A Comparative Study of Sentence Embedding Models' Sensitivity to Syntactic and Lexical Text Modifications

Ali Karimi
University of Science and Technology of Mazandaran
Behshahr, Iran
ali.karimi.research@gmail.com

Jamshid Pirgazi
University of Science and Technology of Mazandaran
Behshahr, Iran
spirgazi@gmail.com

Hani Attar
Faculty of Engineering, Zarqa University, Zarqa, Jordan
College of Engineering, University of Business and Technology, Jeddah, Saudi Arabia
hattar@zu.edu.jo

Ali Ghanbari Sorkhi
University of Science and Technology of Mazandaran
Behshahr, Iran
ali.ghanbari289@gmail.com

Mohamed Hafez
Faculty of Engineering FEQS, INTI-IU University, Nilai, Malaysia
Faculty of Management, Shinawatra University, Pathum Thani, Thailand
mohdahmed.hafez@newinti.edu.my

***Abstract*—Sentence embeddings are fundamental to natural language processing, as they enable models to capture semantic meaning beyond surface-level word similarity. The ability to represent sentences and paragraphs in dense vector spaces facilitates tasks such as semantic search, paraphrase detection, and textual inference. In this work, we present a comparative analysis of eight representative models—paraphrase-distilroberta, msmarco-roberta, paraphrase-mpnet, paraphrase-xlm-r, LaBSE, e5-base, gte-base, and bge-base—evaluated across four benchmark datasets (MRPC, QQP, PAWS, and VISLA). Our experiments highlight strengths and limitations of each model, providing insights into their effectiveness across diverse semantic similarity tasks.**

***Index Terms*—sentence embeddings, retrieval, evaluation**

I. Introduction

Sentence embeddings have emerged as a fundamental representation for a wide range of natural language processing (NLP) tasks, enabling semantic understanding beyond surface-level lexical overlap. They map sentences and paragraphs into dense vector spaces that capture semantic similarity, which is crucial for applications such as information retrieval, semantic search, retrieval-augmented generation (RAG) [1], and text classification [2]–[5]. Early approaches such as the Universal Sentence Encoder demonstrated broad applicability [2], while subsequent models like Sentence-BERT and SimCSE significantly improved embedding quality through fine-tuning and contrastive learning [3], [4].

Despite their success, sentence embedding models often exhibit sensitivity to superficial variations in text. For instance, embeddings may change notably when sentences undergo minor lexical substitutions or syntactic reordering, even though the underlying semantics remain unchanged [6]. Such instability can reduce robustness in downstream tasks such as semantic search and question answering [7], where paraphrases with the same meaning should be mapped to nearby vectors [8]. Moreover, recent studies highlight that embeddings can capture spurious correlations and may fail to generalize across paraphrased or reordered inputs [9]. Addressing these challenges is crucial for improving the reliability of embedding-based retrieval and generation systems.

The sensitivity of sentence embeddings to lexical and syntactic variations is not only a technical curiosity but also a practical concern. In real-world scenarios, users frequently express the same intent with different word choices or sentence structures, and embedding models should map these paraphrases to consistent representations [10]. When embeddings fail to capture semantic equivalence, the performance of downstream systems such as retrieval-augmented generation, dialogue agents, and semantic search can degrade significantly [11]. Furthermore, robustness to paraphrasing is a prerequisite for fairness and reliability, since unstable representations may propagate errors or biases in large-scale applications [12].

In this work, we conduct a systematic study of sentence embedding models with a focus on their sensitivity to lexical and syntactic variations. Specifically, we evaluate a diverse set of models trained with different architectures and objectives, and compare their robustness when facing paraphrased or reordered inputs. Beyond robustness, we analyze the influence of training datasets, model complexity, and the number of parameters on embedding stability. To ensure a fair and meaningful comparison, all selected models from different families are constrained to the same embedding dimensionality of 768.

II. Datasets

A. Hard Comparison

In our study, we employ datasets from two categories: binary classification and three-sentence inference. The first group consists of binary-labeled datasets that evaluate whether

two sentences express the same meaning. The Microsoft Research Paraphrase Corpus contains pairs of sentences extracted from news sources that are annotated by humans as either paraphrases or not, making it a benchmark for paraphrase identification [13]. The Quora Question Pairs dataset consists of question pairs collected from the Quora platform, where the task is to detect whether two questions are semantically equivalent, addressing the challenge of duplicate detection in community forums [14]. Finally, the PAWS dataset introduces challenging paraphrase examples with high lexical overlap but differing meanings, generated through controlled word swapping and back translation, to test model robustness against misleading surface-level similarities [15].

