the informatics scope will be generated by entering the title of the journal, and, most importantly, a simple sentence interpretation will be produced to provide accurate information. This research aims to assess how well the system can overcome its challenges by testing various scientific journals and linguistic variations. This comprehensive evaluation is crucial to identify the strengths and weaknesses of the system to guide further development and improvement.

A. *The preprocessing stage*

1) *Corpus*

A corpus gathers text or language data for linguistic analysis or language technology development. Novel corpus construction approaches aim to maximize targeted resources' coverage by collecting fragments from various sources. In this study, the corpus specifically focuses on English-language journal documents, particularly those related to informatics. The corpus comprises 2485 collected and organized data to create a diverse and comprehensive database. Following the corpus construction process, the next step is the tokenization stage. In the preprocessing stage, involving text cleaning, case folding, stopword removal, and stemming, the text in the corpus will be divided into small units called tokens, such as words, symbols, or other language elements. This process is crucial as it enables the system to quickly comprehend and process the natural language used in journal documents. Tokenization is crucial in preparing natural language data for further Deep Learning processes. Further details regarding the composition and structure of the corpus used will be presented. More details about the composition and structure of the corpus used are shown in Figure 2.



Column1	Column2	Column3	Column4
1	dynam backtrack	occasion need return shallow point search tr	Artificial Intelligence
2	marketori program environ applic distribut multicomm	market price system constitut wellunderstoo	Artificial Intelligence
3	empir analysi search gsat	describ extens studi search gsat approxim pr	Artificial Intelligence
4	difficulti learn logic program cut	real logic programm normal use cut effect lea	Artificial Intelligence
5	softwar agent complet pattern construct user interfac	support goal allow user record retriev inform	Artificial Intelligence
6	decid reason terminolog knowledg represent system	terminolog knowledg represent system tkrss	Artificial Intelligence
7	teleoreact program agent control	formal present comput organ action autonon	Artificial Intelligence
	.....	......	......
2483	java compon vulner experiment classif target osgi platf	osgi platform find grow interest two differ ap	Cryptography and Security
2484	remembr unbear sentienc digit	introduc world vision data endow memori da	Databases
2485	data capsul framework web servic provid flexibl data a	paper introduc notion secur data capsul refe	Cryptography and Security
2486	bypass verifi boot secur chromium	verifi boot interest featur chromium suppose	Cryptography and Security
2487	secur issu android crosslay architectur	secur android recent challeng discoveri numb	Cryptography and Security

Figure 2. the composition and structure of the corpus used

At this stage shows an example dataset with 2487 rows and three column types to be processed. The first column indicates a set of journal titles or papers such as "Dynamic Backtracking," "Market-oriented Programming Environments for the Distributed Multicommodity Flow Problem," "Pattern Matching for Discourse Processing in Information Extraction from Japanese Text," "Solving Multiclass Learning Problems via error-correcting output codes." The second column contains summaries of journals or papers, and the third column contains the primary category, which covers the field of expertise or the scope of informatics. The system will search for the data or words above before preprocessing, and predictions will be made from the results that generate information. From the result, it is crucial to perform text preprocessing.

2) *Data Pre-Processing*

The preprocessing process is a stage that breaks down sentences or documents into pieces of words [23]. This stage plays a crucial role in Natural Language Processing (NLP), transforming text data into manageable units for further analysis. This preprocessing stage aims to obtain the root words by breaking them down into words or subwords. This process facilitates algorithms better to understand the structure and meaning of the text, resulting in more efficient and accurate NLP models. It starts with cleaning text, where data from scientific journal documents in characters other than letters, numbers, and spaces will be cleaned. Then comes the case folding stage, where data from scientific journal documents generally have varied letter sizes and are standardized into lowercase entirely. Next is the tokenization stage, where data from scientific journal documents that are still in the form of sentences will be broken down into individual words or pieces per word, making it easier to filter and convert each word into its root form, and can be vectorized so that words can be processed in machine learning algorithms.

The subsequent step involves the removal of stopwords or filtering the dataset that has been broken down into words to eliminate less useful or meaningful words, such as “right”, “in”, “which”, “with” and similar ones, ensuring that only significant words undergo processing in machine learning. The final stage is stemming, in which every word in each data row is reverted to its root form. Since the dataset processed in this study employs the Indonesian language, affixes are essentially stripped from each word. For instance, "eating" would be transformed into "eat", "hitting", "hit", "learning", into "learn", and so forth.

B. Term Frequency – Inverse Document Frequency

Term weighting is a process conducted to determine the weight of each word in a text. A term or word in a document is assigned a value based on how often the word appears in the text, known as term frequency. The higher the term frequency of a word in a document, the higher its weight, indicating a higher level of relevance. In terms of weighting, it is also essential to consider the term scarcity factor in the document. Words that rarely appear in several documents are considered more critical (uncommon terms) than words that frequently appear in many documents. Term weighting also considers the inverse document frequency factor, which considers the inverse frequency of documents containing that particular word [13].

The TF-IDF (Term Frequency-Inverse Document Frequency) method is a term weighting technique commonly employed as a benchmark for comparison with newer term weighting methods. In this method, the weight of a term (word) in a document can be calculated by multiplying the Term Frequency value with the Inverse Document Frequency. Several formula variations can be applied to Term Frequency (TF), which can be outlined as follows equation (2) measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

- Binary TF (Term Frequency): Focuses on the presence of a word in a document, assigning a value of one if the word is present and zero if not.
- Raw TF (Term Frequency): Assigns the TF value based on the number of word occurrences in a document. For example, if a word appears five times, the TF value will be five.
- Logarithmic TF: Avoids the dominance of documents with few words in the query but high frequency by using a logarithmic scale

$$tf = 1 + \log(\ f_{t,d}\ ) \tag{1}$$

- TF normalization: Applies a comparison between the frequency of a word and the total number of words in a document.

$$tf = 0.5 + 0.5 \times \left(\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}} \right) \tag{2}$$

While Inverse Document Frequency (IDF) is calculated using the formula.

$$idf_j = \log\left(\frac{D}{df_j}\right) \tag{3}$$

where, D = The number of all documents in the collection, df_j = number of documents containing the term $t_{j.}$ The general formula used in TF-IDF is a combination of formulas between raw TF and IDF by multiplying the term frequency (TF) value with the inverse document frequency (IDF) value. Inverse document frequency (IDF) value:

$$w_{ij} = tf_{ij} \times idf_i \tag{4}$$

$$w_{ij} = tf_{ij} \times \log\left(\frac{D}{df_j}\right) \tag{5}$$

Where, W_{ij} = weight of term t_j against document $d_{i,}$ tf_{ij} = number of occurrences of term term t_j in document $d_{i.}$ D = number of all documents in the databased df_j = number of documents containing the term t_j (at least one word term t_j)

Displays the TF values for each term or word, where TF reflects the frequency of occurrence of each term or word in a document. For example, the word "system" appears once in document D1, once in document D2, three times in document D3, and twice in document 4. The TF calculation is carried out similarly for other words in the document. The next stage involves calculating the IDF values.

