International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 08 Aug. 2024, pp: 364-368 www.ijaem.net ISSN: 2395-5252

# Filtering articles based on their abstracts using TF-IDF

Yusra Nuri[1], Edip Senyurek[2]
[1]Student, Department of Computer Engineering Vistula University, Warsaw, Poland.
[2]Head of Department, Department of Computer Engineering Vistula University, Warsaw, Poland.

Date of Submission: 15-08-2024 Date of Acceptance: 25-08-2024

**ABSTRACT**: This paper explores the application of Term frequency - Inverse document frequency (TF-IDF) for the filtering of research articles based on the relevance of their abstracts. Our study aims to improve the accuracy of filtering research articles using Term frequency – Inverse document frequency (TF-IDF), a widely recognized numerical statistic in text mining.

To enhance the filtering process, we present an approach by incorporating similarity estimation into the Term frequency – Inverse document frequency (TF-IDF) framework. The Term frequency component assesses the word's significance in a document. Whereas, the Inverse document frequency part measures how relevant a word is in a collection of documents.

The methodology involves 19 randomly selected research papers. We process the abstracts by removing common stop words, then calculate the TF, IDF, and subsequently the TF-IDF scores. We introduced a weighted balance approach, calculating the mean and standard deviation of TF-IDF values, normalizing the dataset, and deriving z-scores.

Our results show the application of z-scores and normalized TF-IDF values to determine the relevance of documents within our corpus. The findings indicate that while Z-scores and normalized values provide distinct insights, normalization of TF-IDF gives a better accuracy of results.

**KEYWORDS:** TF-IDF, Recommendation Systems, Content based Filtering, Term Frequency, Inverse Document Frequency

## I. INTRODUCTION

In the digital era, the information abundance and numerous research papers published have made the process of retrieving relevant keywords and information a crucial component for searching and categorizing the set of articles according to a specific type [1].

Information retrieval systems serve as crucial gateways, enabling users to extract relevant information through extensive digital libraries. It has become increasingly important in the digital age to classify a huge number of textual data and efficiently extract meaningful insights. One of the most important techniques to classify documents is Term frequency – Inverse document frequency (TF-IDF) which stands for its unique approach to signify the importance of words in a corpora.

This technique not only ranks the significance of terms in documents but also balances the frequency of word appearances within a document Term frequency (TF) against the frequency of occurrences across the entire document Inverse document frequency (IDF). Information retrieval and machine learning models that use text data can be greatly improved by this balance [2].

Term frequency – Inverse document frequency (TF-IDF) is a numerical statistic that is widely used for information retrieval. It calculates how relevant a word is in a document by multiplying two indicators, namely Term frequency (TF) and Inverse document frequency (IDF). How many times the word appeared in the document (TF) and the frequency of the word throughout the document (IDF) [3].

This algorithm not only aids in efficiently finding and categorizing documents but also improves matching words in a query. This approach ensures that words that uniquely define the content of the document are weighted more showing that they are of a bigger significance in that document, while redundant words that appear frequently across documents are weighted less.

A system that helps sort research papers into related subjects using keywords from abstracts was presented by Kim and Gil [1]. This system uses a method known as Latent Dirichlet

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 08 Aug. 2024, pp: 364-368 www.ijaem.net ISSN: 2395-5252

Allocation (LDA) to find these subjects. Then, it uses a method known as k-means clustering utilizing Term frequency – Inverse document frequency (TF-IDF) scores to group papers based on the extracted keywords into related subjects. This makes it easier to find and organize research on similar topics.

Expanding on the foundations laid by Kim and Gil [1], who presented a system for categorizing research papers into related subjects using keywords extracted from abstracts, this paper will employ a similar approach.

## II. BACKGROUND OF THE STUDY

To improve the classification of research papers to help users find their interesting topics easily, Kim and Gil [1] proposed a system that classifies papers using the Term frequency – Inverse document frequency (TF-IDF) and Latent Dirichlet allocation (LDA). This system constructs a representative word dictionary with the keywords that the user inputs and topics extracted by Latent Dirichlet allocation (LDA) and extracts subject words from the abstract of papers based on the keyword dictionary.

Similarly, the effectiveness of machine learning models on spam mail detection was evaluated by Nusrat et al. [4]. In addition, to improve classification accuracy by finding the relation between keywords and the frequency of these words in corpora [5], presented a system called an improved Inverse document frequency (IDF) method. This method addresses the issues of overlooking the distribution of categories and handling uneven datasets ineffectively. Meanwhile, Zhou [6] proposed combining original text features to improve the classification accuracy of Term frequency – Inverse document frequency (TF-IDF).