B. Triplet-based Comparison

The second dataset used in our experiments is the VISLA benchmark, which was recently introduced to evaluate the sensitivity of embedding models to lexical and syntactic alterations. VISLA is provided in two variants: a generic version containing only text and a spatial version that incorporates visual information. Since our focus is on sentence embedding models, we rely exclusively on the generic split. Each entry in the dataset consists of a reference caption along with a positive sample, which is a semantically similar reformulation, and a negative sample, which introduces lexical or structural modifications that alter the meaning. These three columns—caption, positive, and negative—allow systematic evaluation of whether models correctly position semantically related sentences closer in the embedding space while pushing apart unrelated ones. In our analysis, we further examined the statistical distribution of embedding similarities across these pairs, which provided valuable insights into how different models behave when confronted with controlled semantic and lexical perturbations [16].

TABLE I
SPECIFICATIONS OF THE EVALUATED MODELS: NUMBER OF PARAMETERS, PRETRAINING DATASETS, AND NUMBER OF LAYERS.

Model	Parameters (M)	Pretraining Datasets	Layers
par-dis-roberta	82.1	Paraphrase datasets (STS, NLI, SNLI, MNLI, etc.)	6
roberta-base-v3	124.6	Large paraphrase mining and STS datasets	12
par-mpnet	109.5	Paraphrase + NLI datasets	12
par-xlm-r	278.0	Multilingual paraphrase + NLI datasets	12
LaBSE	471.5	Translation pairs (English + 100+ languages)	24
E5-base	109.5	CCPairs, MS MARCO, NLI, web documents	12
GTE-base	109.5	General text embedding corpora (web + NLI)	12
BGE-base-v15	109.5	Large-scale retrieval datasets (MS MARCO, NLI, etc.)	12

III. MODELS

In this study, we selected models from seven distinct families of sentence embeddings to ensure a broad and balanced evaluation. The diversity of these models reflects different design goals and application domains: some are optimized for capturing semantic similarity in English, others are tailored for multilingual understanding, and a subset is fine-tuned with instruction-based objectives to enhance retrieval performance. Importantly, all chosen models are encoder-only architectures (Fig. 1), designed to produce fixed-dimensional sentence representations rather than generating text in an autoregressive

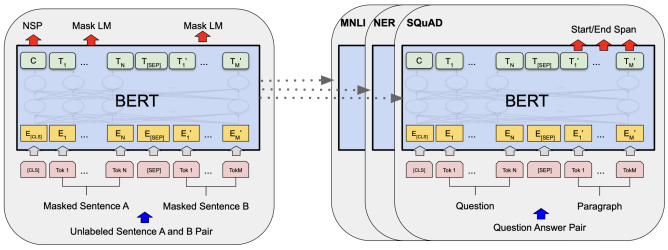

Fig. 1. Illustration of an encoder-only architecture

fashion. Table I summarizes the specifications of these models, including their number of parameters, pretraining datasets, and number of layers. By covering such a wide spectrum, our comparison provides insights into how sentence embedding models behave under various linguistic and task-specific conditions, enabling a more comprehensive understanding of their strengths and limitations.

A. BERT/RoBERTa Family

The first family in our selection is based on the BERT and RoBERTa architectures, which have become the foundation for many sentence embedding models. This group includes paraphrase-distilroberta-base-v1, msmarco-roberta-base-v3, and gte-base. All of them are primarily designed for English and inherit the strong contextual representation capabilities of BERT-style encoders. Within this family, DistilRoBERTa offers a lighter and faster alternative suitable for efficiency-critical scenarios, while models such as RoBERTa-base-v3 trained on MS MARCO provide stronger performance in retrieval tasks. In addition, GTE is designed as a more balanced, general-purpose embedding model that achieves robust results across a broad range of semantic similarity and retrieval benchmarks [3], [17].