TABLE II. TF-IDF calculation result

Term/word	TF						IDF	Bobot (W) = TF*IDF					
	D1	D2	D3	D4	D5	D6		D1	D2	D3	D4	D5	D6
System	1	1	3	2	0	0	$\log\left(\frac{6}{4}\right)$=0.176	0.176	0.176	0.528	0.352	0	0
Analysis	0	1	1	0	1	0	$\log\left(\frac{6}{3}\right)$=0.301	0	0.301	0.301	0	0.301	0
Document	0	0	1	0	0	0	$\log\left(\frac{6}{1}\right)$=0.778	0	0	0.778	0	0	0
similarity	0	0	1	0	0	0	$\log\left(\frac{6}{1}\right)$=0.778	0	0	0.778	0	0	0
Frequency	0	0	1	0	0	0	$\log\left(\frac{6}{1}\right)$=0.778	0	0	0.778	0	0	0
vectorize	0	0	1	0	0	0	$\log\left(\frac{6}{1}\right)$=0.778	0	0	0.778	0	0	0
consistency	0	0	1	0	1	0	$\log\left(\frac{6}{2}\right)$=0.477	0	0	0.778	0	0.778	0

In calculating IDF, it is necessary to determine the values of D and DF, where D is the total number of documents in the dataset. At the same time, DF is the number of documents containing the term (t). As an example, Table II, for the word "system," it is known that its DF value is 3, indicating that the word appears in document 3. Thus, the IDF value for the word "system" is 0.301.

Furthermore, Table II, shows that the word "Document" in document D3 has a TF value of 1 and an IDF value of 0.778. Thus, the calculated weight or TF-IDF value for that word in D3 is 0.778. Once the TF and IDF values are obtained, the next step is to calculate the TF-IDF value or the weight value by multiplying the TF value with the IDF value.

C. *Support Vector Machine (SVM)*

Support Vector Machine (SVM) is a machine learning algorithm for classification and regression tasks. This algorithm belongs to the supervised learning category, requiring labeled data for training. SVM is known for its function to separate multiclass in feature space by finding the best hyperplane and maximizing the margin between them. Here are some key concepts related to SVM [20].

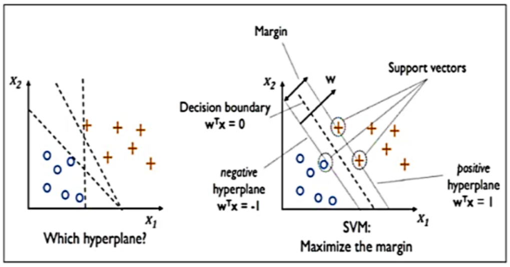


Figure 3. *The hyperplane separates two classes, positive (+1) and negative (-1)*

The hyperplane SVM detects as shown in Fig. 3, Its position is between the two classes. This means that the distance of the hyperplane from the data points differs from the nearest (outer) class, marked by an empty positive circle. The outermost data points closest to the hyperplane in SVM are called support vectors. These support vectors are the most challenging to classify because positions almost overlap with the classes of other objects. Due to crucial nature, SVM only considers these support vectors when searching for the optimal hyperplane.

With the following conditions, a hyperplane with a larger margin distance will be more accurate in classifying data groups than a hyperplane with a more significant margin distance. Classifying data groups than a hyperplane with a smaller margin distance. Distance. In the training stage, Support Vector Machine (SVM) usually looks for the hyperplane with the most significant margin distance, namely the Maximum Marginal Hyperplane (MMH). The hyperplane formula can be written as the equation below:

$$w \,.\, x_i + b = 0 \qquad (6)$$

where, w: weight vector, $x_{i:}$ i – th data, b: bias value. If the b value is considered as an additional weight w_0 (6) can be written back like (7)

$$w_0 + w.x_i = 0 \qquad (7)$$

If every point lies above the hyperplane, (6) and (7) is used. The SVM method divides the dataset into two groups, formulated as equation, where y_i represents the class of the i-th data. The maximum margin can be measured by maximizing the distance value between the hyperplane and its closest point, i.e. $\frac{1}{\|w\|}$. Based on the above assumption, the hyperplane perfectly separates the two classes. However, this phenomenon only sometimes occurs in situations where the hyperplane cannot perfectly separate the classes. The inability to perfectly separate the two classes leads to non-compliance with the constraints imposed by the mathematical formulation of SVM, as described in (6). This condition necessitates additional optimization efforts to address the issue.

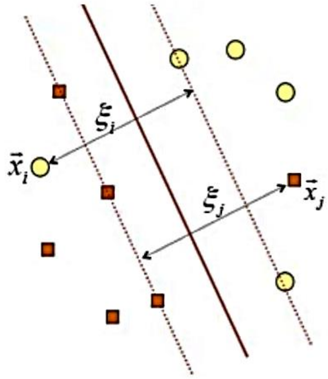


Figure 4. *Modified using soft margin by inserting a slack variable εi(εi>0)*

Generally, it is impossible to separate linearly in the input space in data problems. Soft margin Support Vector Machine (SVM) cannot find a separator in the hyperplane, so high accuracy and good generalization cannot be achieved. Therefore, a kernel is needed to transform the data into a higher-dimensional space called the kernel space, which helps separate data linearly. Generally, frequently used kernel functions include Linear, Polynomial, and Radial Basis Function (RBF) kernels.

The culminating stage involves the application of the Cosine Similarity metric, which meticulously gauges the likeness of text between pairs of documents or passages. An elevated cosine similarity index between two documents manifests substantial resemblance, signifying inherent similarity or relevance. The seamless integration of these sophisticated algorithms facilitates the extraction of granular information. It augments the depth and precision of insights pertinent to the nuanced categorization of topics and the coherent grouping of documents within the academic milieu

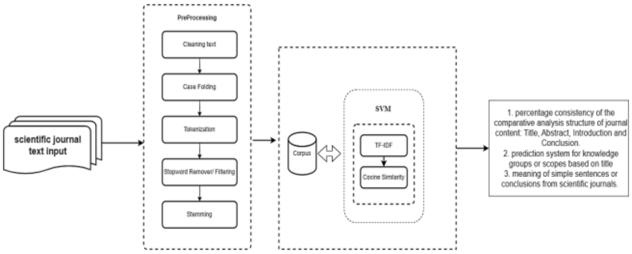

PreProcessing
Cleaning text
Case Folding
Tokenization
Stopword Remover/ Filtering
Stemming
scientific journal text input
Corpus
SVM
TF-IDF
Cosine Similarity
1. percentage consistency of the comparative analysis structure of journal content: Title, Abstract, Introduction and Conclusion.
2. prediction system for knowledge groups or scopes based on title
3. meaning of simple sentences or conclusions from scientific journals.

