Comparative Analysis of TF-IDF and TF-IDF Groups using BERT and GloVe

Alex Conro Manuel[1] and Angelina Naomi C. Sinaga[2]

Information System, Del Institute of Technology, Indonesia
{iss18022, iss18035}@students.del.ac.id

Abstract. Rigidity in finding information and a lot of noise are problems that often occur in knowledge management which has the concept of information retrieval in identifying, explaining and distributing information for use obtained from documents or collections. Currently the development of information retrieval is very interesting to discuss and research, because the application of information retrieval can help overcome some of the problems above. The use of TF-IDF has been widely used because it is simple in the process of calculating keywords or queries and is easy to use in measuring content uniqueness, as well as low-cost computational processes. However, so far the implementation of the TF-IDF group on IR has not been implemented. In this project, we compare the performance of TF-IDF with TF-IDF Group by using word embedding method: BERT and GloVe, where the document grouping process is done using Minibatch K-Means (Cosine Sim). The first stage is text preprocessing which consists of case folding, stopword, tokenization and stemming stages, the second stage is weighting using TF-IDF and TF-IDF Group, the third stage is applying the method used and the last stage is evaluating.

Keywords: information retrieval · BERT · GloVe · Cosine Similarity.

1 Introduction

Information dissemination is currently growing rapidly. Every day people search for information by typing keywords in search engines and want fast and accurate information [1]. Information consists of various categories and is scattered randomly and unclearly [18], which is sometimes difficult for those who need it, moreover, its contents do not necessarily contain important things needed by readers. This information search activity is known as information searching. Information search aims to find the most relevant documents based on keywords in user-generated queries [13]. Basically, the development of this information retrieval system is actually inseparable from the techniques or methods used. There are two jobs in this system, namely prepossessing the data set and applying certain methods to calculate the relevance (similarity) between documents in the prepossessed database [9]. As a result, the system will return a list of documents ordered according to the similarity value to the previously entered query. The solution to this problem is to summarize the text [19].

In this paper, we analyze the comparison of the TF-IDF Group with the TF-IDF using BERT and GloVe. Considering the TF-IDF group and TF-IDF, this comparative analysis aims to compare the output to the performance of text categorization. The TF-IDF weighting consists of 2 factors, namely term frequency (TF) and inverse document frequency (IDF). Term frequency (TF) is a condition in which each term is assumed to have a proportion of importance according to the number of occurrences in the document [17] and inverse document frequency (IDF) is a term weighting method that focuses on paying attention to the occurrence of terms in the entire text collection [14]. There are several parameters that are used as benchmarks to compare the performance of the text categorization, namely precision, recall and f-measure [2]. Mini batch K-means is a version of the standard K-means algorithm in machine learning that uses small, random, and fixed-sized data sets to be stored in memory, and then with each iteration, a random sample of data is collected and used to update the cluster [10]. Global Vectors for Word Representation (Glove) is a word representation to generate word embedding to be used to handle word similarity, word analogy, and named entity recognition [5]. Transformers' Bidirectional Encoder Representations (BERT) is a neural network-based technique for pre training natural language to help understand the context of words in search queries [12]. Based on the above, we first need to combine two concepts for calculation, namely the frequency of occurrence of a word in a particular document and the inverse of the frequency of documents containing that word against the BERT method and the Glove method. Second, calculate the TF-IDF group against the BERT method and the Glove method. Finally, analyze the results from the first and second stages, then make a comparison.

Based on the research of Kamyab et al. [8] proposed a new attention-based model that utilizes CNN with LSTM (named ACL-SA), applies a preprocessor to improve data quality and uses term frequency-inverse document frequency (TF-IDF) feature weighting and Glove's pre trained word embedding approach to extract meaningful information from textual data, use CNN max-pooling to extract contextual features and reduce feature dimensions, also use integrated two-way LSTM to capture long-term dependencies. In the research of Weilong Chen et al. [3] , it focuses on the effect of different contexts to determine the similarity of 2 different words. Their research is based on BERT built with TF-IDF and applies the data collection method (CoSimLex), which covers four languages namely English, Croatian, Slovenian and Finnish. In the model they built word embedding can train the model to predict the similarity of words to understand the meaning of words from different perspectives. Research Jin et al. [7]created a multi-label classification framework for aspect-based sentiment analysis problems in restaurant customer reviews where their processes include text prepossessing, feature extraction using modified BERT and TF-IDF, and fine tuning. The TF-IDF method is used to determine how important the word is in the multi-label classification by calculating the weights. [16].

