Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Based on the Gemini Large-scale Model, Enhanced Accuracy in Semantic Similarity Detection With the Ernie Model

Zihang Li*

Hainan International College, Communication University of China, Beijing, China

*Corresponding author: lzh13121107950@cuc.edu.cn

Abstract. The rapid development of deep learning models has been a hallmark of recent years. As the world moves towards greater intelligence, there is an urgent need for advancements in technologies like question-answering systems, chatbots, and search and recommendation engines, especially within the service and e-commerce sectors. At the heart of these advancements lies the task of detecting semantic similarity in natural language processing. Against this backdrop, this paper examines the accuracy of semantic similarity detection across different deep learning models, focusing specifically on the LSTM, Transformer, and ERNIE models under identical hyperparameters and configuration settings. The study reveals a common challenge among these models in achieving effective data generalization on the LCQMC dataset. To address this, the paper introduces an innovative approach by combining the highest-performing ERNIE model with the Gemini large-scale model and employing data augmentation techniques to enhance accuracy. This strategy increased accuracy from 82% with the ERNIE model alone to 85%.

Keywords: ERNIE, Gemini large-scale model, semantic similarity detection.

1. Introduction

In recent years, the field of artificial intelligence has experienced rapid technological advancements. In this era of increased intelligence, the development and application of Natural Language Processing (NLP) are crucially important for meeting human needs. A fundamental task within NLP is the calculation of semantic similarity, which underpins a variety of application scenarios. This includes question-answering systems, sentiment analysis, text classification, semantic search, and the recommendation of search engine entries and the promotion of articles and products. These applications, all reliant on the computation of textual semantic similarity, underscore its significance and the necessity for its development[1,2].

Moreover, the inherent ambiguity of Chinese semantics, as opposed to English, makes the study of semantic similarity in Chinese corpora more challenging. Yet, the demand for such studies grows with technological advancement and proliferation.

Semantic similarity is broadly categorized into two types: knowledge-based and corpus-based. [3] Owing to the high accuracy, speed, and capability of handling vast amounts of data, semantic similarity calculations performed with deep learning models fall into the corpus-based category. This method utilizes information from large corpora to ascertain semantic similarities and to determine whether the semantics between different texts are alike. The LSTM model, an enhancement over traditional recurrent neural networks, was initially developed to solve the problem of long-term dependencies seen in RNNs. However, due to its exceptional memory capabilities and aptitude for processing lengthy sequences, it has been extensively employed in various NLP tasks, including sentiment analysis. Jonas [4] first applied it to the domain of semantic similarity in 2016, yielding promising results. Following this, the advent of the Transformer model somewhat resolved issues related to context connectivity and proved suitable for semantic similarity tasks [2]. Baidu's subsequent introduction of the ERNIE model has demonstrated notable effectiveness in the NLP field, mainly exhibiting unique advantages in processing Chinese corpora.

Building upon the aforementioned background, this paper employs a Chinese corpus database to explore the semantic similarity task in natural language processing through the LSTM, Transformer, Baidu's ERNIE, and Google's Gemini large-scale models. The aim is to compare the training outcomes of the LSTM, Transformer, and ERNIE models on the same corpus, identify the optimal model, examine the variations across different models during training, and subsequently improve the precision, AUC value, and ROC curve of the chosen model by applying data augmentation techniques with the Gemini large-scale model.

2. Data and Methods

2.1. Data Source

This study's model training dataset is the LCQMC (Large-scale Chinese Question Matching Corpus), a high-quality dataset published by the Social Computing and Information Retrieval Research Center at Harbin Institute of Technology. The LCQMC dataset encompasses 260,068 annotated pairs, selected from high-frequency questions across various domains within Baidu Q&A. In the preprocessing phase, the Wasserstein distance is employed to gauge the distribution differences among samples, aiding in data cleansing and anomaly detection. Moreover, this metric is useful in feature selection and transformation, assisting in identifying features that more effectively differentiate between categories or in reducing feature dimensions, followed by classification and subsequent manual annotation.

