

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

Alex Conro Manuel¹ and Angelina Naomi C. Sinaga²

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

TF-IDF vs TF-IDF Group with GloVe and GloVeBERT

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: GloVe and GloVeBERT, where the document grouping process is done using Minibatch K-Means (Cosine Sim). The first stage is text preprocessing which consists of case folding, stopwords, tokenization and stemming stages, the second stage is weighting using TF-IDF and TF-IDF Group, the third stage is applying the method used, then calculate the cosine similarity and the last stage is evaluating.

Keywords: TF-IDF · TF-IDF Group · BERT · GloVe · Cos Similarity · Evaluation.

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 Glove and GloVeBERT. 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 preprocessing, 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 Glove and GloveBERT. 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 preprocessing 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 Glove and GloveBERT methods. The main contributions are as follows:

1. We propose the Glove and GloveBERT models to make comparisons on the TF-IDF and TF-IDF Group.
2. We Group documents using Mini batch to clustering or grouping document.
3. We calculated cosine similarity to check similar from one document to another document.
4. 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 is an alternative from K-Means algorithm in grouping massive datasets. 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. which is the main vector, w_j is the context vector, b_i , b_j is the scalar bias for the main and context words.

$$W \frac{T}{i} w_j + b_i + b_j = \log(x_{ij}) \quad (1)$$

3. Define a cost function.

$$J = \sum_{i=1}^v \sum_{j=1}^v f(x_{ij}) W \frac{T}{i} w_j + b_i + b_j = \log(x_{ij})^2 \quad (2)$$

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

$$f(x_{ij}) = \left\{ \left(\frac{x_{ij}}{x_{max1}} \right) \text{ if } x_{ij} < XMAX \right\} \text{ otherwise} \quad (3)$$

TF-IDF: The TF-IDF models the TF-IDF Weighted score. using Weighted TF-IDF to rank documents, in both GloVe and GloVeBERT 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:

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \quad (4)$$

$$tf(t, d) = \log(1 + freq(t, d)) \quad (5)$$

$$idf(t, D) = \log\left(\frac{N}{count(d \in D : t \in d)}\right) \quad (6)$$

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:

$$\cos\Theta = \frac{a.b}{\|a\| \cdot \|a\|} \quad (7)$$

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:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