Although the above studies have proven the superiority of each method in calculating the weights, none of them compared the weighting of the TF-IDF with the TF-IDF Group using BERT and Glove. Based on this, we propose a Comparative Analysis of TF-IDF with TF-IDF Group using BERT and Glove to find out what is the special differentiated in calculating the weights. Specifically, we will do text prepossessing which consists of case folding, stop word, tokenization and stemming stages. Then we apply the calculation of the frequency of occurrence of a word in a particular document and inverse the frequency of documents containing the searched word using both methods. Then determine the ranking using cosine similarity and followed by the final analysis of the results obtained in the BERT and Glove methods. The main contributions are as follows:

1. We propose the BERT and Glove models to make comparisons on the TF-IDF and TF-IDF Group.
2. We Group documents using Mini batch K-Means (Cosine Sim). Before grouping the document, do docs expansion (following the concept of query expansion)
3. We conducted experiments on two sets of data sets, namely Spam and BBC News to compare the results of the two applied methods.

The rest of the paper is organized as follows: Section 2 discusses the related works. Then, we present the architecture of the framework in Section 3. Next, Section 4 provides the experiment setup and implementation and experimental results and analysis, and finally, Section 5 concludes this work.

2 Related Works

We first review the important things related to what we are working on, namely the Clustering, Bert, GloVe, TF-IDF and TF-IDF Group,Cosine similarity, and evaluation. **Clustering**: Mini Batch K-means is an alternative from K-means algorithm in grouping massive data sets. The advantage of this algorithm is to reduce computational costs by not using all datasets where each iteration is carried out with a fixed size. This strategy certainly reduces the number of distance calculations per iteration at the expense of lower cluster quality [2]. The purpose of using mini-batch here is for grouping to get the TF-IDF Group value. **BERT**: BERT uses a multi-layer network to capture and Transformers to encode input tokens, and make them a representative model,the token output vector becomes a vector that is used to represent the token sentence [4]. **Glove**: Glove is a method that combines co-occurrence and semantic relationships and is a global matrix factorization method, which represents the occurrence of a word in a document. Where Glove studies the relationship of a word by calculating the frequency of occurrence of the word along with other words in a given corpus. This frequency of occurrence has the potential to encode multiple forms of pronouns and help performance in word analogies. Glove stages are:

1. Collect word co-occurrence statistics in the form of a word co-occurrence matrix.
2. Define soft constraints on word pairs.wi is the main vector, wj is the context vector, bi, bj is the scalar bias for the main and context words.

$$W\frac{T}{i}_{Wj} +_{bi} +_{bj} = log(_{Xij}) \tag{1}$$

3. Define a cost function.

$$J = \sum_{i=1}^{v}\sum_{j=1}^{v} f(_{Xij})W\frac{T}{i}_{Wj} +_{bi} +_{bj} = log\left(_{Xij}\right)^2 \tag{2}$$

f is a weighting function to help prevent the learning of common word pairs. The function is defined as follows:

$$f(_{Xij}) = \left\{ \left(\frac{Xij}{Xmax1}\right) aif_{Xij} < XMAX \right\} otherwise \tag{3}$$

