Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 4 Iss. 4 October 2024, pp: 1309-1319
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Content Classification of the Official Website of the Ministry of Foreign Affairs of the Republic of Indonesia (MoFA RI) using Vector Space Model (VSM)

Prima Bintang Bahtera[1*], Deni Sutendi Kartawijaya[2]

[1,2]Master of Science in Informatics Technology, President University, Indonesia

E-Mail: [1]prima.bahtera@student.president.ac.id, [2]deni.kartawijaya@student.president.ac.id

Received Apr 30th 2024; Revised Jun 7th 2024; Accepted Jul 16 th 2024
Corresponding Author: Prima Bintang Bahtera

Abstract

The official website of the Ministry of Foreign Affairs of the Republic of Indonesia (MoFA RI) is an important platform for disseminating information to a diverse audience. Efficiently categorizing the vast amount of content available on the website is essential for enhancing user experience and optimizing information retrieval. These categories will also become an identifier and topic classification based on the content inside the article. This study presents a systematic approach to content classification of the Official Website of the Ministry of Foreign Affairs of the Republic of Indonesia (MoFA RI) using the Vector Space Model (VSM). The methodology involves preprocessing the text data, constructing a term-document matrix, and implementing cosine similarity to measure the relevance of documents to predefined categories. The study demonstrates the effectiveness of VSM in accurately classifying content, thus facilitating streamlined access to information for users navigating the website. Furthermore, the findings offer insights into enhancing the organization and accessibility of governmental online platforms, contributing to improved user experience and information dissemination.

Keyword: Classification, Cosine Similarity, MoFA RI, TF-IDF, Vector Space Model

1. INTRODUCTION

The Ministry of Foreign Affairs of the Republic of Indonesia (MoFA RI) website serves as a critical repository of information on the nation's foreign policy, international relations, and diplomatic activities. It provides an abundance of resources for researchers, policymakers, and the public seeking insights into Indonesia's diplomacy discourse. However, navigating and extracting specific information from this extensive website can be a time-consuming task due to the sheer volume of content.

Automated content classification techniques are required to address this challenge. In the context of government websites, efficient content classification can significantly improve information accessibility for users. This allows users to quickly locate relevant information based on their needs. Various techniques have been employed for content classification, with the Vector Space Model (VSM) being a well-established and effective approach [1].

Furthermore, VSM represents textual documents as vectors in a high-dimensional space, where each dimension corresponds to a unique term present in the document collection. The weight associated with each term reflects its importance within the document [2]. Documents are then classified based on their similarity to pre-defined topic categories. This approach offers several advantages, including its simplicity, scalability, and effectiveness in handling large document collections. Term Frequency (TF) - Inverse - Document Frequency (IDF) has become one of the most well-known algorithms for text classification. It is also able to combine with other methods like Cosine Similarity to increase performance and achieve better results [3].

Cosine similarity is commonly used in the text classification process because of its computational efficiency and good performance [4] [5]. In order to process the data, the documents need to be weighted based on their frequency and their significance in the overall document by using the TF-IDF algorithm [5]. Cosine similarity is used to measure the degree of similarity, this method is a common method that is often used and combined with the TF-IDF [6].

Building upon prior works, our research focuses on implementing VSM using TF-IDF combined with Cosine Similarity to classify the content of the MoFA RI's official website. We described the implementation process, analyzed the classification results, and evaluated the effectiveness of VSM in categorizing the relevant information from the website. In addition, the MoFA RI website presents unique challenges due to the nature

of its content, which often includes diplomatic language and specific terminology related to international relations. Our research aims to evaluate the effectiveness of VSM in addressing these challenges and contribute to the development of robust content classification systems for government websites.

2. LITERATURE REVIEW

Content classification of governmental websites has gained significant attention in the field of information retrieval and machine learning. Researchers have explored various techniques and methodologies to automate the categorization of content on official websites, aiming to enhance user experience, information accessibility, and overall website usability.

There were many previous kinds of research that examined text classification, the Vector Space Model method, TF-IDF, and cosine similarity. (Mete Eminagaoglu, 2020) [7] proposed a new similarity measure for text classification and information retrieval. They conducted experiments using instance-based algorithms like k-NN and Rocchio with the new similarity measure. The results showed that the proposed similarity measure significantly improved the performance scores within both classifier algorithms, k-NN and Rocchio. Specifically, the new similarity measure led to higher F-score, precision, and recall values compared to other measures or distance metrics used in their experiments. The study indicated that the new similarity measure could be considered and used as an alternative measure for text classification among different languages. It was noted that the measure improved the classification accuracy of instance-based algorithms and vector space models like k-NN and Rocchio. Additionally, the new measure showed potential for being used in document clustering algorithms. The research demonstrated that the proposed similarity measure enhanced the performance of text classification algorithms, particularly k-NN and Rocchio, showcasing its potential for improving accuracy and efficiency in information retrieval tasks.