B. Multilingual Family

The second family in our study consists of multilingual general-purpose models, specifically paraphrase-xlm-r-multilingual-v1 [8] and LaBSE [18]. These models are designed to align sentence representations across different languages, enabling semantic similarity tasks in multilingual settings. The paraphrase-xlm-r-multilingual-v1 model is built on the XLM-RoBERTa architecture and fine-tuned with a paraphrase objective, making it effective for capturing cross-lingual sentence-level meaning. LaBSE, on the other hand, is based on a dual-encoder architecture trained with translation ranking and masked language modeling over 109 languages, providing strong cross-lingual alignment. By including these models in our evaluation, we aimed to examine whether training on multiple languages improves robustness against lexical and syntactic variations, or if such robustness is largely independent of multilingual pre-training.

C. Instruction-tuned Family

The final family in our evaluation focuses on instruction-tuned and retrieval-oriented models, represented by E5-base-

v2 and BGE-base-en-v1.5. Unlike general-purpose encoders, these models are trained with task-specific prompts such as "query:" and "passage:" to explicitly distinguish between different roles in information retrieval. This design encourages the embeddings to capture fine-grained distinctions that are crucial for ranking and retrieval tasks. The E5 model employs weakly-supervised contrastive pre-training on massive text pairs, achieving broad coverage of general retrieval scenarios, while BGE leverages RetroMAE pre-training and large-scale contrastive learning to produce highly competitive embeddings for English retrieval. Including this family allows us to investigate how instruction tuning and retrieval-oriented training objectives impact robustness to lexical and syntactic variations [19], [20].

IV. Similarity and Sensitivity

A. Classifier architecture

We assess the quality and sensitivity of sentence embeddings by training a lightweight neural classifier on top of *fixed* embeddings and measuring classification accuracy under controlled lexical and syntactic perturbations. For each input pair (x_1, x_2), we obtain embeddings $\mathbf{e}_1, \mathbf{e}_2 \in \mathbb{R}^d$ from a pretrained encoder f_θ and build a feature vector by concatenating the raw embeddings, their element-wise absolute difference (to capture symmetric dissimilarity), and their Hadamard product (to capture alignment/interaction), as shown in equation (1). The absolute difference $|\mathbf{e}_1 - \mathbf{e}_2| \in \mathbb{R}^d$ and the element-wise product $\mathbf{e}_1 \odot \mathbf{e}_2 \in \mathbb{R}^d$ both preserve the original dimensionality, and concatenating them with $\mathbf{e}_1$ and $\mathbf{e}_2$ yields a rich yet compact probe of the embedding geometry. A shallow MLP classifier $g_\phi : \mathbb{R}^{4d} \to \{1, \ldots, C\}$ (trained with cross-entropy) then predicts the label; we intentionally keep g_ϕ low-capacity so that performance reflects the separability induced by the embeddings rather than the classifier. This design balances first-order information (raw vectors) with relational cues (difference and interaction), providing a principled way to evaluate how robustly different embedding models encode meaning in the face of lexical and syntactic changes.

$$\begin{aligned} \mathbf{z} \;=\; \big[\, \mathbf{e}_1 \,;\, \mathbf{e}_2 \,;\, |\mathbf{e}_1 - \mathbf{e}_2| \,;\, \mathbf{e}_1 \odot \mathbf{e}_2 \,\big] \in \mathbb{R}^{4d}, \\ \mathbf{e}_k = f_\theta(x_k) \in \mathbb{R}^d \end{aligned} \tag{1}$$

where $[\cdot\,;\,\cdot]$ denotes concatenation and $\odot$ denotes element-wise multiplication.

For the activation function in our neural classifier we employed the Rectified Linear Unit (ReLU) [21], defined as shown in equation (2). For the loss function we adopted the standard Cross-Entropy loss [22], which is widely used in classification tasks since it directly measures the divergence between the predicted probability distribution and the true class distribution. This choice is particularly suitable for our evaluation setup, where the classifier must distinguish fine-grained semantic differences. The mathematical forms of both ReLU and Cross-Entropy are given in equation (2) and (3), respectively.

$$\mathrm{ReLU}(x) = \max(0, x) \tag{2}$$

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} y_c \, \log \hat{y}_c, \tag{3}$$

where $\mathbf{y}$ is the one-hot encoded true label vector and $\hat{\mathbf{y}}$ is the predicted probability distribution over C classes.

