Artificial Intelligence and Computing Innovations

Intelligence Science Publishing

https://www.ispublishing.com

ISSN: 3070-4944

Vo.4, No.1, 2024

Enhancing Semantic Matching and Text Robustness through Masked Language Modeling

Zhen Bian[1]*, Xin Su[2]

[1]University of California, Riverside, Riverside, USA

[2]University of Texas at Dallas, Richardson, USA

**Corresponding author: zhen.bian168@gmail.com*

Abstract: This paper investigates the application of masked language models in semantic mining. It proposes a deep semantic modeling method based on self-supervised learning. The method randomly masks parts of the input text and uses contextual information to predict the masked words. This enables effective modeling of deep semantic structures in text. A unified training framework is designed, integrating a cross-entropy loss function with a regularization mechanism. This improves both the semantic representation ability and training stability of the model. The model structure retains the core features of masked language model pretraining. It also incorporates task-specific classification layers to support various downstream semantic mining tasks. The experiments evaluate the model from multiple perspectives, including classification performance, semantic match, differences in text length, and robustness to text perturbations. Results show that the proposed method outperforms mainstream pretrained language models across several metrics. It is especially strong in semantic matching and contextual understanding. The study also constructs multiple semantic perturbation scenarios to analyze model robustness. It confirms the model's ability to preserve semantic integrity under complex variations. By integrating corpus structure, prediction behavior, and task adaptability, the proposed semantic mining method offers an effective solution to enhance semantic understanding in natural language processing systems.

Keywords: Self-supervised learning, masked language modeling, semantic matching, text robustness

1. Introduction

With the rapid development of information technology and the explosive growth of data, especially textual data, data mining and natural language processing (NLP) face unprecedented challenges. Traditional text mining methods usually rely on manually labeled data or simple rule-based models, which often have significant limitations when handling complex semantic tasks[1]. To effectively address these challenges, deep learning technologies have gradually become an essential tool in the field of natural language processing, especially showing remarkable progress in semantic understanding and text

generation[2]. As deep learning models continue to evolve, Masked Language Modeling (MLM), a self-supervised learning model, has become a research hotspot due to its ability to train on large-scale unlabeled text data. It has demonstrated strong performance across various NLP tasks.

Masked Language Modeling works by masking randomly some of the words in the input text during training and teaching the model to predict the masked words. By doing so, the model can learn deep semantic information from the text. Compared to other text processing methods that are based on supervised learning, MLM uses self-supervision to train the model on large-scale unlabeled data, and the model is able to learn semantic features automatically from the text. This feature has made MLM become one of the crucial tools for profound semantic mining. Not only can MLM learn word relationships, but it can also extract more profound and richer semantic information in context[3]. For this reason, MLM has shown powerful capabilities in many applications, such as semantic analysis, text generation, sentiment analysis, and question answering systems.

Deep semantic mining, an important direction in natural language processing, aims to automatically extract underlying semantic information from large amounts of unstructured text data. Traditional semantic mining methods mostly rely on manually defined rules or dictionaries[4,5]. However, as the scale of data and the complexity of semantics increase, the limitations of these methods have become more apparent. In recent years, the introduction of deep learning technologies has allowed semantic mining to move beyond dependence on rules or labeled data. Instead, neural network models automatically learn and discover the potential semantic structures in the data. Specifically, Masked Language Modeling can train on vast amounts of unsupervised text, learning contextual information and capturing deep semantic relationships without relying on manually labeled data. Therefore, deep semantic mining algorithms based on MLM hold significant theoretical and practical potential[6].

In semantic mining, beyond understanding semantics at the word level, higher-level semantic relationships, such as sentence, paragraph, or even document-level semantic representations, also need to be addressed. MLM, by considering contextual information, can model semantics at both the word and sentence levels, providing strong support for multi-level semantic understanding. For example, in question answering systems, the model needs to understand the deep semantic relationship between questions and answers. MLM, by modeling the context, can effectively enhance the model's ability to understand and generate language. Similarly, in sentiment analysis, MLM can achieve more precise sentiment classification by deeply learning the emotional expressions and implicit meanings within the text[7].