Figure 5. Diagram of the information system built

In this research, by using NLP and SVM approach is crucial in understanding language and achieving better accuracy results. This study is a model of an information system used for students who are continuing studies and need assistance understanding more complex scientific journals.

D. Cosine Similarity

In the case of analyzing the consistency of scientific journal content, combining Natural Language Processing (NLP) and Cosine Similarity is a practical approach. NLP is utilized to parse the text from research titles and simple sentence meanings. At the same time, Cosine Similarity is employed to measure the extent to which the content of the scientific journal is consistent with the field of informatics. Initially, through Natural Language Processing (NLP), the text from research titles and simple sentences in the scientific journal is broken down into vector representations, enabling further analysis. Subsequently, the concept of cosine similarity is utilized to compare the vector representations of the research title with simple sentences from the scientific journal.

Cosine Similarity (CS) yields a score that measures how similar or consistent the content of the research title is with the content of the scientific journal within the field of informatics. The higher the score, the more consistent the content is with the predetermined field of informatics. The results can be presented as a consistency percentage, indicating how well the scientific journal content aligns with the targeted field of informatics.

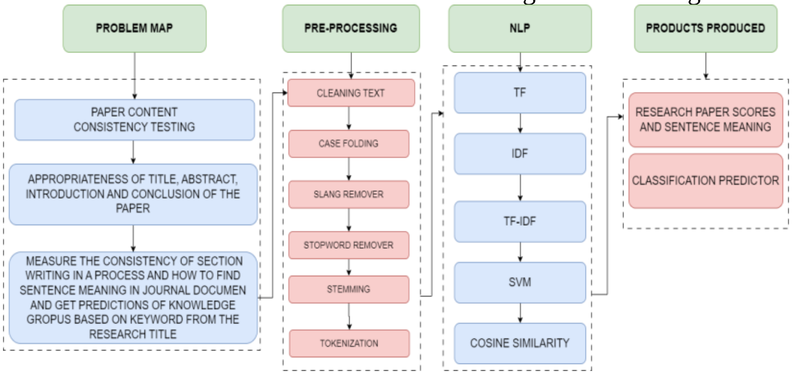

PROBLEM MAP
PRE-PROCESSING
NLP
PRODUCTS PRODUCED
PAPER CONTENT CONSISTENCY TESTING
APPROPRIATENESS OF TITLE, ABSTRACT, INTRODUCTION AND CONCLUSION OF THE PAPER
MEASURE THE CONSISTENCY OF SECTION WRITING IN A PROCESS AND HOW TO FIND SENTENCE MEANING IN JOURNAL DOCUMEN AND GET PREDICTIONS OF KNOWLEDGE GROPUS BASED ON KEYWORD FROM THE RESEARCH TITLE
CLEANING TEXT
CASE FOLDING
SLANG REMOVER
STOPWORD REMOVER
STEMMING
TOKENIZATION
TF
IDF
TF-IDF
SVM
COSINE SIMILARITY
RESEARCH PAPER SCORES AND SENTENCE MEANING
CLASSIFICATION PREDICTOR

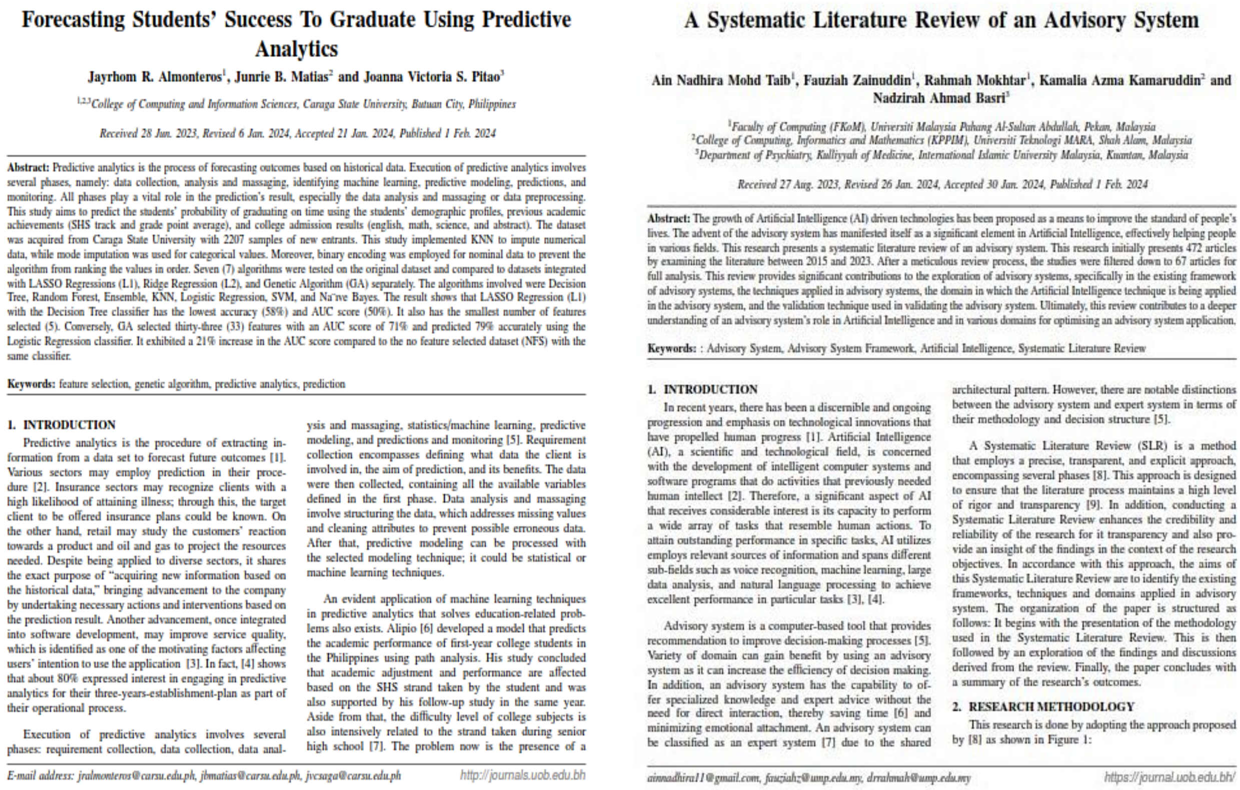
Figure 6. Design framework of thought system of Text Similarity

Fig. 6 depicts the road mapping of our research. The researcher proposed mapping issues as follows: in the initial column, the researcher outlined the mapping of issues to assess content consistency in the research paper. This analysis the coherence of content in the research paper's title, abstract, introduction, and conclusion. The issue addressed in this research is how to measure the consistency between the content of the research paper and the existing journal content in the research paper and draw conclusions about the relevance of content in the research paper[23]. The mapping of issues explains that the challenge lies in identifying consistency between the words used in the title, abstract, introduction, and conclusion. In the final mapping a statement aims to add an information system on how a system provides information about a field of study by entering the title or keywords from the related research title to be investigated to make it more relevant.