2.2. Methods and Models

LSTM Model: An enhancement of the RNN (Recurrent Neural Network), the LSTM model introduces three gates: the forget gate, output gate, and input gate. These gates enable the selective forgetting of irrelevant information, optimizing pertinent information retention for subsequent cycles. [5]

Transformer Model: Introduced by Google, this model employs a self-attention mechanism, with its distinct multi-head attention mechanism facilitating improved learning of word relationships and contextual understanding. It can process complex sentences and offers significant parallelization capabilities and reduced training durations. [2]

ERNIE Model: Building upon the Transformer model's architecture and incorporating the self-attention mechanism, the ERNIE model introduces a continuous semantic integration strategy for enhanced semantic comprehension. It captures sentence semantics more efficiently during pre-training and incorporates structured knowledge. [6]

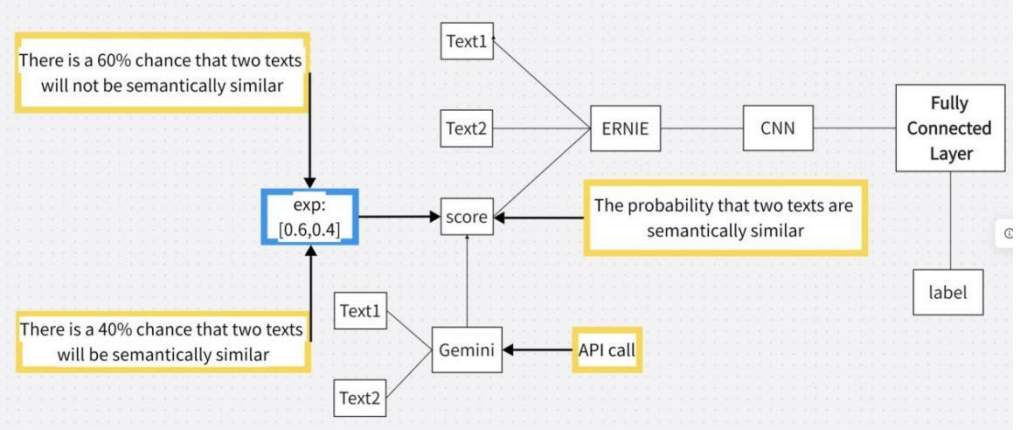

Fig 1. Innovative Model

Innovative Algorithm Model Based on ERNIE: This model, proposed in the paper, begins with training the ERNIE model, followed by the LSTM and Transformer models, to compare the accuracy of each on the dataset. The study then employs the newly proposed model for training, juxtaposing

its performance with the previous results. As outlined in Fig.1, the model leverages the Gemini large model to predict text data's semantic similarity, offering a score range from 0 to 1, where 0 signifies complete dissimilarity and 1 indicates complete similarity (e.g., [0.6,0.4] indicates a 0.6 probability of dissimilarity and a 0.4 probability of similarity). This approach provides an additional feature layer from which the ERNIE model can learn. Given its superior accuracy and AUC values on this dataset, the ERNIE model serves as the foundation for optimization. The refined dataset is input into the ERNIE model for feature extraction, then trained within a CNN layer and through a fully connected layer to obtain a predicted label value. This predicted label is used to compute the AUC value and generate the ROC curve.

2.3. Evaluation Metrics

This study employs accuracy as the primary metric to evaluate the effectiveness of a model directly. To address potential imbalances in the sample test set distribution, precision, recall, and the F1 score are utilized to provide a more nuanced understanding of model performance. Considering that semantic similarity detection is fundamentally a binary classification task, the models developed herein are binary classifiers. Accordingly, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) value are introduced as indicators of exemplary performance.

Accuracy: Determined by the ratio of correctly predicted observations to the total observations in the test dataset.

Precision: The ratio of true positive predictions to the total number of positive predictions made.

Recall: The ratio of true positive predictions to the total number of actual positives in the test dataset.

F1 Score: The harmonic mean of precision and recall, integrating both metrics. The F1 score increases with the precision and recall, indicating better model performance.

ROC Curve: A graphical plot that assesses the performance of binary classification models by illustrating the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR).

AUC Value: The measure of the Area under the ROC curve, gauging the model's classification efficacy. A closer AUC value to 1 denotes superior model performance.