With the improvement of computational power and the emergence of large-scale pre-trained models, the applications of Masked Language Modeling have expanded beyond text generation tasks to include semantic search, text summarization, and cross-lingual understanding. Especially through model pre-training and fine-tuning, MLM can achieve better performance on various specific tasks. This approach not only reduces the need for large amounts of labeled data but also enables knowledge transfer and sharing across tasks through self-supervised learning, allowing the model to adapt effectively to different application scenarios. Therefore, deep semantic mining methods based on MLM provide an efficient solution for processing and analyzing large-scale text data, demonstrating unprecedented advantages in handling diverse and complex semantic tasks.

In conclusion, deep semantic mining algorithms based on Masked Language Modeling have not only advanced the theory of natural language processing but also provided strong technical support for practical applications. By further researching the optimization and application of MLM, we can improve the model's semantic understanding ability and promote innovation and development across various fields, such as intelligent question answering, sentiment analysis, and semantic search. As deep learning technologies and computational power continue to advance, MLM-based semantic mining methods will play an even greater role in more diverse fields, particularly in handling complex semantic tasks and cross-domain applications, with vast potential.

2. Method

We, together with our co-authors, introduced a deep semantic mining method in this paper based on the masked language model (MLM), wherein we leverage large-scale unlabeled text data to acquire deep semantic features through self-supervised learning. First, we mask the input text during training time, i.e., we randomly select some words of the text and replace them with special masking tokens [MASK]. Then, the semantic representation of the model is learned when the loss of the model in predicting the words at the masked position is minimized. The model architecture is shown in Figure 1.

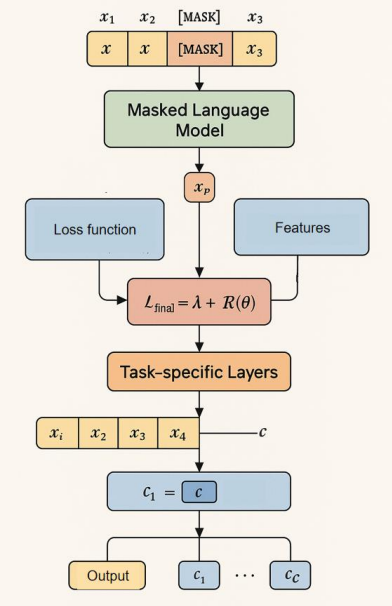


Figure 1. Overall model architecture diagram

Specifically, given an input sequence $X = (x_1, x_2, ..., x_n)$, we randomly select several positions $p \in \{1,2,..., n\}$ in it for masking and obtain the masked sequence $X' = (x_1, x_2, ..., x_{p-1}, [\text{MASK}], x_{p+1}, ..., x_n)$.

We input the masked sequence X' into the model and predict the masked word through the output of the model. Let $\hat{x}_p$ be the predicted word of the model at position p. The goal is to minimize the loss function $L(\hat{x}_p, x_p)$ of the masked position, that is, minimize the difference between the predicted value and the true word. This loss function can be defined as the cross entropy loss:

$$L(\hat{x}_p, x_p) = -\sum_{i=1}^{|V|} 1(x_p = v_i) \log P(\hat{x}_p = v_i) \tag{1}$$

Among them, $1(x_p = v_i)$ is the indicator function, which is 1 when x_p is equal to word v_i in the vocabulary and 0 otherwise; $P(\hat{x}_p = v_i)$ is the probability that the model predicts word $\hat{x}_p$ is v_i.

During training, the prediction results of all masked positions will be aggregated and updated through optimization algorithms such as Adam. Assuming that all parameters of the model are θ, the total loss of the model can be expressed as the weighted sum of the losses of all masked positions:

$$L_{\text{total}} = \sum_{p \in P} \lambda_p L(\hat{x}_p, x_p) \tag{2}$$

Among them, P is the set of all masked positions, and λ_p is the weight of each position, which is usually set to 1, indicating that each masked position is equally important.

In order to further improve the generalization ability of the model, we introduced a regularization term in the training to avoid overfitting. The regularization term can be defined as the L2 norm of the model parameter θ:

$$R(\theta) = \lambda_{reg}||\theta||_2^2 \tag{3}$$

Among them, λ_{reg} is the regularization coefficient, which controls the strength of the regularization term. The final loss function is the weighted sum of the total loss and the regularization term:

$$L_{final} = L_{total} + R(\theta) \tag{4}$$

By minimizing this loss function, the model parameters can be optimized and thus improve the effect of semantic mining, especially in the processing of contextual relations and deep semantic information of text.