Another research by (Isa et al., 2008) [8] proposed a hybrid classification approach that combines the simplicity of the Bayes formula as a vectorizer with the effectiveness of text classifiers based on the vector space model, such as Support Vector Machine (SVM) and Self Organizing Map (SOM). This hybrid approach aims to enhance classification accuracy compared to using the traditional naïve Bayes method alone. The results of their experiments conducted show that the hybrid classification approaches, specifically the naïve Bayes-SVM hybrid approach, outperform the pure naïve Bayes classifiers in terms of classification accuracy. The naïve Bayes-SVM hybrid approach demonstrated a significant improvement in classification accuracy compared to the pure naïve Bayes model, especially when using different ranking schemes and specialized techniques. However, they noted that the naïve Bayes-SOM hybrid approach did not perform as well as expected, particularly when dealing with high-dimensional datasets like the 20 Newsgroups dataset. The limitations of the SOM algorithm in handling high-dimensional data, where the Euclidean distance measure becomes inaccurate for dimensions above four, contributed to the reduced classification accuracy in their scenario. The proposed hybrid classification approach combining the Bayes formula for vectorization with SVM or SOM classifiers at the back end shows promise in improving text document classification accuracy, with the SVM hybrid approach proving to be more effective in certain scenarios compared to the SOM hybrid approach.

Research conducted by Vahora et al., 2011 [9] they used the fusion of the Vector space model with the Naïve Bayes method for text classification. The Naïve Bayes classifier is utilized to build a vector of words found in the training set of documents and calculate probabilities for each category (e.g., Spam and Non-Spam). The approach involves creating hash tables for each category with word occurrences and total word counts to calculate probabilities for each word in a category. The experimental results showed the success rates of the spam classification system during the training and testing phases. The results indicate that as the size of the training set increases, the success ratio also increases, reaching a saturation point of around 85%. By providing personalized vectors for each user with specific word lists, the performance of spam detection can be significantly improved. Overall, the fusion of the Vector space model with Naïve Bayes provides reasonable accuracy in classifying spam mail, with the research achieving nearly 85% accuracy in spam classification.

Castells et al., 2007 [10] employed the utilization of an ontology-based knowledge base (KB) to enhance information retrieval over a large document corpus. The key steps in their approach include: Construction of a detailed and densely populated conceptual space in the form of an ontology-based KB; combining the flexibility and generality of an Information Retrieval (IR) model with the expressiveness and detail of a structured relational model; leveraging ontology-based approach for enhanced inferencing capabilities and semantic data integration; testing the system on a corpus of 145,316 documents from the Convolutional Neural Networks (CNN) website using the KIM domain ontology and KB, with manual additions for completeness; comparing the performance of the ontology-based retrieval model with a conventional keyword-only search using the Jakarta Lucene library; Evaluating the results based on a set of 20 manually prepared queries and subjective metrics on a scale from 0 to 5. The results of the experiments indicate that the semantic retrieval algorithm outperforms keyword-based search in certain cases, particularly when the query involves complex conditions that are not easily expressed through keywords alone. The study demonstrates improvements in retrieval effectiveness with the ontology-based approach, showcasing the potential for qualitative enhancement

over traditional keyword-based search methods. The precision and recall of the semantic search are shown to be promising, especially in regions of the ontology with high completeness in terms of instances and annotations. Overall, the study presented a novel approach to information retrieval by integrating ontology-based knowledge representation with traditional IR models, showing promising results in terms of retrieval effectiveness and the ability to handle complex search queries.

Hasugian, et al,. 2021 [3] used the utilization of TF-IDF and cosine similarity algorithms for classifying documents based on abstracts from scientific journals. The process involves calculating the weight of each document using TF-IDF and determining the similarity between documents using cosine similarity. The TF-IDF algorithm is used to assign weights to terms in documents based on their frequency and importance in the corpus. This helps in identifying key terms that differentiate documents and aid in classification. Cosine similarity, on the other hand, measures the similarity between two documents based on the angle between their feature vectors in a high-dimensional space. Higher cosine similarity values indicate greater similarity between documents. The results of the research show that the classification process based on TF-IDF and cosine similarity algorithms yielded promising outcomes. The study successfully classified documents into different topics, such as Data Mining, Decision Support Systems, Expert Systems, Artificial Neural Networks, Image Processing, and Cryptography based on the abstract descriptions. The calculated similarity values between test data and training data also provided insights into the dominant topics and the proximity of test data to existing documents. Overall, the research demonstrates the effectiveness of using TF-IDF and cosine similarity algorithms in classifying documents based on abstracts from scientific journals. The combination of these algorithms helps in extracting meaningful patterns, identifying relevant topics, and organizing documents into coherent groups for knowledge discovery and information retrieval purposes.

Another research by Singh et al., 2021 [11] employed three different methods to estimate the semantic similarity between two news articles on the same topic/event in Hindi and English. The methods used were: Cosine Similarity with TF-IDF vectors, Jaccard Similarity with TF-IDF vectors, and Bag of Words Euclidean Distance. Among these methods, the Cosine Similarity with TF-IDF vectors showed the best results with greater accuracy, recall, and F-measure scores of 81.25%, 100%, and 76.92%, respectively. The other two methods also showed promising results but could potentially be improved with the use of the Doc2Vec model. The results of the experiment indicated that the Cosine Similarity method using TF-IDF vectors outperformed the other methods in terms of accuracy and recall. This suggests that the Cosine Similarity method is effective in measuring the similarity between news articles in different languages and can be a valuable tool for content analysis and comparison.

