International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 18, No. 3, 2025

Study on the Classification of Chinese Black Humor Texts based on Transformer, Word2Vec, and BERT Techniques

Yingjie Zhao*

School of Foreign Languages, Beijing Institute of Technology, Beijing, 100081, China
*jasonzhao140@gmail.com

Abstract: With the rapid development of information technology, the number of Chinese black humor texts has surged. Efficiently and accurately classifying their sentiment orientation has become a key research focus in natural language processing (NLP). This study delves into the performance of three models-BERT, Transformer, and Word2Vec-on the task of Chinese black humor text classification. By collecting and organizing relevant corpora from online forums, black humor expressions, and GitHub repositories, two types of datasets were constructed: single-sentence and question-answer datasets. Comparative analyses of the three models reveal that the BERT model, leveraging its bidirectional encoding mechanism and pre-training capabilities, outperforms others in understanding nuanced emotional tendencies and complex contextual relationships in Chinese black humor texts. This research not only provides novel solutions for Chinese black humor text classification but also enhances our understanding of pre-trained language models in domain-specific applications.

Keywords: Transformer, BERT, Word2Vec, Black Humor, Sentiment Classification.

1. Introduction

With the rapid development of the internet and big data technologies, society has entered an era of information explosion. As the main carrier of information, the amount of text has surged, covering fields such as news, social media, and literature. Black humor is typically defined as an artistic form of humor that satirizes and mocks death, illness, and all topics traditionally considered serious or taboo through irony, exaggeration, and absurdity. This form of humor often reveals a cold and indifferent attitude toward human suffering and tragedy. For instance, in Image 1, the distinction between black humor and regular humor is quite apparent:

The context of regular humor in the dialogue demonstrates a relaxed and pleasant interaction between colleagues. Colleague A humorously remarks, with a wry smile, about the long hours spent in the office due to overtime. Colleague B responds playfully, mentioning the company's provision of free dinner as a small perk for working late. Such exchanges not only alleviate the stress brought on by overtime but also reflect an optimistic and friendly attitude. In contrast, the context of black humor presents a much sharper and more ironic tone. Colleague A, with a deadpan expression, introduces the absurd scenario that working overtime until the early hours would entitle one to a free luxury coffin, directly pointing to the extreme phenomenon of overtime culture. Colleague B responds with a cold laugh, further pushing the absurd scenario to an even more extreme point. This response intensifies the satirical undertone, critically exposing the inherent absurdities and irrationalities embedded within the culture of excessive overtime.

Through this form of ironic expression, audiences are guided to grasp the impermanence and absurdity of life within the laughter. However, with the overwhelming volume of Chinese black humor texts, the challenge of efficiently and accurately classifying their nature-so as to better support applications such as information retrieval, sentiment analysis, and content recommendation-has emerged as a critical issue in the field of Natural Language Processing (NLP) that requires urgent attention [1].

Table 1. ordinary humor and black humor

Type	Q	A
Ordinary Humor	Colleague A (wry smile): "With the way things are going, are we running the office tonight?"	Colleague B (jokingly): "Yeah, but at least the company is offering free dinners. Consider it a little bonus for our overtime grind."
Black Humor	Colleague A (deadpan): "Did you hear? The company's new policy now offers a complimentary luxury coffin for anyone working until midnight."	Colleague B (cold laugh): "Wonderful. At least our overtime comes with a parting gift. Maybe we'll even survive the morning rush to the afterlife."

2. Research Background

Black humor addresses serious topics by highlighting the absurdity of human existence and contradictions within social norms. It often treats serious, sorrowful, or taboo subjects in a flippant manner. With the increasing prevalence of black humor in online texts, the demand for its automatic recognition has gradually risen. Traditional text classification methods struggle to capture the subtle nuances of its semantics and emotional inclinations. The rise of deep learning and pre-trained language models (such as BERT and Word2Vec) has provided new solutions for text classification [2].

2.1. Research Significance and Cultural Impact

Theoretically, investigating the performance of different

pre-trained language models in classifying black humor texts deepens our understanding of model efficacy and optimization strategies. In practical terms, accurate classification enhances content recommendation, sentiment analysis, and social media monitoring systems. Additionally, accounting for cultural variations is crucial during the classification and recognition process.

