The Seybold Report

ISSN 1533-9211

DOI: 10.5281/zenodo.10209822

CONSISTENCY AND STRUCTURE ANALYSIS OF SCHOLARLY PAPERS USING BASED ON NATURAL LANGUAGE PROCESSING

SITTI MAWADDAH UMAR

Department of Informatics, Hasanuddin University, Gowa, South Sulawesi, Indonesia.
Email: umarsm21d@student.unhas.ac.id

INGRID NURTANIO*

Department of Informatics, Hasanuddin University, Gowa, South Sulawesi, Indonesia.
*Corresponding Author Email: ingrid@unhas.ac.id

ZAHIR ZAINUDDIN

Department of Informatics, Hasanuddin University, Gowa, South Sulawesi, Indonesia.
Email: zahir@unhas.ac.id

Abstract

This research presents a comprehensive similarity analysis of the consistency of authors in crafting papers and providing simple conclusions or meanings in a journal. Machine learning technique are employed to assess the similarity and interpretation of these sentences. The study attempts to mine text data, making it more structured and easily understood, introducing an approach to identifying relevant author consistency in an extensive collection by utilizing text analysis and the understanding of the meanings of new words using Natural Language Processing. In the interim, the weight analysis was conducted through the validation of TF-IDF and Cosine Similarity. By conducting an in-depth analysis across the corpus dataset consisting of 60 to 150 journal documents, this research utilizes classification patterns, preprocessing patterns, similarity calculations, and interpretation of results. This study is able to provide information about how consistent researchers are in writing assembled journals. The results underscore the effectiveness of NLP in processing natural language, enhanced by the incorporation of TF-IDF and Cosine Similarity, which refine the representation of relevance in journal content.

Keyword: Text Mining, Natural Language Processing, TF-IDF, Corpus, Cosine Similarity.

INTRODUCTION

This research presents a comprehensive comparative analysis of author consistency in crafting and summarizing academic papers. It is observed that approximately 75% of the monthly journals published in Indonesia amount to 1 million papers (Batra et al., 2020) .Academic papers serve as valuable references for research and are often published to fulfill academic or professional requirements. In writing research papers, objectivity and the use of empirical data are crucial (Kupiyalova et al., 2020). Authors are expected to structure their papers per the proposed research topic consistently. In assessing author consistency, a clear and reader-friendly writing style is necessary.

Xiaofan's research indicates that 62% of researchers need help determining the consistency of their writing in research papers, with 38% opting to await peer review results from publishers (Xiaofan Lin, 2017) Furthermore, adherence to discipline-specific terminology and conventions in every paper is essential (Rahmawati & Khodra, 2017). A well-constructed paper

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

not only requires consistency in content but also coherence between sections, including the title, abstract, introduction, and conclusion (Huynh et al., 2020). The interrelatedness of paper sections ensures systematic assessment by the system, eliminating the inefficiency of manual evaluations (Xia et al., 2019) Once consistency is established across paper sections, the final output includes values and simple interpretations (Mardatillah et al., 2021). The research uses a corpus dataset and LsaSummarize to identify the fundamental meaning of unstructured text. (Gunasinghe et al., 2014) employ StandFordCoreNLP, WordNet, and Syntactic Similarity, using the vector space model, to assess sentence meaning based on text similarity. This research utilizes Python programming and responsive web design (Liu et al., 2018). In achieving consistency and simple interpretation, deep text analysis is conducted, involving NLP preprocessing stages, including case folding, tokenization, stop word removal, parsing, and stemming (Zhang et al., 2020) Accurate descriptions improve consistency (Akhter et al., 2020)

Cosine Similarity measures the similarity of words or text within research articles or papers on subtopics/descriptions within research topics, resulting in an efficient system. The proposed text vectorizer employs TF-IDF to minimize assessment errors. The entire system is affected by the relevance of individual words or texts to detect author consistency in relevant paper creation (Gunasinghe et al., 2014). Weight extraction calculates values based on scientific relevance. The word similarity system can be used as a reference for accurate output (Ridwang et al., 2020a) Cosine similarity measures word frequency similarity. (Rinartha & Surya Kartika, 2019)

METHODOLOGY

1. Natural Language Processing

Natural Language Processing (NLP), also known as computational linguistics, has established itself as a research field that involves computational models and processes to address practical issues in understanding and manipulating human language (Mia & Latiful Hoque, 2019). Since natural language is continually evolving, it is challenging to establish explicit rules for computers (de Oliveira et al., 2021). NLP places a greater emphasis on extracting contextual information from sentences (Kamsties et al., 2001) in documents and individual words within sentences (Kilany et al., 2018).

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

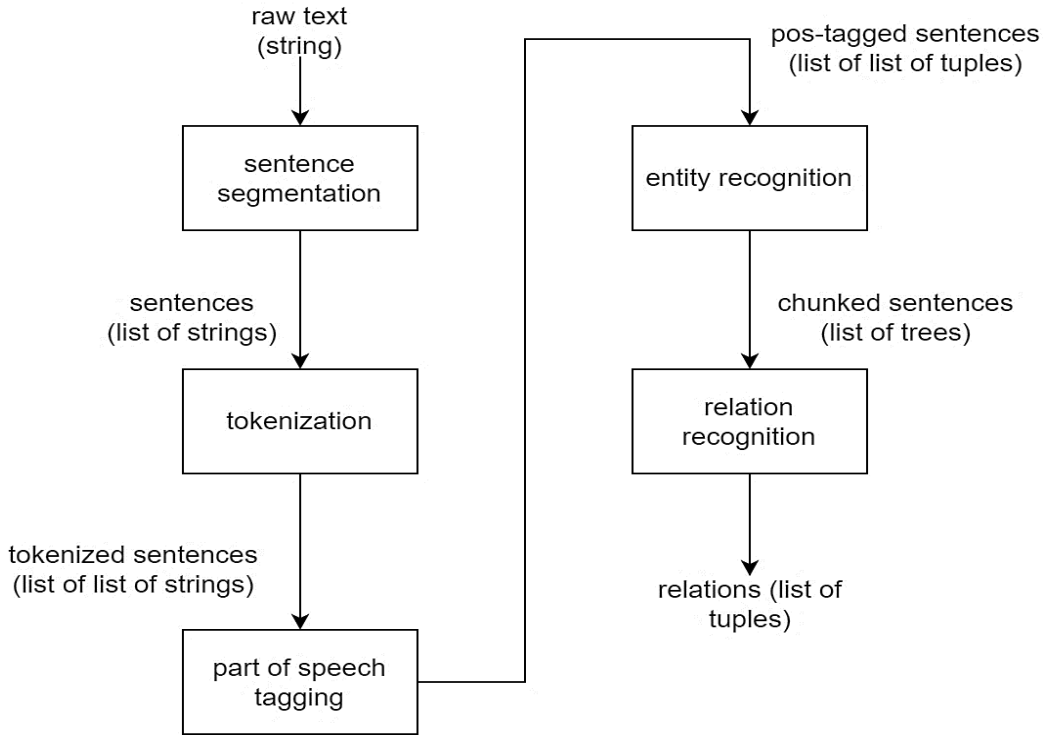

Figure 1: Sequnce Diagram Natural Language Processing

The primary focus is the mechanism of sentence mining for identifying keywords that serve as representations of features in a specific document (Zeng et al., 2017).The initial step involves extracting text, documents into sentences, which are then further segmented into words with specific numbering, a process known as tokenization (Amanullah et al., 2019) as illustrated in the system design above figure 1.

