CENTIVE
VOL. 3, NO. 1, PP.001-010, DEC 2023

E-ISSN : 3031-8548

Conference on Electrical Engineering, Informatics, Industrial Technology, and Creative Media 2023

The Visualization of the Vector Space Model in Searching for Immigration News in the East Nusa Tenggara Region

Juventinho Jose M. De Araujo [#1], Natalia Magdalena R. Mamulak [#2], Alfry Aristo Jansen Sinlae [#3]

[# 1-3] *Department of Computer Science, Faculty of Engineering, Widya Mandira Catholic University*
San Juan Street, Penfui, Kupang

[1] araujojuventinho@gmail.com
[2] mamulaknatalia@unwira.ac.id
[3] alfry.aj@unwira.ac.id

Received on dd-mm-yyyy, revised on dd-mm-yyyy, accepted on dd-mm-yyyy

Abstract

The Immigration Office is a public service agency involved in various activities, many of which are documented and published in the form of news articles. The sheer volume of these published articles can create challenges when trying to locate specific ones. One approach to improve search efficiency is through ranking, a subfield of information retrieval. Information retrieval involves the process of finding materials, typically documents, within an unstructured dataset, often consisting of text, to fulfill information requirements from a large collection. One technique for document retrieval is the utilization of the Vector Space Model (VSM). VSM employs principles from linear algebra, particularly the vector space, to develop a document model for conducting searches for the desired documents. A column vector representation is used to transform input documents. Another key concept is measuring the proximity between two vectors by calculating the angle they form and then sorting the data from the smallest to the largest angle. This establishes the ranking order, from the most relevant to the least relevant documents. Among the weighting algorithms, the tf-idf algorithm stands out, as it considers the frequency of word occurrences in each online document and the frequency of online documents containing the word. This study elucidates the visualization of the VSM in the search for documents related to immigration in the East Nusa Tenggara region.

Keywords: Immigration, Information Retrieval, TF-IDF, Vector Space Model, Visualization

Corresponding Author:
Alfry Aristo Jansen Sinlae
Department of Computer Science, Faculty of Engineering, Widya Mandira Catholic University
San Juan Street, Penfui, 85361, Kupang, East Nusa Tenggara, Indonesia
Email: alfry.aj@unwira.ac.id

I. INTRODUCTION

With the rapid advancement in computer technology usage, both in companies and in the field of education, an increasing number of documents are in digital format [1]. Such a large volume of digital documents necessitates a mechanism for users to search for or retrieve documents that meet their needs quickly and easily. Without this, the obtained information may not align with the information sought [2].

One way to retrieve desired information from a large collection of documents is through ranking. Ranking is a branch of information retrieval commonly employed in document retrieval, information filtering, online ad placement, and other applications [3].

A method for ranking documents involves using the Vector Space Model (VSM). The Vector Space Model is a system used to measure the similarity between a document and a query. VSM employs the concept of linear algebra, specifically vector spaces. Using this concept, we can determine the proximity between two vectors by calculating the angle formed between them, and then arrange the data in ascending order of the angles. This indicates the ranking of data from the most relevant to the least relevant [4].

The Vector Space Model can be applied to search a vast number of short articles or news articles. One institution frequently covered in the daily news is the Immigration Office. Immigration is a public service operating under the Ministry of Law and Human Rights (Kemenkumham) in Indonesia. In the East Nusa Tenggara (NTT) region, there are three branches of the Immigration Office: Class I Kupang, Class II Maumere, and Class II Atambua. As a public service entity, immigration offices engage in numerous activities, which are usually documented and published in news articles. These news articles are disseminated through official websites and mass media.

Research using the Vector Space Model method that has been conducted includes Lake (2015) on an information retrieval system for abstract thesis text documents using the tf-idf weighting method and measuring similarity using the vector space model. This study addressed how to assist computer science students at Widya Mandira Catholic University in finding the desired thesis documents. The result was an information retrieval system application capable of helping students search for thesis documents within the Computer Science program based on desired queries [5].

Maarif (2015) on the application of the TF-IDF algorithm for searching scientific works This research tackled how to apply the TF-IDF algorithm for searching scientific works. The outcome was an application for scientific works, including the weight values of each found scientific work [6].