4. RESULT AND DISCUSSION

In the Results and Discussion section, we examined 2485 journal samples out of 2237 samples ready to be submitted to the publication service provider. This action was taken to evaluate the extent of similarity and consistency among the authors in composing manuscripts. and to gain a comprehensive understanding of the paper's content.

This challenge often hinders efforts to ensure that all parts of academic work are interconnected and represent a consistent understanding of the topic. The system can identify and analyze the consistency of authors in composing papers using techniques such as Vectorizer and Cosine Similarity. As a result, the system optimizes checking the alignment between the paper's title, abstract, introduction, and conclusion.



Forecasting Students' Success To Graduate Using Predictive Analytics

Jayrhom R. Almonteros[1], Junrie B. Matias[2] and Joanna Victoria S. Pitao[3]

[1,2,3]*College of Computing and Information Sciences, Caraga State University, Butuan City, Philippines*

Received 28 Jun. 2023, Revised 6 Jan. 2024, Accepted 21 Jan. 2024, Published 1 Feb. 2024

Abstract: Predictive analytics is the process of forecasting outcomes based on historical data. Execution of predictive analytics involves several phases, namely: data collection, analysis and massaging, identifying machine learning, predictive modeling, predictions, and monitoring. All phases play a vital role in the prediction's result, especially the data analysis and massaging or data preprocessing. This study aims to predict the students' probability of graduating on time using the students' demographic profiles, previous academic achievements (SHS track and grade point average), and college admission results (english, math, science, and abstract). The dataset was acquired from Caraga State University with 2207 samples of new entrants. This study implemented KNN to impute numerical data, while mode imputation was used for categorical values. Moreover, binary encoding was employed for nominal data to prevent the algorithm from ranking the values in order. Seven (7) algorithms were tested on the original dataset and compared to datasets integrated with LASSO Regressions (L1), Ridge Regression (L2), and Genetic Algorithm (GA) separately. The algorithms involved were Decision Tree, Random Forest, Ensemble, KNN, Logistic Regression, SVM, and Na¨ıve Bayes. The result shows that LASSO Regression (L1) with the Decision Tree classifier has the lowest accuracy (58%) and AUC score (50%). It also has the smallest number of features selected (5). Conversely, GA selected thirty-three (33) features with an AUC score of 71% and predicted 79% accurately using the Logistic Regression classifier. It exhibited a 21% increase in the AUC score compared to the no feature selected dataset (NFS) with the same classifier.

Keywords: feature selection, genetic algorithm, predictive analytics, prediction

1. INTRODUCTION

Predictive analytics is the procedure of extracting information from a data set to forecast future outcomes [1]. Various sectors may employ prediction in their procedure [2]. Insurance sectors may recognize clients with a high likelihood of attaining illness; through this, the target client to be offered insurance plans could be known. On the other hand, retail may study the customers' reaction towards a product and oil and gas to project the resources needed. Despite being applied to diverse sectors, it shares the exact purpose of "acquiring new information based on the historical data," bringing advancement to the company by undertaking necessary actions and interventions based on the prediction result. Another advancement, once integrated into software development, may improve service quality, which is identified as one of the motivating factors affecting users' intention to use the application [3]. In fact, [4] shows that about 80% expressed interest in engaging in predictive analytics for their three-years-establishment-plan as part of their operational process.

Execution of predictive analytics involves several phases: requirement collection, data collection, data analysis and massaging, statistics/machine learning, predictive modeling, and predictions and monitoring [5]. Requirement collection encompasses defining what data the client is involved in, the aim of prediction, and its benefits. The data were then collected, containing all the available variables defined in the first phase. Data analysis and massaging involve structuring the data, which addresses missing values and cleaning attributes to prevent possible erroneous data. After that, predictive modeling can be processed with the selected modeling technique; it could be statistical or machine learning techniques.

An evident application of machine learning techniques in predictive analytics that solves education-related problems also exists. Alipio [6] developed a model that predicts the academic performance of first-year college students in the Philippines using path analysis. His study concluded that academic adjustment and performance are affected based on the SHS strand taken by the student and was also supported by his follow-up study in the same year. Aside from that, the difficulty level of college subjects is also intensively related to the strand taken during senior high school [7]. The problem now is the presence of a

E-mail address: jralmonteros@carsu.edu.ph, jbmatias@carsu.edu.ph, jvcsaga@carsu.edu.ph http://journals.uob.edu.bh

A Systematic Literature Review of an Advisory System

Ain Nadhira Mohd Taib[1], Fauziah Zainuddin[1], Rahmah Mokhtar[1], Kamalia Azma Kamaruddin[2] and Nadzirah Ahmad Basri[3]

[1]*Faculty of Computing (FKoM), Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Malaysia*
[2]*College of Computing, Informatics and Mathematics (KPPIM), Universiti Teknologi MARA, Shah Alam, Malaysia*
[3]*Department of Psychiatry, Kulliyyah of Medicine, International Islamic University Malaysia, Kuantan, Malaysia*

Received 27 Aug. 2023, Revised 26 Jan. 2024, Accepted 30 Jan. 2024, Published 1 Feb. 2024

Abstract: The growth of Artificial Intelligence (AI) driven technologies has been proposed as a means to improve the standard of people's lives. The advent of the advisory system has manifested itself as a significant element in Artificial Intelligence, effectively helping people in various fields. This research presents a systematic literature review of an advisory system. This research initially presents 472 articles by examining the literature between 2015 and 2023. After a meticulous review process, the studies were filtered down to 67 articles for full analysis. This review provides significant contributions to the exploration of advisory systems, specifically in the existing framework of advisory systems, the techniques applied in advisory systems, the domain in which the Artificial Intelligence technique is being applied in the advisory system, and the validation technique used in validating the advisory system. Ultimately, this review contributes to a deeper understanding of an advisory system's role in Artificial Intelligence and in various domains for optimising an advisory system application.