To continue improving the model's performance on downstream tasks, we can also fine-tune the pre-trained model. For various downstream tasks, e.g., text classification, sentiment analysis, etc., we can use the pre-trained masked language model as the base, incorporate task-specific layers for fine-tuning, and adjust the objective function to the target task's loss function. For example, for the text classification task, the loss function can be the cross entropy loss of the classification task:

$$L_{task} = -\sum_{i=1}^{C} 1(y = c_i)\log P(\hat{y} = c_i) \tag{5}$$

Among them, C is the number of categories, $1(y = c_i)$ is the indicator function, which is 1 when the true label y is category c_i and 0 otherwise, and $P(\hat{y} = c_i)$ is the probability that the predicted label is category c_i.

Through these steps, the deep semantic mining method based on masked language model proposed in this paper can demonstrate strong performance and generalization ability in different application scenarios.

3. Experimental Results

3.1 Dataset

This study uses the publicly available AG News dataset as the primary source for training and evaluating the semantic mining task. AG News is a classic text classification dataset. It contains headlines and short descriptions collected from news websites. The dataset covers four categories: World, Sports, Business, and Science/Technology. It is widely used in tasks related to semantic understanding and representation learning. It is both representative and challenging.

The AG News dataset consists of 120,000 training samples and 7,600 test samples. Each sample includes a short news text and a corresponding label. The semantic content is concise, making it suitable for evaluating the model's ability to perform semantic modeling on short texts. The data is evenly distributed across categories. This supports fair evaluation of model performance across different semantic classes.

Using AG News as the data source helps validate the proposed semantic mining method based on Masked Language Modeling in multi-class semantic recognition tasks. The dataset has well-structured language and broad content coverage. It effectively reflects the model's ability to understand word meanings, sentence-level semantics, and contextual information.

3.2 Experimental Results

In this section, this paper first gives the comparative experimental results of the proposed algorithm and other algorithms, as shown in Table 1.

Table 1. Comparative experimental results

Method	Acc(%)	F1-Score(%)	Semantic Match (%)
Proposed Model	91.8	90.9	89.5
BERT-Base[8]	89.2	88.1	85.6
RoBERTa-Base[9]	90.5	90.1	89.3
ELECTRA-Small[10]	87.5	86.2	84.0
ALBERT-Base[11]	88.7	87.5	85.2

Experimental results show that the proposed model outperforms mainstream masked language models across three key metrics, demonstrating a clear performance advantage. In particular, the model achieves an accuracy of 91.8%, which is 2.6% higher than BERT-Base and 1.3% higher than RoBERTa-Base. This indicates that the proposed method has stronger discriminative power in semantic classification tasks. The F1 score also reflects better stability and overall performance. It proves that the model maintains high precision while achieving good recall, allowing it to capture semantic information more comprehensively.

In terms of semantic match, the proposed model also leads with a score of 89.5%, higher than RoBERTa-Base at 89.3% and BERT-Base at 85.6%. This result suggests that the model has a stronger ability to understand contextual semantic structures. It can more accurately identify hidden semantic links in the text. Although RoBERTa performs closely, it still falls slightly behind. This shows that the proposed method is more effective in capturing deeper semantic information.

Compared with lightweight models such as ELECTRA and ALBERT, the proposed algorithm shows clear improvements in all metrics. This indirectly suggests that while lightweight models offer advantages in efficiency, they still lag behind in the accuracy and depth of semantic representation. In contrast, the proposed method improves semantic modeling quality by using a more effective training objective and masking strategy, while keeping the model complexity manageable.

In conclusion, the experiment confirms that the semantic mining method based on masked language modeling achieves higher accuracy and better semantic understanding across multiple dimensions. The proposed model performs well on various metrics, indicating its strength not only in word prediction tasks but also in constructing complete semantic representations. It models contextual information and semantic relationships more effectively. This provides a technical foundation for high-level semantic understanding and proves the method's practical value in complex semantic mining scenarios.