2.2. Challenges in Automatic Recognition

Automatic recognition of black humor involves a deep understanding of language, including wordplay, subtle contextual shifts, and non-verbal semantic elements such as irony and exaggeration. Traditional methods struggle to capture the complexity of black humor. Moreover, black humor texts often contain complex emotions, such as humor, sadness, or horror, which further complicates the classification process.

2.3. Research Objectives and Algorithm Application

This study aims to conduct an in-depth and accurate classification of Chinese black humor texts based on Transformer, BERT, and Word2Vec models, establishing a classification system for positive, neutral, and negative emotional tendencies. The specific objectives include:

(1) Clearly define the standards and characteristics of positive, neutral, and negative emotional tendencies in Chinese black humor texts:

1) Positive black humor mocks everyday dilemmas or minor misfortunes in a mild, non-offensive manner, aiming to alleviate tension or provide a lighthearted laugh without touching on sensitive topics;2) Neutral black humor addresses serious or sensitive issues in a more direct and candid manner, using humor to reveal social phenomena or human weaknesses without overly stimulating or hurting their feelings; 3) Negative black humor, on the other hand, mocks severe misfortunes, disasters in an extreme, offensive, or cruel way. It is likely to provoke controversy or criticism, touching on moral boundaries or emotional taboos, and is often considered inappropriate or offensive.

(2) Conduct a comprehensive comparison of the performance differences among models such as Word2Vec and BERT in the task of classifying Chinese black humor texts. The evaluation will be based on multiple dimensions, including accuracy, recall, and F1 score, to assess the models' performance and provide a foundation for model optimization.

2.4. Research Contributions

This study is expected to make contributions in the following areas:

(1) Establish a systematic classification framework for the sentiment orientation of black humor texts.

(2) Conduct an in-depth evaluation of the model performance of Word2Vec and BERT to provide a basis for model selection and optimization.

(3) Propose effective feature extraction and fusion strategies, incorporating attention mechanisms, to enhance classification accuracy and generalization ability.

3. Related Research

Black humor is a style of humor that addresses serious topics in a satirical manner, first proposed by André Breton (Breton, 1940). Its research has become increasingly significant in the areas of humor classification and automatic detection. For example, in 2018, Vikram Ahuja et al. introduced a new framework for humor classification based on patterns, themes, and topics, emphasizing the importance of sentiment analysis [3]. Peng-Yu Chen et al. demonstrated the effectiveness of combining convolutional neural networks with high-speed networks for automatic humor detection[4]. Mihai Samson and Daniela Gifu, using deep learning techniques, investigated humor detection and offensiveness rating tasks in social media texts, finding that fine-tuning pre-trained BERT models yielded the best performance [5].

In the field of large models, Tomas Mikolov proposed the Word2Vec model in 2013, demonstrating the model's ability to capture syntactic and semantic relationships between words[6]. The Transformer-based pre-trained model BERT, introduced by Jacob Devlin in 2018, significantly enhanced the performance of natural language processing (NLP) tasks through deep bidirectional pre-training [2].

In the area of text classification, Joseph Lilleberg et al. proposed a method combining Word2Vec and tf-idf to improve the accuracy and efficiency of text classification [7]. Sidan M and Dongsu L introduced a weighted Word2Vec-based approach for text classification, which significantly improved classification accuracy[8]. Sirui Li optimized the performance of the BERT model for Chinese text classification tasks [9]. Building on this work, Qing Yu et al. proposed the BERT-BiGRU model, which addressed the complexities and diversity inherent in Chinese text classification[10]. Thomsen et al. found that fine-tuning the BERT model yielded excellent results in humor recognition and scoring tasks, revealing the advantages of neural network embeddings in humor detection [11]. In 2024, Yinghui Li et al. introduced the FLUB dataset, designed to evaluate the performance of large language models (LLMs) in understanding and interpreting black humor texts in Chinese, such as "dark jokes" and "wordplay" [12].

