Semantic Similarity Detection: From Traditional Models to Advanced SBERT

Jingyi Zhang, Ruoxi Jin

Boston University, Boston, USA

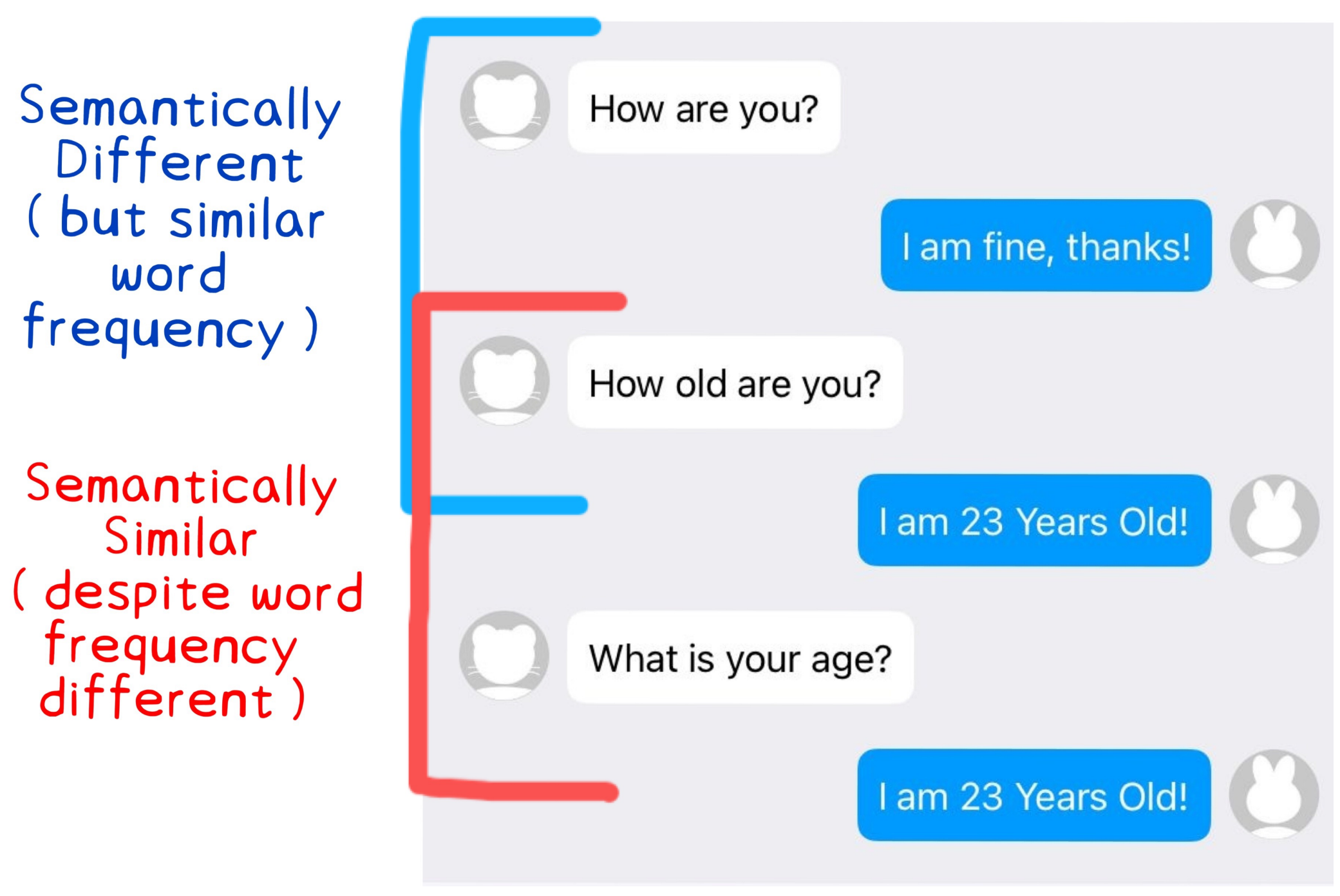
Introduction

This project focuses on measuring semantic similarity between sentence pairs (e.g., ”A plane is taking off.” and ”An airplane is taking off.”), with a similarity score output (e.g., 0.85). This can be used for tasks like information retrieval, question answering, machine translation evaluation, and plagiarism detection.