2. Vectorize with TF-IDF

The TF-IDF (Term Frequency Inverse Document Frequency) algorithm is a weighted statistical method commonly used in information retrieval and data mining (Mz et al., 2023). It assesses the importance of words within the text (Gunawan et al., 2017) or a corpus. The significance of words increases proportionally with their frequency of appearance in the text but decreases inversely with their frequency of appearance in the corpus (ZHANG Guangrong1, 2019). Hence, the main idea behind the TF-IDF algorithm is that if a word or phrase appears frequently within an article (high TF value) and rarely in other articles (low DF value, high IDF value) (Lahitani et al., 2016), it is considered to represent the article effectively and can be used for classification purposes (Yao et al., 2019). For example text:

A.	Jupiter is the largest plane	5 number of words in a document
B.	Mars is the fourth planet form the sun	8 number of words in a document

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

The initial step involves creating a vocabulary of unique words and calculating TF for each document. TF will be higher for words that frequently appear in documents and lower for words that are infrequent in documents. In TF-IDF, the TF value of a word is calculated as follows:

$$tf_{ij} = n\frac{n_{ij}}{\Sigma_k n_{kj}} \quad (1)$$

The first stage in calculating TF-IDF, n_{ij} represents the frequency of the occurrence of term t_i in document d_j, and the denominator is the sum of the frequencies of all words in document d_j. The IDF value of term ti is determined by dividing the total number of documents by the number of documents containing term t_i, followed by taking the logarithm of the result.

$$idf_i = log\frac{|D|}{|\{j:t_i \in d_j\}|} \quad (2)$$

The next stage, |D| represents the total number of all documents, and |{j: t_i}| represents the number of documents that contain the term t_i in the text (the number of documents in which the term $n_{ij} \neq 0$). It is important to note that if the term t_i does not exist in the corpus, it can result in a denominator of zero. To avoid division by zero, it is common practice to add 1 to the denominator. It can be seen from table 3 that is presented, there are the results of the calculations for the IDF stage before entering the next calculation stage.

$$TF - IDF_{ij} = tf_{ij} \times idf_i \quad (3)$$

Table 1: Result for text from TF-IDF

Words	TF (for A)	TF (for B)	IDF	TFIDF (A)	TFIDF (B)
Jupiter	1/5	0	ln(2/1) = 0.69	0.138	0
Is	1/5	1/8	ln(2/2) = 0	0	0
The	1/5	2/8	ln(2/2) = 0	0	0
largest	1/5	0	ln(2/1) = 0.69	0.138	0
Planet	1/5	1/8	ln(2/2) = 0	0.138	0
Mars	0	1/8	ln(2/1) = 0.69	0	0.086
Fourth	0	1/8	ln(2/1) = 0.69	0	0.086
From	0	1/8	ln(2/1) = 0.69	0	0.086
Sun	0	1/8	ln(2/1) = 0.69	0	0.086

After applying TF-IDF in the table 1, text in documents A and B can be represented as TF-IDF vectors with dimensions equal to the vocabulary. The corresponding values for each word represent the importance of that word in the respective document.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

3. Cosine Similarity

Cosine similarity is used to compare the similarity between documents, in this case, comparing a query with training documents (Pattnaik & Nayak, 2019). When calculating cosine similarity, the first step is to perform the dot product between the query and the document and then sum it. Afterward, it involves multiplying the length of the document by the squared length of the query, followed by calculating the square root. Next, the result of the dot product is divided by the result of the product of the document lengths and the query. The formula can be expressed as follows:

$$cosSim(d_j, q_k = \frac{\sum_{i=1}^{n}(tdij \times tiqk)}{\sqrt{\sum_{i=1}^{n} tdij^2} \times \sum_{t=1}^{n} tqik^2} \quad (4)$$

Explanation

cosSim: Represents the degree of similarity between a specific document and a query.

tdij : Stands for the i-th term in the vector for document j.

tqik : Stands for the i-th term in the vector for query k.

n : Denotes the total number of unique terms in the dataset.

Here are the manual steps for calculating the Cosine Similarity algorithm:

a) Begin by defining three queries: the answer query (D), the critical answer query (Q), and the combined query (Queries).

b) Remove any stop words or symbols that do not affect the evaluation from all three queries. This includes punctuation marks like periods, commas, and exclamation points.

c) Eliminate common stop words from all three queries. Stop words are common words that are often used in queries but do not carry significant meaning, such as "and," "if," "in," "but," "however," and so on.

These steps help preprocess the queries to prepare them for Cosine Similarity calculation, where you compare the similarity between the queries.

4. Corpus

Based on the state of the art that we compiled, a corpus is a well-structured collection of digital texts that allows for systematic analysis and data retrieval. A corpus can be constructed based on international journal references by gathering articles from these journals and storing them in a database (Adil et al., 2019). Corpora are used for various purposes, especially in language and linguistic analysis (Roul et al., 2017) In this research, the LsaSummarize corpus is utilized to help condense text into a coherent form.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

It provides accurate and representative data:

a) **Accurate and Representative Data**: Corpora offers precise and representative data about how language is used in specific contexts.

b) **Developing Natural Language Processing (NLP) Systems**: Corpora serves as a crucial data source for developing natural language processing (NLP) tools used for automating tasks like machine translation, sentiment analysis, and text classification (Manalu et al., 2017).

Collecting articles can be done manually or through web scraping (Ridwang et al., 2020b). Then, journal articles can be selected randomly or based on specific criteria (Shin et al., 2018) such as publication period, citation count, or popularity (Villanueva et al., 2022).

5. Consistency Analysis

In this research context, "consistency" refers to how various elements or parts of a scientific paper maintain alignment, coherence, and good relationships. It includes upholding uniformity in writing style, structure, terminology usage, and cohesion among elements such as the title, abstract, introduction, conclusion, and other paper sections. The consistency model in this research may involve natural language processing (NLP) analysis and techniques to measure consistency in scientific papers. In other words, the research aims to identify whether authors of scientific papers maintain the quality of consistency throughout their papers or if there is significant variation or inconsistency between these sections.