The methods used by (Sintia et al., 2021) [12] for improving the accuracy of product codefication are Cosine Similarity and Weighted TF-IDF. Cosine Similarity is a method for calculating similarity by using keywords from the code of goods. It measures the cosine of the angle between two vectors, which helps in determining how similar two documents are based on their content. TF-IDF is a technique that assigns weights to terms in a document based on their frequency and importance in the corpus. It stands for Term Frequency-Inverse Document Frequency and is used to reflect how important a word is to a document in a collection or corpus. The results of the research show that the application of TF-IDF weighting and Cosine Similarity has successfully increased the accuracy in the search for goods codification. By using these methods, the search process in the SiPaGa application became more accurate than before. The combination of Cosine Similarity calculations and TF-IDF weighting after entering keywords for search led to improved accuracy in product codification. The research findings indicate that the highest value obtained through Cosine Similarity represents the codification of the item most similar to the keyword entered, thus enhancing the search process.

Another previous research by Swe et al., 2011 [13] used the methods of constructing a Content Structure Tree (CST) based on the Document Object Model (DOM) tree of web pages. The CST helps separate main content blocks from other blocks, and a cosine similarity measure is used to evaluate and rank the nodes in the CST to identify the most informative blocks. The results of the proposed algorithm are promising as it aims to extract relevant information from web pages efficiently. By utilizing the CST and cosine similarity measure, the algorithm can effectively identify and extract the main content blocks from web documents, providing a structured approach to information extraction. The effectiveness of the results can be assessed based on the accuracy of extracting relevant information compared to noisy blocks such as navigation panels, advertisements, etc. Further evaluation metrics or comparisons with existing methods could provide insights into the performance and efficiency of the proposed algorithm in extracting main content from web pages.

The Vector Space Model (VSM) has been widely employed in information retrieval tasks, including text classification and document clustering [8]. In VSM, documents and queries are represented as vectors in a high-dimensional space, where the cosine similarity between vectors measures their relevance. VSM-based approaches have since been applied to various domains, showcasing their effectiveness in content classification tasks [14].

TF-IDF is a statistical measure commonly used to evaluate the importance of a term within a document corpus. Term frequency (TF) measures the frequency of a term within a document, while inverse document

frequency (IDF) is the inverse frequency of documents that contain the word [2]. The complete formula of the TF-IDF is shown below:

$$tf_{t,d} = term\ t\ occurence\ in\ the\ document\ d \quad (1)$$

$$idf_t = \log(\frac{N}{df_t}) \quad (2)$$

$$tf\ idf_{t,d} = tf_{t,d}\ x\ idf_t \quad (3)$$

Description:

d	: document of-d
t	: the word of-t from the keyword
tfidf	: weight of the d document to the word of -t
tf	: amount of words that must be found in the document
idf	: Inversed Document Frequency
N	: Total amount of Documents
df	: amount of documents containing the word that we search

The Cosine Similarity method is a method for calculating the level of similarity between two or more objects or documents. It calculates the cosine of the angle between two vectors, representing their similarity in orientation within the vector space. Manning et al. (2008) highlighted the effectiveness of cosine similarity in information retrieval tasks, particularly for document ranking and relevance assessment. The advantage of Cosine Similarity is that the process of calculating the level of similarity between documents can be done quickly [2]. The formula of the Cosine Similarity can be seen below:

$$\cos\alpha = \frac{A.B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i\ x\ B}{\sqrt{\sum_{i=1}^{n} (A_i)^2}\ x\ \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (4)$$

Description:

A	: vector A
B	: vector B
A.B	: dot product between vector A and vector B
|A|	: the length of vector A
|B|	: the length of vector B
|A| |B|	: cross product between |A| and |B|

The text preprocessing stage holds a crucial role in transforming raw data into a more usable format. It involves some steps to eliminate incompleteness, inconsistencies, and irregularities within the data [15]. The steps in the text preprocessing stage are [16]:

1. Case folding
   Case Folding is the process of converting every capital letter into lowercase in a sentence. Case folding is conducted because not all documents are consistent with the use of capital or lowercase letters.

2. Tokenization
   Tokenizing is the process of cutting the input string in accordance with each word that composes a sentence. During this process, some characters are also considered as word separators, such as whitespace, enter, and period [17].

3. Stopwords Removal
   Within this step, we remove unnecessary or meaningless words. Usually, the words that are used as stopwords are stored in an array or text file. If the word that appears is the same as the word in the stop list, then the word will be removed from the document.

4. Stemming
   Stemming is the process of converting words into its basic form. The process carried out is like removing the suffix from each word.

Depending on the language of the dataset, some libraries can assist in the text preprocessing stage. The common library that can be employed for the process is WordNet. Then, for Bahasa Indonesia, we can use the Sastrawi library [18].

3. RESEARCH METHODOLOGY

This section will outline the methodology employed for classifying the content of the MoFA RI website using the Vector Space Model (VSM). In this case, the researcher will focus on implementing the TF-IDF as Feature Extraction and Cosine Similarity as the classifier method.