Wisnu et al. (2015) on the design of an information retrieval (IR) system for finding the main ideas in English-language article texts using the vector space model weighting. This research focused on how to utilize information retrieval in text mining to discover the main ideas in English-language article texts. The result was an automatic main idea search system that helps readers better understand the content of articles [7].

Mustain et al. (2015) on an application for searching thesis journal papers using the vector space model. This study addressed how to implement information retrieval to find abstracts of thesis publication manuscripts like the user's input query using the vector space model. The result was an information retrieval system using the vector space model successfully applied to the search for abstracts of Computer Science thesis papers for the years 2011–2012, which significantly facilitated and sped up the search process [8].

Based on previous research on information retrieval systems, this study refers to Lake's research on an information retrieval system for abstract thesis documents using the tf-idf weighting method and measuring similarity using the vector space model because this study uses the same method and weighting. Lake's research developed an information retrieval system capable of displaying thesis documents within the Computer Science program. This study will provide a detailed explanation of the vector space model and tf-idf weighting in document retrieval based on a query, which will be visualized in graphical form.

Information retrieval is the process of finding the necessary information in an information storage and retrieval system. Information retrieval systems require user information needs, structured information or data containing organized information, and appropriate search strategies to retrieve relevant documents. An information retrieval system consists of three main components: input, processor, and output. According to Manning et al. (2009), information retrieval (IR) is about finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from a large collection (usually stored on a computer) [9].

The Vector Space Model is a model used to measure the similarity between documents and queries, representing each document in a collection as a point in space. It is typically used in information filtering, information retrieval, indexing, and ranking relevant documents. If there are n different words forming the word vocabulary or term index, these words will form a vector space with a dimension of n [10].

The index is a language used in a conventional book to find information based on words or terms that refer to a page. By using an index, information seekers can easily find the information they are looking for. In an information retrieval system, this index is used to represent information within a document [10].

The indexing stages include: Phrasing: This process transforms the document into a collection of terms by removing all punctuation characters in the document and converting the term collection to lowercase; Stop word Removal: This process eliminates stop words such as 'but,' 'is,' 'while,' and so on;

Stemming: The process of removing or reducing a word to its base form; Term Weighting and Inverted File: This process assigns weights to terms [11].

The TF-IDF (Term Frequency - Inverse Document Frequency) method is a way to give weight to the relationship between a word (term) and a document. This method combines two concepts for weight calculation: the frequency of a word appearing in a specific document and the inverse frequency of documents containing the word. The Term Frequency (tf) formula is based on the word's frequency in a document [11].

Due to the large number of immigration-related news articles, the vector space model can be applied to search for immigration-related news articles for visualization purposes. Based on the background provided, the problem at hand is how to visualize the Vector Space Model method for searching immigration-related news articles in the East Nusa Tenggara region. The research objective is to visualize the Vector Space Model method in the search for immigration-related news articles in the East Nusa Tenggara region (NTT).

II. RESEARCH METHOD

The development of a product requires a development model, and the stages in software development are often referred to as Software Development Life Cycle (SDLC). The model used in this information retrieval system is the SDLC waterfall model. This model is illustrated, as shown in the following diagram:

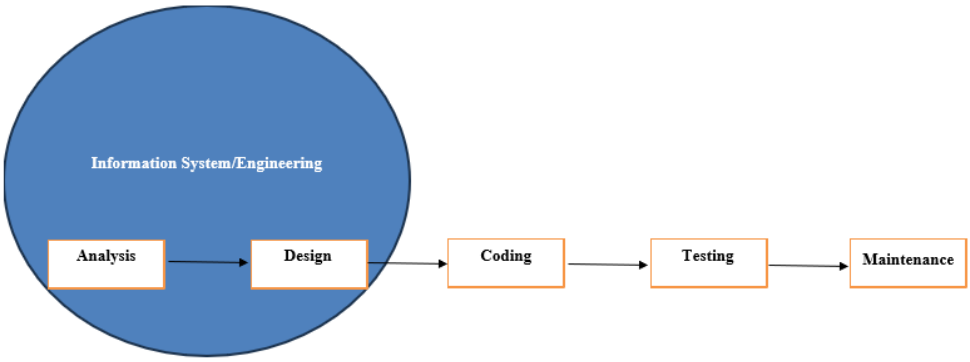

Fig. 1. Waterfall Model

Based on figure 1, in the analysis phase, an examination is conducted on the concept of information retrieval, analyzing the methods or algorithms employed, the programming language used, and suitable data for conducting searches using the vector space model method. In this phase, data used in the system is also collected, specifically news articles related to immigration in the East Nusa Tenggara region. These news articles will be utilized in the search process using the vector space model method.