Keywords: : Advisory System, Advisory System Framework, Artificial Intelligence, Systematic Literature Review

1. INTRODUCTION

In recent years, there has been a discernible and ongoing progression and emphasis on technological innovations that have propelled human progress [1]. Artificial Intelligence (AI), a scientific and technological field, is concerned with the development of intelligent computer systems and software programs that do activities that previously needed human intellect [2]. Therefore, a significant aspect of AI that receives considerable interest is its capacity to perform a wide array of tasks that resemble human actions. To attain outstanding performance in specific tasks, AI utilizes employs relevant sources of information and spans different sub-fields such as voice recognition, machine learning, large data analysis, and natural language processing to achieve excellent performance in particular tasks [3], [4].

Advisory system is a computer-based tool that provides recommendation to improve decision-making processes [5]. Variety of domain can gain benefit by using an advisory system as it can increase the efficiency of decision making. In addition, an advisory system has the capability to offer specialized knowledge and expert advice without the need for direct interaction, thereby saving time [6] and minimizing emotional attachment. An advisory system can be classified as an expert system [7] due to the shared architectural pattern. However, there are notable distinctions between the advisory system and expert system in terms of their methodology and decision structure [5].

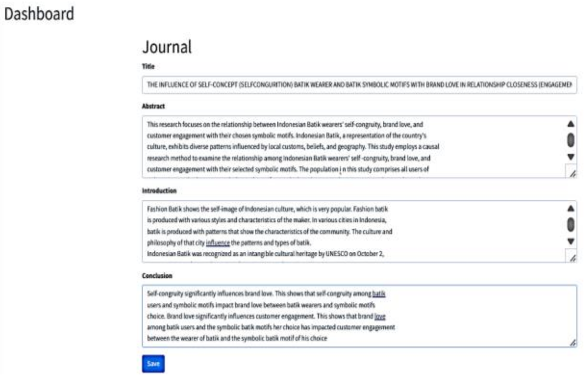
A Systematic Literature Review (SLR) is a method that employs a precise, transparent, and explicit approach, encompassing several phases [8]. This approach is designed to ensure that the literature process maintains a high level of rigor and transparency [9]. In addition, conducting a Systematic Literature Review enhances the credibility and reliability of the research for it transparency and also provide an insight of the findings in the context of the research objectives. In accordance with this approach, the aims of this Systematic Literature Review are to identify the existing frameworks, techniques and domains applied in advisory system. The organization of the paper is structured as follows: It begins with the presentation of the methodology used in the Systematic Literature Review. This is then followed by an exploration of the findings and discussions derived from the review. Finally, the paper concludes with a summary of the research's outcomes.

2. RESEARCH METHODOLOGY

This research is done by adopting the approach proposed by [8] as shown in Figure 1:

ainnadhira11@gmail.com, fauziahz@ump.edu.my, drrahmah@ump.edu.my https://journal.uob.edu.bh/

Figure 7. Illustrates the input model of textual data extracted from journal document.



Dashboard

Journal

Title

THE INFLUENCE OF SELF-CONCEPT (SELFCONGURITION) BATIK WEARER AND BATIK SYMBOLIC MOTIFS WITH BRAND LOVE IN RELATIONSHIP CLOSENESS (ENGAGEMEI

Abstract

This research focuses on the relationship between Indonesian Batik wearers' self-congruity, brand love, and customer engagement with their chosen symbolic motifs. Indonesian Batik, a representation of the country's culture, exhibits diverse patterns influenced by local customs, beliefs, and geography. This study employs a causal research method to examine the relationship among Indonesian Batik wearers' self -congruity, brand love, and customer engagement with their selected symbolic motifs. The population i n this study comprises all users of

Introduction

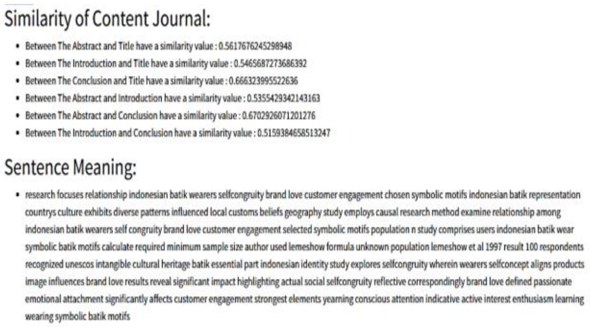
Fashion Batik shows the self-image of Indonesian culture, which is very popular. Fashion batik is produced with various styles and characteristics of the maker. In various cities in Indonesia, batik is produced with patterns that show the characteristics of the community. The culture and philosophy of that city influence the patterns and types of batik.
Indonesian Batik was recognized as an intangible cultural heritage by UNESCO on October 2,

Conclusion

Self-congruity significantly influences brand love. This shows that self-congruity among batik users and symbolic motifs impact brand love between batik wearers and symbolic motifs choice. Brand love significantly influences customer engagement. This shows that brand love among batik users and the symbolic batik motifs her choice has impacted customer engagement between the wearer of batik and the symbolic batik motif of his choice

Figure 8. Dashboard system (Document 1)

The image above, number 8, shows the system interface, which displays a dashboard for entering text content such as title, abstract, introduction, and conclusion. This dashboard assesses the uniformity of text in research papers before the process is carried out to produce appropriate values and precise interpretations.



Similarity of Content Journal:

- Between The Abstract and Title have a similarity value : 0.5617676245298948
- Between The Introduction and Title have a similarity value : 0.5465687273686392
- Between The Conclusion and Title have a similarity value : 0.666323995522636
- Between The Abstract and Introduction have a similarity value : 0.5355429342143163
- Between The Abstract and Conclusion have a similarity value : 0.6702926071201276
- Between The Introduction and Conclusion have a similarity value : 0.5159384658513247

Sentence Meaning:

- research focuses relationship indonesian batik wearers selfcongruity brand love customer engagement chosen symbolic motifs indonesian batik representation countrys culture exhibits diverse patterns influenced local customs beliefs geography study employs causal research method examine relationship among indonesian batik wearers self congruity brand love customer engagement selected symbolic motifs population n study comprises users indonesian batik wear symbolic batik motifs calculate required minimum sample size author used lemeshow formula unknown population lemeshow et al 1997 result 100 respondents recognized unescos intangible cultural heritage batik essential part indonesian identity study explores selfcongruity wherein wearers selfconcept aligns products image influences brand love results reveal significant impact highlighting actual social selfcongruity reflective correspondingly brand love defined passionate emotional attachment significantly affects customer engagement strongest elements yearning conscious attention indicative active interest enthusiasm learning wearing symbolic batik motifs

Figure 9. Illustrates the comparative similarity of content journal and sentence meaning (Document 1)

In Fig. 8 and Fig. 9 show that we used document one with the results of the consistency analysis comparison based on journal content.