4. Transformer

The Transformer model is a sequence-to-sequence model based on the attention mechanism, proposed by Vaswani et al. in 2017. Its core component is the self-attention mechanism, which allows the model to simultaneously focus on different positions within a sequence, thereby capturing long-range dependencies. The Transformer employs a multi-head attention mechanism, utilizing multiple self-attention layers, with each layer focusing on different subspaces of the sequence [13].

4.1. Encoder:

The encoder maps the input sequence to a high-dimensional space.

4.1.1. Input Embedding:

Convert input data into high-dimensional vector representations.

4.1.2. Multi-Head Attention:

Computes attention weights in parallel through multiple heads, learning representations in different subspaces, and concatenates the outputs before passing them through a linear layer to produce the final result.

4.1.3. Add & Norm:

Incorporates residual connections with layer normalization. It adds the input directly to the sublayer output to help alleviate vanishing or exploding gradients, then normalizes

this result to stabilize learning [14].

4.1.4. Feed-Forward Network:

The network consists of two linear transformations and a ReLU activation function, which are applied independently to each position, providing additional non-linear transformation capability and enhancing the model's expressive power[13].

4.2. Decoder:

The decoder predicts the next output based on the encoder's output and the previously generated sequence. Its structure is similar to that of the encoder, but with the addition of a masking mechanism to prevent information leakage. Key components include:

4.2.1. Masked Multi-Head Attention:

The masking operation preserves the autoregressive property, preventing information leakage by ensuring that predictions are based solely on previously known output positions. This maintains a left-to-right causal relationship during the prediction process.

4.2.2. Cross-Attention:

The decoder utilizes queries from the decoder and keys and values from the encoder's output to achieve effective information transfer, thereby enhancing parallelism and training efficiency[13].

4.3. Output Processing:

The output embedding transforms the decoder's output vectors into symbolic representations of the target sequence, while the Softmax layer converts them into a probability distribution, thereby generating the final prediction.

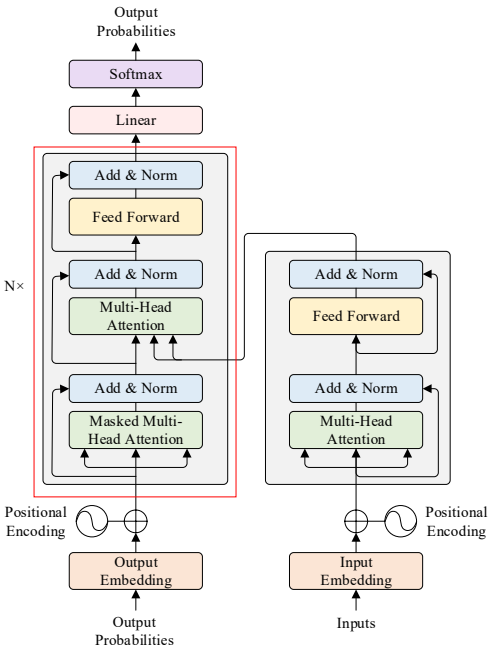

Figure 1. The structure of Transformer Model

The Transformer model captures global dependencies through the self-attention mechanism, which supports highly parallelizable computation. However, its self-attention mechanism has a computational complexity of $O(n^2)$, which can lead to performance degradation when handling long sequences [13]. To mitigate this issue, Transformer-XL introduces the caching of hidden states and relative positional encodings, demonstrating superior performance in long-text tasks [15,16]. It particularly excels in long-sequence modeling and computational efficiency [17].

5. Word2Vec

Word2Vec, a lightweight neural network introduced by Tomas Mikolov in 2013, consists of two core architectures: the Continuous Bag of Words (CBOW) model and the Skip-gram model[6], as illustrated in the figure below:

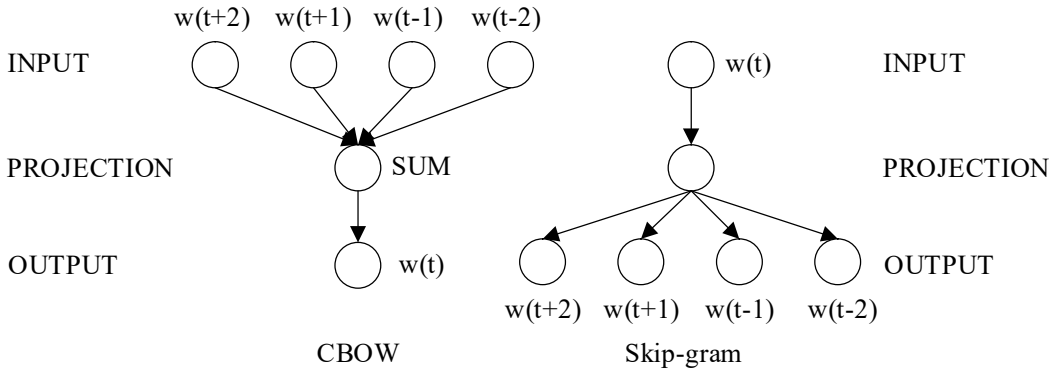

Figure 2. The structure of Word2Vec Model

5.1. Continuous Bag-of-Word (CBOW)

The model predicts the target word based on the context words. It is particularly suited for smaller datasets. The input layer converts the one-hot encoded context words into word vectors, typically achieved using an embedding matrix. The hidden layer sums these vectors, capturing the contextual information needed for predicting the target word. The output layer then maps the hidden layer's vector to the vocabulary size using a weight matrix and applies the softmax function to generate the probability distribution for the target word, thereby enabling effective semantic learning and prediction.

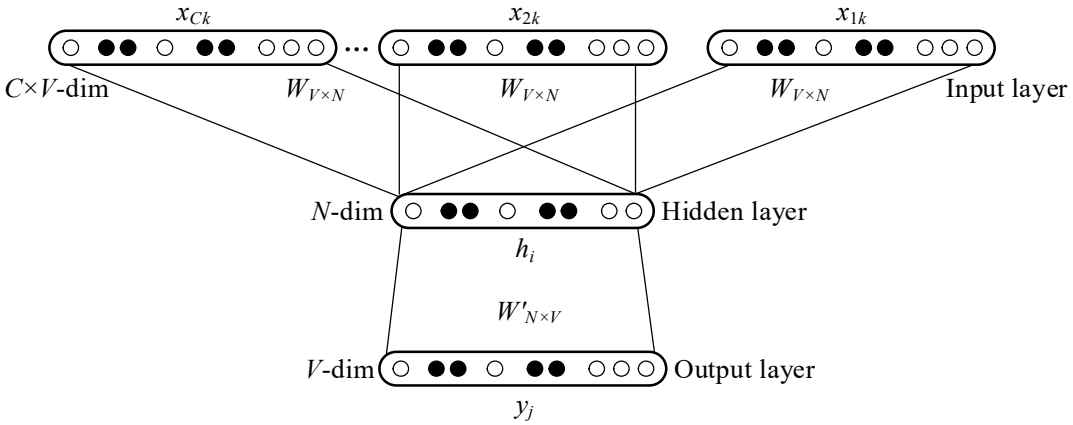

Figure 3. The Structure of the Continuous Bag-of-Word Model

5.2. Skip-gram Model

The Skip-gram model [6] predicts the context words based on a given center word, making it particularly suitable for large-scale datasets. The input layer converts the one-hot encoded words into embedding vectors, typically using an embedding matrix. The hidden layer directly processes these vectors to capture the semantic properties of the words. The output layer then applies the softmax function to predict the probability distribution of context words from the embedding vectors in the hidden layer. The model efficiently learns the semantic relationships between words using a gradient descent optimization algorithm.