TF-IDF: The TF–IDF models the TF–IDF Weighted score. using Weighted TF–IDF to rank documents, in both Glove and BERT Glove methods we calculate the similarity between these user and candidate queries, then re-rank the documents and get similar TF–IDF scores. TF–IDF gives the same score regardless of different semantic information, we aim to compare the results of both methods and datasets. TF-IDF is used to calculate the relevance of a word in a particular document by multiplying two matrices between the frequency of words in a document with the frequency of word documents in a set of documents [15], calculated as follows:

$$tfidf(t, d, D) = tf(t, d), idf(t, D) \tag{4}$$

$$tf(t, d) = log(1 = freq(t, d)) \tag{5}$$

$$idf(t, D) = log(\frac{N}{count(d\epsilon D : t\epsilon d)}) \tag{6}$$

Cosine Similarity: The Cosine similarity score represents the scenario that the user input exactly matches the document; the semantic score represents the scenario that the user wants to search for some relevant document [11].The formula for calculating cosine similarity is as follows:

$$cos\Theta = \frac{a.b}{\|a\| \,.\, \|a\|} \tag{7}$$

Evaluation: A model certainly gives the results of the prediction probability from the model. There are many useful metrics to test the model's ability, but in this study, accuracy is used, while the accuracy formula is as follows: **Accuracy**: The Accuracy formula considers the number of True Positive and True Negative elements in the numerator and the sum of all the confusion matrix entries in the denominator. True Positive and True Negative are elements correctly classified by the model and they are on the main diagonal of the confusion matrix, while the denominator is also considers all elements of the main diagonal that are misclassified by the model [6].There is a formulation as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

3 Framework

In this section, we search for TF-IDF values and TF-IDF Group values using the BERT and GloVe methods, where the document grouping process is carried out using Minibatch K-Means (Cosine Sim), and an overview is shown in Figure 1 and Figure 2.