In the NLP research context, a consistency model may involve the application of algorithms to check and compare different elements in the paper, such as the title, abstract, introduction, and conclusion, to measure the extent to which they are consistent in the use of words, phrases, or language. This model can provide valuable insights into the quality and effectiveness of scientific communication in papers and enable authors or editors to improve if discrepancies or lack of consistency are found in their writing.

6. Structure Analysis

The structural analysis model in this research refers to the framework or methodology used to analyze and evaluate the content structure in scientific papers, particularly in terms of the relationships between various elements in the paper. It includes ways to identify and measure the extent to which various elements in the paper are interconnected and consistent, such as the title, abstract, introduction, and conclusion.

The structural analysis model in this NLP research may include steps such as the following:

1. Text Extraction: Text from various parts of the paper, such as the title, abstract, introduction, and conclusion, is extracted from the source documents.
2. Preprocessing: This text may go through preprocessing stages, including cleaning the text of irrelevant characters, changing letter case to lowercase (or vice versa), removing common words, and performing stemming or word lemmatization.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

3. Similarity Measurement: Algorithms like TF-IDF and Cosine Similarity measure how similar or related these elements are. It can help identify potential inconsistencies or discrepancies in the use of words or terms across the paper.
4. Relevance Analysis: This model will evaluate the relationships between elements like title and abstract, abstract and introduction, introduction and conclusion, and then check content relevance and consistency.
5. Results and Visualization: The analysis results may be presented in numeric values or visualizations that indicate the extent of consistency and relevance between elements in the paper.

This structural analysis model helps understand how different elements in the paper are interrelated and whether they consistently present information and arguments. It can assist researchers, authors, or editors in improving the quality of scientific communication in papers.

RESULTS AND DISCUSSION

In the Results and Discussion section, we reached out to 50 journal samples out of 120 samples ready to be submitted to the publication service provider. It was done to assess the similarity and consistency of the authors in composing papers and to obtain a straightforward interpretation of the paper's content.

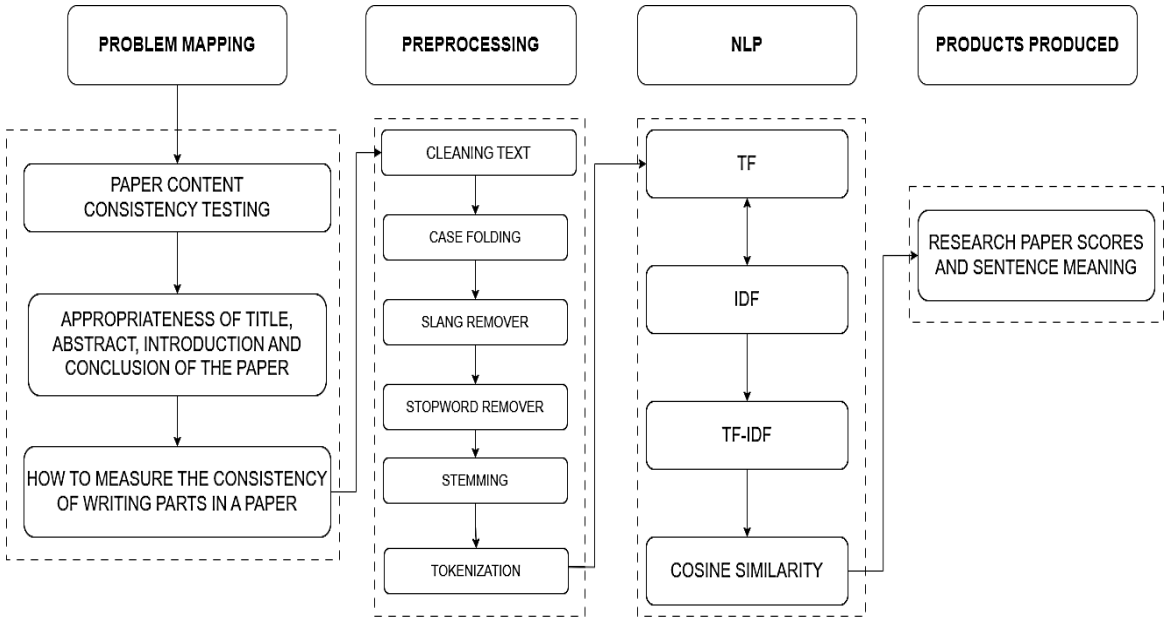

Figure 2: Design Framework of thought System of Text Similarity

Figure 2. depicts about our mapping road The research proposed by the researcher has problem mapping as follows: in the initial column, the researcher outlines the problem mapping for assessing content consistency in research papers. This assessment analyzes the alignment of content in the title, abstract, introduction, and conclusion of the research paper. The problem

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

addressed in this study 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 (Ridwang et al., 2020c). In the problem mapping, it is explained that the challenge lies in identifying the consistency between words used in the content of the paper's title, abstract, introduction, and conclusion.

This challenge often hinders efforts to ensure that all sections of the academic work are interconnected and represent a consistent understanding of the topic. The system is capable of identifying and analyzing 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 title, abstract, introduction, and conclusion of a paper.

Sentence Similarity Measurement for Bengali Abstractive Text Summarization

Abu Kaisar Mohammad Masum, Dept. of CSE, Daffodil International University, Dhaka, Bangladesh, mohammad15-6759@diu.edu.bd
Sheikh Abujar, Dept. of CSE, Daffodil International University, Dhaka, Bangladesh, sheikh.cse@diu.edu.bd
Raja Tariqul Hasan Tusher, Dept. of CSE, Daffodil International University, Dhaka, Bangladesh, tusher.cse@diu.edu.bd
Fahad Faisal, Dept. of CSE, Daffodil International University, Dhaka, Bangladesh, fahad.cse@diu.edu.bd
Syed Akhter Hossain, Dept. of CSE, Daffodil International University, Dhaka, Bangladesh, aktarhossain@daffodilvarsity.edu.bd

Abstract— Text summarization is a massive research area in natural language processing. It reduces the larger text and provided the prime meaning of a text document. Find the meaning of the larger text needed of a proper text analysis which gives a better text summarizer. Abstractive text summarizer gives a summary which can present or not present in the text document. The machine produces a text summary after learning from the human given summary. Sentence similarity is a way to judge a better text summarizer. It is exploring the similarity between sentences or words. This paper we discuss several methods of sentence similarity and proposed a method for identifying a better Bengali abstractive text summarizer. We used human given summary and machine response summary sentences for similarity measurement where both sentences contain a Bengali short text. There are several approaches to English sentences similarity measurement, and we applied some of the approaches for similarity measure for our Bengali text which give a satisfying result. For our given methods we collect data from online and social media and create a summary of those texts. After creating a summary pre-processing this text and generate a summary from our abstractive text summarization model. All summary sentence similarity measurement cases using the method provided an effective value and optimal result.