Furthermore, Khan et al. [7] discussed a method for summarizing a large number of texts to help users extract the information of the text easily. This involves taking large relevant information and creating a short version in a way that reflects the idea understandably. Umadevi [8] also discusses the challenge of extracting meaningful data from the massive amounts of digital data, which makes up 80% of the data. As a solution, the paper proposed text mining for discovering significant patterns in text documents using Term frequency – Inverse document frequency (TF-IDF) and cosine similarity metrics for document comparison.

The application of the Term frequency – Inverse document frequency (TF-IDF) algorithm for assessing keyword relevance in document corpora was further explored by Das et al. [9]. They developed a web and mobile-based application for managing official letters, aiming to improve public services. This used the Term frequency – Inverse document frequency (TF-IDF) for letter classification [10]. Alternatively, the paper introduces the Semantic sensitive Term frequency-inverse document frequency (STF-IDF), enhancing keyword extraction from informal documents in a healthcare social medial corpus by combining it with traditional TF-IDF [11].

SciBert has been introduced to improve the processing of scientific literature and demonstrate its improvement in different Natural language processing (NLP) tasks [12]. Similarly, a method has been proposed for accurately classifying crisis-related data on social networks, using contextual representations like (Embeddings from language models – ELMo) demonstrating higher accuracy over traditional classifiers [13].

Addressing the vastly increasing online data, the paper [7] focuses on extractive-based summarization using K-means clustering and Term frequency – Inverse document frequency (TF-IDF), efficiently summarizing documents while reflecting on the central idea. This range of studies showcases ongoing advancements in Natural language processing (NLP) and their significant implications for improving efficiency in information retrieval across various domains.

## III. METHODOLOGY

The methodology involved randomly selecting 20 research papers, from which 19 of them were used as our sample corpora and one abstract for testing as an active abstract. The study followed the several steps defined below:

1. Prepared the data by separating the words in each of the abstracts by using an index of space characters.
2. Assembled these separated words into a single column,
3. Eliminated common words that have little value in terms of relevance such as ‘a’, ‘the’, ‘are’, ‘with’, ‘for’ etc.
4. These steps were implemented to all abstracts that we have chosen randomly, including the active abstract that we used for testing
5. Calculated Term frequency (TF) values by using Equation (1)

$$TF = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (1)$$

where $n_{i,j}$ is the number of occurrences of the word $t_i$, represents ith term, in document $d_j$.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 08 Aug. 2024, pp: 364-368 www.ijaem.net ISSN: 2395-5252

represent jth document. The denominator $\sum_k n_{k,j}$ is the total number of occurrences of words in the document $d_j$. k represents the number of words.

6. Calculated Inverse document frequency (IDF) by using Equation (2)

$$IDF = \frac{1+|D|}{1+|d \in D : t \in D|} \quad (2)$$

where D is the total number of documents in the corpus and $|d \in D : t \in D|$ is the number of documents where the term t occurs. It is common to adjust the numerator and denominator by adding +1 because if the term isn't in the corpus, this will lead to a division by zero.

7. Calculated Term frequency – Inverse document frequency (TF-IDF) by multiplying Term frequency (TF) and Inverse document frequency (IDF)

$$TF - IDF = TF \, x \, IDF \quad (3)$$

In the above calculations, Term frequency (TF) measures the frequency of words in a single document. Whereas Inverse document frequency (IDF) measures the frequency of words across the entire document.

We aimed to predict the relevance of our active abstract for testing to our corpus, for this, we used weighted balance (WB). The steps to calculate WB included:

1. computing the mean

$$\bar{x} = \frac{\sum x}{N} \quad (4)$$

where $\bar{x}$ is the average of variable x. $\sum x$ is the sum of x values. $N$ is the total number of values

2. standard deviation (STD) of the Term frequency – Inverse document frequency (TF-IDF) values

$$STD_i = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} \quad (5)$$

where $\sum (x - \bar{x})^2$ is the squared value of the deviation of each value from the mean. The denominator $n - 1$ is the number of values in the sample.

3. normalizing the dataset

$$N_i = x_i - \bar{x}_i \quad (6)$$

here, to normalize we subtract the normal values from the sample means.