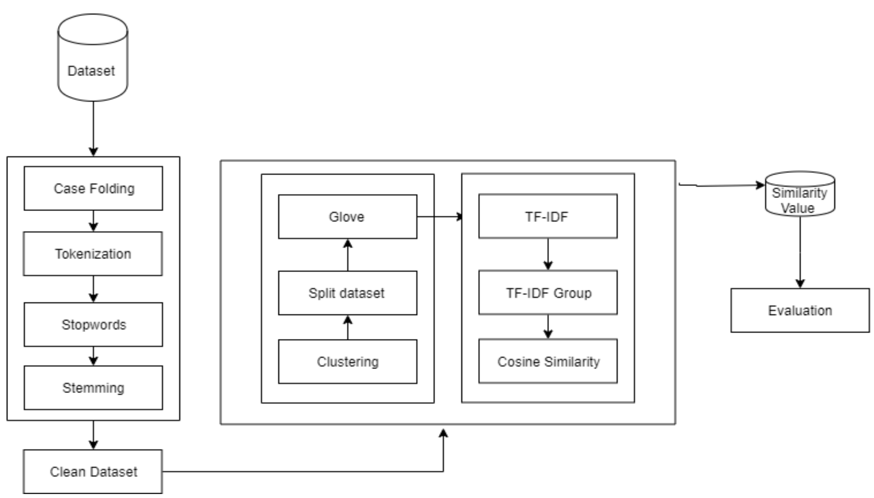


Fig. 1: Framework Glove

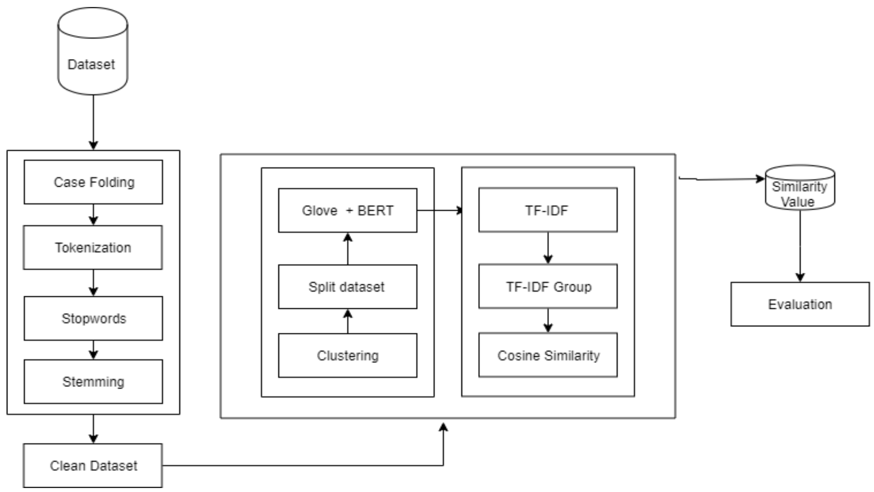
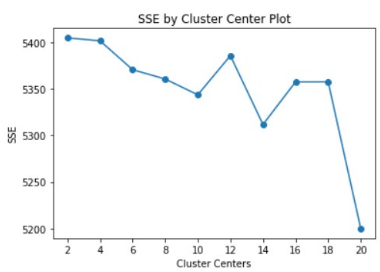
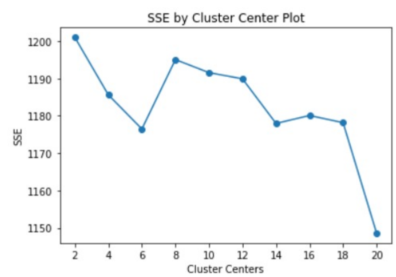


Fig. 2: Framework GloveBERT



(a) Dataset Spam Cluster.



(b) Dataset BBC News Cluster.

Fig. 3: Clustering Plot.

In this study, we created a model to compare the value of TF-IDF with the TF-IDF Group by using the Glove and BERT methods. The stages start with pre-processing which consists of case folding, tokenization, stpowords and stemming. This research also applies clustering to group words from sentences, specifically the clustering used is mini batch, this algorithm will partition the data set into n clusters. With the application of clustering, it is hoped that it can help predict the value of the TF-IDF group.
After the clustering process continues with the split dataset. From the split dataset, word embedding will be performed with Glove for figure 1, or after word embedding is inputted into the BERT model as in figure 2. Next, a weighting scheme will be carried out using TF-ID and checking for similarity with cosine similarity. To get this similarity value, a search will be carried out on one of the documents. The last step is to calculate the evaluation of

the model using the accuracy evaluation metric, which method is good or not.

4 Experimental Evaluation

In this section, we discuss the evaluation of BERT and Glove by conducting comprehensive experiments on two data sets namely the spam data set and the BBC News Data set and show which model performs better on which data set..

4.1 Experimental Setup

We conducted experiments on two publicly available data sets, namely the Spam Data set and the BBC News Data set, which are the data sets commonly used for information retrieval. The Spam Data set is a collection of SMS tagged messages that have been collected for SMS Spam research. It contains a set of SMS messages in English from 5,574 messages, marked as ham (legitimate) or spam. While the BBC News data set is a collection of RSS feeds from BBC News which consists of several columns such as title, date and description.

Table 1: Data set

Data set	Items	Size(KB)
Spam	5574	492
BBC News	1816	585

To find out how the performance of the model, it is proposed to calculate with accuracy. Accuracy is defined as the level of closeness between the predicted value and the actual value.

4.2 Evaluation Methodology

The results of the system certainly need to be measured. Evaluation means-certainty that is very commonly used as a reference in determining accuracy of information. To calculate accuracy use following formula:

$$Accuracy(Acc) = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

In addition, to calculate tf-idf, the following formula is used:

$$tf - idf(t, d, D) = tf(t, d), idf(t, D) \tag{10}$$

and to calculate the tf-idf group as follows:

$$tf - idfgroup(t, dlust, D) = tf(t, dclust), idf(t, D) \tag{11}$$

4.3 Evaluation on Effectiveness

The effectiveness comparison that we compare to Glove method and the Glove BERT by performing the same two dataset ,to the same target set. Table describes the comparison results of Glove and Glove BERT,in terms of TF-IDF VS TF-IDF Group ,Cosine similarity, and accuracy.

Table 2: Effectiveness comparison with for *Spam and BBC* Dataset in terms of cosine similarity and accuracy (the lower the value, the better).