We evaluate baseline models (Bag-of-Words and TF-IDF) and an advanced model (Sentence-BERT) using metrics such as F1-Score, Accuracy, Pearson, and Spearman Correlation. By comparing predicted scores with original human annotations, we analyze each model’s performance and reliability and conclude that SBERT model performs excellent, demonstrating its effectiveness for these applications.

Motivation

Semantic similarity plays a critical role in natural language processing (NLP), impacting tasks such as text comparison, machine translation, and information retrieval. Traditional methods for measuring similarity often rely on surface-level comparisons, such as word frequency or keyword matching. While these methods can be effective for simpler tasks, they often fall short when deeper semantic understanding is required. A more effective model for semantic similarity is required.


Semantically Different (but similar word frequency)
Semantically Similar (despite word frequency different)

How are you?

I am fine, thanks!

How old are you?

I am 23 Years Old!

What is your age?

I am 23 Years Old!

Figure 1: This is an example shows shortage of traditional methods, if we simply use word frequency or keyword matching, the system will judge ”How are you” and ”How old are you” as similar,but these are two totally different sentences in real meaning . By contrast, ”how old are you” and ”what is your age” have no same word but their semantic meaning are same.

Research Goals

Our primary goal is to explore and advance methods for measuring semantic similarity between sentences. Includes:

- Compare the performance of baseline models (BoW and TF-IDF) with Sentence-BERT (SBERT) using the STS Benchmark dataset.
- Analyze model predictions against human-annotated scores to assess reliability and effectiveness.
- Highlight the practical applications of semantic similarity in tasks like information retrieval, machine translation, and question answering.