3. Results and Discussion

3.1. Analysis of the Training Process through Visualization

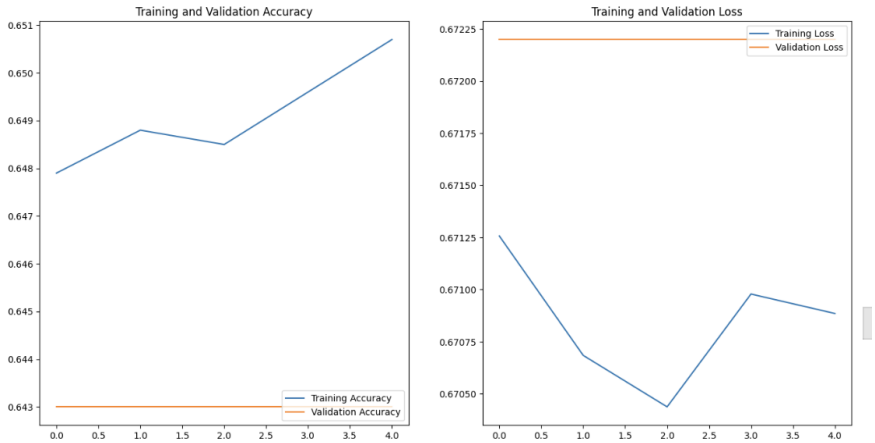

Fig.2 LSTM

As depicted in Fig.2, concerning the LSTM model, the training and validation accuracies appear relatively constant, suggesting that the model may not be extracting substantial learning from the training data or learning at a sluggish pace. The nearly static validation accuracy, being lower than

the training accuracy, implies potential issues with the model's ability to generalize to unseen data effectively.

A significant reduction in training loss is observed, indicating positive learning progress. However, a peak in the validation loss could signal overfitting or anomalies within the data or learning process.

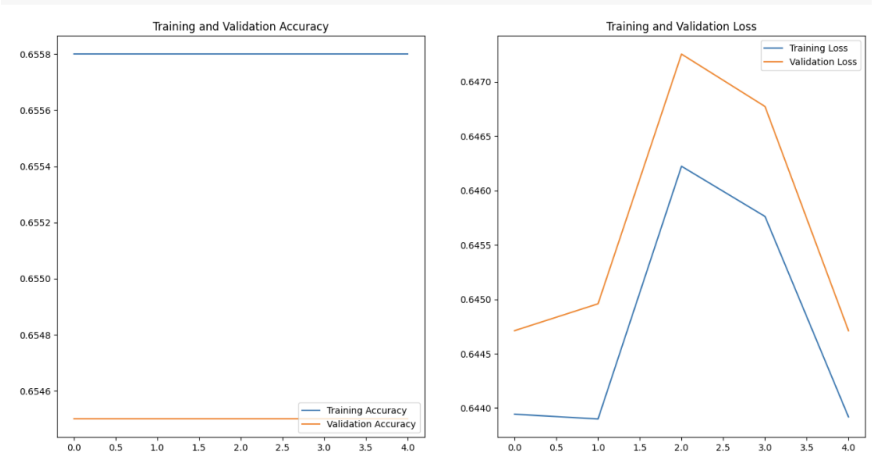

Fig.3 ERNIE

According to Fig.3， it's similar to LSTM models; the training accuracy is increasing, but the validation accuracy remains unchanged. This pattern typically indicates overfitting, where the model learns the training data well but fails to generalize.

The training loss decreases slightly and then sharply increases, then sharply decreases which is atypical and indicates an unstable training process. The validation loss also shows the same pattern, indicating that the model's performance on the validation set is unstable.