Method	Dataset	Cosine Similarity	Accuracy
Glove	Spam	0.710	0.977
Glove BERT	Spam	0.677	0.954
Glove	BBC News	0.683	0.058
Glove BERT	BBC News	0.736	0.0010

Configuration Parameter. In this study, we did not configure parameters, we used parameters and their values are default values. Based on table 2 above.

TF-IDF VS TF-IDF Group. The picture below contains a comparison between the TF-IDF and the TF-IDF Group on the two approaches.

term	TF-IDF	TF-IDF-Group
9	0.13	[0.26 0.26 0.26 ... 0.26 0.26 0.26]
1	0.09	[0.18 0.18 0.18 ... 0.18 0.18 0.18]
0	0.11	[0.22 0.22 0.22 ... 0.22 0.22 0.22]
3	0.13	[0.26 0.26 0.26 ... 0.26 0.26 0.26]
7	0.07	[0.14 0.14 0.14 ... 0.14 0.14 0.14]
2	0.1	[0.2 0.2 0.2 ... 0.2 0.2 0.2]
5	0.09	[0.18 0.18 0.18 ... 0.18 0.18 0.18]
8	0.1	[0.2 0.2 0.2 ... 0.2 0.2 0.2]
4	0.08	[0.16 0.16 0.16 ... 0.16 0.16 0.16]
6	0.1	[0.2 0.2 0.2 ... 0.2 0.2 0.2]

Fig. 4: TF-IDF and TF-IDF Group Glove Dataset 1

term	TF-IDF	TF-IDF-Group
jeremy	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
bowen	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
	0.006482982171799027	[0.06482982 0.06482982 0.06482982 ... 0.06482982 0.06482982 0.06482982]
frontline	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
irpin	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
,	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
residents	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
came	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
russian	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
fire	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
trying	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
flee	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
.	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]

Fig. 5: TF-IDF and TF-IDF Group Glove Dataset 2

term	TF-IDF	TF-IDF-Group
ok	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
lar	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
...	0.0003589375448671931	[0.00071788 0.00071788 0.00071788 ... 0.00071788 0.00071788 0.00071788]
joking	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
wif	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
u	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
oni	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]

Fig. 6: TF-IDF and TF-IDF Group GloveBERT Dataset 1

term	TF-IDF	TF-IDF-Group
jeremy	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
bowen	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
	0.006482982171799027	[0.06482982 0.06482982 0.06482982 ... 0.06482982 0.06482982 0.06482982]
frontline	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
irpin	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
,	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
residents	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
came	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
russian	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
fire	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
trying	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
flee	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
.	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]

Fig. 7: TF-IDF and TF-IDF Group GloveBERT Dataset 2

5 Conclusion

In this paper, we propose a comparison of the TF-IDF with the TF-IDF Group using Glove method, from the results of the experiments we carried out:

1. In both datasets, the Cosine Similarity value and the accuracy value are quite high.
2. Combining Glove with BERT gives the effect of decreasing the value of cosine similarity and decreasing the accuracy of the model.
3. The TF-IDF value in Dataset 1 increased when the TF-IDF Group was performed using the GloVe and GloveBERT methods.
4. The TF-IDF value in Dataset 2 has the same value as the TF-IDF Group using the GloVe and GloveBERT methods.

6 Contribution

This section describes the division of tasks in this project.

Table 3: STBI Contribution Teori

Nama	**Task**
Alex Conro Manuel	Mengerjakan Related Works Mengerjakan Frameworks Mengerjakan Experimental Setup Mengerjakan Evaluation Methodology Mengerjakan Evaluation on Effectiveness Mengerjakan Conclusion
Angelina Naomi C. Sinaga	Mengerjakan Abstrak Mengerjakan Introduction Mengerjakan Frameworks Mengerjakan Evaluation on Effectiveness Mengerjakan Conclusion