Evaluation Standard

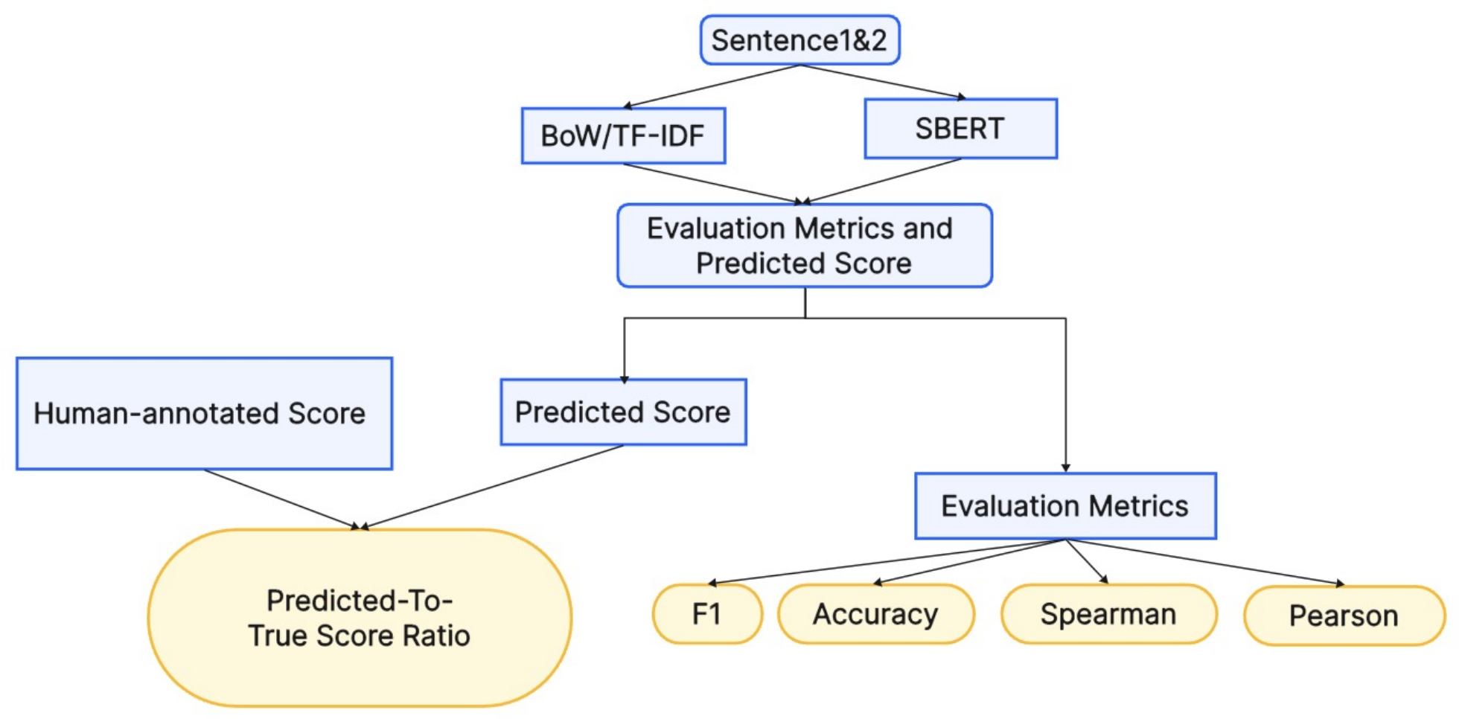


Figure 2: Evaluation Method Process. For judging model performance we use: (1) Metrics include F1-Score, Accuracy, Spearman Correlation, and Pearson Correlation. (2) Predicted scores to True scores ratio.

Model Details

We use BoW and TF-IDF as baseline models, and use SBERT as advanced final model.

Bag-of-Words (BoW) BoW represents sentences as word frequency vectors. Given two sentence vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, their similarity is computed using cosine similarity:

$$\text{Cosine Similarity} = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|}$$

TF-IDF (Term Frequency-Inverse Document Frequency) TF-IDF refines Bag-of-Words by weighting terms based on their importance in the corpus, calculated as:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \cdot \log\left(\frac{N}{\text{DF}(t)}\right)$$

where $\text{TF}(t, d)$ is the term frequency of term t in document d, N is the total number of documents, and $\text{DF}(t)$ is the document frequency of t. Cosine similarity is then applied to the weighted vectors.

Advanced Model: Sentence-BERT (SBERT) Sentence-BERT (SBERT) performs sentence embedding and similarity computation through the following steps:

1. Embedding generation. SBERT uses a BERT model to create embeddings for input sentences S_1 and S_2:

$$e_1 = \text{BERT}(S_1), \quad e_2 = \text{BERT}(S_2)$$

2. Similarity Computation. Semantic similarity is measured using cosine similarity, same as in the BoW model.

3. Loss Function. Cosine similarity serves as the target, optimizing the model with this loss function:

$$L = \frac{1}{N} \sum_{i=1}^{N} (\text{Cosine Similarity}(e_{1i}, e_{2i}) - y_i)^2$$

where y_i represents the human-annotated similarity score.

4. Semantic Classification. For specific tasks, sentence embeddings are fed into a fully connected layer to perform classification or regression tasks:

$$\hat{y} = \sigma(W \cdot e + b)$$

where σ is the activation function, and W and b are learnable parameters.