Apart from the data, supporting reference materials are also gathered for the research process. These references include topics related to information retrieval introduction, determination of indexing and tf-idf weighting, and visualization techniques used in the vector space model method.

In design phase, the design of the input and output interfaces for the system used in the vector space model search experiment is carried out. Additionally, the design of the database used within the system is also conducted. In coding stage, the program code is developed to conduct the experimental search for news using the vector space model method.

In the testing stage, a system test is conducted to search for news that has been input into the database using the created system. It is not uncommon for software to change once it has been delivered to the user. Changes may occur due to errors that were not detected during the testing phase. The maintenance phase can involve revisiting the development process, starting from analyzing specifications to making changes to existing software. However, this phase is not meant for creating entirely new software.

III. RESULTS AND DISCUSSION

To facilitate the search for desired news articles, an information retrieval system using the vector space model method is required. The weighting used in this method is the term frequency-inverse document frequency weighting. The stages of this research can be seen in Figure 1.

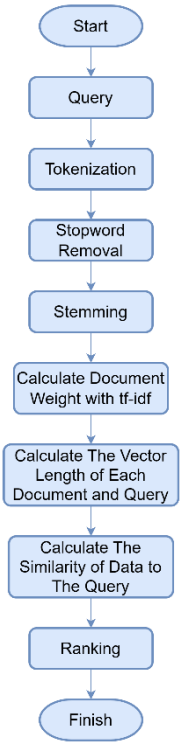

Fig. 2. Research stages

Based on Figure 1, the initial step in completing this research involves the tokenization process. During this stage, each word within every document is extracted and separated into individual tokens. Additionally, stopwords are removed, and each word is converted to lowercase during this phase.

The second step involves the stemming process. During this stage, each token is transformed into its base form. Subsequently, the tokens are converted into terms.

The third step is the indexing process. In this stage, the existing terms are indexed into a table to calculate the frequency of occurrence of a term in a document, obtained from the previously transformed tokens into their base forms. After obtaining the term frequency, the next step is to assign weights to each term using the term frequency-inverse document frequency (tf-idf) weighting. The formula for calculating the weight is as follows:

$$W_{ij} = tf_{ij} * idf_j \; ; \; idf_j = \log\left(\frac{N}{df_j}\right) \quad (1)$$

Where: W_{ij} is the weight of term *j* in document *i*; tf_{ij} is the term frequency of term *j* in document *i*; idf_j is the inverse of the document frequency value for term *j*; N is the total number of documents; and df_j is the number of documents containing term *j*.

After obtaining the index table, the next step is to calculate the length of the vector for each document. The vector length for document D1, denoted as |D1|, is the square root of the sum of the squared weights for each term present in document D1. This step is repeated for documents D2 to D20.

After performing the calculations, the results are obtained as shown in the following table:

Table 1. Cosine similarity

No	CoSim	Dok_ID
1	0.304081	1
2	0.211513	14
3	0.188451	4
4	0.135469	13
5	0.131694	2
6	0.0300581	10

The next step is to calculate the similarity between the input query and each document (D1–D20) using cosine similarity.

The simplified formula for calculating cosine similarity (S(Q,D)) between query (Q) and a document (D) is as follows:

$$S\ (Q, D) = \frac{\sum_t (W_{t,Q} . W_{t,D})}{\sqrt{\sum_t \left(w_{t,Q}^2\right)} . \sqrt{\sum_t (w_{t,D}^2)}} \quad (2)$$

Where: $w_{t,Q}$ is the weight of term t in the query Q; $w_{t,D}$ is the weight of term t in the Document D; $\sum$ is the symbol for summation; and $\sqrt{}$ is square root.

Before calculating the similarity between the query and the document, the query also undergoes processes such as tokenization, stemming, and other steps similar to those applied to the document.