The complete step of the method that has been used in this research is shown in the figure below:

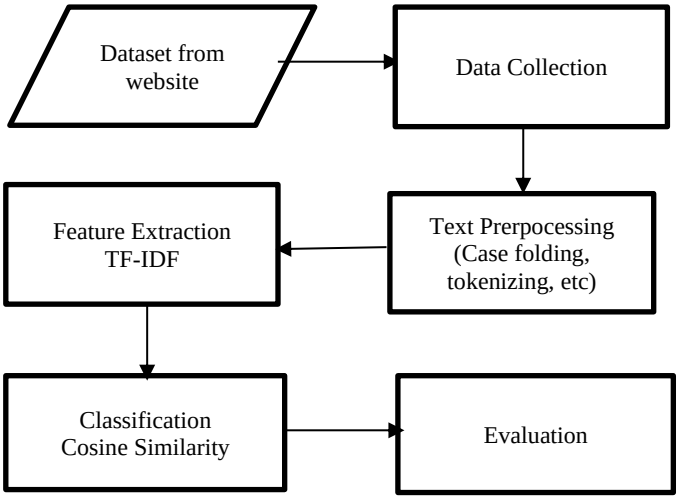

Figure 1. Methodology of Content Classification

3.1. Data Collection

The dataset used in this research is the text data gathered from the Official Website of the Ministry of Foreign Affairs of the Republic of Indonesia (MoFA RI). The dataset uses articles from website that represent some topics related to diplomatic discourse, and each article consists of 100 up to more than 500 words. To comply with the research objectives, the contents collected from the website were then saved into an Excel file. The total amount of data that has been used for the research is 60 articles, divided into 3 (three) categories. Then, split the dataset for the training and testing process with composition 30:10 [18].

For the training process, we added an additional column called Topic, which will be used as the main category of the documents. In this research, we used three topics: politik, ekonomi, and pelindungan to train the data.

Table 1. Sample of the dataset

Topic	Title	Content
Pelindungan	Pemulangan WNI Perempuan Pekerja Migran dari Singapura	Surabaya, Indonesia - Kementerian Luar Negeri telah memfasilitasi Pemulangan WNI/PMI a.n. Siti Zulaikhoh (SZ) dari Singapura di Bandara Juanda, Surabaya (17/4). Pemulangan tersebut dilakukan sebagai respon atas kasus KDRT yang dilakukan suami SZ terhadap Anak Kandungnya di Situbondo, Jawa Timur. Kasus kekerasan terhadap anak tersebut beredar luas di media sosial dan menjadi perhatian publik. Kementerian Sosial RI langsung berkoordinasi dengan Kemlu untuk segera memulangkan ibu korban yang masih bekerja di Singapura. Setibanya di Bandara Juanda, Surabaya, SZ diserahterimakan oleh Kasubdit Kawasan Asia Tenggara, Dit. PWNI, Kementerian Luar Negeri kepada Direktur Rehabilitasi Sosial Korban Bencana dan Kedaruratan, Kementerian Sosial untuk mendapatkan penanganan lebih lanjut. BP3MI dan Disnaker Jawa Timur turut pula memfasilitasi pemulangan tersebut. Selanjutnya PMI Siti akan diantar menuju Sentra Mahatmiya, Tabanan, Bali untuk mendapatkan proses Rehabilitasi Sosial yang dibutuhkan bersama dengan Anak dan Suaminya.

3.2. Text Preprocessing

The data needs to be cleaned up before continuing to the next process. Since the language that has been used in the article is Bahasa Indonesia, then we use the Sastrawi library for the Text Preprocessing stage. In this stage, there are some steps that had to be followed in order to get clean data, which are:

1. Case Folding

   During this step, all of the letters from the dataset will be converted to lowercase. See the example as table 2.

Table 2. Example of the Case Folding Process

Before Case Folding	After Case Folding
Surabaya, Indonesia - Kementerian Luar Negeri telah memfasilitasi Pemulangan WNI/PMI a.n. Siti Zulaikhoh (SZ) dari Singapura di Bandara Juanda, Surabaya (17/4). Pemulangan tersebut dilakukan sebagai respon atas kasus KDRT yang dilakukan suami SZ terhadap Anak Kandungnya di Situbondo, Jawa Timur.	surabaya, indonesia - kementerian luar negeri telah memfasilitasi pemulangan wni/pmi a.n. siti zulaikhoh (sz) dari singapura di bandara juanda, surabaya (17/4). pemulangan tersebut dilakukan sebagai respon atas kasus kdrt yang dilakukan suami sz terhadap anak kandungnya di situbondo, jawa timur.

2. Tokenization
   The words in each document will be tokenized to become a single word and converted into an array. Here are the examples of the result from the tokenization process:

   [‘surabaya’, ‘indonesia’, ‘kementerian’, ‘luar’, ‘negeri’, ‘telah’, … , ‘timur’]

3. Stopwords Removal
   With the reference from the Sastrawi library, we can remove the stopwords from the dataset, so the weighting process will exclude unnecessary words and only focus on the related keywords. The stopwords sample from the Sastrawi library are:

   ['ada', 'adalah', 'agak', 'agar', 'akan', 'amat', 'anda', 'antara', 'anu', 'apakah', 'apalagi', 'atau', 'bagaimanapun', 'bagi', 'bahwa', ….]

4. Stemming
   During this process, the words will be transformed into their basic form by removing the suffix and/or prefix. The basic form will use the Sastrawi library as the reference. Here is an example of the stemming process:

   kandungnya → kandung

3.3. Feature Extraction

In this research, the feature extraction process will utilize the TF-IDF technique to represent each document as a numerical vector. This is the stage of calculating word weighting to calculate the frequency of occurrence of each word in the test document in each document in the dataset [19]. The TF-IDF weighting stages can be seen in Figure 2.