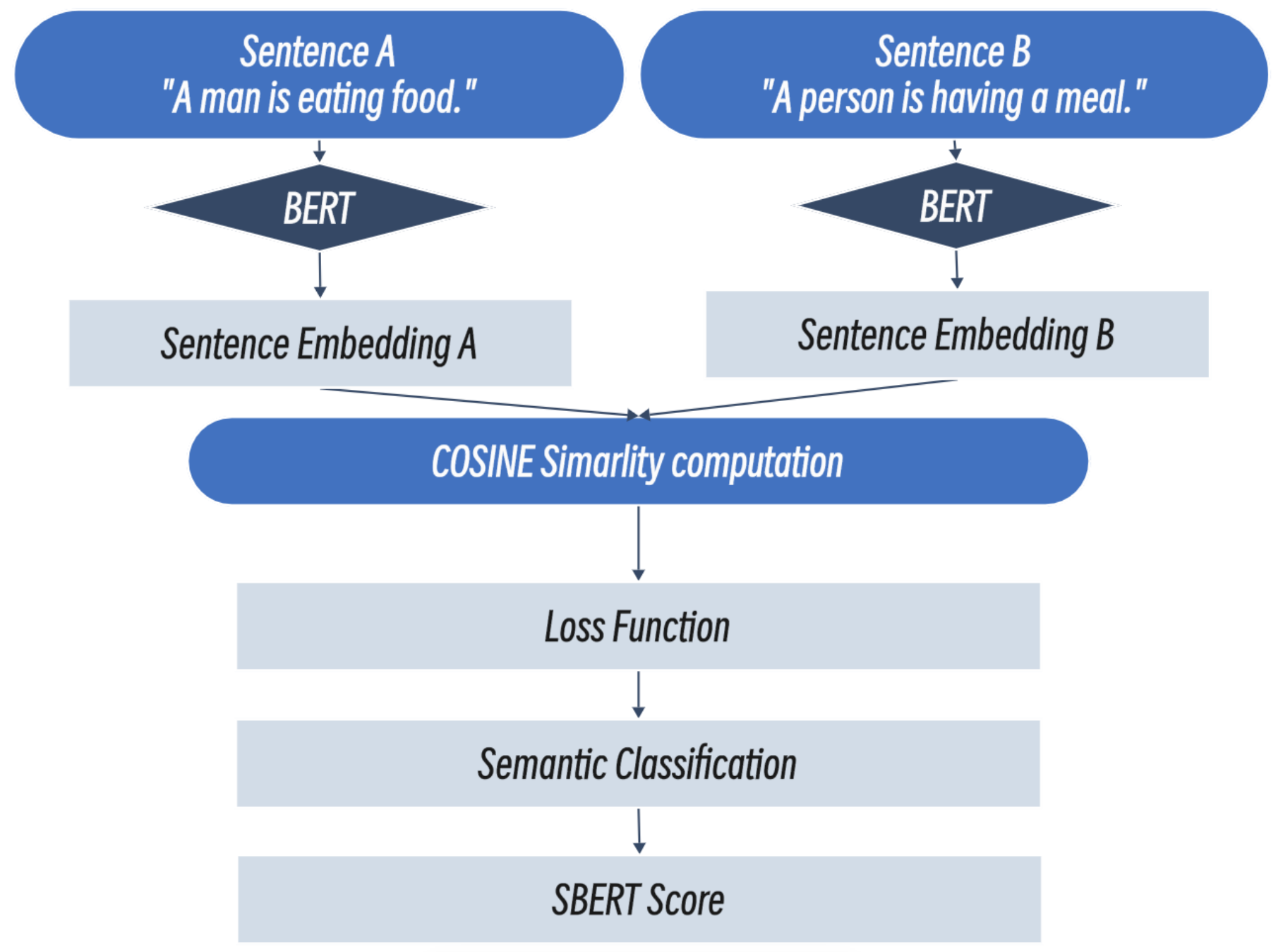


Figure 3: SBERT flow chart.

Experiments & Results

Figure 4: Distribution of BoW /real score ratios. The X-axis represents the ratio of the Bow predicted score to the original human-generated (true) score, while the Y-axis shows the counts of each ratio (eg 100 means this ratio appears 100 times). Left: validation set; Right: test set.

