Cosine Similarity-Based Evidences Selection for Fact Verification Using SBERT on the FEVER Dataset

Harya Gusdevi*[1], Arief Setyanto[2], Kusrini[3], Ema Utami[4]
[1,2,3,4] Dept. of Informatics Doctorate, Universitas Amikom Yogyakarta
[1,2,3,4]Jurusan/Program Studi, Fakultas Ilmu Komputer, Universitas Klabat, Airmadidi
*[1]deviharya@students.amikom.ac.id, [2]arief_s@amikom.ac.id, [3]kusrini@amikom.ac.id, [4]ema.u@amikom.ac.id

Abstract

The spread of misinformation on digital platforms has emphasized the urgent need for automated fact verification systems. However, selecting the most semantically relevant evidence to support or refute a claim remains a challenge, especially within the widely used FEVER dataset. Traditional approaches like TF-IDF often fall short in capturing the contextual meaning between claims and evidence. This study addresses the problem by comparing TF-IDF with Sentence-BERT (SBERT) in measuring semantic similarity. The novelty of this research lies in embedding both claims and evidence using SBERT, then calculating cosine similarity to quantify their semantic relevance. Before embedding, standard preprocessing steps were applied, including tokenization, stemming, lowercasing, and stopword removal. A quantitative approach is used to compute cosine similarity between claim-evidence pairs using both TF-IDF and SBERT embeddings. Similarity analysis, distribution statistics, and t-tests are conducted to evaluate the methods. The results show that SBERT achieves higher similarity with the "SUPPORTS" category (0.65) and stronger negative similarity with "NOT ENOUGH INFO" (-0.90), compared to TF-IDF (0.49 and -0.62, respectively). SBERT also demonstrates more stable score distributions and significantly higher t-test values across all label comparisons, indicating stronger semantic discrimination. These findings confirm that SBERT outperforms TF-IDF in identifying the most relevant evidence. The new dataset generated can serve as a foundation for future fact verification model development.

Keywords— Semantic Similarity, Cosine Similarity, Fact Verification, Evidence Selection

1. INTRODUCTION

The rapid development of digital technology today has significantly impacted the dissemination of news information, emphasizing the importance of fact verification to ensure the accuracy of circulating content [1], [2], [3]. Fact verification is the process of analyzing a claim or news item to assign a label that reflects its veracity. In general, the outcome of this process can be categorized into three labels: supported if the claim is proven true, refuted if the claim is proven false, and not enough info if the claim lacks sufficient evidence. This labeling scheme was introduced by Thorne et al., who developed the Fact Extraction and VERification (FEVER) Dataset [4], a large-scale dataset comprising 185,445 claims sourced from Wikipedia articles and manually verified by annotators. The FEVER Dataset has been widely used in various studies aimed at developing automated fact verification systems [5], [6]. The release of the FEVER Dataset has also encouraged the development of other datasets such as FEVEROUS [7], KILT [8], and SciFact [9]. Figure 1 presents a graph showing the number of fact verification dataset usages in research by year.

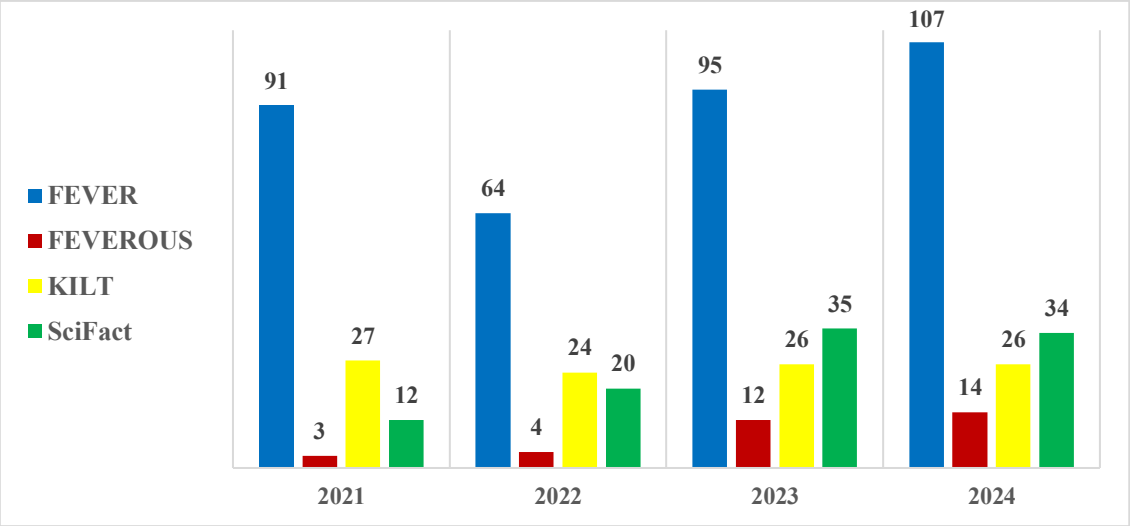

Figure 1. Frequency of Fact Verification Dataset Usage [7], [8], [9], [10]

Figure 1 shows that the FEVER Dataset remains the most widely used dataset in fact verification research throughout the 2021-2024 period. Although there was a significant decline in 2022, its usage sharply increased again in 2023 and reached 107 studies in 2024. This trend indicates that despite the development of various fact verification datasets, FEVER continues to serve as the primary benchmark in this research domain. However, the FEVER Dataset has limitations in evidence presentation structure, where each claim may be supported by more than one piece of evidence. The accuracy of models trained on the FEVER Dataset remains relatively low, with an average of 73% [11], [12], [13], [14], [15], [16], compared to other NLP tasks such as sentiment analysis, which generally achieve around 80% [17], [18], [19], [20], [21], [22]. One of the contributing factors is that models tend to rely more on the quantity of available evidence rather than its quality and relevance. As a result, models often experience shortcut learning, where predictions are made based on the number of evidences without truly understanding the semantic relationship between the claim and the supporting information. This suggests that challenges in fact verification stem not solely from model performance, but also from how the dataset is organized and labeled. Moreover, the gold evidence structure in the FEVER dataset does not explicitly indicate which sentence is the most relevant among the available ones, leading to ambiguity in evidence selection and reducing the reliability of automatic classification systems.