Keywords— *Text Summarization, Bengali Abstractive Summarization, Sentence Similarity, Text Pre-processing, Natural Language Processing.*

I. Introduction

After the invention of Machine translation, text summarization is a vital topic of natural language processing research area. Extractive and Abstractive text summarization are two major parts of text summarizations. The extractive rule contains short and important content of a large text document which can present or not present in the document. The abstractive rule gives a new summary based on the given summary, and this is not necessary to a given word or the summary is present in the text document. So for a better abstractive text summarizer, need to give an understandable summary, which helps for a machine to generate an automatic new summary. Sentence similarity can estimate the closeness between two sentences or words. Since abstractive text summarizer generate a new text so the similarity measure of human and machine summary is important for better abstractive text summarizer.

Bengali text processing is different from other languages for its structure. There is some limitation in the preprocessing stage of Bengali text such as remove an unwanted character, space remove, stop word remove, etc. So the best way to analysis of Bengali text is using Unicode conversion [3]. Removing unnecessary things provide a pure Bengali text for analysis. When measuring sentence similarity two Bengali sentences or text are required. All sentences are made by the combination of some word. So measuring the similarity of the sentence means measuring the word similarity where each word varies a specific meaning. Given sentences are provided whether the sentence is equal or not equal in semantically meaning.

The current similarity measure algorithms are used to measure semantic and word orders. As a dataset uses lexical and corpus data [5] for finding the semantic and word order measurement. The lexical dataset is used human opinion data and corpus builds any kinds of text dataset. Measuring texts show meaningful representation and Interrelation between sentences or texts. A text contains sequences of words, and each word contains a piece of useful information. In a sentence, distinct words are used for joint words in the lexical dataset. The noteworthiness of a word is weighted by utilizing data content got from a corpus.

For text summarization at first, machine read from human summary and then its response some word or text for correspondune text. In automatic text summarization, some of

A Survey on NLP based Text Summarization for Summarizing Product Reviews

Ravali Boorugu, Assistant System Engineer, Tata Consultancy Services Limited, Hyderabad., ravali.boorugu@gmail.com
Dr. G. Ramesh, Associate Professor, Department of CSE, GRIET, Hyderabad., ramesh680@gmail.com

Abstract—No one can imagine life without a smartphone and internet nowadays. It has become essential for people of all age groups. With an increase in the usage of internet and smartphones, there has been a steady increase in online shopping too. Everyone wishes to get their products delivered at their home without any hassle. How to detect which products are genuine and pick the best among the unlimited options at the same price? Every user looks at the reviews before ordering anything online. Nevertheless reading those long reviews is not easy for everyone. Therefore, there must be something that can reduce the long reviews to short sentences of limited words depicting the same meaning. Text Summarization can come in hand in this aspect. Many NLP researchers are interested in Text Summarization. This paper is a survey on the various types of text summarization techniques starting from the basic to the advanced techniques. According to this survey, seq2seq model along with the LSTM and attention mechanism is used for increased accuracy.

Keywords— Text Summarization, Product Review Summarization, NLP Techniques.

I. INTRODUCTION

There has been a continuous increase in the number of internet users every year. With an increase in Internet users, comes a great deal of information that gets stored online every second. There is a need for summarizing this data without losing the original meaning of the data. Thus the process of Text Summarization comes into the picture with its benefits spread over different fields such as Machine Learning, Natural Language Processing, Artificial Learning, Semantics etc.,

Online Shopping has become a common thing these days as a wide variety of products are available at a single place. The ease of ordering a product and getting the product delivered directly to home at a convenient date and time has attracted many people. Along with these, the discount offers being offered by online shopping sites are making the people stick to online ordering. Everyone refers to the product reviews before buying a product. Then they can conclude which is the best product to buy among the different products available. Suppose a user needs to buy a laptop. Then he must go through different kinds of laptops available at his budget, make a note of different reviews for each product and choose the best among the available laptops. This is a tedious and time taking task. In addition to this, some users' reviews are so long that the user could get the actual meaning of it only after closely going through the review. Thus there is a need for minimizing the review to a shorter representative sentence which depicts the same meaning as the whole content.

This is the area where Text Summarization comes into picture with a great deal of benefits that could help us in choosing the best product from the whole lot. Text Summarization methods are broadly categorized into different types as shown in the figure below.

Figure 3: Shows the Model Input of Text from Journal Documents

In figure 3 is shows an image of a journal document that wants to be processed to determine how consistently the journal has been compiled. In the journal, a simple conclusion will be made explaining the contents of the journal studied.

ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

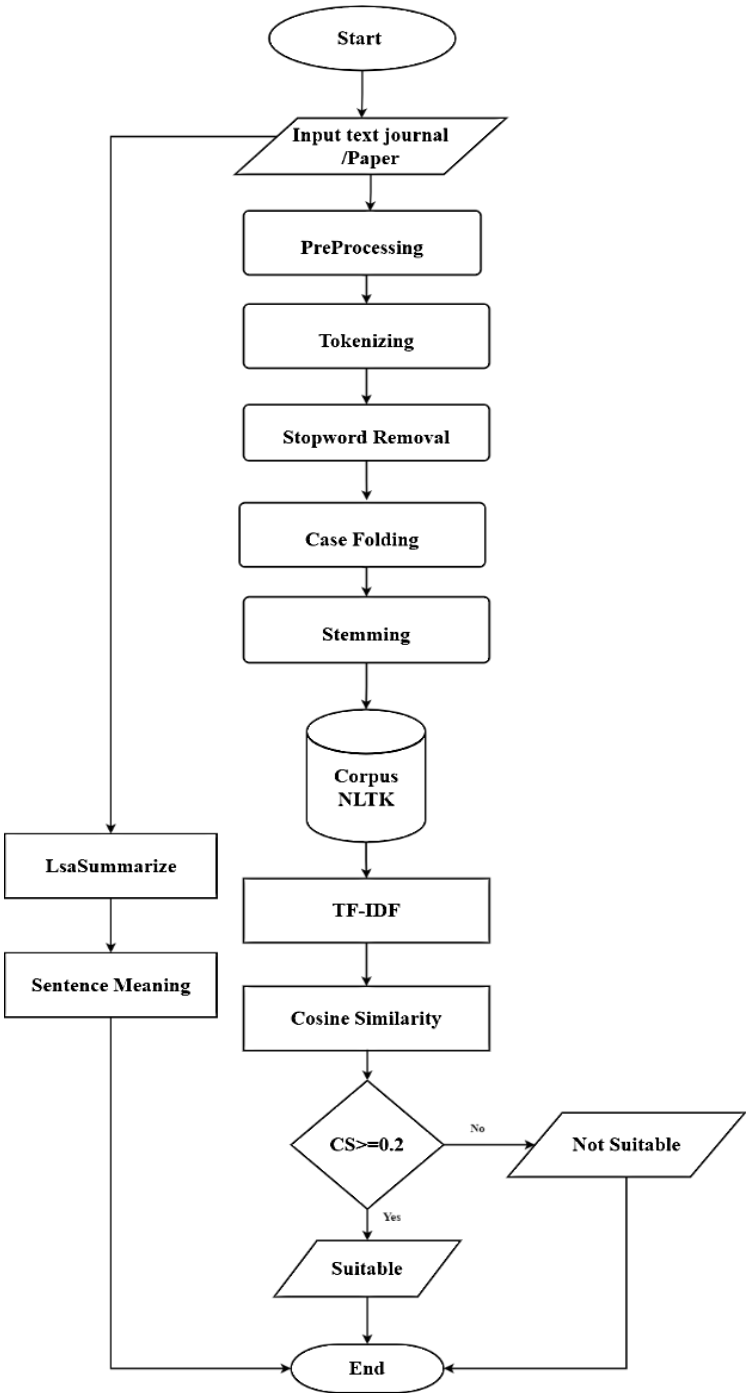

Figure 4: Proposed Design System

Figure.4 explaining the system function used for text preprocessing, the following steps are employed: Removing non-letter characters or spaces from the text. Converting all letters in the text to lowercase (Wibowo et al., 2017). Tokenizing the text into words. Removing stop words

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

from the text. Applying stemming, which involves transforming words into their base forms. The `compute_cosine_similarity` function calculates the cosine similarity between two texts. Firstly, this function uses Scikit-learn's CountVectorizer to convert text into vectors based on word frequencies (Raza et al., 2021). Then, the function calculates the cosine similarity between the two resulting text vectors. The cosine similarity result is between 0 and 1, where a higher value indicates a more significant similarity between two texts (Pohl et al., 2019). Both functions are used to process text and measure similarity between different texts. However, before using these functions, it is essential to import the required libraries such as re, nltk, stopwords, and nltk. Tokenize, nltk.stem, nltk—corpus, stemmer, CountVectorizer, and cosine_similarity (Masum et al., 2019).

The text processing involves several stages of preprocessing, including: Removing non-letter characters or spaces from the text. Converting all letters in the text to lowercase. Tokenizing the text into words (Kuang & Davison, 2018). The described functions and preprocessing steps are essential for text analysis and similarity measurement:

a) **Text Cleaning**: This is the process of cleaning and processing raw text to make it easier to analyze or process. It is often required in text analysis, natural language processing, or data processing tasks involving text. The process includes: Correcting typos or spelling errors in the text. Removing excessive numbers and punctuation (e.g., "Text!!!!"). Dealing with text full of emojis, emoticons, usernames, and links (if the text is from platforms like Twitter or Facebook). Handling parts of the text that are outside of English. Managing data that contains a mixture of more than one language. Dealing with hyphenated words or data containing abbreviated words (e.g., text processing). Addressing word repetition (e.g., "Data"). Text cleaning is crucial to prepare text data for meaningful analysis and interpretation.

b) **Case Folding**: Case folding is one of the text processing steps that involves converting all letter characters in the text to either lowercase or uppercase.

c) **Slang Remover:** A tool or process used to eliminate or replace slang words with appropriate or standard words in a language. This is useful in text analysis and natural language processing to ensure that the analyzed or processed text is not influenced by slang terms.

d) **Stop word Remover**: A process used to remove stop words (common words that often do not carry specific meaning) from the text.

e) **Stemming**: A process in text processing that involves removing word suffixes to return words to their base or root form.

f) **Tokenizing**: The process of breaking down text into discrete units called tokens. Tokens can be words, phrases, sentences, or even characters, depending on the level of processing desired. Tokenization is a crucial step in text processing and natural language processing as it helps break down text into smaller units for further analysis.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

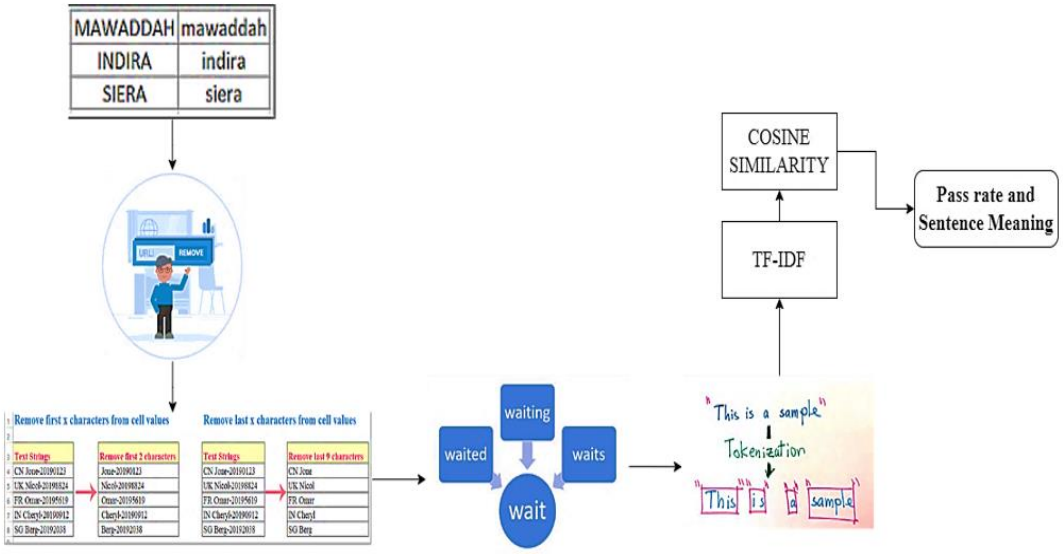

Figure 5: Preprocessing of Stage NLP

Figure 5. our explanation about preprocessing stages. After preprocessing, the results include the similarity values for each paper's content, and a simple interpretation is generated to summarize the paper's content that serves as the research object. The output contains values representing the relevance comparison between the title and abstract, introduction and title, conclusion and title, abstract and introduction, abstract and conclusion, and introduction and conclusion.