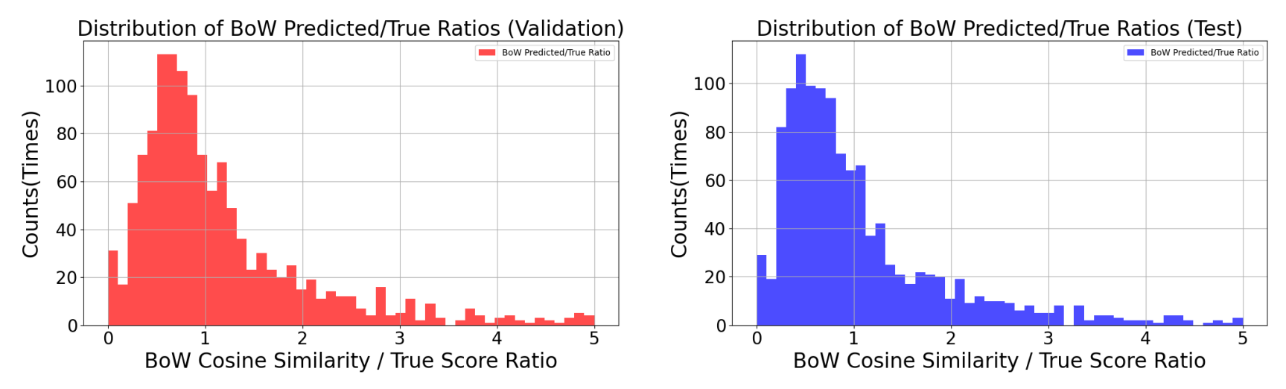


Figure 5: Distribution of TF-IDF/real score ratios.

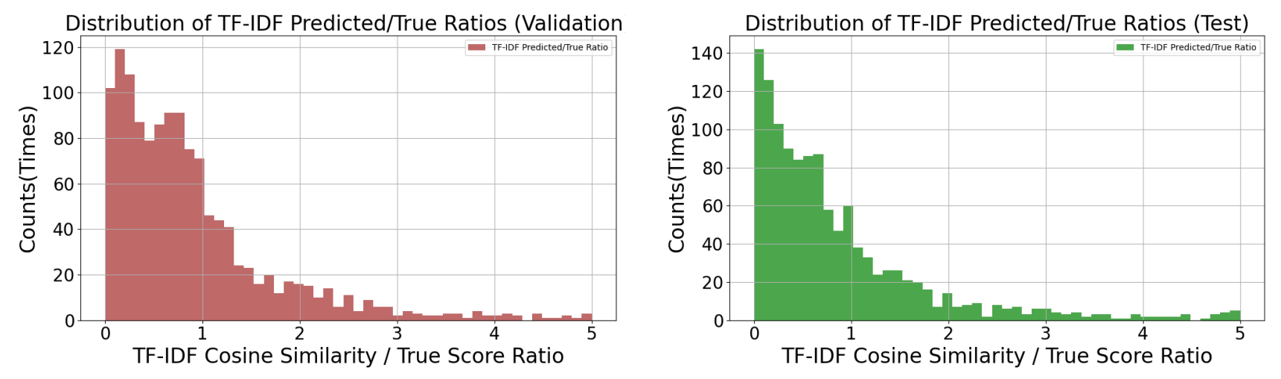


Figure 6: Distribution of SBERT/real score ratios.