Semantic relationships between sentences have been implemented in several studies, particularly in selecting the most relevant sentences for question answering and text summarization tasks. In question answering, cosine similarity models utilizing Sentence-BERT (SBERT) have been used to measure semantic relevance and identify the most appropriate answers from a set of available questions [23]. In text summarization, semantic relationships are employed to identify core sentences that can represent the overall content of a document using cosine similarity [24], [25]. However, the application of semantic similarity measurement in the context of fact verification has not been widely explored in previous research.

Furthermore, selecting the most relevant evidence for a given claim is a crucial step in fact verification systems. Moreover, the use of TF-IDF, which relies solely on frequency-based word representation [26] without capturing deeper semantic meaning, has limitations in identifying conceptual relationships between words in claims and evidence. Previous studies that applied TF-IDF and transformer-based approaches to fact verification include the work of Jiang Y. et al. [27], who developed a verification pipeline consisting of document retrieval, sentence selection, and claim verification stages. However, most prior works have not fully addressed the semantic gap introduced by TF-IDF's reliance on surface-level lexical matching. As a result, these approaches tend to overlook semantically relevant evidence expressed using different wording, leading to suboptimal evidence selection. In that study, the utilization of transformer models was

still limited to the final classification stage and did not specifically focus on measuring the semantic similarity between claims and evidence. Moreover, the approach did not explicitly evaluate cosine similarity scores as the basis for selecting the most relevant evidence. Therefore, the use of cosine similarity as a method for measuring semantic similarity is a necessary approach for selecting the most appropriate evidence for each claim.

Based on the limitations of the FEVER dataset and the challenges in fact verification, this study aims to measure the semantic relationship between claims and evidence in the FEVER Dataset using SBERT and cosine similarity, which leverages TF-IDF. The semantic similarity results are used to select the most relevant evidence for each claim. The novelty of this research lies in the integration of SBERT-based sentence embeddings with cosine similarity to optimize evidence selection within the FEVER dataset. The optimal distribution benchmark in this study is determined based on three main indicators: (1) a high average cosine similarity value, indicating strong semantic proximity between claims and evidences; (2) a low standard deviation, reflecting score consistency within each label category; and (3) T-Test results demonstrating statistically verifiable differences in the distribution across the SUPPORTS, REFUTES, and NOT ENOUGH INFO label categories. These three indicators are used to assess the quality of semantic representations produced by each method in the evidence selection process. Through this approach, the study produces a new dataset containing claims, the most relevant evidence, labels, and similarity scores, which can serve as a foundation for developing fact verification models in future research.

2. RESEARCH METHODS

This study aims to optimize the selection of the most relevant evidence in the FEVER Dataset by measuring the semantic relationship between claims and evidence using SBERT and cosine similarity. The stages of this research are illustrated in Figure 2.

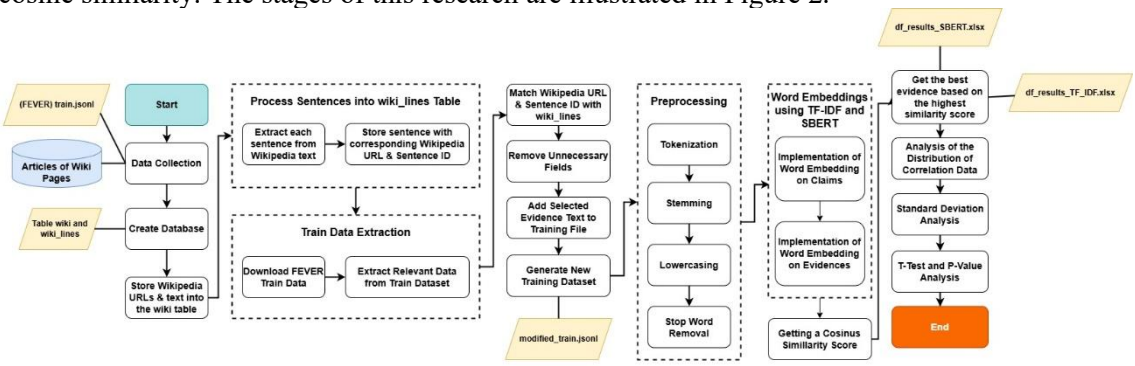

Figure 2. Research Method for Calculating Cosine Similarity on the FEVER Dataset

2.1. Data Collection

At this stage, data collection was carried out by retrieving wiki page documents containing Wikipedia articles that serve as the source of evidence for claim verification. The FEVER Dataset is a large-scale dataset widely used in fact verification research, and it is publicly accessible via https://fever.ai/dataset/fever.html.

This structure consists of three main elements: “id”, which contains the unique identifier of the Wikipedia article; “text”, which stores the full content of the article as a string; and “lines”, which contains each sentence in the article, separated into individual evidence units. Each sentence is marked with an index number and separated by a tab character (\t). The structure and example content of the wiki page documents are shown in Figure 3.

{
"id" = "1998_All-Ireland_Senior_Hurling_Championship"

"text" = "The All-Ireland Senior Hurling Championship of 1998 -LRB- known for sponsorship reasons as the Guinness Hurling Championship 1998 -RRB- was the 112th staging of Ireland 's premier hurling competition . Offaly won the championship , beating Kilkenny 2-16 to 1-13 in the final at Croke Park , Dublin ."

"lines" = "0\tThe All-Ireland Senior Hurling Championship of 1998 -LRB- known for sponsorship reasons as the Guinness Hurling Championship 1998 -RRB- was the 112th staging of Ireland 's premier hurling competition .\tIreland\tIreland\thurling\thurling\n1\tOffaly won the championship , beating Kilkenny 2-16 to 1-13 in the final at Croke Park , Dublin .\tOffaly\tOffaly GAA\tKilkenny\tKilkenny GAA\tfinal\t1998 All-Ireland Senior Hurling Championship Final\tCroke Park\tCroke Park\tDublin\tDublin\n2\t"
}

Figure 3. Example Content of the FEVER Dataset

The FEVER dataset consists of 185,445 human-annotated claims categorized into three classes: SUPPORTS, REFUTES, and NOT ENOUGH INFO, with each class having a relatively balanced distribution. Evidence sentences are taken from a separate wiki_pages file, which contains pre-processed articles sourced from Wikipedia. Each claim is linked to one or more sentences from these articles that either support, refute, or provide insufficient information regarding the claim.