For optimizing the neural classifier we adopted the AdamW optimizer [23], which is a decoupled variant of Adam that corrects the interaction between weight decay and gradient updates. AdamW combines the benefits of adaptive learning rates with proper regularization, making it particularly effective in preventing overfitting while ensuring stable convergence. This choice is well suited for our experimental setup, as we train lightweight classifiers on top of fixed embeddings and require both efficiency and robustness. The update rule of AdamW is expressed in equation (4).

$$\theta_{t+1} = \theta_t - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_t \right), \tag{4}$$

where $\hat{m}_t$ and $\hat{v}_t$ denote the bias-corrected first and second moment estimates of the gradient, η is the learning rate, λ is the weight decay coefficient, and ϵ is a small constant for numerical stability.

B. Results on MRPC

On the MRPC dataset, which consists of relatively short sentence pairs where the task is to determine whether they convey the same meaning, the retrieval-oriented models achieved the highest accuracies. In particular, E5-base (0.769) and BGE-base-v15 (0.763) performed best due to their strong ability to capture overall semantic similarity, as illustrated in Figure 2.

Models such as LaBSE and par-XLM-R reached moderate scores (around 0.75), reflecting their multilingual training objectives that prioritize cross-lingual alignment rather than fine-grained monolingual paraphrase detection.

By contrast, RoBERTa-base-v3 obtained the lowest performance (0.730), which can be attributed to its broader paraphrase-mining focus without task-specific fine-tuning for semantic retrieval. Overall, the results indicate that models optimized for retrieval and sentence-level semantic similarity are better suited for MRPC.

C. Results on PAWS

The PAWS dataset was specifically designed to challenge models with adversarial examples that exhibit high lexical overlap but divergent meanings due to word order or syntactic structure. As shown in Figure 3, paraphrase-focused models such as par-MPNet (0.651) and RoBERTa-base-v3 (0.650) outperformed retrieval-based models. Their stronger sensitivity to syntactic variation allows them to better distinguish subtle semantic differences introduced by reordering or scrambling words.

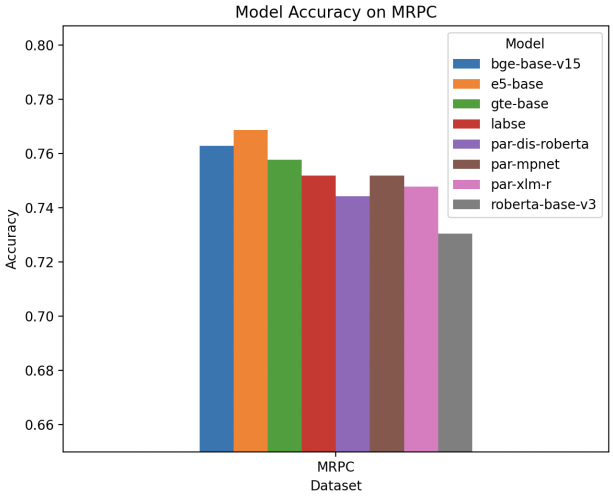

Fig. 2. Accuracy of different sentence embedding models on the MRPC dataset.

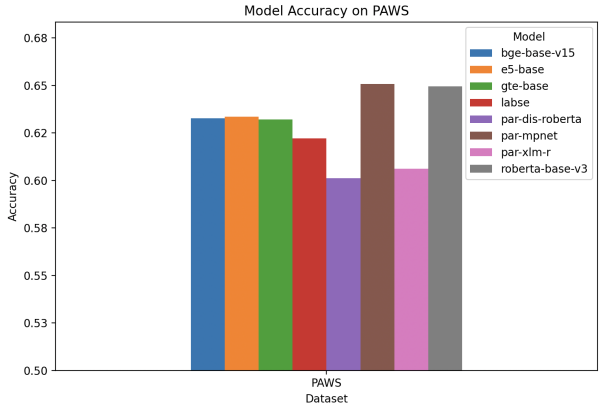

Fig. 3. Accuracy of different sentence embedding models on the PAWS dataset.

In contrast, retrieval models such as E5-base, BGE-base-v15, and GTE-base, while effective in capturing overall semantic similarity, tend to overemphasize surface lexical overlap, which results in lower performance on PAWS.

Multilingual models like LaBSE (0.622) and par-XLM-R (0.606) performed less competitively, reflecting their primary focus on cross-lingual alignment rather than fine-grained monolingual syntactic sensitivity.