4. calculated z-scores matrix of a dataset

$$Z - score_i = \frac{N_i}{STD_i} \quad (7)$$

here, the division of normalized value by standard deviation will give z-score value.

5. Finally, we calculated similarity estimation

$$W_{a,b} = \frac{\sum N_{ai} * N_{bi}}{STD_{ai} * STD_{bi}} \quad (8)$$

the weighted value between a and b is the summation of the product $N_{ai} * N_{bi}$ divided by the product of STD. Where $N_{ai}$ indicates normalized value of an active testing article and $N_{bi}$ indicates normalized value of sample article. $STD_{ai}$ indicates standard variation of active testing articles and $STD_{bi}$ indicates the standard variation of the sample article.

The final step involved applying our methodology to the dataset of 20 abstracts, focusing particularly on the active abstract selected for testing. By comparing normalized datasets and z-scores, we were able to assess the effectiveness of our approach in identifying relevant documents and terms within our corpus.

## IV. EXPERIMENT RESULTS

Table 1 has document numbers ordered slightly differently based on Z-score and table 2 has the Normalized values for the documents. A Z-score is measured in terms of standard deviations from the mean. This involves taking away the smallest value from every data point and then dividing it by the dataset's range.

**Table 1: Z-score Values**

| | DOCUMENT NUMBER | Z-SCORE VALUE | | DOCUMENT NUMBER | Z-SCORE VALUE |
|---|---|---|---|---|---|
| **1** | 8 | 25.97072 | **11** | 1 | 15.2848 |
| **2** | 13 | 24.3833 | **12** | 10 | 14.99364 |
| **3** | 12 | 23.88685 | **13** | 5 | 14.93913 |
| **4** | 11 | 21.60607 | **14** | 16 | 14.00257 |
| **5** | 4 | 21.19413 | **15** | 6 | 12.79975 |

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 08 Aug. 2024, pp: 364-368 www.ijaem.net ISSN: 2395-5252

| | | | | | | |
|---|---|---|---|---|---|---|
| **6** | 18 | 21.1693 | | **16** | 2 | 12.79975 |
| **7** | 15 | 20.71309 | | **17** | 14 | 10.79958 |
| **8** | 17 | 17.0425 | | **18** | 19 | 10.45156 |
| **9** | 7 | 15.82606 | | **19** | 9 | 6.648057 |
| **10** | 3 | 15.56065 | | | | |

The comparison was observed between the Z-score of a document and its normalized value. The top five documents were checked from Z-score and normalized and seen that four of them were the same.

**Table 2: Normalized Values**

| | **DOCUMENT NUMBER** | **NORMALIZED VALUE** | | **DOCUMENT NUMBER** | **NORMALIZED VALUE** |
|---|---|---|---|---|---|
| **1** | 13 | 135102.3 | **11** | 1 | 62803.61 |
| **2** | 15 | 130439.1 | **12** | 10 | 55650.77 |
| **3** | 8 | 124429.2 | **13** | 3 | 51053.5 |
| **4** | 11 | 119809 | **14** | 14 | 47235.5 |
| **5** | 4 | 116275 | **15** | 5 | 40163.42 |
| **6** | 18 | 113001.9 | **16** | 19 | 35393.38 |
| **7** | 12 | 75005.42 | **17** | 6 | 34157.48 |
| **8** | 17 | 71597.01 | **18** | 2 | 33376.66 |
| **9** | 16 | 70451.67 | **19** | 9 | 29445.31 |
| **10** | 7 | 70322.7 | | | |

Document numbers 4, 8, 11, 12, and 13 are in the top 5 for Z-Score, whereas, document numbers 4, 8, 11, 13, and 15 are in the top-5 for normalized value. The different documents from each top-5 are document number 12 and document number 15. To see which one was more accurate and more related to the abstract that is mentioned as a target document, the documents were read again. Of those two documents that were compared, the most relevant one was found to be Document 15, and because of that normalization is giving better results.

Normalized values of each document are shown on the first graph in Figure 1. The bars are ordered from highest to lowest which suggests ranking based on normalized values. On the other hand, the second graph shows the Z-scores for each document similarly presented in descending order. The purpose of such a graph is to provide a visual comparison between two different statistical measures.