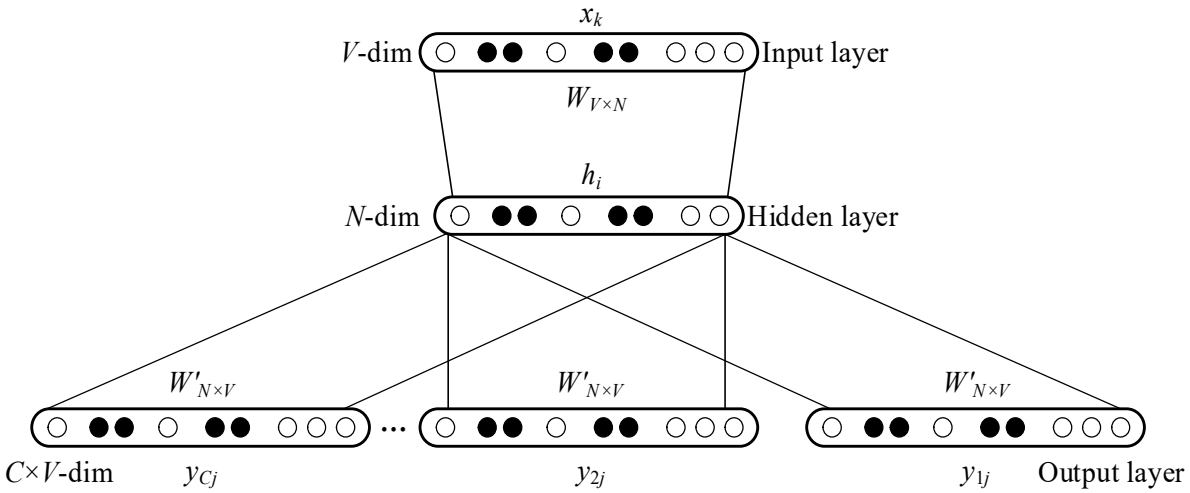

Figure 4. The Structure of the Skip-gram Model

The effectiveness of Word2Vec lies in its ability to capture both syntactic structures and semantic relationships between words, addressing the challenge of extracting high-quality word embeddings from large datasets.

To enhance training efficiency, the Skip-gram model typically employs negative sampling and hierarchical softmax to approximate the softmax computation. Negative sampling reduces the computational burden by randomly selecting non-relevant examples, thereby simplifying similarity calculations in large datasets. Hierarchical softmax, on the other hand, uses a Huffman tree to convert the multi-class classification problem into binary classification, significantly reducing the computational complexity from O(N) to approximately O(logN), thus accelerating the training process [6].

6. Bert

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model based on a multi-layer Transformer encoder, introduced by Google in 2018. Its core functionality is to capture bidirectional contextual information from text, demonstrating exceptional performance in a variety of natural language processing tasks[2].

BERT takes preprocessed text sequences as input, tokenizing the text into tokens using the WordPiece method. It then generates the corresponding Token, Segment, and Position Embeddings. These embeddings are subsequently passed through multiple Transformer encoder layers, where they undergo processing to extract text features. The Transformer encoder layers capture complex dependencies between tokens through the self-attention mechanism [13] and by incorporating non-linear activation functions (such as ReLU) in the feed-forward neural network[18]. Additionally, residual connections and layer normalization are employed to prevent gradient vanishing [14], enhancing the model's convergence and stability, which are applied to the outputs of both the self-attention and feed-forward networks in BERT .