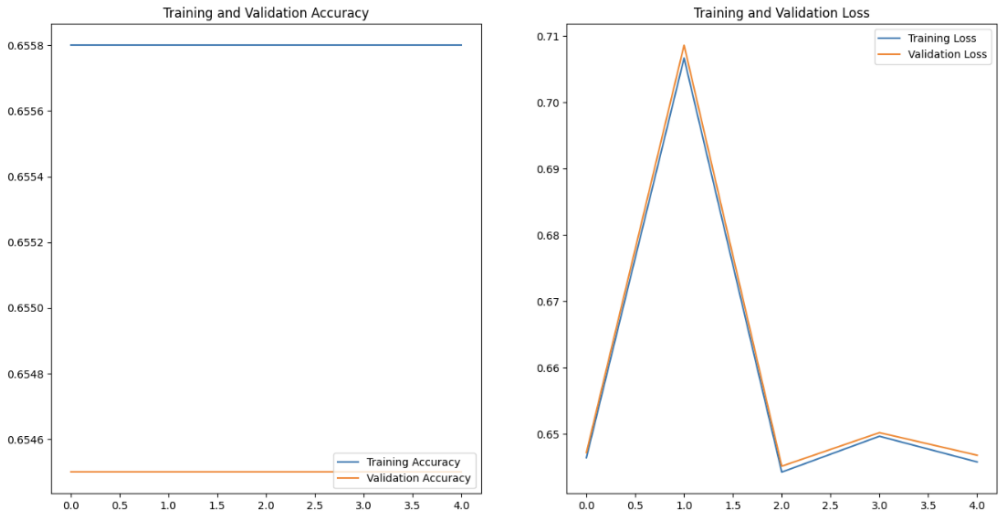

Fig.4 Transformer

In the case of the Transformer model, while the training accuracy shows an upward trend and maintains a high overall rate (Fig. 4), the validation accuracy experiences lesser fluctuations compared to the training dataset. This pattern often indicates good overall training efficiency but suggests some level of data overfitting; the model learns the training data well but struggles to generalize effectively.

A continual decline in training loss points towards a smooth training trajectory for both training and validation loss. However, the pattern of validation loss—initially decreasing, then increasing, and decreasing again—highlights the model's unstable performance on the validation set.

Once more, an increase in training accuracy over time is noted, yet without corresponding improvements in validation accuracy, aligning with the signs of overfitting.

The training loss markedly decreases, demonstrating the model's learning capabilities. However, the drastic fluctuation in validation loss—first increasing then decreasing—suggests potential issues with the model's learning strategy or the characteristics of the validation set.

Overall, all models demonstrated enhancements in training accuracy and decreased loss, indicating effective learning of the training data patterns. However, a lack of significant progress in validation accuracy for any model might suggest a tendency toward overfitting the training data.

Regarding stability, the pronounced fluctuations in validation loss shown by the ERNIE and Transformer models indicate these models' heightened sensitivity to specific training cycles or data batches. This sensitivity could stem from an overly aggressive learning rate or the inclusion of non-representative outlier data within the training set.

Hence, it becomes clear that the LSTM and Transformer models exhibit suboptimal performance, possibly due to various factors, including insufficient data generalization across models. Courtney Corley and colleagues have previously introduced a methodology for enhancing semantic similarity detection by integrating word and word index metrics, [7] leading to the proposition of augmenting data with additional indicators. This approach inspired the development of the innovative model presented in this paper.

3.2. Comparison of LSTM, Transformer, ERNIE, and the Innovative Model

Table 1 The comparison of statistical metrics

Model	Accuracy	Precision	Recall	F1-score
LSTM	0.643	0.411	0.499	0.330
Transformer	0.654	0.538	0.511	0.410
ERNIE	0.788	0.838	0.785	0.779
Innovative model	0.857	0.865	0.856	0.856

The comparison of statistical metrics presented in Table 1 reveals that the ERNIE model exhibits superior performance among the three foundational models. The learning rate and the number of epochs were uniform across models to standardize the training process. Considering the extensive time required to train 260,000 text pairs, the study limited the training set to 20,000 text pairs and the test set to 2,000 text pairs for each model. The LSTM and Transformer models demonstrated less than optimal performance and lacked high accuracy, which suggests that specific learning rates might improve model precision. Nonetheless, under a uniform learning rate, the innovative model introduced in this study outperforms the three foundational models in terms of accuracy and other metrics, thereby demonstrating its potential to enhance the accuracy of semantic similarity detection.

Fig. 5 and Fig. 6 illustrate the ROC curves and their respective AUC values for each model.