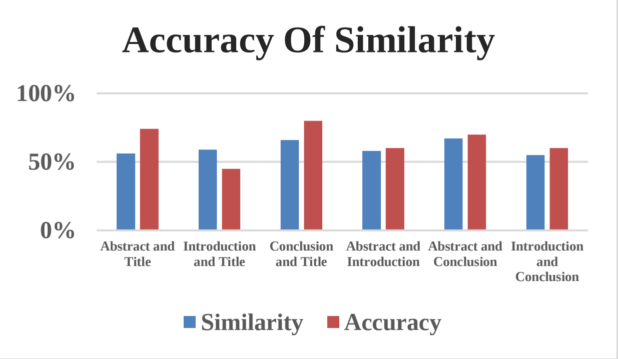
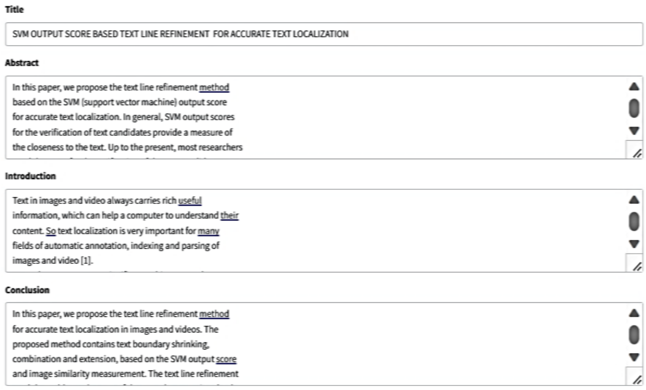


Figure 10. Result accuracy of journal content similarity.

In Fig. 9 and Fig. 10 above, the displayed results depict the accuracy of relevant content similarity obtained by the system. High similarity values are observed among various sections:The abstract and title exhibit a similarity value of 0.56%. The introduction and title show a value of 0.54%. The conclusion and title have a value of 0.66%.

Additionally, the abstract and introduction display a value of 0.53%, while the abstract and conclusion demonstrate 0.67%, and the introduction and conclusion show 0.51%. These findings indicate reasonable and relevant similarity across each section of the content. The straightforward interpretation of the research paper suggests a comprehensive explanation of the essential aspects of the research journal.



Title

SVM OUTPUT SCORE BASED TEXT LINE REFINEMENT FOR ACCURATE TEXT LOCALIZATION

Abstract

In this paper, we propose the text line refinement method based on the SVM (support vector machine) output score for accurate text localization. In general, SVM output scores for the verification of text candidates provide a measure of the closeness to the text. Up to the present, most researchers

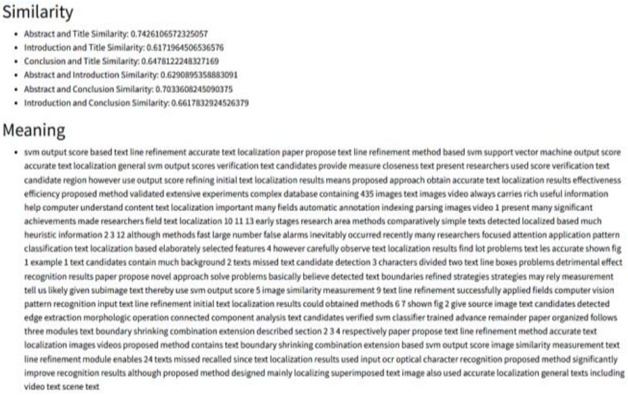
Introduction

Text in images and video always carries rich useful information, which can help a computer to understand their content. So text localization is very important for many fields of automatic annotation, indexing and parsing of images and video [1].

Conclusion

In this paper, we propose the text line refinement method for accurate text localization in images and videos. The proposed method contains text boundary shrinking, combination and extension, based on the SVM output score and image similarity measurement. The text line refinement

Figure 11. Dashboard system (Document 2)



Similarity

- Abstract and Title Similarity: 0.7426106572325057
- Introduction and Title Similarity: 0.6171964506536576
- Conclusion and Title Similarity: 0.6478122248327169
- Abstract and Introduction Similarity: 0.6290895358883091
- Abstract and Conclusion Similarity: 0.7033608245090375
- Introduction and Conclusion Similarity: 0.6617832924526379

Meaning

- svm output score based text line refinement accurate text localization paper propose text line refinement method based svm support vector machine output score accurate text localization general svm output scores verification text candidates provide measure closeness text present researchers used score verification text candidate region however use output score refining initial text localization results means proposed approach obtain accurate text localization results effectiveness efficiency proposed method validated extensive experiments complex database containing 435 images text images video always carries rich useful information help computer understand content text localization important many fields automatic annotation indexing parsing images video 1 present many significant achievements made researchers field text localization 10 11 13 early stages research area methods comparatively simple texts detected localized based much heuristic information 2 3 12 although methods fast large number false alarms inevitably occurred recently many researchers focused attention application pattern classification text localization based elaborately selected features 4 however carefully observe text localization results find lot problems text les accurate shown fig 1 example 1 text candidates contain much background 2 texts missed text candidate detection 3 characters divided two text line boxes problems detrimental effect recognition results paper propose novel approach solve problems basically believe detected text boundaries refined strategies strategies may rely measurement tell us likely given subimage text thereby use svm output score 5 image similarity measurement 9 text line refinement successfully applied fields computer vision pattern recognition input text line refinement initial text localization results could obtained methods 6 7 shown fig 2 give source image text candidates detected edge extraction morphologic operation connected component analysis text candidates verified svm classifier trained advance remainder paper organized follows three modules text boundary shrinking combination extension described section 2 3 4 respectively paper propose text line refinement method accurate text localization images videos proposed method contains text boundary shrinking combination extension based svm output score image similarity measurement text line refinement module enables 24 texts missed recalled since text localization results used input ocr optical character recognition proposed method significantly improve recognition results although proposed method designed mainly localizing superimposed text image also used accurate localization general texts including video text scene text

Figure 12. Illustrates the comparative similarity of content journal and sentence meaning (Document 2)

In Fig. 12, there is a significant difference in the results from the processing of previous journal manuscripts; in the processing of the journal manuscripts above, obtaining similarity results between abstract and titles with a percentage of 74%, introduction and title get similarity results of 62%, for conclusions and titles with 65% results, abstract and introduction with similarity 62.90%, then for abstract and conclusion get similarity 70.33%, and finally for introduction and conclusion get a similarity of 66.17%. The percentage values were obtained from Fig. 9, providing different percentage results. This proves that the consistency of each author varies, but with this system, this consistency can be detected quickly.