2.2. Create Database

At this stage, a database was designed to store and manage Wikipedia articles and individual sentences used as evidence in the fact verification process. At this stage, a database was designed to store and manage Wikipedia articles and individual sentences used as evidence in the fact verification process. Figure 4 shows the database structure, where the wiki table contains the full text of each Wikipedia article, and the wiki_lines table stores the individual sentences from those articles. Each article can have multiple sentences, which is represented by a one-to-many relationship between the wiki and wiki_lines.

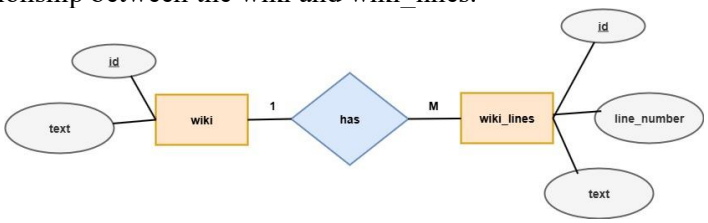

Figure 4. Entity-Relationship Diagram of the Modified FEVER Dataset

The database consists of two main tables, namely wiki and *wiki_lines*, which have a one-to-many (1:M) relationship. The wiki table is used to store entire Wikipedia articles, while the *wiki_lines* table stores the sentence index and the corresponding sentence text. The table structure designed in this study is shown in Figure 5.

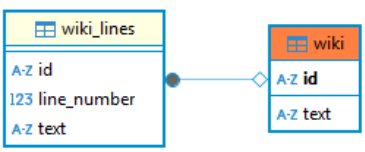

Figure 5. Schema Relationship of the Modified FEVER Dataset Tables

Furthermore, the attributes of the wiki and *wiki_lines* tables are presented in Table 1.

Table 1. Modified FEVER Dataset Table Attributes

Table	Attributes	Description
wiki	Id	Primary Key, a unique identifier for each Wikipedia article.
	text	Stores the full content of the corresponding Wikipedia article.
wiki_lines	id	Foreign Key referring to the wiki.id, indicating the source article of each sentence.
	line_number	The sequential number of the sentence within the Wikipedia article.
	text	The text content of each sentence is extracted from the Wikipedia article.

2.3. Store Wikipedia URLs and Text into a Table

At this stage, the data from *wiki_pages* obtained from the FEVER Dataset is stored in a database to facilitate evidence retrieval. The stored data includes the Wikipedia URL and the article text, which are saved in the wiki table and serve as sources of evidence for fact verification. Subsequently, each sentence and its position within the article are stored in the *wiki_lines* table.

2.4. Data Extraction

The next step is to extract the claim ID, label, claim, and evidence data (Wikipedia URL and Sentence ID) from the training data, which can be downloaded from https://fever.ai/download/fever/train.jsonl. Train Data Extraction refers to the process of extracting the training data provided by the FEVER dataset, which includes claims and their associated labels. This step involves downloading the original train jsonl file and selecting relevant entries that are then matched with corresponding evidence texts extracted from Wikipedia. The structure of the replicated dataset is presented in Table 2.

Table 2. Attributes of the FEVER Dataset Training Data

No	Attribute	Description
1	id	A unique identifier for each claim in the dataset.
2	Label	The claim label indicates whether the claim is SUPPORTS, REFUTES, or NOT ENOUGH INFO.
3	Claim	The text of the claim is to be verified.
4	Evidences	A list of evidence related to the claim, represented as tuples: [Wikipedia URL, Sentence ID, Text Evidences].
5	Annotation ID	The annotation ID is used for internal debugging and evaluation purposes, not publicly released.
6	Evidences ID	An ID referencing specific evidence for internal debugging and evaluation purposes, not publicly released.
7	Wikipedia URL	The URL of the Wikipedia page serving as the source of evidence is found in the wiki-pages document.
8	Sentence ID	The ID of the sentence in the Wikipedia page that serves as the source of evidence found in the wiki-pages document.

2.5. Match Wikipedia URL and Sentence ID, Remove Unnecessary Fields, and Add Selected Evidence Text to Training File

At this stage, a matching process is performed between the Wikipedia URL and the Sentence ID from the training dataset and the data stored in the *wiki_lines* table. This process aims to link the claims in the FEVER Dataset with the relevant evidence from the corresponding Wikipedia articles. The matching is carried out through the following steps:

a. Extract the Wikipedia URL and Sentence ID from the training dataset.

b. Search for the corresponding entry in the *wiki_lines* table based on the combination of Wikipedia URL and Sentence ID.
c. Retrieve the evidence text from the *wiki_lines* table to be paired with the corresponding claim.

The matching process is used to ensure that each claim in the dataset is linked to relevant evidence, which is essential in the FEVER Dataset replication stage. Subsequently, unused fields such as *annotation_id* and *evidences_id* are removed, as they are not required for the fact verification process. An example of the replicated FEVER dataset can be seen in Figure 5. In Figure 6, once the matching of the Wikipedia URL and Sentence ID is complete, the selected evidence text is added to the training file. This stage ensures that every claim in the FEVER Dataset is matched with the correct evidence in a format better suited for fact verification. The output of this process is a new training file named *modified_train.jsonl,* which contains cleaned and reformatted data where each claim is linked to the corresponding evidence text.