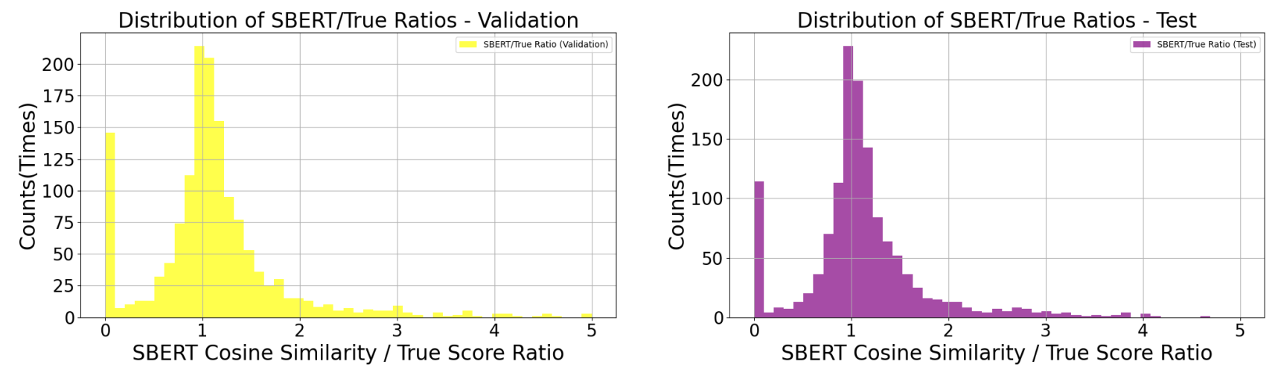
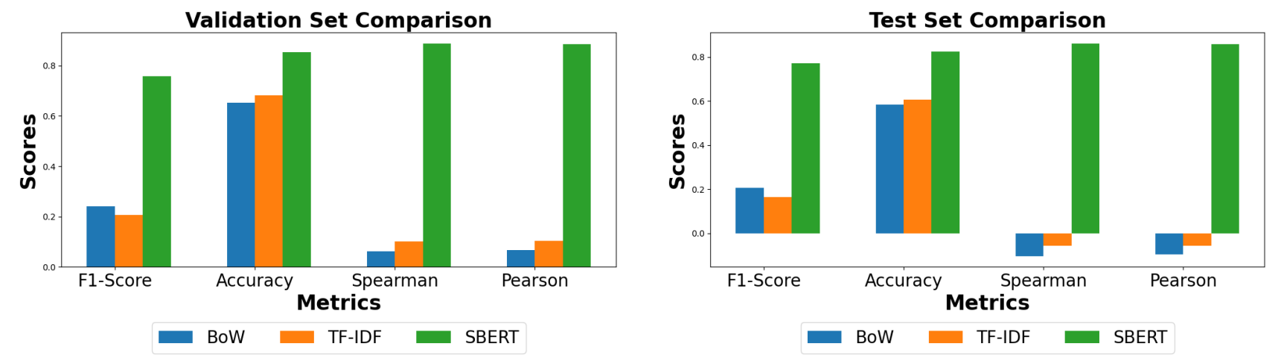


Figure 7: F1/Accuracy/Spearman/Pearson score comparison between BOW/TF-IDF/SBERT.



The results clearly show SBERT’s superior performance and reliability over baseline models, as summarized below:

1. Evaluation Metrics:SBERT significantly outperforms BoW and TF-IDF across all four metrics (F1-Score, Accuracy, Spearman Correlation, and Pearson Correlation). For example, SBERT achieves an F1-Score of 0.84 and an Accuracy of 0.87, compared to BoW (0.29, 0.65) and TF-IDF (0.24, 0.61).

2. Prediction-to-True Score Ratio: SBERT’s predicted-to-true score ratios are highly concentrated around 0.9-1.1 on the validation set, indicating minimal deviation from true scores. In contrast, BoW and TF-IDF exhibit more dispersed distributions, with extreme values in the tail regions, highlighting their lack of stability in capturing sentence similarity.

Conclusion

The experimental results highlight SBERT’s superiority over traditional models (BoW and TF-IDF) in sentence similarity tasks. SBERT consistently outperformed the baselines across all evaluation metrics, demonstrating its ability to capture deeper semantic relationships and contextual meaning.

This robustness and superior performance make SBERT an ideal choice for complex and resource-constrained scenarios, where accurate semantic understanding is critical. In contrast, while BoW and TF-IDF are computationally simple and interpretable, their reliance on surface-level features limits their applicability in tasks requiring nuanced semantic analysis.