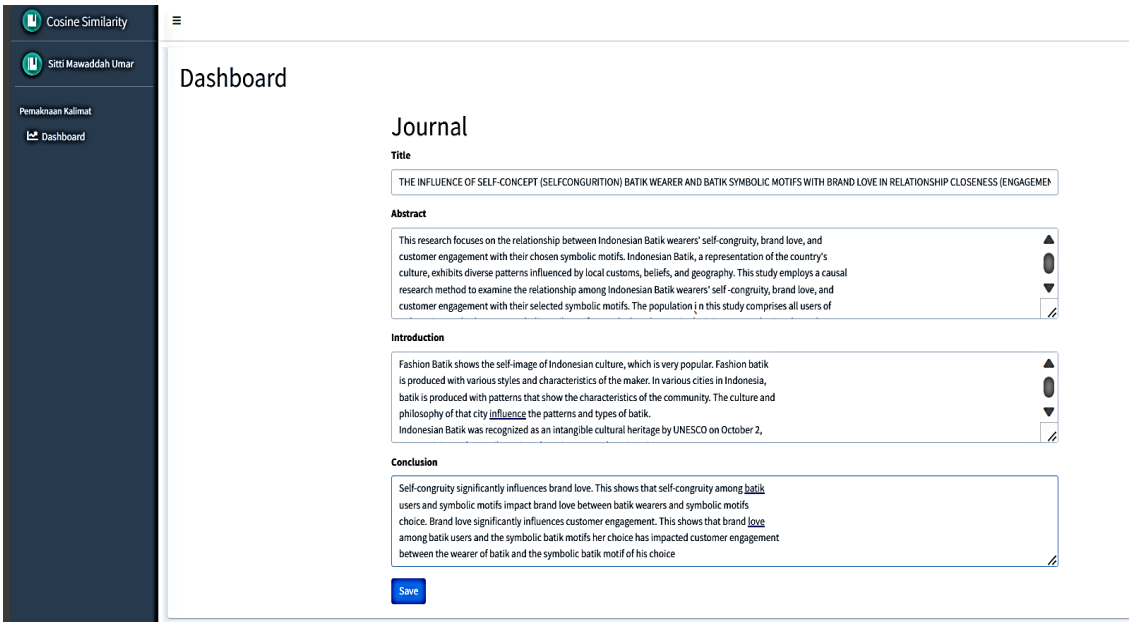

Figure 6: The System Dashboard Display

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

The above image figures.6 the system dashboard display represents the dashboard interface for entering text content such as the title, abstract, introduction, and conclusion. It is used to determine the similarity of text within research papers before processing to generate relevant values and simple 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 7

The figure above shows the results of the comparison value of each TF-IDF and Cosine Similarity match between the journal content and the results of the inferred simple meaning obtained from the NLP process.

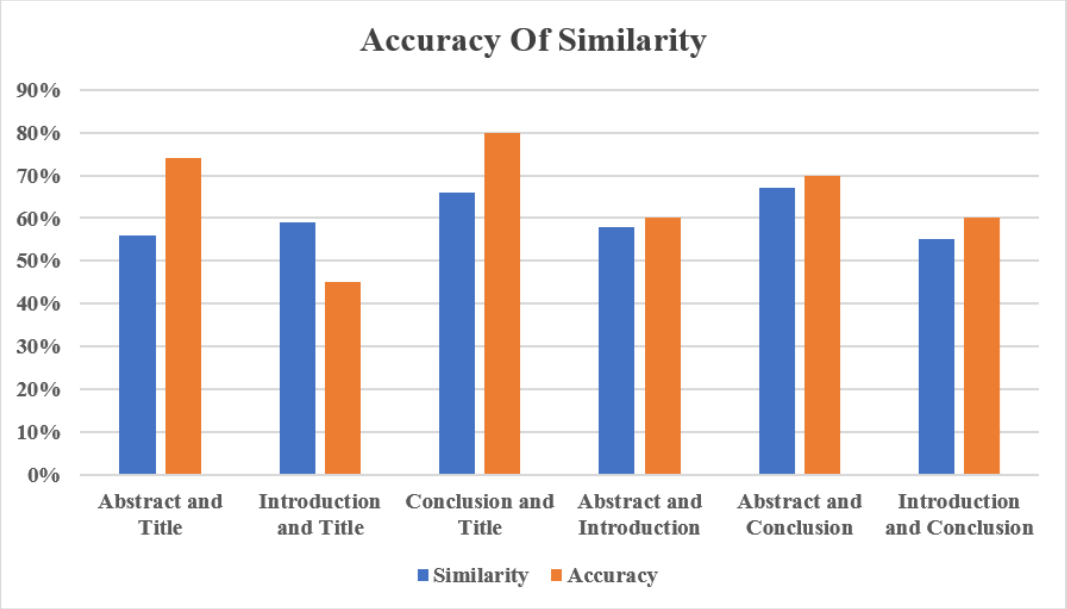

Figure 8: Accuracy Result of similarity of content journal

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

In figure 7 above, is the result of the system and figure 8 the above image shows the accuracy result about relevant content similarity results are obtained, and they have high similarity values with the following comparison values: abstract and title have a similarity value of 0.56%, introduction, and title receive a value of 0.54%, and conclusion and title are 0.66%. The abstract and introduction have a value of 0.53%. The Abstract and conclusion are 0.67%, and the introduction and conclusion are 0.51%. The results indicate that the obtained similarity is reasonable and relevant for each section of the content. The results of the straightforward interpretation of the research paper suggest that the paper explains essential aspects of the research journal.

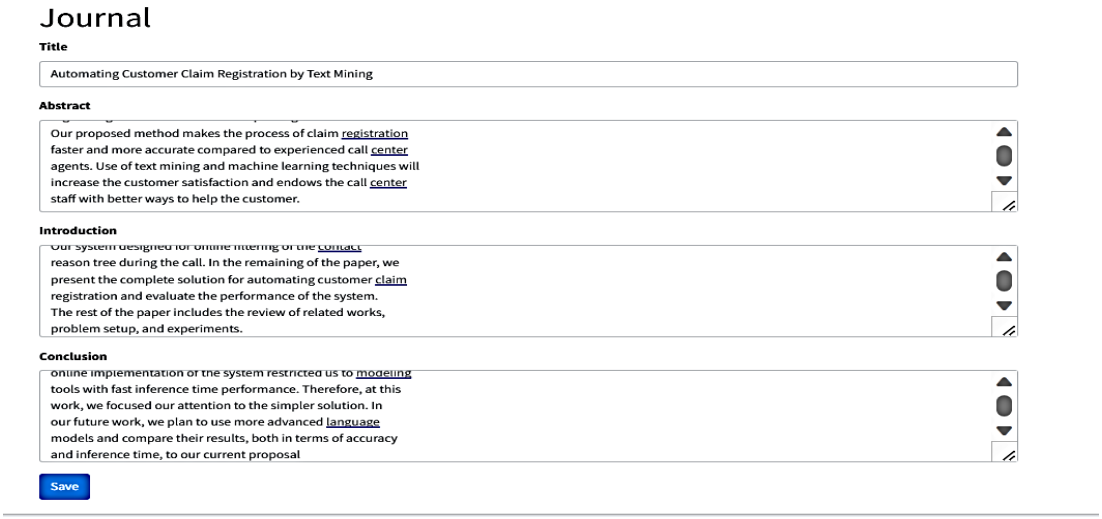

Figure 9: The Text Content before Processing Refers to the Raw

From the image above figure 9. It displays a dashboard interface for inputting text into the text boxes before undergoing the preprocessing stage.