{
"id" = 75397,

"**label**" = "SUPPORTS"

"claim" = "Nikolaj Coster-Waldau worked with the Fox Broadcasting Company"
"evidences" = [
["Nikolaj_Coster-Waldau", 7, "He then played Detective John Amsterdam in the short-lived Fox television series New Amsterdam -LRB- 2008 -RRB- , as well as appearing as Frank Pike in the 2009 Fox television film Virtuality , originally intended as a pilot .
["Fox_Broadcasting_Company", 0, "The Fox Broadcasting Company -LRB- often shortened to Fox and stylized as FOX -RRB- is an American English language commercial broadcast television network that is owned by the Fox Entertainment Group subsidiary of 21st Century Fox"]
]
}

Figure 6. FEVER Dataset Modification Result

2.6. Preprocessing

The preprocessing stage aims to clean and standardize the claim and evidence texts before they are used in semantic analysis. This process begins with tokenization, which involves splitting the text into individual word units or tokens. Next, stemming is applied to reduce words to their root forms, thereby minimizing word variation. To ensure consistency, all text is converted to lowercase through a lowercasing process, eliminating distinctions between uppercase and lowercase letters. Additionally, stop word removal is performed to eliminate common words that do not carry significant meaning in semantic analysis. These preprocessing techniques were chosen to reduce lexical variability and standardize the input text before embedding. Stemming was preferred over lemmatization due to its faster performance and lower computational cost, which is important considering the large size of the FEVER dataset. Additionally, SBERT as a contextual embedding model is inherently robust to minor morphological variations, making stemming a sufficient and efficient choice in this context.

2.7. Word Embeddings and Getting a Cosine Similarity Score using TF-IDF and SBERT

At this stage, a word embedding process is conducted using Term Frequency-Inverse Document Frequency (TF-IDF) and Sentence-BERT (SBERT) to convert the claims and evidence in the dataset into numerical representations in the form of vectors. In this study, we used the mpnet-base model from the Sentence Transformers library due to its strong performance in

semantic similarity benchmarks and its balanced trade-off between accuracy and efficiency. This pre-trained model was selected to generate context-aware sentence embeddings that capture both syntactic and semantic relationships between claims and evidence. In SBERT, the word embedding process begins by applying an embedding to the claims, where each claim is converted into a vector that represents the semantic meaning of the text. The SBERT-based similarity process is illustrated in Figure 7.

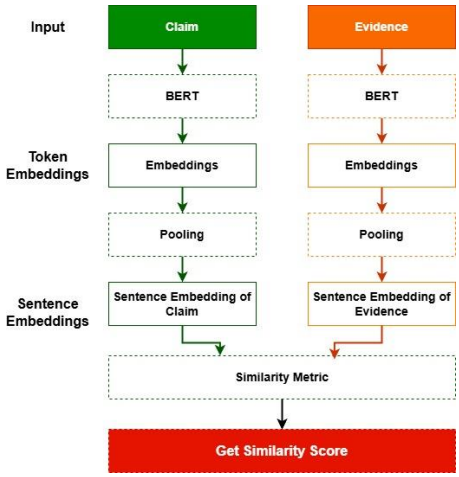

Figure 7. Claim and Evidence Similarity Process using SBERT

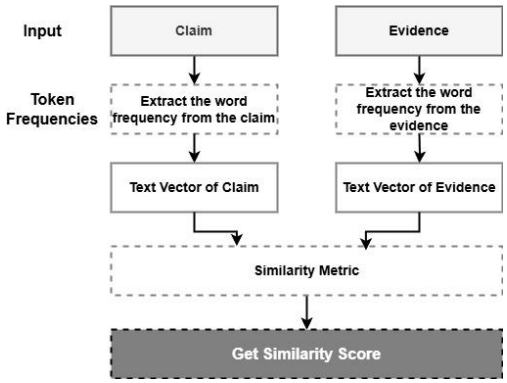

Figure 8. Claim and Evidence Similarity Process using TF-IDF

Subsequently, word embedding is performed on the evidence sentences to obtain their corresponding numerical representations within the dataset. After the embeddings are generated, cosine similarity is calculated between each claim and its corresponding evidence to measure the degree of semantic similarity between them. The result of this stage is the selection of the most relevant evidence based on the highest similarity score, reflecting the closest semantic match to the claim. This selected evidence is then used in the semantic relationship analysis within the fact verification process.

Furthermore, in Figure 8, TF-IDF is also employed as a baseline approach to calculate textual similarity between claims and evidence. In this approach, each sentence is first transformed into a numerical vector based on the frequency of words it contains, weighted by how rare or common the words are across the dataset, which is measured using inverse document frequency. Specifically, the process begins by extracting the word frequency from each claim and evidence sentence, and then converting these into TF-IDF vectors. After both vectors are obtained, a similarity metric such as cosine similarity is applied to measure the alignment between the two texts. The evidence with the highest similarity score is selected as the best match for the claim.

In this study, cosine similarity scores between claim-evidence pairs are used to analyze the semantic relationship across different label categories in the FEVER dataset: SUPPORTS, REFUTES, and NOT ENOUGH INFO. These scores are not used to predict labels directly, but rather to observe how the similarity distributions vary among categories. This analysis helps evaluate the effectiveness of SBERT embeddings in capturing semantic relevance and provides insights into their potential as a basis for evidence selection and future classification strategies.

2.8. Analysis of the Distribution of Similarity Data

This stage aims to analyze the distribution of Cosine Similarity scores based on the labels in the FEVER dataset, namely SUPPORTS, REFUTES, and NOT ENOUGH INFO.

$$\min(X) = \min\{x_1, x_2, x_3, \ldots x_4\} \quad (1)$$

$$\max(X) = \max\{x_1, x_2, x_3, \ldots x_4\} \quad (2)$$

$$\mu = \frac{1}{N}\sum_{i=1}^{n} x_i \quad (3)$$

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{, if n is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{, if n is even} \end{cases} \quad (4)$$

The analysis is carried out by calculating the minimum (min), maximum (*max*), mean (μ), and median ($\tilde{x}$) values of the Cosine Similarity for each category. The data distribution is computed using Equations 1 through 4, where X represents the set of Cosine Similarity values from the dataset, and n is the total number of data points within each category.

2.9. Data Standard Deviation Analysis

In this study, standard deviation is used to measure the extent of variation or dispersion of Cosine Similarity values within the dataset. A low standard deviation indicates that the data points are closely clustered around the mean (μ), while a high standard deviation suggests that the data is more widely spread. In the context of this research, standard deviation is analyzed to understand the variability of semantic similarity between claims and evidence across each label category (SUPPORTS, REFUTES, NOT ENOUGH INFO). To compute the standard deviation (s), the variability is first calculated as shown in Equation 5, followed by the calculation of the standard deviation using Equation 6.

$$variance = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \quad (5)$$

$$standard\ deviation = \sqrt{variance} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \quad (6)$$

Where variance is a measure of data dispersion that indicates how far individual values in the dataset deviate from the μ, N is the total number of data points in the analyzed category, and x_i is the Cosine Similarity score of the *i-th* claim in the dataset. Each claim has a Cosine Similarity score corresponding to the selected evidence, which is then used in the distribution analysis.