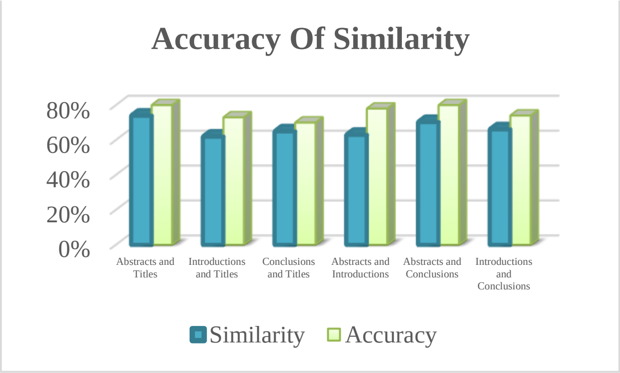


Figure 13. Result accuracy of journal content similarity.

The chart above presents the similarity results of journal manuscripts across various sections and the associated accuracy rates. Here is a detailed explanation for each row in the table: Abstract and Title: This indicates that the abstract of the journal has a similarity of 74% with its title, and the system classifies this similarity with an accuracy rate of 80%. Introduction and Title: This indicates that the introduction section has a similarity of 62% with its title, and the system classifies this similarity with an accuracy rate of 73%. Conclusion and Titles: This indicates that the conclusion section has a similarity of 65% with its title, and the system classifies this similarity with an accuracy rate of 70%. Abstract and Introductions: This indicates that the abstract has a similarity of 63% with the introduction section, and the system classifies this similarity with an accuracy rate of 78%. Abstract and Conclusion: This indicates that the abstract has a similarity of 70.33% with the conclusion section, and the system classifies this similarity with an accuracy rate of 80%. Introductions and Conclusions:

This indicates that the introduction section has a similarity of 66.17% with the conclusion section, and the system classifies this similarity with an accuracy rate of 74%. Thus, the table provides an overview of how similar the content is across various journal sections and how accurately the system classifies this similarity.

TABLE IV An illustration of the cosine similarity calculation process for the initial document of the journal-title content

Term	Cosine similarity		Result
	Document 1	Document 2	$(\omega_Q(t_i)x\omega_D(t_i))$
Natural"	0,200347	0	0
"Language"	0,200347	0	0
"Processing"	0.260658	0.174822	0.045573
"(NLP)"	0.260658	0.174822	0
"is"	0.260658	0.174822	0
"a"	0,200347	0	0.045573
"computerized"	0.260658	0.174822	0.045573
"way"	0.260658	0.174822	0.045573
"of"	0.260658	0.174822	0.045573
"analyzing"	0,200347	0.174822	0.045573
"texts."	0,200347	0	0
"NLP"	0,200347	0	0
"involves"	0,200347	0	0
"the"	0,200347	0	0
"acquisition"	0.260658	0.174822	0.045573
"of"	0,200347	0	0
knowledge"	0.260658	0	0
"on"	0.260658	0	0
"how"	0.260658	0	0
"a"	0.260658	0	0
"person"	0.260658	0	0
"understands"	0.260658	0	0
"and"	0.260658	0.174822	0.045573
$\sum T\ C$ (D1)			**0.460201**

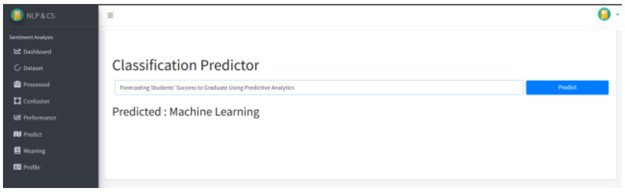
TABLE V An instance of the cosine similarity computation process for the initial document of the journal's abstract content.

Term	Cosine similarity		Result
	Document 1	Document 2	$\omega_Q(t_i)x\omega_D(t_i))$
"Semantic"	0	0.328253	0
"search"	0.252303	0.252303	0.063657
"Using"	0.252303	0.252303	0.063657
"Natural"	0.252303	0.252303	0.063657
"Language"	0.252303	0.252303	0.063657
"Processing"	0.252303	0.252303	0.063657
$\sum T\ C$ (D2)			**0.280201**

c (Q.D) = $\sum_{r=1}^{M} \omega_Q (t_i) x \omega_D (t_i)$

=**0.460201 + 0.280201 = 0.740402**

The results from Tab. 4 and Tab. 5 above depict the textual similarity outcomes processed using the cosine similarity formula. The overall result of Document 1 indicates a score of 0.460201, while the calculated result for Document 2 shows the same score of 0.280201. These values are obtained from the similarity between Document 1 and Document 2, calculated to achieve uniformity, resulting in an overall accuracy or correspondence value of 0.7404, or, equivalently, 74% similarity in the tested text. These results exemplify the computational analysis of the above content, utilized to compare each journal's content against one another. However, the testing outcomes do not significantly deviate from the manual analysis results conducted without the system. This system helps discern our writing consistency for the paper being composed.



Classification Predictor

Forecasting Students' Success to Graduate Using Predictive Analytics

Predicted : Machine Learning

Figure 14. Classification predictor page

The classification prediction display indicates that a journal manuscript can be classified into the field or scope of informatics by entering the title of the scientific journal manuscript. This result indicates that the system can recognize and associate the topic of the journal manuscript with the context or discipline of informatics

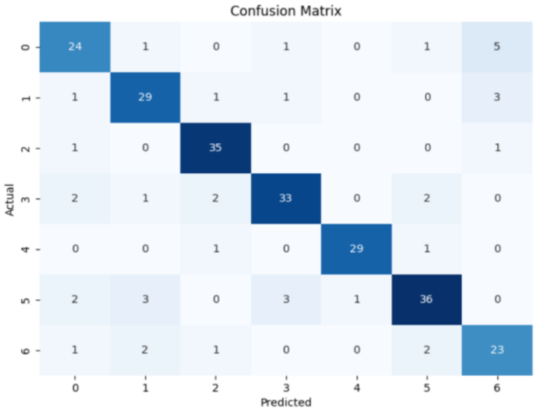


Figure 15. Scope prediction confusion matrix

This research resulted in an accuracy of 83.94% using the sciences of artificial intelligence, computer vision and pattern recognition, cryptography and security, databases, digital libraries, information retrieval, and machine learning.

TABLE VI The accuracy result of the machine learning processing model

Scope	Confusion Matrix			
	Precision	Recall	F1-Score	Support
Artificial Intelligence	0.77	0.75	0.76	32
Computer vision and pattern recognition	0.81	0.83	0.82	35
Cryptography and security	0.88	0.95	0.91	37
Database	0.87	0.82	0.85	40
Information retrieval	0.86	0.80	0.83	45
Machine learning	0.72	0.79	0.75	29

V. Conclusion

Based on the system testing results, it has been proven that the combination of NLP, TF-IDF, and Cosine Similarity can generate essential words that are easy to understand and have simple sentence meanings. This is supported by a score of 0.740402, which indicates the system's ability to provide sufficiently accurate information. Additionally, the classification results used to predict titles based on fields of study with keywords achieved an accuracy score of 83.94%. The dataset comprised 2237 journal documents, of which 2485 were utilized for sampling, achieving the highest accuracy at 85.20%. These results were derived from the testing of journals with the highest value among all tests. The findings of this research provide us with a clear and in-depth analysis, enabling us to analyze the main research topic regarding journal writing consistency independently.