3 Framework

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

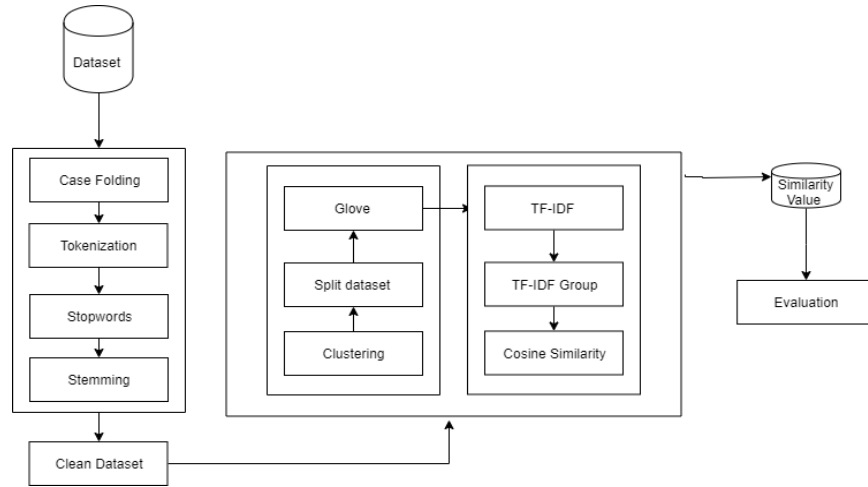


Fig. 1: Framework Glove

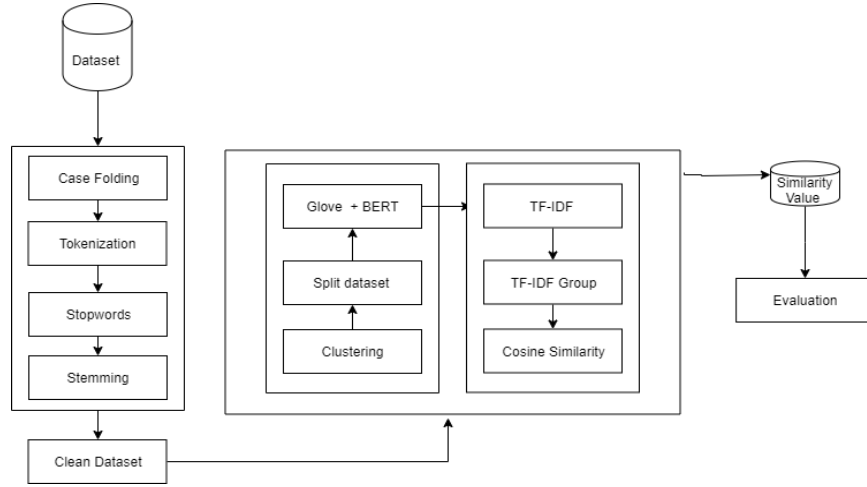


Fig. 2: Framework GloVeBERT

In this study, we created a model to compare the value of TF-IDF with the TF-IDF Group by using the GloVe and GloVeBERT methods. The stages start with pre-processing which consists of case folding, tokenization, stopwords 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, TF-IDF, and Cosine Similarity of Glove and GloVeBERT by conducting experiments on two datasets namely the Spam Dataset and the BBC News Dataset and show which model performs better.

4.1 Experimental Setup

We conducted experiments on two publicly available datasets, namely the Spam Dataset and the BBC News Dataset, which are the datasets commonly used for information retrieval. The Spam Dataset 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 dataset is a collection of RSS feeds from BBC News which consists of several columns such as title, date and description.

Table 1: Dataset

Dataset	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. beside accuracy, TF-IDF calculations and TF-IDF calculations are also carried out comparison to see which one is better and also calculate cosine similarity to see the similarities between documents in the dataset.

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:

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

In addition, according to the purpose and title of this research, which is to compare the TF-IDF with the TF-IDF Group, it is of course necessary to calculate the TF-IDF and also the TF-IDF Group, with the following formula:

$$tf - idf(t, d, D) = tf(t, d), idf(t, D) \quad (10)$$

$$tf - idf_{group}(t, d_{lust}, D) = tf(t, d_{clust}), idf(t, D) \quad (11)$$

4.3 Evaluation on Effectiveness

The effectiveness comparison that we compare to GloVe method and the GloVeBERT by performing the same two dataset, to the same target set. Table 2 and Table 3 describes the comparison results of Glove and GloVeBERT, in terms of TF-IDF VS TF-IDF Group, Cosine Similarity, and accuracy.

Table 2: Effectiveness comparison with for *Spam* Dataset in terms of Cosine Similarity and accuracy (the higher the value, the better).

Method	Cosine Similarity	Accuracy
GloVe	0.738	0.977
GloVeBERT	0.677	0.954

Based on Table 2, it can be seen that the accuracy value has a higher value than the cosine similarity value in the GloVe and GloVeBERT methods.

Meanwhile, from Table 3, it can be seen that the BBC-News dataset has a higher Cosine similarity value than the accuracy value, either using the GloVe method or the GloVeBERT method.

Table 3: Effectiveness comparison with for *BBC News* Dataset in terms of cosine similarity and accuracy (the higher the value, the better).

Method	Cosine Similarity	Accuracy
GloVe	0.774	0.058
GloVeBERT	0.736	0.0010

Effectiveness TF-IDF VS TF-IDF Group. The tables below contains a comparison between the TF-IDF and the TF-IDF Group on GloVe and GloVeBERT.

Table 4: Comparison of the effectiveness of TF-IDF and TF-IDF Group on Spam Datasets using GloVe and GloVeBERT.