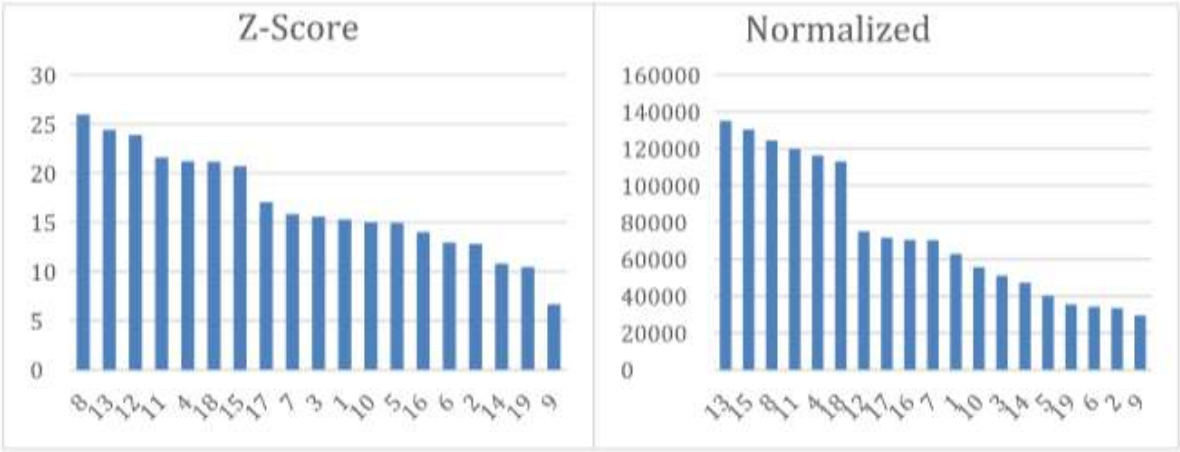


**Figure 1. Comparison of Normalized Value (left) and Z-score (right)**

## V. CONCLUSION

This study has presented an examination of the Term frequency - Inverse document frequency (TF-IDF) technique with similarity estimation to improve filtering research articles based on the relevance of their abstracts. Through

methodology, the significance of TF-IDF in isolating important terms within documents was shown but also how similarity estimation can strengthen this process by identifying the related articles. As a result, it might be concluded that for TF-IDF, normalization implementation is better than Z-Score.

Future work will be implemented on more abstracts than it is implemented in this study. To get faster results a program will be developed by using Microsoft Visual Studio C# programming.

## REFERENCES

[1]. S. Kim, J. Gil., Research paper classification system based on TF-IDF and LDA Schemes. Hum. Cent. Inf. Sci, (2019).

[2]. W. Scott. TF-IDF from scratch in python on a real world dataset. Towards Data Science, (2019).

[3]. S. Qaiser, R. Ali. Text Mining: use of TF-IDF to examine the relevance of words to documents. International journal of Computer Applications, (2018).

[4]. Euna, N. J., Hossain, S. M. M., Anwar, M. M., & Sarker, I. H. (2024). Content-based spam email detection using an N-gram machine learning approach. In Applied Intelligence for Industry 4.0 (pp. 123-138). CRC Press.

[5]. L. Xiang. Application of an improved TF-IDF method in literary text classification. Advances in multimedia, (2022), 1-10.H. Zhou. Research of Text classification based on TF-IDF and CNN-LSTM. Advances in multimedia., (2022), 1-10.

[6]. R. Khan, Y. Qian, S. Naeem. Extractive based text summarization using k-means and TF-IDF. International journal of information engineering and electronic business, (2019), 33-44.

[7]. D. Umadevi. Document comparison based on TF-IDF metric. International Research Journal of Engineering and technology(IRJET), (2020).

[8]. M. Das, S. Kamalanathan, A. Pja, A comparative study on TF-IDF feature weighting method and its analysis using unstructured dataset. COLINS-2021, 5th International conference on computational Linguistics and Intelligent systems, (2021), 1-10.

[9]. M. Artama, N. Sukajaya, G. Indrawan. Classification of official letters using TF-IDF method., Journal of physics: Conference series. (2020).

[10]. A.Jalilifard, V. F. Carida, A. F. Mansano, R. S. Cristo, F. Penhorate, C. Fonseca. Semantic sensitive TF-IDF to Determine Word Relevance. Advances in computing and network communication, (2021).

[11]. S. Madichetty, S. Muthukumarasamy, Improved classification of crisis-related data on Twitter using contextual representations, Procedia computer science, (2020).

[12]. Beltagy, K. Lo, A. Cohan, SciBERT: A Pretrained Language Model for scientific text, Association for computational Linguistics, (2019), 3615−3620.