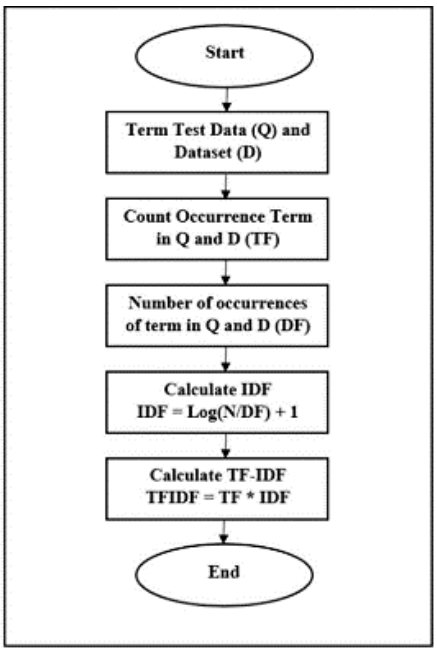

Figure 2. The step of weighting TF-IDF

In this stage, we calculate the TF-IDF score for each term in the document corpus, considering both term frequency and inverse document frequency. Then, we construct a term-document matrix where each row represents a document, and each column represents a term, with the cell values indicating the TF-IDF scores.

Table 3. Term Frequency (TF) Matrix

yogyakarta	yohanes	yordania	york	zealand	zeki	zona	zulaikhoh	angel
0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0
0	2	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0

Table 4. TF-IDF Matrix

yogyakarta	yohanes	yordania	york	zealand	zeki	zona	zulaikhoh	angel
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.111586	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.090646	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.076611	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.090531	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

3.4. Vector Space Model

In this process, we implement the Vector Space Model to represent documents and queries as vectors in a high-dimensional space. Then we compute the cosine similarity between document vectors to measure their similarity and relevance. We need to define a set of predefined categories or topics based on the thematic structure of MoFA RI's website. The cosine similarity metric is employed to measure the similarity between document vectors and category vectors. The cosine similarity between two vectors reflects the cosine of the angle between them in the high-dimensional space. Documents with higher cosine similarity to a particular category vector are classified as belonging to that category. The stages of the Cosine Similarity calculation can be seen in Figure 3.

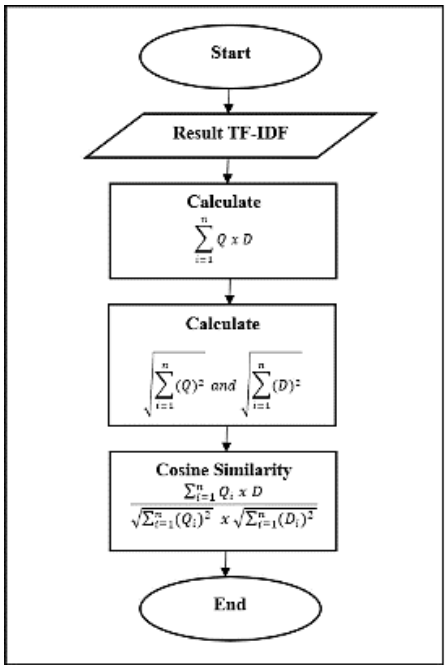

Figure 3. The stage of calculating the similarity with cosine similarity

Figure 4 illustrates the comparing similarity level of documents using the cosine degree concept, where vector coordinates as the documents that are compared and the cosine degree between vectors is the similarity degree. Based on the cosine principle, if cosine 0o is 1 and less than 1 to the value of another angle, then the value of the similarity of the two vectors is said to be similar when the value of the cosine similarity is 1 [6].

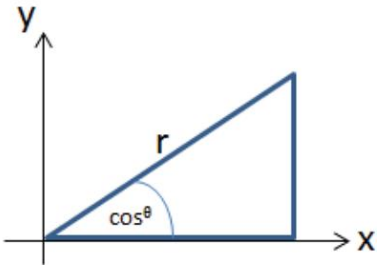

Figure 4. Cosine degree for similarity concept

Table 5. Cosine Similarity Matrix

	Dokumen 0	Dokumen 1	Dokumen 2	Dokumen 3	Dokumen 4	Dokumen 5	Dokumen 6	Dokumen 7	Dokumen 8	Dokumen 9	Label Train
0	0.080451	0.042243	0.171948	0.060968	0.055841	0.024293	0.067608	0.036289	0.09058	0.042476	Pelindungan
1	0.088731	0.130128	0.080677	0.128912	0.072921	0.091654	0.055544	0.064439	0.07024	0.142542	Politik
2	0.104985	0.104495	0.070006	0.151294	0.059427	0.150468	0.060999	0.051906	0.088618	0.101067	Ekonomi
3	0.063301	0.063744	0.050178	0.083058	0.041153	0.064142	0.055769	0.045081	0.067531	0.072165	Ekonomi
4	0.085072	0.108638	0.078785	0.104051	0.108558	0.072533	0.080916	0.038651	0.153942	0.080945	Pelindungan
5	0.09741	0.117908	0.203293	0.070318	0.06304	0.053842	0.103073	0.060824	0.120646	0.057423	Pelindungan
6	0.105861	0.129205	0.082159	0.12888	0.067415	0.063942	0.106077	0.037712	0.143952	0.073936	Pelindungan
7	0.173478	0.068084	0.108864	0.046521	0.083613	0.042951	0.116851	0.047953	0.132792	0.039052	Pelindungan
8	0.094931	0.10066	0.098963	0.121139	0.064671	0.059731	0.064846	0.118465	0.080222	0.078868	Politik
…	…	…	…	…	…	…	…	…	…	…	…
25	0.088803	0.116958	0.067015	0.089793	0.070088	0.060934	0.058132	0.039874	0.071282	0.126654	Politik
26	0.088612	0.142258	0.096058	0.085472	0.056603	0.04576	0.068702	0.127835	0.09993	0.069496	Politik
27	0.109324	0.078745	0.176965	0.057351	0.07858	0.038004	0.156404	0.052261	0.10847	0.042285	Pelindungan
28	0.079005	0.046174	0.0703	0.042886	0.032224	0.038066	0.067144	0.059589	0.058924	0.03541	Pelindungan
29	0.100636	0.094438	0.062989	0.123287	0.078162	0.105594	0.048478	0.055236	0.071067	0.137231	Ekonomi
Label Test	Pelindungan	Politik	Pelindungan	Ekonomi	Pelindungan	Ekonomi	Pelindungan	Politik	Pelindungan	Ekonomi	NaN