Table 2: Example of Cosine Similarity Processing Step for the First Document of Journal-Title Content

Term	Cosine similarity		Hasil ($\omega_Q(t_i) \, x \, \omega_D(t_i)$)
	Document 1	Document 2	
Natural"	0,200347	0	0
"Language"	0,200347	0	0
"Processing"	0.260658	0.174822	0.035025
"(NLP)"	0.260658	0.174822	0
"is"	0.260658	0.174822	0
"a"	0,200347	0	0.035025
"computerized"	0.260658	0.174822	0.035025
"way"	0.260658	0.174822	0.035025
"of"	0.260658	0.174822	0.035025
"analyzing"	0,200347	0.174822	0.035025
"texts."	0,200347	0	0

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

"NLP"	0,200347	0	0
"involves"	0,200347	0	0
"the"	0,200347	0	0
"acquisition"	0.260658	0.174822	0.035025
"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	0.035025
$\sum T\ C$ (D1)			

Table 3: Example of Cosine Similarity Processing Stage for the First Document of Journal Abstract Content

Term	Cosine similarity		Hasil ($\omega_Q(t_i)\ x\ \omega_D(t_i)$)
	Document 1	Document 2	
"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**

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

$$= 0.310201 + 0.280201 = \mathbf{0.5904}$$

The results of Tables 2 and 3 above show the results of text similarity p sed using the cosine formula of text similarity. The overall result of Document 1 shows the result of 0.310201, and the calculation result of Document 2 shows the same result of 0.280201. The value is obtained from the similarity between Document 1 and Document 2, which is calculated to get the same value so that when the results of D1 and D2 are summed up, the accuracy or match value is 0.5904. If the result is made in percentage form, the similarity value between the two documents is 59%. A similarity above 20% indicates that the correspondence between the content of the paper is accurate and relevant. The sentence interpretations are created in a simple and reader-friendly manner. However, if the representation results are below 0.2%, it suggests that the content of the research journal is not relevant and has low similarity.

CONCLUSION

This research unveiled the complex dynamics of effectiveness and efficiency in understanding natural language and assessing author consistency in journal writing. It leveraged standard NLP algorithms combined with TF-IDF and Cosine Similarity to analyze the relevance between journal content sections. The study involved a dataset of 60 documents out of 150 sample

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

documents, demonstrating different trends compared to another research. By performing comparative analysis using TF-IDF, a similarity percentage of 0.2% was achieved, indicating that results fall below 0.2%, they are considered less relevant, and author consistency needs to be improved. Performance in determining consistency with TF-IDF requires initial NLP preprocessing, resulting in structured text that is more effective in conveying meaning in a very understandable manner. However, the analysis results still need to be more accurate, which is due to the dependence on the quantity of text processed in the TF-IDF stage, estimated to take around 5 seconds per system run. When input texts have similarities, the relevance ranges from 50% to 68%. It emphasizes the ability of TF-IDF and Cosine Similarity algorithms to produce accurate results.

Furthermore, the sentence interpretations generated by NLP achieved an accuracy rate of 87% in the preprocessing performed within the system. The projected training duration for a dataset containing content journals has the potential for improved efficiency, focusing on faster processes related to text and natural language processing, which teaches machines to generate easily understandable natural language for readers. It is a text-mining approach using journal content prepared for submission to publication service providers and for assessing the consistency of published journals.

Reference

1) Adil, S. H., Ebrahim, M., Ali, S. S. A., & Raza, K. (2019). Identifying Trends in Data Science Articles using Text Mining. *2019 4th International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST)*, 1–7. https://doi.org/10.1109/ICEEST48626.2019.8981702

2) Akhter, M. P., Jiangbin, Z., Naqvi, I. R., Abdelmajeed, M., Mehmood, A., & Sadiq, M. T. (2020). Document-Level Text Classification Using Single-Layer Multisize Filters Convolutional Neural Network. *IEEE Access*, *8*, 42689–42707. https://doi.org/10.1109/ACCESS.2020.2976744

3) Amanullah, R. F., Utami, E., & Sunyoto, A. (2019). Citation Detection on Scientific Journal Using Support Vector Machine. *2019 International Conference on Information and Communications Technology (ICOIACT)*, 549–553. https://doi.org/10.1109/ICOIACT46704.2019.8938522

4) Batra, P., Chaudhary, S., Bhatt, K., Varshney, S., & Verma, S. (2020). A Review: Abstractive Text Summarization Techniques using NLP. *2020 International Conference on Advances in Computing, Communication & Materials (ICACCM)*, 23–28.

5) https://doi.org/10.1109/ICACCM50413.2020.9213079

6) de Oliveira, N. R., Pisa, P. S., Lopez, M. A., de Medeiros, D. S. v., & Mattos, D. M. F. (2021). Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges. *Information*, *12*(1), 38. https://doi.org/10.3390/info12010038

7) Gunasinghe, U. L. D. N., De Silva, W. A. M., de Silva, N. H. N. D., Perera, A. S., Sashika, W. A. D., & Premasiri, W. D. T. P. (2014). Sentence similarity measuring by vector space model. *2014 14th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 185–189. https://doi.org/10.1109/ICTER.2014.7083899

8) Gunawan, D., Amalia, A., & Charisma, I. (2017). Clustering articles in bahasa Indonesia using self-organizing map. *2017 International Conference on Electrical Engineering and Informatics (ICELTICs)*, 239–244. https://doi.org/10.1109/ICELTICS.2017.8253288

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

9) Huynh, H. T., Duong-Trung, N., Ha, X. S., Quynh Thi Tang, N., Huynh, H. X., & Quoc Truong, D. (2020). Automatic Keywords-based Classification of Vietnamese Texts. *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 1–3. https://doi.org/10.1109/RIVF48685.2020.9140761

10) Kamsties, E., Berry, D. M., Paech, B., Kamsties, E., Berry, D. M., & Paech, B. (2001). *Detecting Ambiguities in Requirements Documents Using Inspections*. http://www.vuse.vanderbilt.edu/

11) Kilany, R., Ammar, R., & Rajasekaran, S. (2018). A correlation-based algorithm for classifying technical articles. *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 050–053. https://doi.org/10.1109/ISSPIT.2011.6151534

12) Kuang, S., & Davison, B. D. (2018). Numeric-Attribute-Powered Sentence Embedding. *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 623–626. https://doi.org/10.1109/BigComp.2018.00110

13) Kupiyalova, A., Satybaldiyeva, R., & Aiaskarov, S. (2020). Semantic search using Natural Language Processing. *2020 IEEE 22nd Conference on Business Informatics (CBI)*, 96–100. https://doi.org/10.1109/CBI49978.2020.10065

14) Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment. *2016 4th International Conference on Cyber and IT Service Management*, 1–6. https://doi.org/10.1109/CITSM.2016.7577578