This paper also presents an experiment to analyze the performance of the masked language model in short and long text semantic modeling. The experimental results are shown in Figure 2.

The experimental results in the figure show that masked language models generally perform better on long texts than on short texts. This is especially clear in semantically sensitive metrics such as Semantic Match and F1-Score. It indicates that the model can capture more contextual information from long texts. This helps in learning more complete semantic representations, which improves understanding and abstraction of textual meaning. In contrast, short texts contain sparse information, leading to more limited semantic associations learned by the model.

In terms of accuracy, the performance gap between long and short texts is also present. Although the difference is not large, it still shows that the model performs better when provided with richer input information. This aligns with the mechanism of masked language models, which depend on the structure of the context. The prediction of masked positions relies on the integrity and coherence of the surrounding content. Long texts offer extended context and more semantic cues, which support deeper semantic modeling.

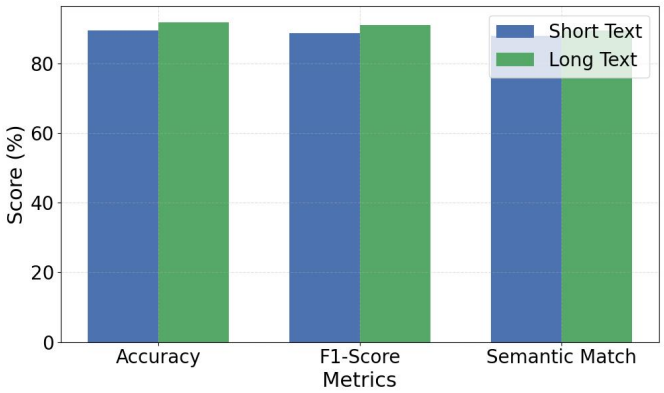


Figure 2. Experimental analysis of the performance of masked language models in semantic modeling of short and long texts

The improvement in F1-Score further indicates that the model not only makes more accurate decisions on long texts but also maintains better stability in recall. This is especially important for semantic mining tasks. In many real-world scenarios, it is not enough to simply make correct predictions. It is also essential to identify all relevant semantic features and patterns in the text. The multi-level information structure in long texts helps the model capture implicit semantics and adapt to complex language structures.

Overall, the results confirm the difference in semantic modeling performance of masked language models when dealing with texts of different lengths. The ability to utilize context is a key factor in improving performance in semantic mining tasks. Long texts provide more complete context, which helps the model discover deeper semantic relations. Short texts, on the other hand, require efficient modeling of limited semantic clues. These findings offer theoretical guidance for choosing input structures and modeling strategies in different application scenarios.

Finally, this paper also presents a semantic preservation experiment on the robustness of the model under different text perturbations, and the experimental results are shown in Figure 3.

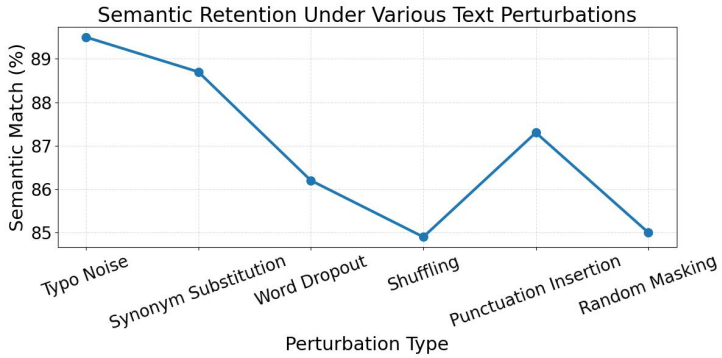


Figure 3. Semantic preservation experiment on model robustness under different text perturbations

The experimental results show that masked language models exhibit varying levels of semantic retention when facing different types of text perturbations. For mild perturbations such as typo noise and synonym substitution, the model maintains high semantic match scores, reaching 89.5% and 88.7%, respectively. This indicates strong semantic robustness in handling local lexical variations. Such robustness comes from the model's ability to capture lexical redundancy and semantic generalization through context modeling.

However, as the intensity of perturbation increases, the model's performance declines. In particular, semantic match scores drop significantly to 86.2% under word dropout and 84.9% under word shuffling.