2.10. T-Test and P-Value Analysis

This stage aims to test the significance of differences in Cosine Similarity scores among the SUPPORTS, REFUTES, and NOT ENOUGH INFO categories in the FEVER dataset. A *T-test* is used to determine whether there is a statistically significant difference in the distribution of Cosine Similarity scores between two groups of data ($\bar{X}$), for example, the mean Cosine Similarity scores of SUPPORTS and REFUTES. Furthermore, the *p-value* is used to assess whether the observed difference is statistically significant or merely due to chance. The T-test calculation is presented in Equation 7, and the p-value computation is shown in Equation 8.

$$T - test = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (7)$$

$$p - value = P(T > |t|) \quad (8)$$

This is computed using the formula shown in Equation (7), where s represents the sample variances within each group, and n denotes the number of data points in each respective group. Furthermore, Equation (8) represents the probability function (P) that a t-distributed (T), under the null hypothesis of no difference, is greater than the absolute value of the observed t-statistic. In this context, t represents the resulting test statistic that quantifies the standardized difference

between the two group means based on their variances and sample sizes. A small p-value (typically less than 0.05) indicates that the difference between the two groups is statistically significant, suggesting a meaningful variation in semantic similarity patterns across the dataset labels being compared.

3. RESULT AND DISCUSSION

This section presents the results of the Cosine Similarity distribution analysis, standard deviation statistical test, as well as the T-test and p-value analysis to evaluate the semantic relationship between claims and evidence in the FEVER Dataset. The discussion includes the interpretation of statistical values and their implications for the effectiveness of the SBERT and Cosine Similarity approach in selecting the most relevant evidence. The results of this study demonstrate that SBERT-based evidence selection yields stronger semantic similarity compared to TF-IDF, as reflected in the similarity score distributions in Figures 7 and 8, as well as the statistical analysis presented in Tables 3 and 4. The SBERT-based approach not only produces higher semantic similarity but also contributes to the construction of a higher-quality dataset that is well-suited for training deep learning-based fact verification models. The optimal distribution indicators are derived from the highest cosine similarity values, low standard deviation, and statistically validated differences across labels as evidenced by the T-Test [28].

The similarity score measurement in this study is used to analyze the data distribution within the modified FEVER dataset, in which the labels are already provided by the original dataset. The results are then analyzed to observe how the distribution of cosine similarity values differs across label categories (SUPPORTS, REFUTES, and NOT ENOUGH INFO), as well as to assess how effectively the cosine similarity method captures semantic meaning based on the closeness between claim and evidence sentences. This approach evaluates the potential of similarity scores as a basis for decision-making or threshold determination in future classification stages.

3.1. Result

This study analyzes the relationship between cosine similarity scores and claim categories in the FEVER Dataset. Based on the results shown in Figure 9, the TF-IDF method yields a cosine similarity score of only 0.49 for the SUPPORTS category, indicating that this method is not yet optimal in capturing the semantic closeness between claims and supporting evidence. This result is attributed to the fundamental nature of TF-IDF, which models text as vectors based on the relative frequency of terms within the entire corpus. Such representations are sparse and lack contextual information between words, causing semantically identical sentences that use synonyms or paraphrases to be considered dissimilar [29]. As a result, TF-IDF is inadequate for capturing deep semantic relationships that are essential in fact verification tasks. For the REFUTES category, the similarity score is even lower, at 0.05, suggesting that TF-IDF is less effective in distinguishing claims that contradict the evidence. Meanwhile, the NOT ENOUGH INFO category yields a score of -0.62, implying that even in the absence of evidence, TF-IDF tends to assign a relatively high semantic similarity, which complicates the process of identifying claims that lack sufficient supporting information.

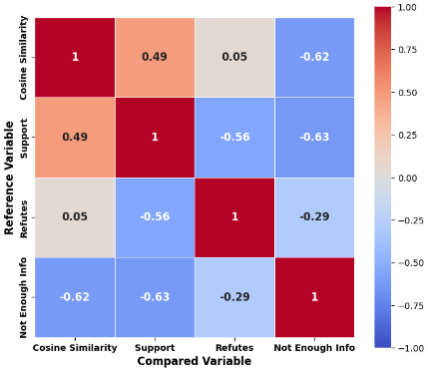

Figure 9. Cosine Similarity using TF-IDF

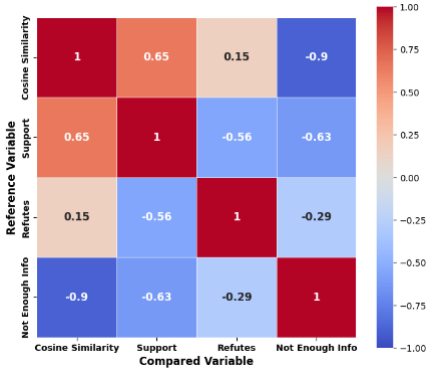

Figure 10. Cosine Similarity using SBERT

In contrast, as illustrated in Figure 10, the SBERT-based method demonstrates better performance in measuring semantic similarity. The cosine similarity score for the SUPPORTS category increases to 0.65, indicating SBERT's stronger ability to capture the semantic relationship between claims and the supporting evidence. For the REFUTES category, the score reaches 0.15, which, although still relatively low, is higher than that of TF-IDF, indicating an improvement in distinguishing contradictory claims. The most notable result appears in the NOT ENOUGH INFO category, where the cosine similarity score sharply decreases to -0.90, suggesting that SBERT is more effective in identifying claims that lack evidence by assigning them very low semantic similarity scores. These findings confirm that the SBERT-based approach offers superior capability in capturing semantic similarity between claims and evidence compared to the TF-IDF-based method. The transformer-based contextual embeddings employed by SBERT yield sharper cosine similarity scores that reflect deeper semantic relationships between claim and evidence pairs. SBERT encodes entire sentences into fixed-dimensional dense representations by incorporating positional encoding, self-attention, and both syntactic and semantic relationships between words within a sentence. Through this mechanism, SBERT is better equipped to interpret the meaning of sentences, even when surface forms differ, in contrast to TF-IDF [30]. Furthermore, to evaluate the degree of semantic similarity between claims and evidence, a distribution analysis of cosine similarity scores is presented in Table 3.