15) Liu, C., Sheng, Y., Wei, Z., & Yang, Y.-Q. (2018). Research of Text Classification Based on Improved TF-IDF Algorithm. *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)*, 218–222. https://doi.org/10.1109/IRCE.2018.8492945

16) Manalu, S. R., Willy, & Priatna, W. S. (2017). Development of review rating and reporting in open journal system. *2017 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 842–845. https://doi.org/10.1109/ECTICon.2017.8096370

17) Mardatillah, U., Zulfikar, W. B., Atmadja, A. R., Taufik, I., & Uriawan, W. (2021). Citation Analysis on Scientific Articles Using Cosine Similarity. *2021 7th International Conference on Wireless and Telematics (ICWT)*, 1–4. https://doi.org/10.1109/ICWT52862.2021.9678402

18) Masum, A. K. M., Abujar, S., Tusher, R. T. H., Faisal, F., & Hossain, S. A. (2019). Sentence Similarity Measurement for Bengali Abstractive Text Summarization. *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–5. https://doi.org/10.1109/ICCCNT45670.2019.8944571

19) Mia, Md. R., & Latiful Hoque, A. S. Md. (2019). Question Bank Similarity Searching System (QB3S) Using NLP and Information Retrieval Technique. *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, 1–7. https://doi.org/10.1109/ICASERT.2019.8934449

20) Mz, L. F., Tahir, Z., & Suyuti, A. (2023). Development of Software Cost Estimation and Resource Allocation Using Natural Language Processing, Cosine Similarity and Function Point. *2023 International Conference on Digital Applications, Transformation & Economy (ICDATE)*, 1–6. https://doi.org/10.1109/ICDATE58146.2023.10248788

21) Pattnaik, S., & Nayak, A. K. (2019). Summarization of Odia Text Document Using Cosine Similarity and Clustering. *2019 International Conference on Applied Machine Learning (ICAML)*, 143–146. https://doi.org/10.1109/ICAML48257.2019.00035

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

22) Pohl, A., Cosenza, B., & Juurlink, B. (2019). Portable Cost Modeling for Auto-Vectorizers. *2019 IEEE 27th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, 359–369.

23) https://doi.org/10.1109/MASCOTS.2019.00046

24) Rahmawati, D., & Khodra, M. L. (2017). Automatic multilabel classification for Indonesian news articles. *2017 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 1–6. https://doi.org/10.1109/ICAICTA.2015.7335382

25) Raza, G. M., Butt, Z. S., Latif, S., & Wahid, A. (2021). Sentiment Analysis on COVID Tweets: An Experimental Analysis on the Impact of Count Vectorizer and TF-IDF on Sentiment Predictions using Deep Learning Models. *2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, 1–6. https://doi.org/10.1109/ICoDT252288.2021.9441508

26) Ridwang, Ilham, A. A., Nurtanio, I., & Syafaruddin. (2020a). Image search optimization with web scraping, text processing and cosine similarity algorithms. *2020 IEEE International Conference on Communication, Networks and Satellite (Comnetsat)*, 346–350.

27) https://doi.org/10.1109/Comnetsat50391.2020.9328982

28) Ridwang, Ilham, A. A., Nurtanio, I., & Syafaruddin. (2020b). Image search optimization with web scraping, text processing and cosine similarity algorithms. *2020 IEEE International Conference on Communication, Networks and Satellite (Comnetsat)*, 346–350.

29) https://doi.org/10.1109/Comnetsat50391.2020.9328982

30) Ridwang, Ilham, A. A., Nurtanio, I., & Syafaruddin. (2020c). Image search optimization with web scraping, text processing and cosine similarity algorithms. *2020 IEEE International Conference on Communication, Networks and Satellite (Comnetsat)*, 346–350.

31) https://doi.org/10.1109/Comnetsat50391.2020.9328982

32) Rinartha, K., & Surya Kartika, L. G. (2019). Scientific Article Clustering Using String Similarity Concept. *2019 1st International Conference on Cybernetics and Intelligent System (ICORIS)*, 13–17. https://doi.org/10.1109/ICORIS.2019.8874879

33) Roul, R. K., Sahoo, J. K., & Arora, K. (2017). Modified TF-IDF Term Weighting Strategies for Text Categorization. *2017 14th IEEE India Council International Conference (INDICON)*, 1–6. https://doi.org/10.1109/INDICON.2017.8487593

34) Shin, J., Kim, Y., Yoon, S., & Jung, K. (2018). Contextual-CNN: A Novel Architecture Capturing Unified Meaning for Sentence Classification. *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 491–494. https://doi.org/10.1109/BigComp.2018.00079

35) Villanueva, A., Atibagos, C., De Guzman, J., Dela Cruz, J. C., Rosales, M., & Francisco, R. (2022). Application of Natural Language Processing for Phishing Detection Using Machine and Deep Learning Models. *2022 International Conference on ICT for Smart Society (ICISS)*, 01–06. https://doi.org/10.1109/ICISS55894.2022.9915037

36) Wibowo, M. D., Nurtanio, I., & Ilham, A. A. (2017). Indonesian sign language recognition using leap motion controller. *2017 11th International Conference on Information & Communication Technology and System (ICTS)*, 67–72. https://doi.org/10.1109/ICTS.2017.8265648

37) Xia, L., Luo, D., Zhang, C., & Wu, Z. (2019). A Survey of Topic Models in Text Classification. *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 244–250. https://doi.org/10.1109/ICAIBD.2019.8836970

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10209822

38) Xiaofan Lin. (2017). Text-mining based journal splitting. *Seventh International Conference on Document Analysis and Recognition, 2016. Proceedings.*, 1075–1079.

39) https://doi.org/10.1109/ICDAR.2003.1227822

40) Yao, L., Pengzhou, Z., & Chi, Z. (2019). Research on News Keyword Extraction Technology Based on TF-IDF and TextRank. *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, 452–455. https://doi.org/10.1109/ICIS46139.2019.8940293

41) Zeng, J., Ge, J., Zhou, Y., Feng, Y., Li, C., Li, Z., & Luo, B. (2017). Statutes Recommendation Based on Text Similarity. *2017 14th Web Information Systems and Applications Conference (WISA)*, 201–204. https://doi.org/10.1109/WISA.2017.52

42) ZHANG Guangrong1, W. B. H. Y. (2019). Real-Valued Conditional Restricted Boltzmann Machines with Tag for Recom-mendation Algorithm. *Journal Of Frontiers Of Compiter Science And Technology*, *Vol.13*(1), 138–146.

43) Zhang, W., Liu, F., Zhang, Z., Liu, S., & Huang, Q. (2020). Commodity Text Classification Based E-Commerce Category and Attribute Mining. *Proceedings - 3rd International Conference on*

44) *Multimedia Information Processing and Retrieval, MIPR 2020*, 105–108. https://doi.org/10.1109/MIPR49039.2020.00028