As an example, a news search was conducted with the query "sosialisasi penerbitan dpri," resulting in an index table for the query displayed in Table 2. The steps to accomplish this are as follows:

a. Tokenization: Break down the query into individual terms. For example:
   "sosialisasi"
   "penerbitan"
   "dpri"
b. Stemming: Convert each term to its base or root form.
   "sosialisasi" becomes ***"sosialisasi"***
   "penerbitan" becomes ***"terbit"***
   "dpri" remains ***"dpri"*** (assuming it's not a stemmable term)
c. Calculate Weights: Assign weights to each term based on a method such as tf-idf.
d. Create Index Table: Create a table with columns for terms, documents, and weights. Populate this table with the calculated weights for each term in each relevant document.

Table 2. Index for queries

TERM	TF	DF	IDF	TF-IDF
sosialisasi	1	1	0,69897	0,69897
terbit	1	1	0,69897	0,69897
dpri	1	1	1	1

$$|\ Q\ | = \sqrt{0{,}69897^2 + 0{,}69897^2 + 1^2}$$

$$= \sqrt{0{,}488559067 + 0{,}488559067 + 1}$$

$$= \sqrt{1{,}977118134}$$

$$= 1{,}40610033$$

The documents whose terms intersect with the terms from the query are as follows:
"sosialisasi" : D1, D2, D4, D14
"terbit" : D1, D10, D13, D14
"dpri" : D1, D4
Therefore, the similarity calculation is only performed for documents D1, D2, D4, D10, D13, and D14. Other documents are not calculated because there is no similarity (similarity values are equal to zero).

$$(Q, D1) = \left(TF.IDF_{sosialisasi\ dalam\ Q} * TF.IDF_{sosialisasi\ dalam\ D1}\right) + \left(TF.IDF_{terbit\ dalam\ Q} * TF.IDF_{terbit\ dalam\ D1}\right) + \left(TF.IDF_{dpri\ dalam\ Q} * TF.IDF_{dpri\ dalam\ D1}\right) /\ (|\ Q\ | * |\ D1\ |)$$

$$= \frac{(0{,}69897 * 0{,}69897) + (0{,}69897 * 1{,}39794001) + (1 * 3)}{(1{,}40610033 * 9{,}98257565)}$$

$$= \frac{4{,}4656772}{14{,}0365029} = 0{,}31814742$$

$$S(Q, D2) = \left(TF.IDF_{sosialisasi\ dalam\ Q} * TF.IDF_{sosialisasi\ dalam\ D2}\right) + \left(TF.IDF_{terbit\ dalam\ Q} * TF.IDF_{terbit\ dalam\ D2}\right) + \left(TF.IDF_{dpri\ dalam\ Q} * TF.IDF_{dpri\ dalam\ D2}\right) /\ (|\ Q\ | * |\ D2\ |)$$

$$= \frac{(0{,}69897 * 4{,}19382003) + (0{,}69897 * 0) + (1 * 0)}{(1{,}40610033 * 13{,}67205401)}$$

$$= \frac{2{,}93135440}{19{,}2242797} = 0{,}15248189$$

$$S(Q,D4) = \left(TF.IDF_{sosialisasi\ dalam\ Q} * TF.IDF_{sosialisasi\ dalam\ D4}\right) + \left(TF.IDF_{terbit\ dalam\ Q} * TF.IDF_{terbit\ dalam\ D4}\right) + \left(TF.IDF_{dpri\ dalam\ Q} * TF.IDF_{dpri\ dalam\ D4}\right) / \ (|\ Q\ |*|\ D4\ |)$$

$$= \frac{(0{,}69897 * 2{,}79588002) + (0{,}69897 * 0) + (1 * 3)}{(1{,}40610033 * 15{,}90691751)}$$

$$= \frac{4{,}95423627}{22{,}366722} = 0{,}22150033$$

$$S(Q,D10) = \left(TF.IDF_{sosialisasi\ dalam\ Q} * TF.IDF_{sosialisasi\ dalam\ D10}\right) + \left(TF.IDF_{terbit\ dalam\ Q} * TF.IDF_{terbit\ dalam\ D10}\right) + \left(TF.IDF_{dpri\ dalam\ Q} * TF.IDF_{dpri\ dalam\ D10}\right) / \ (|\ Q\ |*|\ D10\ |)$$

$$= \frac{(0{,}69897 * 0) + (0{,}69897 * 0{,}69897) + (1 * 0)}{(1{,}40610033 * 10{,}15867498)}$$