4. RESULT AND EVALUATION

4.1. Result

In this research, we conducted 4 (four) scenarios in order to check the performance and the result of the classification using TF-IDF and Cosine Similarity. From this experiment, it will find out the effect of the stemming and stopwords in the classification process. The scenarios are:

1. Without Stemming, include the stopwords

Table 6. Weight Matrix of the Scenario 1

yogyakarta	yohanes	yordania	york	zealand	zeki	zona	zulaikhoh	ángel
0	0	0	0	0	0	0	0.111586	0
0	0	0	0	0	0	0	0	0
0	0.090646	0	0	0	0	0	0	0
0	0	0	0	0	0.076611	0	0	0
0	0	0	0	0	0	0.090531	0	0

Table 7. Cosine Similarity Matrix of the Scenario 1

	Dokumen 0	Dokumen 1	Dokumen 2	Dokumen 3	Dokumen 4	Dokumen 5	Dokumen 6	Dokumen 7	Dokumen 8	Dokumen 9	Label Train
0	0.080451	0.042243	0.171948	0.060968	0.055841	0.024293	0.067608	0.036289	0.09058	0.042476	Pelindungan
1	0.088731	0.130128	0.080677	0.128912	0.072921	0.091654	0.055544	0.064439	0.07024	0.142542	Politik
2	0.104985	0.104495	0.070006	0.151294	0.059427	0.150468	0.060999	0.051906	0.088618	0.101067	Ekonomi
3	0.063301	0.063744	0.050178	0.083058	0.041153	0.064142	0.055769	0.045081	0.067531	0.072165	Ekonomi
4	0.085072	0.108638	0.078785	0.104051	0.108558	0.072533	0.080916	0.038651	0.153942	0.080945	Pelindungan
5	0.09741	0.117908	0.203293	0.070318	0.06304	0.053842	0.103073	0.060824	0.120646	0.057423	Pelindungan
Label Test	Pelindungan	Politik	Pelindungan	Ekonomi	Pelindungan	Ekonomi	Pelindungan	Politik	Pelindungan	Ekonomi	NaN

2. Without Stemming, exclude the stopwords

Table 8. Weight Matrix of the Scenario 2

yogyakarta	yohanes	yordania	york	zealand	zeki	zona	zulaikhoh	ángel
0	0	0	0	0	0	0	0.114971	0
0	0	0	0	0	0	0	0	0
0	0.093579	0	0	0	0	0	0	0
0	0	0	0	0	0.078523	0	0	0
0	0	0	0	0	0	0.093137	0	0

Table 9. Cosine Similarity Matrix of the Scenario 2

	Dokumen 0	Dokumen 1	Dokumen 2	Dokumen 3	Dokumen 4	Dokumen 5	Dokumen 6	Dokumen 7	Dokumen 8	Dokumen 9	Label Train
0	0.05697	0.026965	0.14505	0.035092	0.034501	0.007264	0.044418	0.007299	0.068185	0.017298	Pelindungan
1	0.056087	0.097225	0.046742	0.10007	0.047192	0.068776	0.038315	0.03395	0.042578	0.114567	Politik
2	0.063493	0.075027	0.03482	0.124152	0.033809	0.122843	0.03846	0.030448	0.050045	0.074423	Ekonomi
3	0.041394	0.037836	0.021945	0.058557	0.017498	0.040838	0.037379	0.025342	0.042998	0.04593	Ekonomi
4	0.066887	0.086655	0.054845	0.078639	0.091326	0.043661	0.062529	0.020782	0.124533	0.056585	Pelindungan
5	0.063155	0.09189	0.168103	0.045476	0.038088	0.025928	0.077287	0.031819	0.08319	0.0352	Pelindungan
Label Test	Pelindungan	Politik	Pelindungan	Ekonomi	Pelindungan	Ekonomi	Pelindungan	Politik	Pelindungan	Ekonomi	NaN

3. With Stemming, include the stopwords

Table 10. Weight Matrix of the Scenario 3

yogyakarta	yohanes	yordania	york	zealand	zeki	zona	zulaikhoh
0	0	0	0	0	0	0	0.11637
0	0	0	0	0	0	0	0
0	0.09441	0	0	0	0	0	0
0	0	0	0	0	0.08146	0	0
0	0	0	0	0	0	0.096643	0