Table 3. Analysis of the Distribution of Cosine Similarity Scores

Method	Label	Min	Max	Mean	Median	Standard Deviation
TF-IDF	SUPPORTS	0.000	1.000	0.283	0.257	0.170
	REFUTES	0.000	1.000	0.218	0.191	0.149
	NOT ENOUGH INFO	0.000	0.000	0.000	0.000	0.000
SBERT	SUPPORTS	0.025	1.000	0.655	0.675	0.145
	REFUTES	-0.016	0.987	0.5688	0.5804	0.144
	NOT ENOUGH INFO	0.000	0.000	0.000	0.000	0.000

Based on the results presented in Table 3, the SBERT model shows higher mean and median values compared to TF-IDF, particularly for the SUPPORTS and REFUTES labels. This indicates that the context-based representations produced by SBERT are more effective in capturing the semantic relationship between claims and evidence. For example, the average cosine similarity score for the SUPPORTS label using SBERT is 0.655, which is significantly higher than 0.283 obtained with TF-IDF. Similarly, for the REFUTES label, SBERT achieves a mean score of 0.5688, while TF-IDF only reaches 0.218. Meanwhile, for the NOT ENOUGH INFO label, both methods yield a cosine similarity score of 0.000, indicating that there is no meaningful

similarity between the claim and the provided evidence, consistent with the nature of this category. In terms of standard deviation analysis, TF-IDF shows slightly higher deviation scores (Support: 0.170, Refutes: 0.149) compared to SBERT (Support: 0.145, Refutes: 0.144). For example, the similarity results that serve as the basis for deriving the minimum, maximum, mean, median, and standard deviation values can be found in Table 4.

Table 4. Best Similarity Examples from the Replicated FEVER Dataset

Method	Claim	Best Evidences	Best Similarity
TF-IDF	Roman Atwood builds his career as a content creator.	His popularity stems from his regular vlogs sharing life updates with viewers.	0
	The show Stranger Things is set in the small town of Bloomington, Indiana.	The debut season, set in 1980s Hawkins, centers on the search for a missing boy by his family, friends, and local law enforcement. Paranormal occurrences multiply, and a girl with psychic abilities emerges as pivotal to the investigation.	0.075
	System of a Down disbanded briefly, leaving fans in suspense about their status.	-	0
SBERT	Roman Atwood is a content creator.	The creator also operates the 'RomanAtwood' channel as an outlet for his prank-related content.	0.4903
	Stranger Things is located in the fictional town of Bloomington, Indiana.	The debut season, set in 1980s Hawkins, centers on the search for a missing boy by his family, friends, and local law enforcement. Paranormal occurrences multiply, and a girl with psychic abilities emerges as pivotal to the investigation.	0.5337
	System of a Down entered a period of indefinite hiatus, leaving their future uncertain.	-	0

In Table 4, these results suggest that cosine similarity scores from TF-IDF are more dispersed, while those from SBERT are more tightly clustered and stable, reflecting the model's consistency in detecting semantic similarity. These findings reinforce the argument that SBERT is more capable of capturing the overall meaning of a sentence rather than relying solely on word frequency, as TF-IDF does. Consequently, SBERT demonstrates stronger potential for fact verification tasks that depend on semantic similarity between claims and evidence. The SUPPORTS data under TF-IDF shows an inability to perform semantic analysis, as indicated by a similarity score of 0.000. In contrast, SBERT can capture semantic relations between sentences, achieving a similarity score of 0.4903. Following this analysis, a two-sample t-test was conducted to examine whether the differences in cosine similarity scores across labels are statistically significant for both TF-IDF and SBERT methods. The similarity scores used in the T-test analysis were derived from the full set of cosine similarity values calculated for each claim and evidence pair within the replicated FEVER dataset. These scores were grouped based on their respective label categories (SUPPORTS, REFUTES, and NOT ENOUGH INFO) and computed separately using both the TF-IDF and SBERT methods. The results are presented in Table 5.

Table 5. T-Tes dan P-Value Testing

Method	Comparison	T-Test	P-Value
TF-IDF	SUPPORTS vs REFUTES	57,198	0.000
	SUPPORTS vs NOT ENOUGH INFO	315,061	0.000
	REFUTES vs NOT ENOUGH INFO	278,393	0.000
SBERT	SUPPORTS vs REFUTES	88.096	0.000
	SUPPORTS vs NOT ENOUGH INFO	851,035	0.000
	REFUTES vs NOT ENOUGH INFO	741,183	0.000

Based on the T-test results, it can be concluded that there is a statistically significant difference (*p-value* < 0.001) between each pair of labels for both the TF-IDF and SBERT methods. Additionally, the t-statistic values produced by SBERT are consistently higher than those of TF-IDF. These findings indicate that SBERT is more effective in distinguishing fact categories based on cosine similarity scores. A higher t-statistic value indicates that the difference between group means is large relative to the variation within each group, suggesting a stronger separation between categories such as SUPPORTS, REFUTES, and NOT ENOUGH INFO based on their cosine similarity distributions.

3.2. Discussion

The results of this study highlight the differences between the TF-IDF and SBERT methods in measuring semantic similarity between claims and evidence in the FEVER dataset. Although TF-IDF is a commonly used method for measuring textual similarity, the findings of this research reveal its limitations in the context of fact verification. This limitation has also been acknowledged in previous studies, where transformer-based models outperformed TF-IDF in both fact verification [27] and fact-checking tasks [31]. In the present study, the low average cosine similarity scores and high standard deviations observed in the SUPPORTS and REFUTES categories indicate that TF-IDF lacks the sensitivity to effectively distinguish between claims that are supported by evidence and those that are not. This limitation arises because TF-IDF relies solely on word frequency and does not consider the semantic context between words in a sentence. As a result, two semantically similar sentences using different words may be considered dissimilar by TF-IDF. The high-dimensional vectors produced by TF-IDF tend to hinder the model's ability to generalize, and when word occurrences are sporadic across documents, the resulting representations become increasingly unstable. The similarity distribution also becomes more dispersed, as indicated by the high standard deviation in cosine similarity scores across labels [29], [30]. This demonstrates that TF-IDF is not only weak in capturing semantic similarity but also inconsistent in measuring the relationship between claims and evidence. Given these limitations, TF-IDF is less suitable for automated fact verification systems that require cross-sentence semantic understanding and precise measurement of contextual support.