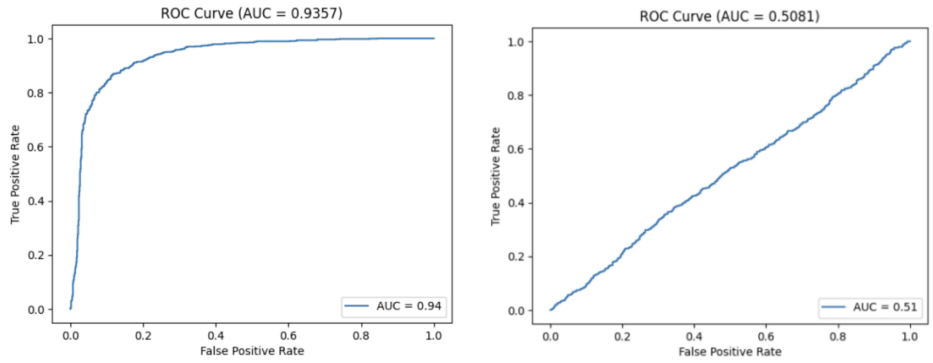

Fig. 5(a). Innovative model;(b)： LSTM

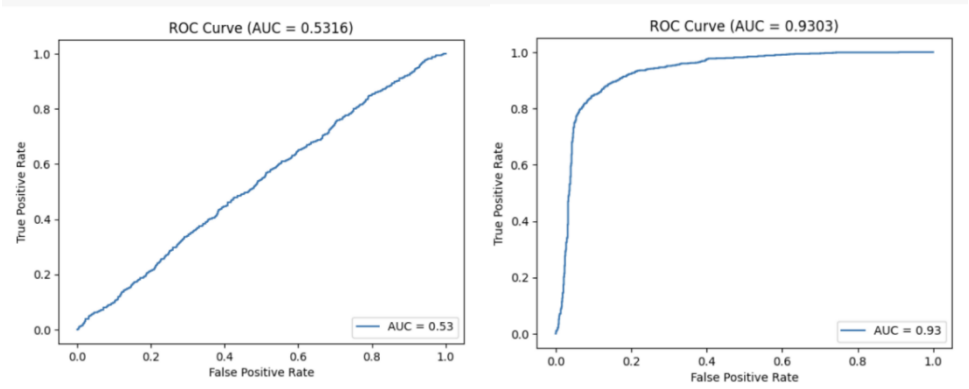

Fig.6 (a): Transformer;(b): ERNIE

The AUC value, which represents the Area under the ROC curve, indicates a binary classification model's performance. In this context, it is observed that the LSTM(Fig.5(b)) and Transformer(Fig. 6(a)) models yield poor results, which achieve an AUC value of approximately0.51 and 0.53 indicating average classification effectiveness. Conversely, the ERNIE model (Fig.6(b))achieves an AUC value of approximately 0.93, signifying its proficient learning and feature extraction capabilities from the LCQMC Chinese corpus, thereby resulting in superior classification performance. Building upon this, the proposed innovative model reaches an AUC value of 0.935, indicating a marginal improvement. This achievement underscores that the innovative model surpasses the foundational ERNIE, LSTM, and Transformer models across a spectrum of evaluation metrics.

4. Conclusion

To address semantic similarity challenges, this study examined three foundational deep learning models—LSTM, Transformer, and ERNIE. Under stringent control of variables, it was observed that all three models struggled with comprehensively understanding the LCQMC Chinese corpus. The LSTM and Transformer models notably underperformed, achieving merely around 64% accuracy. Their AUC values, deduced from ROC curves, were approximately 0.5, signifying mediocre classification effectiveness. In contrast, the ERNIE model showed relative success, reaching an accuracy of about 82% and an AUC value of 0.93 under identical hyperparameters and conditions. Nonetheless, the training phase highlighted opportunities for the ERNIE model to refine its data learning capabilities further. As a response, this paper introduced an innovative model that leverages both the ERNIE and Gemini large-scale models, employing data augmentation to enhance precision. The resulting innovative model attained an accuracy of 85% and an AUC value of 0.935, thereby not only exhibiting superior classification performance but also showcasing an improvement in accuracy.

References

[1] Zhou, S.K. Research and Application of a Short Text Semantic Similarity Model Based on Deep Learning (Master's thesis, North University of China),2023.

[2] Ding, Q., Chi, H.Y., Yan, X., Xu, G.Y., & Deng, Z.Y. Calculation of Question Semantic Similarity Based on the Transformer Model. Computer Engineering and Design, 2023,(03), 887-893. doi:10.16208/j.issn1000-7024.2023.03.034.

[3] Wang, C.L., Yang, Y.H., Deng, F., & Lai, H.Y. Review of Text Similarity Calculation Methods. Information Science, 2019,(03), 158-168. doi:10.13833/j.issn.1007-7634.2019.03.026.

[4] Mueller, J., & Thyagarajan, A. Siamese Recurrent Architectures for Learning Sentence Similarity.

[5] Meenakshi, D., & Mohamed Shanavas, A.R. A Novel Shared Input-Based LSTM for Semantic Similarity Prediction. JAIT, 2022,4, 387-392.

[6] Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. ERNIE: Enhanced Language Representation with Informative Entities. 2019,,arXiv preprint arXiv:1905.07129.

[7] Corley, C.D., & Mihalcea, R. Measuring the Semantic Similarity of Texts. In Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment 2005,pp. 13-18.