$$= \frac{0{,}48855907}{14{,}2841162} = 0{,}03420296$$

$$S(Q,D13) = \left(TF.IDF_{sosialisasi\ dalam\ Q} * TF.IDF_{sosialisasi\ dalam\ D13}\right) + \left(TF.IDF_{terbit\ dalam\ Q} * TF.IDF_{terbit\ dalam\ D13}\right) + \left(TF.IDF_{dpri\ dalam\ Q} * TF.IDF_{dpri\ dalam\ D13}\right) / \ (|\ Q\ |*|\ D13\ |)$$

$$= \frac{(0{,}69897 * 0) + (0{,}69897 * 2{,}09691001) + (1 * 0)}{(1{,}40610033 * 6{,}81585650)}$$

$$= \frac{1{,}46567720}{9{,}58377808} = 0{,}15293313$$

$$S(Q,D14) = \left(TF.IDF_{sosialisasi\ dalam\ Q} * TF.IDF_{sosialisasi\ dalam\ D14}\right) + \left(TF.IDF_{terbit\ dalam\ Q} * TF.IDF_{terbit\ dalam\ D14}\right) + \left(TF.IDF_{dpri\ dalam\ Q} * TF.IDF_{dpri\ dalam\ D14}\right) / \ (|\ Q\ |*|\ D14\ |)$$

$$= \frac{(0{,}69897 * 0{,}69897) + (0{,}69897 * 2{,}7958802) + (1 * 0)}{(1{,}40610033 * 7{,}11884059)}$$

$$= \frac{2{,}44279533}{10{,}0098041} = 0{,}24404027$$

The final step is to sort the results of the cosine similarity calculations from largest to smallest. The documents will be displayed as shown in Table 3.

Table 3. The result of calculating the similarity of each document

NO	COSIM	DOK_ID
1	0,31814742	1
2	0,24404027	14
3	0,22150033	4
4	0,15293313	13
5	0,15248189	2
6	0,03420296	10

Because the similarity between documents 1, 14, 4, 13, 2, and 20 with the query is zero, they should not be returned to the user (they do not appear in the ranking list).

After manually calculating the query "sosialisasi penerbitan dpri," the system is implemented. The search results for the query "sosialisasi penerbitan dpri" are shown in the following image:

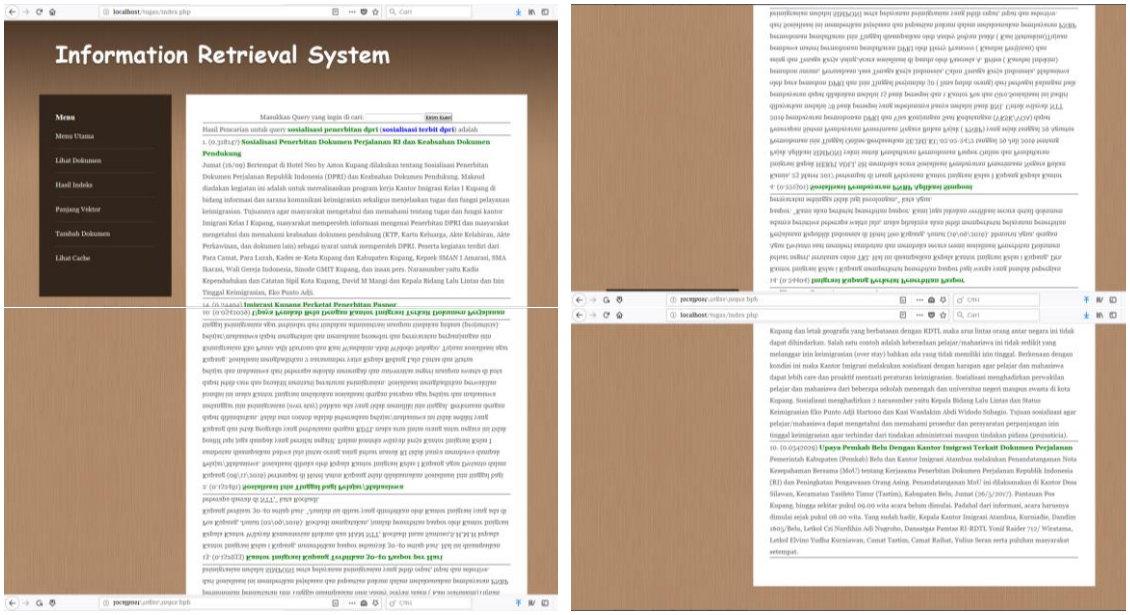

Fig 3. Search results against the query " sosialisasi penerbitan dpri"

The results obtained from the system are shown in table 4.