D. Results on QQP

The QQP dataset, which is large and diverse in question-answering style, favored both retrieval-oriented and paraphrase-trained models. As shown in Figure 4, the best results were obtained by par-MPNet (0.894) and BGE-base-v15 (0.892). Other models, including E5-base and GTE-base, also performed strongly, with scores in a narrow range between 0.863 and 0.884. This small performance gap can be attributed to the scale and diversity of QQP, which closely resembles the kinds of data used to fine-tune retrieval models, while also rewarding the fine-grained semantic sensitivity of paraphrase-oriented approaches. Overall, the results suggest that both retrieval-based and paraphrase-based models are well suited for QQP, with only minor differences in performance across model families.

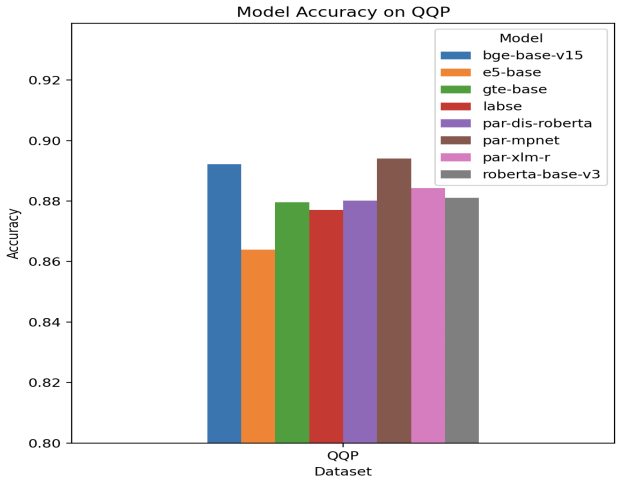

Fig. 4. Accuracy of different sentence embedding models on the QQP dataset.

E. Results on VISLA

As shown in Figure 5, the VISLA dataset consists of a central anchor sentence paired with two additional sentences: one semantically positive (paraphrastic) and one negative (semantically divergent but lexically overlapping). A model is expected to assign higher cosine similarity between the anchor and the positive sentence compared to the negative one. Models such as LaBSE, RoBERTa-base-v3, and par-dis-RoBERTa failed to meet this requirement, as their mean similarities for positives and negatives largely overlapped. This limitation can be attributed to their higher reliance on lexical overlap rather than capturing the overall sentence meaning, which makes them vulnerable to adversarial examples with similar word usage.

In contrast, par-MPNet achieved the largest gap between the means of positive and negative pairs, making it the most effective model in this evaluation. Its negative scores also showed a wide variance, including very low similarity values, indicating the model's strong ability to capture semantic distinctions and to push semantically divergent sentences further apart. par-XLM-R also demonstrated reasonable variance among negative samples; however, the mean gap between positive and negative sentences was not as pronounced, suggesting that while the model can sometimes separate divergent meanings, its cross-lingual training focus reduces its monolingual discrimination power in this context.

Finally, models such as GTE-base, BGE-base-v15, and E5-base produced slightly higher mean similarities for positives than negatives, but both distributions were narrow and tightly centered around their means. This indicates that while these retrieval-oriented models capture overall semantic similarity, they lack the fine-grained discriminative ability to strongly

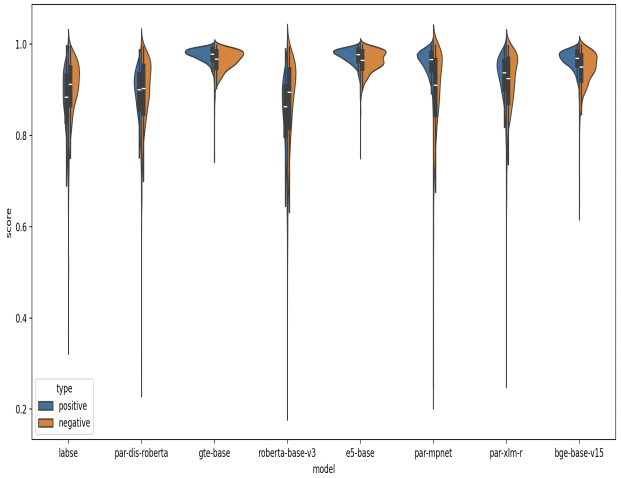

Fig. 5. Distribution of cosine similarity scores for positive and negative sentence pairs in the VISLA dataset across different models.

separate adversarial negative samples from true paraphrases. In other words, their embeddings tend to cluster semantically related and adversarially misleading examples too closely, limiting their effectiveness on VISLA.