References

[1] L. Yao, Z. Pengzhou, and Z. Chi, “Research on News Keyword Extraction Technology Based on TF-IDF and TextRank,” pp. 452–455, 2019.

[2] U. Mardatillah, W. B. Zulfikar, A. R. Atmadja, I. Taufik, and W. Uriawan, “Citation Analysis on Scientific Articles Using Cosine Similarity,” pp. 1–4, 2021.

[3] N. K. Seong, J. H. Lee, J. B. Lee, and P. H. Seong, “Retrieval methodology for similar NPP LCO cases based on domain specific NLP,” ScienceDirect Access, vol. 55, no. 2, pp. 421–431, 2023.

[4] Xiaofan Lin, “Text-mining based journal splitting,” IEEE Access, pp. 1075–1079, 2017.

[5] M. Kikuchi, M. Yoshida, and K. Umemura, “Journal Name Extraction from Japanese Scientific News Articles,” pp. 143–148, 2018.

[6] H. Setiadi, R. Saptono, R. Anggrainingsih, and R. Andriani, “Recommendation Feature of Scientific Articles on Open Journal System Using Content-based Filtering,” pp. 1–6, 2019.

[7] Y. Du, W. Liu, X. Lv, and G. Peng, “An improved focused crawler based on Semantic Similarity Vector Space Model,” vol. 36, pp. 392–407, 2015.

[8] M. Hanifi, H. Chibane, R. Houssin, and D. Cavallucci, “Problem formulation in inventive design using Doc2vec and Cosine Similarity as Artificial Intelligence methods and Scientific Papers,” *Eng Appl Artif Intell*, vol. 109, p. 104661, 2022.

[9] C. Zheng, H. Fan, R. Singh, and Y. Shi, “A domain expertise and word-embedding geometric projection based semantic mining framework for measuring the soft power of social entities,” *IEEE Access*, vol. 8, pp. 204597–204611, 2020.

[10] A. K. M. Masum, S. Abujar, R. T. H. Tusher, F. Faisal, and S. A. Hossain, “Sentence Similarity Measurement for Bengali Abstractive Text Summarization,” IEEE Access, pp. 1–5, 2019.

[11] N. Kapoor, S. Vishal, and K. K. S., “Movie Recommendation System Using NLP Tools,” pp. 883–888, 2020.

[12] A. Shetty, D. Makati, M. Shah, and S. Nadkarni, “Online Product Grading using Sentimental Analysis with SVM,” IEEE Access, pp. 1079–1084, 2020.

[13] Jin-Liang Yao, Yan-Qing Wang, Lu-Bin Weng, and Yi-Ping Yang, “Locating text based on connected component and SVM,” IEEE Access, pp. 1418–1423, 2007.

[14] Md. R. Mia and A. S. Md. Latiful Hoque, “Question Bank Similarity Searching System (QB3S) Using NLP and Information Retrieval Technique,” IEEE Access, pp. 1–7, 2019.

[15] N. R. de Oliveira, P. S. Pisa, M. A. Lopez, D. S. V. de Medeiros, and D. M. F. Mattos, “Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges,” *Information*, vol. 12, no. 1, p. 38, Jan. 2021.

[16] H. T. Huynh, N. Duong-Trung, X. S. Ha, N. Quynh Thi Tang, H. X. Huynh, and D. Quoc Truong, “Automatic Keywords-based Classification of Vietnamese Texts,” pp. 1–3, 2020.

[17] search optimization with web scraping, text processing and cosine similarity algorithms,” IEEE Access, pp. 346–350, 2020.

[18] D. Sudigyo, A. A. Hidayat, R. Nirwantono, R. Rahutomo, J. P. Trinugroho, and B. Pardamean, “Literature study of stunting supplementation in Indonesian utilizing text mining approach,” *Procedia Comput Sci*, vol. 216, pp. 722–729, 2023.

[19] N. Jahan, R. U. Haquey, A. K. Saha, M. F. Mridha, and M. A. Hamid, “Identification of expectancy, proximity, and compatibility of the bengali language,” pp. 349–354, 2019.

[20] A. Kupiyalova, R. Satybaldiyeva, and S. Aiaskarov, “Semantic search using Natural Language Processing,” pp. 96–100, 2020.

[21] P. Beyranvand and T. Aytekin, “Automating Customer Claim Registration by Text Mining,” pp. 1-5, 2020.

[22] Z. Liu, J. Zhu, X. Cheng, and Q. Lu, “Optimized Algorithm Design for Text similarity Detection Based on Artificial Intelligence and Natural Language Processing,” *Procedia Comput Sci*, vol. 228, pp. 195–202, 2023.

[23] C. Leite da Silva, L. May Petry, V. Freitas, and C. Friedrich Dorneles, “Mining Journals to the Ground: An Exploratory Analysis of Newspaper Articles,” IEEE Access, pp. 78–83, 2019.

Sitti Mawaddah Umar and a short biography A student from Makassar, Indonesia, is currently studying at Hasanuddin University. He completed his undergraduate studies in informatics engineering at Dipanegara University, Makassar, Indonesia, in 2021. His academic interests primarily revolve around Natural Language Processing (NLP).

Ingrid Nurtanio received the bachelor degree in Electrical Engineering from Hasanuddin University, Makassar, Indonesia in 1986. She received her Master of Technology from Hasanuddin University, Makassar, Indonesia in 2002. She received her doctoral degree from Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia in 2013. Her research interest are in Digital Image Processing, Computer Vision and Intelligent System. Currently, she is the staff of Department of Informatics, Faculty of Engineering, Hasanuddin University. She is a member of IAENG and member of IEEE.

Zahir Zainuddin holds a Doctor of Computer Engineering from Bandung Institute of Technology, Indonesia, in 2004. He also received his B.Sc. in Electrical Engineering Department, Hasanuddin University, Indonesia, in 1988 and his M.Sc. (Computer Engineering) from Florida Institute of Technology USA in 1995. He is an associate professor at the Department of Informatics at Hasanuddin University in Indonesia. His research includes Computer Systems, intelligent systems, computer vision, and smart cities. He has published over 60 papers in international journals and conferences. In 1989, he was a JSPS research fellow at the Tokyo Institute of Technology.