Table 4. System calculation results

No	CoSim	Dok_ID
1	0.304081	1
2	0.211513	14
3	0.188451	4
4	0.135469	13
5	0.131694	2
6	0.0300581	10

IV. CONCLUSION

Based on the designed system and conducted testing, several conclusions can be drawn. Firstly, the representation of documents involves preprocessing steps, such as the removal of non-essential words (stopwords), followed by indexing using term frequency-inverse document frequency (Tf/Idf) weighting. The resulting document representation is stored in an index table. The process of calculating document similarity with a user query mirrors the preprocessing steps applied to documents. The similarity calculation is performed using the vector space model formula. In testing, when the query "sosialisasi penerbitan dpri" is entered, the system returns six documents ranked by their similarity levels. Document ID 1 exhibits the highest similarity at 0.318147, followed by Document ID 14 (0.24404), Document ID 4 (0.221501), Document ID 13 (0.152933), Document ID 2 (0.152481), and Document ID 10 (0.0342029).

ACKNOWLEDGMENT

The authors would like to thanks The Immigration Office Class I Kupang has been instrumental in providing essential data for the smooth progress of this research.

REFERENCES

[1] Y. K. Dwivedi *et al.*, “Opinion Paper: ‘So what if ChatGPT wrote it?’ Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy,” *Int. J. Inf. Manage.*, vol. 71, p. 102642, Aug. 2023, doi: 10.1016/J.IJINFOMGT.2023.102642.

[2] A. Haleem, M. Javaid, M. A. Qadri, and R. Suman, “Understanding the role of digital technologies in education: A review,” *Sustain. Oper. Comput.*, vol. 3, pp. 275–285, Jan. 2022, doi: 10.1016/J.SUSOC.2022.05.004.

[3] E. Yulianti and L. Rahadianti, “Determining subject headings of documents using information retrieval models,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 23, no. 2, pp. 1049–1058, Aug. 2021,

doi: 10.11591/ijeecs.v23.i2.pp1049-1058.
[4] A. F. Waliyyul Haq, E. Carinia, S. Supian, and S. Subiyanto, "Detecting Similarities in Posts Using Vector Space and Matrix," *Int. J. Glob. Oper. Res.*, vol. 1, no. 3, pp. 103–108, 2020, doi: 10.47194/ijgor.v1i3.53.
[5] Louisa Kristofora Lake, "Information Retrieval System Dokumen Teks Abstrak Skripsi dengan Metode Pembobotan Tf-Idf dan Pengukuran Similarity Menggunakan Vector Space Model," Universitas Katolik Widya Mandira, 2015.
[6] A. A. Maarif, "Penerapan Algoritma TF-IDF Untuk Pencarian Karya Ilmiah," Universitas Dian Nuswantoro, 2015.
[7] D. W. B. and A. Hetami, "Perancangan Information Retrieval (IR) Untuk Pencarian Ide Pokok Teks Artikel Berbahasa Inggris dengn Pembobotan Vector Space Model," *J. Ilm. Teknol. Inf. Asia*, vol. 9, no. 1, pp. 53–59, 2015, [Online]. Available: https://jurnal.stmikasia.ac.id/index.php/jitika/article/view/118
[8] Mustain, "Aplikasi Pencarian Jurnal Skripsi Menggunakan Metode Vector Space Model (Model Ruang Vector)," Universitas Muhammadiyah Gresik, 2013.
[9] S. S. Sonawane, P. N. Mahalle, and A. S. Ghotkar, "Information Retrieval," in *Information Retrieval and Natural Language Processing: A Graph Theory Approach*, Singapore: Springer Singapore, 2022, pp. 81–94. doi: 10.1007/978-981-16-9995-5_4.
[10] D. Munteanu, "Vector space model for document representation in information retrieval," *Ann. Dunarea Jos*, pp. 43–44, 2007.
[11] M. Abdullah and M. G. I. Al Zamil, "The Effectiveness of Classification on Information Retrieval System (Case Study)," 2018, [Online]. Available: http://arxiv.org/abs/1804.00566