Table 4: STBI Contribution Code

Nama	**Task**	**Dataset**	**Method**
Alex Conro Manuel	Mengerjakan Implementasi Code	Spam	BERT
	Mengerjakan Implementasi Code	Spam	BERTGlove
Angelina Naomi Sinaga	Mengerjakan Implementasi Code	BBC News	BERT
	Mengerjakan Implementasi Code	BBC News	BERTGlove

Link gitbub : https://tinyurl.com/TF-IDFandGroupBERTandGlove

References

1. https://balitbangsdm.kominfo.go.id/upt/bandung/?mod=publikasiamp;action=dlamp;cid=10amp;pub$_id =$ 5
2. Béjar Alonso, J.: K-means vs mini batch k-means: A comparison (2013)
3. Chen, W., Yuan, X., Zhang, S., Wu, J., Zhang, Y., Wang, Y.: Ferryman at semeval-2020 task 3: bert with tfidf-weighting for predicting the effect of context in word similarity. In: Proceedings of the fourteenth workshop on semantic evaluation. pp. 281–285 (2020)
4. Fimoza, D., et al.: Analisis sentimen terhadap film indonesia dengan pendekatan bert (2021)
5. Fitri, M.: perancangan sistem temu balik informasi dengan metode pembobotan kombinasi tf-idf untuk pencarian dokumen berbahasa indonesia. JUSTIN (Jurnal Sistem dan Teknologi Informasi) **1**(1), 80–85 (2013)
6. Grandini, M., Bagli, E., Visani, G.: Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756 (2020)
7. Jin, Z., Lai, X., Cao, J.: Multi-label sentiment analysis base on bert with modified tf-idf. In: 2020 IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-CN). pp. 1–6. IEEE (2020)
8. Kamyab, M., Liu, G., Adjeisah, M.: Attention-based cnn and bi-lstm model based on tf-idf and glove word embedding for sentiment analysis. Applied Sciences **11**(23), 11255 (2021)
9. Kesuma, H.W.A.: Penerapan metode tf-idf dan cosine similarity dalam aplikasi kitab undang-undang hukum dagang (2016)
10. Khatri, A., et al.: Sarcasm detection in tweets with bert and glove embeddings. arXiv preprint arXiv:2006.11512 (2020)
11. Marwah, D., Beel, J.: Term-recency for tf-idf, bm25 and use term weighting. In: Proceedings of the 8th International Workshop on Mining Scientific Publications. pp. 36–41 (2020)
12. Melita, R.: Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim). B.S. thesis, Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta (2018)
13. Merdekawan, R., et al.: Sistem Penelusuran Katalog Perpustakaan Menggunakan Metode Rocchio Relevance Feedback. Ph.D. thesis, Universitas Kanjuruhan Malang (2014)
14. Putra, J.W.G.: Pengenalan konsep pembelajaran mesin dan deep learning. Tokyo. Jepang (2019)
15. Qaiser, S., Ali, R.: Text mining: use of tf-idf to examine the relevance of words to documents. International Journal of Computer Applications **181**(1), 25–29 (2018)
16. Rachman, F.P., Santoso, H., Djajadi, A.: Machine learning mini batch k-means and business intelligence utilization for credit card customer segmentation. International Journal of Advanced Computer Science and Applications **12**(10) (2021)
17. Saadah, M.N., Atmagi, R.W., Rahayu, D.S., Arifin, A.Z.: Sistem temu kembali dokumen teks dengan pembobotan tf-idf dan lcs. Jurnal Ilmiah Teknologi Informasi **11**(1), 19–22 (2013)
18. Widyasanti, N.K., Putra, I.K.G.D., Rusjayanthi, N.K.D.: Seleksi fitur bobot kata dengan metode tfidf untuk ringkasan bahasa indonesia. Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi) pp. 119–126 (2018)
19. Winarti, T., Setiawan, W., Iswoyo, Pudjiastuti, E.: Deteksi kemiripan dokumen bahasa indonesia dengan menggunakan model ruang vektor,

http://repository.usm.ac.id/files/research/G001/20191127114354-Deteksi-Kemiripan-Dokumen-Bahasa-Indonesia-Dengan-Menggunakan-Model-Ruang-Vektor.pdf