	Term	TF-IDF	TF-IDF Group
GloVe	free	0.0001794	0.0003589
	entry	0.0003589	0.0007178
		0.001076	0.002153
	wkly	0.0001794	0.0003589
	comp	0.0001794	0.0003589
	win	0.0001794	0.0003589
	fa	0.0003589	0.0007178
	cup	0.0001794	0.0003589
	final	0.0001794	0.0003589
	tkts	0.0001794	0.0003589
GloVeBERT	free	0.0001794	0.0003589
	entry	0.0003589	0.0007178
		0.001076	0.002153
	wkly	0.0001794	0.0003589
	comp	0.0001794	0.0003589
	win	0.0001794	0.0003589
	fa	0.0003589	0.0007178
	cup	0.0001794	0.0003589
	final	0.0001794	0.0003589
	tkts	0.0001794	0.0003589

Based on Table 4, it can be seen that the value for TF-IDF in the Spam Dataset has the same value in the GloVe and GloVeBERT methods, as well as the value for TF-IDF Group. From the comparison of the TF-IDF with the TF-IDF Group, the two methods used in this study have a higher score on the TF-IDF Group.

Table 5: Comparison of the effectiveness of TF-IDF and TF-IDF Group on BBC News Datasets using GloVe and GloVeBERT.

	Term	TF-IDF	TF-IDF Group
GloVe	jeremy	0.0008103	0.008103
	bowen	0.0008103	0.008103
		0.0008103	0.008103
	frontline	0.0008103	0.008103
	irpin	0.0008103	0.008103
	,	0.0008103	0.008103
	residents	0.0008103	0.008103
	came	0.0008103	0.008103
	russian	0.0008103	0.008103
	trying	0.0008103	0.008103
	flee	0.0008103	0.008103
	.	0.0008103	0.008103
GloVeBERT	jeremy	0.0008103	0.008103
	bowen	0.0008103	0.008103
		0.0008103	0.008103
	frontline	0.0008103	0.008103
	irpin	0.0008103	0.008103
	,	0.0008103	0.008103
	residents	0.0008103	0.008103
	came	0.0008103	0.008103
	russian	0.0008103	0.008103
	trying	0.0008103	0.008103
	flee	0.0008103	0.008103
	.	0.0008103	0.008103

Based on Table 5, it can be seen that the TF-IDF value in the BBC News dataset has the same value in the GloVe and GloVeBERT methods, as well as the TF-IDF Group in the GloVe and GloVeBERT methods which have the same value.

5 Conclusion

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

1. In the comparison of TF-IDF with TF-IDF Group, the value of TF-IDF Group is higher than TF-IDF in the GloVe and GloVeBERT approaches for the Spam dataset.

2. In the comparison of TF-IDF with TF-IDF Group the value of TF-IDF has the same value as TF-IDF Group on the GloVe approach and the GloVe-BERT approach for the BBC News dataset.
3. The value of Cosine Similarity in both datasets is quite high, this indicates that one document has similarities to the other in both datasets.
4. Combining Glove with BERT has a decreasing effect on both the cosine similarity value and the model accuracy value.
5. Unlike the accuracy values in the Spam dataset in both models, the accuracy values from the Glove and GloVeBERT models in the BBC News dataset are low.

6 Contribution

This section describes the division of tasks in this project.

Table 6: 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 7: STBI Contribution Code

Nama	Task	Dataset	Method
Alex Conro Manuel	Mengerjakan Implementasi Code	Spam	GloVe
	Mengerjakan Implementasi Code	Spam	GloVeBERT
Angelina Naomi Sinaga	Mengerjakan Implementasi Code	BBC News	GloVe
	Mengerjakan Implementasi Code	BBC News	GloVeBERT

Link gitubub : <https://tinyurl.com/TF-IDF-vs-TF-IDF-Group>

References

1. https://balitbangsdm.kominfo.go.id/upt/bandung/?mod=publikasi&action=dl&cid=10&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., Rusjyanthi, 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>