These results suggest that the model's reliance on contextual coherence is weakened when facing structural disruptions. It becomes difficult for the model to reconstruct a complete semantic chain. Although the model has some language understanding ability, it remains sensitive to syntactic alterations and struggles to recover heavily damaged semantic relationships.

It is worth noting that under punctuation insertion, a relatively non-structural perturbation, the model's semantic retention improves to 87.3%. This suggests that the model shows robustness to surface-level, non-lexical changes. It can still accurately understand the semantic content without changes to syntactic structure. In contrast, random masking leads to another drop in semantic match, indicating that without specific contextual guidance, the model tends to produce less precise interpretations.

Overall, the experiment highlights differences in the semantic retention and robustness of masked language models under various perturbation conditions. The model shows strong tolerance to minor or surface-level noise. However, it still encounters performance limitations when dealing with structural disruptions or damaged contextual integrity. These findings provide a clear direction for improving semantic mining algorithms and enhancing their robustness in real-world scenarios.

4. Conclusion

This study focuses on the application of masked language models in semantic mining tasks. It proposes a semantic modeling method that integrates deep language understanding. Through a self-supervised masking prediction mechanism, the method captures deep semantic relationships between contextual elements. This improves the model's text understanding ability across multiple dimensions. With structured design and optimized objective functions, the model not only performs well in standard semantic classification tasks but also demonstrates strong context modeling and semantic matching capabilities. It provides both theoretical and practical foundations for handling complex semantic tasks.

The study further analyzes how the model performs under different text structure conditions. It examines the differences between short and long texts, as well as the model's robustness under various semantic perturbations. The results reveal the model's sensitivity to the richness of semantic cues and its adaptability to structural variations in language. These findings offer empirical evidence for improving generalization and stability. In addition, the proposed method does not rely on manual annotation. It has good transferability and flexibility, making it suitable for a wide range of natural language processing tasks.

The proposed method can be applied to real-world tasks such as text classification, intelligent question answering, sentiment analysis, and semantic search. It is particularly effective in tasks involving complex data distributions or hidden semantic relationships. The context-awareness and semantic generalization capabilities of masked language models provide strong technical support for building efficient and stable semantic mining systems. This also opens new directions for knowledge extraction and text understanding in data-driven applications. The method shows strong potential for extension in multilingual processing and cross-domain semantic modeling.

Future research may further optimize the model in areas such as modeling depth, pre-training strategies, and multi-task learning. It can also explore the model's potential in complex tasks such as multimodal semantic mining, low-resource language processing, and semantic reasoning. With the continued growth of large-scale corpora and computational resources, masked language models will play an increasing role in real-world applications. They will continue to advance natural language processing in areas like intelligent search, knowledge graph construction, and human-computer interaction.

5. References

[1] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//International Conference on Machine Learning. PMLR, 2020: 1597-1607.

[2] Yoon J, Zhang Y, Jordon J, et al. VIME: Extending the success of self-and semi-supervised learning to tabular domain[C]//Advances in Neural Information Processing Systems. 2020, 33: 11033-11043.

[3] Azizi S, Mustafa B, Ryan F, et al. Big self-supervised models advance medical image classification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 3478-3488.

[4] Zhang F, Hua B S, Ye M, et al. Self-supervised learning on point clouds via reconstructing space[C]//Advances in Neural Information Processing Systems. 2021, 34: 12942-12953.

[5] Sinha K, Jia R, Hupkes D, et al. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little[J]. arXiv preprint arXiv:2104.06644, 2021.

[6] Bao H, Dong L, Wei F, et al. Unilmv2: Pseudo-masked language models for unified language model pre-training[C]//International conference on machine learning. PMLR, 2020: 642-652.

[7] Ghazvininejad M, Levy O, Liu Y, et al. Mask-predict: Parallel decoding of conditional masked language models[J]. arXiv preprint arXiv:1904.09324, 2019.

[8] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.

[9] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[10] Clark K, Luong M T, Le Q V, et al. Electra: Pre-training text encoders as discriminators rather than generators[J]. arXiv preprint arXiv:2003.10555, 2020.

[11] Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv:1909.11942, 2019.