Table 11. Cosine Similarity Matrix of the Scenario 3

	Dokumen 0	Dokumen 1	Dokumen 2	Dokumen 3	Dokumen 4	Dokumen 5	Dokumen 6	Dokumen 7	Dokumen 8	Dokumen 9	Label Train
0	0.105161	0.052459	0.199588	0.065019	0.059474	0.030348	0.076119	0.046603	0.110632	0.054998	Pelindungan
1	0.113008	0.150816	0.093424	0.172084	0.09906	0.116619	0.06482	0.081093	0.100894	0.167603	Politik
2	0.120145	0.124691	0.091966	0.183376	0.07551	0.164673	0.073104	0.061728	0.112115	0.114742	Ekonomi
3	0.11195	0.086636	0.065989	0.113563	0.045386	0.076459	0.072558	0.052528	0.102814	0.101103	Ekonomi
4	0.129684	0.128221	0.114872	0.134086	0.115234	0.084998	0.107056	0.04322	0.203842	0.097308	Pelindungan
5	0.159289	0.140867	0.237756	0.086212	0.078249	0.076075	0.131193	0.063574	0.16187	0.090952	Pelindungan
Label Test	Pelindungan	Politik	Pelindungan	Ekonomi	Pelindungan	Ekonomi	Pelindungan	Politik	Pelindungan	Ekonomi	NaN

4. With Stemming, exclude the stopwords

Table 12. Weight Matrix of the Scenario 4

yogyakarta	yohanes	yordania	york	zealand	zeki	zona	zulaikhoh
0	0	0	0	0	0	0	0.119586
0	0	0	0	0	0	0	0
0	0.097695	0	0	0	0	0	0
0	0	0	0	0	0.084123	0	0
0	0	0	0	0	0	0.099939	0

Table 13. Cosine Similarity Matrix of the Scenario 4

	Dokumen 0	Dokumen 1	Dokumen 2	Dokumen 3	Dokumen 4	Dokumen 5	Dokumen 6	Dokumen 7	Dokumen 8	Dokumen 9	Label Train
0	0.08003	0.03463	0.174478	0.041257	0.041703	0.009183	0.051344	0.01327	0.088676	0.030949	Pelindungan
1	0.075196	0.113555	0.059883	0.146658	0.06863	0.09406	0.046077	0.042663	0.072675	0.141749	Politik
2	0.07329	0.091331	0.055664	0.155909	0.045339	0.136432	0.049321	0.036816	0.070929	0.087519	Ekonomi
3	0.088672	0.060631	0.038074	0.084994	0.020381	0.049205	0.050145	0.031252	0.074969	0.07726	Ekonomi
4	0.111474	0.101181	0.085422	0.108419	0.096103	0.0583	0.088307	0.021751	0.172728	0.070355	Pelindungan
5	0.123645	0.102462	0.192495	0.059343	0.050047	0.043154	0.102513	0.031277	0.12026	0.068165	Pelindungan
Label Test	Pelindungan	Politik	Pelindungan	Ekonomi	Pelindungan	Ekonomi	Pelindungan	Politik	Pelindungan	Ekonomi	NaN

4.2. Evaluation

In the evaluation phase, the performance of the recommendation system will be assessed. In this research, performance evaluation metrics like Mean Average Precision (MAP), Precision@ K, and R-precision are employed. Mean Average Precision (MAP) is a comprehensive metric that evaluates the overall effectiveness of a ranking model. It takes into account both precision (proportion of relevant documents retrieved) and the order in which they are retrieved [20]. For each category in the MoFA RI website classification, we calculate the Average Precision (AP). AP represents the average precision at all retrieval positions where a relevant document is retrieved. To calculate AP, we consider the following:

1. Precision at each position where a relevant document is retrieved.
2. Number of relevant documents for a particular category.

MAP is then obtained by averaging the AP values across all categories. A higher MAP value indicates a model that retrieves relevant documents with higher precision and at higher ranks within the retrieved list. Precision@K (P@K) measures the precision of the retrieved documents at a specific cut-off point (K) in the ranked list. Precision@K measures the precision of the top K retrieved documents for each query. It assesses

the relevance of the top-ranked documents returned by the retrieval system. The evaluation process will calculate the precision at K by dividing the number of relevant documents among the top K retrieved documents by K.

R-Precision focuses on retrieving all relevant documents for a particular category. It is defined as the number of relevant documents retrieved divided by the total number of relevant documents in the collection. This metric is useful for assessing the model's ability to comprehensively capture all the relevant information for a specific category. It evaluates the precision of the retrieval system at the depth of relevant documents.

Table 14. Evaluation Matrix from all scenarios

Scenario	MAP	P@K	R-Precision
1	0.8456263227513225	[0.5, 0.4, 0.6, 0.7, 0.4, 0.8, 0.7, 0.5, 0.7, 0.6]	[0.5, 0.4, 0.6, 0.7, 0.4, 0.8, 0.7, 0.5, 0.7, 0.6]
2	0.8443667328042327	[0.6, 0.4, 0.7, 0.7, 0.5, 0.8, 0.7, 0.6, 0.8, 0.6]	[0.6, 0.4, 0.7, 0.7, 0.5, 0.8, 0.7, 0.6, 0.8, 0.6]
3	0.8762839191232048	[0.6, 0.5, 0.6, 0.7, 0.5, 0.8, 0.7, 0.4, 0.7, 0.6]	[0.6, 0.5, 0.6, 0.7, 0.5, 0.8, 0.7, 0.4, 0.7, 0.6]
4	0.8894644746787603	[0.6, 0.5, 0.6, 0.7, 0.5, 0.7, 0.7, 0.5, 0.8, 0.6]	[0.6, 0.5, 0.6, 0.7, 0.5, 0.7, 0.7, 0.5, 0.8, 0.6]

From Table 14, it can be concluded that the stemming and stopwords affected the precision level of the model. The table shows that using the basic form of the words and removing the stopwords can increase the precision level.