In contrast, the SBERT-based approach demonstrates significantly better performance. While TF-IDF accounts solely for individual word frequencies, SBERT captures holistic relationships between words within a sentence, including word order, grammatical structure, and semantic dependencies [32]. The model is trained using a triplet loss strategy [33], which enables the mapping of semantically similar sentences into proximate vector representations, while pushing dissimilar sentences further apart. In the context of fact verification, SBERT is capable of identifying those two sentences using paraphrases, synonyms, or structural variations that can still convey equivalent meaning, an ability that traditional statistical methods do not possess. Furthermore, SBERT produces low-dimensional dense embeddings, which not only accelerate inference processes but also improve the stability and consistency of similarity measurements. This is evident from the lower standard deviation observed in cosine similarity score distributions across labels, indicating more robust model behavior. SBERT also demonstrates clearer separation margins across label categories such as SUPPORTS, REFUTES, and NOT ENOUGH INFO, making it particularly well-suited for classification tasks that rely on thresholding or confidence-based scoring.

These findings reinforce the argument that pre-trained language model approaches such as SBERT are superior in capturing complex semantic relationships between texts. Furthermore, the *T-test* statistical results confirm that the differences in the distribution of cosine similarity scores across label categories are statistically significant for both TF-IDF and SBERT. The substantially higher *t-statistic* values observed in SBERT indicate that this model is not only more accurate but also more consistent and decisive in distinguishing between supported, refuted, and insufficiently evidenced claims.

4. CONCLUSION

This study aims to measure the semantic relationship between claims and evidence in the FEVER Dataset using a cosine similarity approach based on two text representation methods: TF-IDF and SBERT. The analysis reveals that the TF-IDF approach has limitations in distinguishing semantic proximity between claims and evidence. This is reflected in the generally low cosine similarity scores and unstable value distributions across different claim categories. In contrast, SBERT demonstrates a more accurate ability to capture semantic context, producing more consistent similarity score distributions and significant differences across categories, as confirmed by the results of the T-test. By integrating SBERT-based sentence embeddings with cosine similarity, this study successfully optimizes the process of selecting the most relevant evidence for a given claim in the FEVER Dataset. By applying SBERT with cosine similarity, this study not only successfully measured the semantic similarity between claims and evidence but also optimized the process of selecting the most relevant evidence. Another contribution is the construction of a new dataset consisting of claims, evidence pairs, labels, and semantic similarity scores. This dataset holds significant potential as a foundation for training and evaluating deep learning-based fact verification models. Furthermore, this research produces a new dataset containing claims, selected evidence, labels, and semantic similarity scores. This dataset can serve as a valuable foundation for the development of deep learning-based fact verification models in future studies.

However, this study also has several limitations. The cosine similarity scores generated by SBERT do not always reflect consistent clustering of labels. Claims with supporting or refuting evidence may sometimes result in low similarity scores due to linguistic variation, such as the use of different vocabulary or sentence structures, even when semantically relevant. Additionally, the method is highly dependent on the quality and coverage of the pre-trained SBERT model, which may limit generalization to other domains. The computation cost of embedding large-scale claim-evidence pairs is also non-trivial. As a result, determining appropriate similarity thresholds for each label category must take into account distributional statistics such as mean and median similarity scores.

For future research, this approach can be extended by evaluating the integration of various SBERT model variants to assess the robustness and consistency of similarity score distributions across different architectures, such as bert-base-nli-mean-tokens, distilroberta-base, or roberta-large-nli-stsb-mean-tokens. Additionally, exploring hybrid similarity scoring methods that combine SBERT with traditional statistical weighting schemes, or testing different embedding aggregation strategies, may yield further insights into semantic matching performance. Future studies may also consider transforming this exploratory similarity analysis into a fully supervised classification pipeline by applying threshold-based or learned decision boundaries using the constructed dataset. Furthermore, the newly generated dataset consisting of claim-evidences-similarity-label tuples can be leveraged to train and evaluate deep learning-based fact verification models in a more end-to-end manner.

REFERENCES

[1] M. K. H. Al Asy ari and M. Rahman, “Technology: Technological Advances and Changes in Human Lifestyles in a Socio-Cultural Perspective,” *Proceeding International Conference on Science and Engineering*, vol. 3, pp. 721–730, Apr. 2020, doi: 10.14421/icse.v3.592.

[2] P. C. Verhoef *et al.*, “Digital transformation: A multidisciplinary reflection and research agenda,” *J Bus Res*, vol. 122, pp. 889–901, Jan. 2021, doi: 10.1016/j.jbusres.2019.09.022.

[3] C. López-Marcos and P. Vicente-Fernández, “Fact checkers facing fake news and disinformation in the digital age: A comparative analysis between Spain and United Kingdom,” *Publications*, vol. 9, no. 3, 2021, doi: 10.3390/publications9030036.

[4] A. Krishna, S. Riedel, and A. Vlachos, "ProoFVer: Natural Logic Theorem Proving for Fact Verification," Aug. 2021, [Online]. Available: http://arxiv.org/abs/2108.11357

[5] B. Portelli, J. Zhao, T. Schuster, G. Serra, and E. Santus, "Distilling the Evidence to Augment Fact Verification Models," 2020.

[6] S. Subramanian and K. Lee, "Hierarchical Evidence Set Modeling for Automated Fact Extraction and Verification." [Online]. Available: https://github.com/

[7] R. Aly *et al.*, "FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.05707

[8] F. Petroni *et al.*, "KILT: a Benchmark for Knowledge Intensive Language Tasks," Sep. 2020, [Online]. Available: http://arxiv.org/abs/2009.02252

[9] D. Wadden *et al.*, "Fact or Fiction: Verifying Scientific Claims," Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.14974

[10] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a Large-scale Dataset for Fact Extraction and VERification," pp. 809–819, 2018, doi: 10.18653/v1/n18-1074.