V. Conclusion

In this work, we conducted a comprehensive evaluation of eight sentence embedding models drawn from diverse architectural families and training objectives. By examining their performance on datasets with different characteristics, we demonstrated that retrieval-oriented models such as E5-base and BGE-base-v15 excel in tasks where global semantic similarity is dominant (e.g., MRPC, QQP), while paraphrase-focused models such as par-MPNet and RoBERTa-base-v3 show superior sensitivity to syntactic variations, as observed in PAWS. Multilingual models like LaBSE and par-XLM-R exhibited competitive but not leading performance in strictly monolingual tasks, reflecting their broader cross-lingual training objectives. The VISLA dataset analysis further highlighted the limitations of models that rely heavily on lexical overlap, and emphasized the strength of par-MPNet in distinguishing semantically divergent sentences with high variance in negative examples. Overall, our findings confirm that no single model is universally optimal, and that their strengths are closely aligned with their training strategies and design goals.

For future research, several promising directions can be explored. First, hybrid or ensemble embeddings could be developed by combining complementary models to balance semantic robustness with syntactic sensitivity. Second, instruction-tuned embedding models could be investigated under varied prompting schemes, enabling deeper insights into how different instruction formulations affect embedding space. Finally, metaheuristic optimization algorithms can be applied to automatically discover effective instructions and configurations for embedding models, thereby advancing the adaptability and generalization of sentence embeddings across diverse downstream applications.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[2] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[3] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[4] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," *arXiv preprint arXiv:2104.08821*, 2021.

[5] H. Attar, Y. Sharrab, S. Smadi, and S. Jbaily, "Comparative analysis of machine learning techniques for arabic text classification using the sanad dataset," in *2025 11th International Conference on Mechatronics and Robotics Engineering (ICMRE)*. IEEE, 2025, pp. 60–67.

[6] M. Antoniak and D. Mimno, "Evaluating the stability of embedding-based word similarities," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 107–119, 2018.

[7] M. Essam, M. A. Deif, H. Attar, A. Alrosan, M. A. Kanan, and R. Elgohary, "Decoding queries: An in-depth survey of quality techniques for question analysis in arabic question answering systems," *IEEE Access*, 2024.

[8] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[9] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.

[10] C. Zhou, C. Qiu, L. Liang, and D. E. Acuna, "Paraphrase identification with deep learning: A review of datasets and methods," *IEEE Access*, 2025.

[11] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International conference on machine learning*. PMLR, 2020, pp. 11 328–11 339.

[12] Z. Shi and M. Huang, "Robustness to modification with shared words in paraphrase identification," *arXiv preprint arXiv:1909.02560*, 2019.

[13] B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Third international workshop on paraphrasing (IWP2005)*, 2005.

[14] L. Sharma, L. Graesser, N. Nangia, and U. Evci, "Natural language understanding with the quora question pairs dataset," *arXiv preprint arXiv:1907.01041*, 2019.

[15] Y. Zhang, J. Baldridge, and L. He, "Paws: Paraphrase adversaries from word scrambling," *arXiv preprint arXiv:1904.01130*, 2019.

[16] S. H. Dumpala, A. Jaiswal, C. Sastry, E. Milios, S. Oore, and H. Sajjad, "Visla benchmark: Evaluating embedding sensitivity to semantic and lexical alterations," *arXiv preprint arXiv:2404.16365*, 2024.

[17] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "Ms marco: A human-generated machine reading comprehension dataset," 2016.

[18] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," *arXiv preprint arXiv:2007.01852*, 2020.

[19] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J.-Y. Nie, "C-pack: Packed resources for general chinese embeddings," in *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, 2024, pp. 641–649.

[20] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei, "Text embeddings by weakly-supervised contrastive pre-training," *arXiv preprint arXiv:2212.03533*, 2022.

[21] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[22] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," in *International conference on Machine learning*. pmlr, 2023, pp. 23 803–23 828.

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.