5. CONCLUSION AND FUTURE WORK

This study's result highlights that the deployment of the Vector Space Model (VSM) algorithm has been shown to be effective for content classification on the official website of the MOFA RI. The TD-IDF and Cosine Similarity methods could be used to calculate the degree of similarity in the content of the MoFA RI website in order to classify based on given topics or categories. The research has also shown that utilizing VSM in the efficient retrieval and categorization of information can improve user experience and help with information dissemination. Nevertheless, the preprocessing stage affected the precision level of the result, so it needs to be considered an important parameter during the process.

For future works, we need to improve the performance of the recommended system. Future research should focus on expanding and enriching the reference dataset with comprehensive and pertinent information. We should also consider exploring the integration of the VSM algorithm with other machine learning techniques, such as deep learning models, ensemble methods, or any other system, to leverage their complementary strengths and enhance classification accuracy for content classification on the MoFA RI website. This will enable the system to generate more precise and tailored recommendations.

REFERENCES

[1] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Survey*, vol. 34, pp. 1–47, Mar. 2002, [Online]. Available: www.ira.uka.de/bibliography/Ai/automated.text.

[2] Christopher. D. Manning, P. Raghavan, and H. Schutze, *An Introduction to Information Retrieval*, Online Edition. Cambridge: Cambridge University Press, 2009.

[3] P. M. Hasugian, J. Manurung, Logaraz, and U. Ram, "IMPLEMENTATION OF TF-IDF AND COSINE SIMILARITY ALGORITHMS FOR CLASSIFICATION OF DOCUMENTS BASED ON ABSTRACT SCIENTIFIC JOURNALS," *JURNAL INFOKUM*, vol. 9, no. Juni, Jun. 2021.

[4] K. Park, J. S. Hong, and W. Kim, "A Methodology Combining Cosine Similarity with Classifier for Text Classification," *Applied Artificial Intelligence*, vol. 34, no. 5, pp. 396–411, Apr. 2020, doi: 10.1080/08839514.2020.1723868.

[5] M. Umadevi, "DOCUMENT COMPARISON BASED ON TF-IDF METRIC," *International Research Journal of Engineering and Technology*, 2020, [Online]. Available: www.irjet.net

[6] A. Rizqi Lahitani, A. Erna Permanasari, and N. Akhmad Setiawan, "Cosine Similarity to Determine Similarity Measure: Study Case in Online Essay Assessment," 2016. doi: 10.1109/CITSM.2016.7577578.

[7] M. Eminagaoglu, "A new similarity measure for vector space models in text classification and information retrieval," *J Inf Sci*, vol. 48, no. 4, pp. 463–476, Aug. 2022, doi: 10.1177/0165551520968055.

[8] D. Isa, L. Hong, V. P. Kallimani, and R. Rajkumar, "Text Document Pre-Processing Using the Bayes Formula for Classification Based on the Vector Space Model," 2008.

[9] S. Vahora, M. Hasan, and R. Lakhani, "Novel Approach: Naïve Bayes with Vector Space Model for Spam Classification," *IEEE*, 2011.

[10] P. Castells, M. Ferná Ndez, and D. Vallet, "An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval," 2007.
[11] R. Singh and S. Singh, "Text Similarity Measures in News Articles by Vector Space Model Using NLP," *Journal of The Institution of Engineers (India): Series B*, vol. 102, no. 2, pp. 329–338, Apr. 2021, doi: 10.1007/s40031-020-00501-5.
[12] Sintia, S. Defit, and G. Widi Nurcahyo, "PRODUCT CODEFICATION ACCURACY WITH COSINE SIMILARITY AND WEIGHTED TERM FREQUENCY AND INVERSE DOCUMENT FREQUENCY (TF-IDF)," 2021.
[13] S. S. Nyein, *Mining Contents in Web Page Using Cosine Similarity*. University of Computer Studies, Mandalay, 2011.
[14] M. Hay, W. Oo, and P. Pa, "Myanmar News Retrieval in Vector Space Model using Cosine Similarity Measure," 2020.
[15] A. Hiro Juni Permana and A. Toto Wibowo, "Movie Recommendation System Based on Synopsis Using Content-Based Filtering with TF-IDF and Cosine Similarity," *Intl. Journal on ICT*, vol. 9, no. 2, pp. 1–14, 2023, doi: 10.21108/ijoict.v9i2.747.
[16] D. Meidelfi, I. Rahmayuni, T. Hidayat, and D. Chandra, "TF-IDF Implementation for Similarity Checker on The Final Project Title," 2021.
[17] M. Artama, I. N. Sukajaya, and G. Indrawan, "Classification of official letters using TF-IDF method," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Jun. 2020. doi: 10.1088/1742-6596/1516/1/012001.
[18] A. Prasetyo, B. D. Septianto, G. F. Shidik, and A. Z. Fanani, *Evaluation of Feature Extraction TF-IDF in Indonesian Hoax News Classification*. IEEE, 2019.
[19] C.-Z. Liu, Y.-X. Sheng, Z.-Q. Wei, and Y.-Q. Yang, "Research of Text Classification Based on Improved TF-IDF Algorithm," 2018.
[20] Y. Yue, T. Finley, F. Radlinski, and T. Joachims, "A Support Vector Method for Optimizing Average Precision," *SIGIR*, 2007.