[11] T. Schuster, A. Fisch, and R. Barzilay, "Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.08541

[12] A. Krishna, S. Riedel, and A. Vlachos, "ProoFVer: Natural Logic Theorem Proving for Fact Verification," Aug. 2021, [Online]. Available: http://arxiv.org/abs/2108.11357

[13] Y. Liu, C. Zhu, and M. Zeng, "Modeling Entity Knowledge for Fact Verification," in *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 50–59. doi: 10.18653/v1/2021.fever-1.6.

[14] B. Zhu, X. Zhang, M. Gu, and Y. Deng, "Knowledge Enhanced Fact Checking and Verification," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 29, pp. 3132–3143, 2021, doi: 10.1109/TASLP.2021.3120636.

[15] Y. Du, A. Bosselut, and C. D. Manning, "Synthetic Disinformation Attacks on Automated Fact Verification Systems," Feb. 2022, [Online]. Available: http://arxiv.org/abs/2202.09381

[16] Z. Chen *et al.*, "A syntactic evidence network model for fact verification," *Neural Networks*, vol. 178, Oct. 2024, doi: 10.1016/j.neunet.2024.106424.

[17] M. T. Uliniansyah *et al.*, "Twitter dataset on public sentiments towards biodiversity policy in Indonesia," *Data Brief*, vol. 52, p. 109890, Feb. 2024, doi: 10.1016/j.dib.2023.109890.

[18] A. A. Firdaus, A. Yudhana, I. Riadi, and Mahsun, "Indonesian presidential election sentiment: Dataset of response public before 2024," *Data Brief*, vol. 52, Feb. 2024, doi: 10.1016/j.dib.2023.109993.

[19] A. Athar, S. Ali, M. M. Sheeraz, S. Bhattachariee, and H.-C. Kim, "Sentimental Analysis of Movie Reviews using Soft Voting Ensemble-based Machine Learning," no. March, pp. 01–05, 2022, doi: 10.1109/snams53716.2021.9732159.

[20] Z. Chen *et al.*, "How does the perception of informal green spaces in urban villages influence residents' complaint Sentiments? a Machine learning analysis of Fuzhou City, China," *Ecol Indic*, vol. 166, Sep. 2024, doi: 10.1016/j.ecolind.2024.112376.

[21] N. A. Rakhmawati, M. I. Aditama, R. I. Pratama, and K. H. U. Wiwaha, "Analisis Klasifikasi Sentimen Pengguna Media Sosial Twitter Terhadap Pengadaan Vaksin

COVID-19," *Journal of Information Engineering and Educational Technology*, vol. 4, no. 2, pp. 90–92, 2020, doi: 10.26740/jieet.v4n2.p90-92.

[22] C. J. Varshney, A. Sharma, and D. P. Yadav, "Sentiment analysis using ensemble classification technique," *2020 IEEE Students' Conference on Engineering and Systems, SCES 2020*, no. July, 2020, doi: 10.1109/SCES50439.2020.9236754.

[23] A. Condor, M. Litster, and Z. Pardos, "Automatic short answer grading with SBERT on out-of-sample questions."

[24] J. Opitz and A. Frank, "SBERT studies Meaning Representations: Decomposing Sentence Embeddings into Explainable Semantic Features," Jun. 2022, [Online]. Available: http://arxiv.org/abs/2206.07023

[25] T. P. Rinjeni, A. Indriawan, and N. A. Rakhmawati, "Matching Scientific Article Titles using Cosine Similarity and Jaccard Similarity Algorithm," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 553–560. doi: 10.1016/j.procs.2024.03.039.

[26] O. A. Resta, A. Aditya, and F. E. Purwiantono, "Plagiarism Detection in Students' Theses Using The Cosine Similarity Method," *SinkrOn*, vol. 5, no. 2, pp. 305–313, May 2021, doi: 10.33395/sinkron.v5i2.10909.

[27] Y. Jiang, S. Bordia, Z. Zhong, C. Dognin, M. Singh, and M. Bansal, "HoVer: A Dataset for Many-Hop Fact Extraction And Claim Verification," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.03088

[28] S. Lestari, M. Z. Dj, and U. Hasanah, "THE CORRELATION BETWEEN READING INTERNATIONAL JOURNAL ARTICLES ON ENRICHING THE UNIVERSITY EFL STUDENTS' ACADEMIC VOCABULARY," 2023.

[29] F. Lan, "Research on Text Similarity Measurement Hybrid Algorithm with Term Semantic Information and TF-IDF Method," *Advances in Multimedia*, vol. 2022, 2022, doi: 10.1155/2022/7923262.

[30] A. D. Susanto, S. Andrian Pradita, C. Stryadhi, K. E. Setiawan, and M. Fikri Hasani, "Text Vectorization Techniques for Trending Topic Clustering on Twitter: A Comparative Evaluation of TF-IDF, Doc2Vec, and Sentence-BERT," in *2023 5th International Conference on Cybernetics and Intelligent Systems, ICORIS 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICORIS60118.2023.10352228.

[31] P. J. Verschuuren, J. Gao, A. van Eeden, S. Oikonomou, and A. Bandhakavi, "Logically at Factify 2: A Multi-Modal Fact Checking System Based on Evidence Retrieval techniques and Transformer Encoder Architecture," Jan. 2023, doi: https://doi.org/10.48550/arXiv.2112.09253.

[32] A. Martín, J. Huertas-Tato, Á. Huertas-García, G. Villar-Rodríguez, and D. Camacho, "FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference," *Knowl Based Syst*, vol. 251, Sep. 2022, doi: 10.1016/j.knosys.2022.109265.

[33] Y. Chu, H. Cao, Y. Diao, and H. Lin, "Refined SBERT: Representing sentence BERT in manifold space," *Neurocomputing*, vol. 555, Oct. 2023, doi: 10.1016/j.neucom.2023.126453.