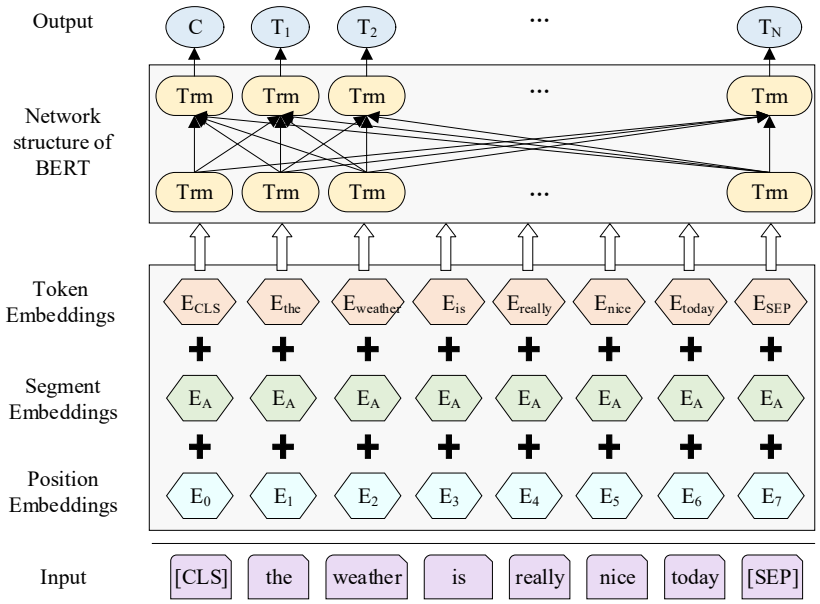

Figure 5. The Structure of the BERT Model

7. Dataset Description and Preprocessing

7.1. Data Scale and Specifications

This study involves the collection and organization of data from Baidu Tieba (including forums such as "Black Humor Tieba" (黑色幽默吧) and "Ruòzhì Ba" (弱智吧)), online black humor statements, and GitHub repositories related to black humor. The Chinese black humor texts collected can be broadly categorized into two types: single-sentence and question-answer formats.

Single-Sentence refers to the organization of individual Chinese black humor sentences as separate units. Each sentence possesses its own independent semantic structure and logic of black humor. This type does not include contextual settings and is not situated within a dialogue environment. In textual modality, it is manifested as discrete, single-dimensional linguistic information fragments. The dataset comprises 2,353 texts, totaling 60,142 characters, with an average length of 25.56 characters per sentence. Question-Answer refers to the organization of groups of Chinese black humor sentences as units, characterized by contextual associations. They are structured in an interactive question-and-answer format, where the question and answer are interdependent and responsive to each other, forming a dual-dimensional linguistic modality based on conversational scenarios. The semantic connotations and effects of black humor rely on the dynamic interactions between the question and answer. Compared to single-sentence corpora, question-answer corpora possess more specific contextual and emotional directional properties. The dataset includes 23 texts, totaling 916 characters, with an average length of 39.82 characters per entry.

7.2. Dataset Preprocessing

7.2.1. Data Modality

The data is presented in Excel text format, facilitating efficient processing and analysis.

7.2.2. Data Processing

(1) Data Cleaning

Removal of irrelevant characters (e.g., HTML tags, special symbols), stop words, and excess punctuation to reduce noise and inconsistencies.

(2) Formatting

Text across the dataset is standardized to ensure uniformity for processing. This includes adjusting text to a consistent font size and style, and streamlining punctuation usage. The Excel documents categorize data into several primary columns:

- "TEXT": Collection of single-sentence or question-answer corpora.
- "Negative", "Neutral", "Positive": Designate the sentiment classification of Chinese black humor in the sentences.
- "LEN": Denotes the number of characters in each sentence unit.
- "Q" and "A": Represents the collection of question-answer corpora.

(3) Handling Missing Values: Missing data is either filled in or removed to ensure the integrity of the dataset.

(4) Feature Extraction

For Word2Vec, it is necessary to construct a vocabulary and generate word vectors that capture semantic relationships between words. In contrast, for BERT and Transformer models, feature extraction is performed directly using their pre-trained models, leveraging their ability to understand complex linguistic patterns without the need for manual feature engineering.

(5) Text Tokenization and Initial Labeling

For Chinese text, appropriate tokenization tools, such as jieba, are employed to segment the text, as Chinese text does not have natural whitespace separation like English. This step provides a foundational structure for subsequent feature extraction and model training. Furthermore, experts from various professional backgrounds are consulted, and the judgments of researchers and some basic large models are referenced to categorize sentences in both corpora into three classes: negative, neutral, and positive

8. Advantages, Disadvantages, and Suitability of Large Models

Introduced in 2017 by Vaswani et al., the Transformer model has significantly advanced natural language processing with its ability to effectively manage long-range dependencies and its excellent scalability, despite being computationally intensive and less effective on imbalanced

datasets compared to RNNs [13]. BERT, an evolution of the Transformer developed by Devlin et al. in 2018, utilizes a bidirectional architecture to deeply understand the context from both directions of a sentence. This capability is crucial for processing complex texts such as black humor, as it allows BERT to grasp subtle emotional expressions like irony and reversal, which are often pivotal in sentiment analysis[2]. Although BERT delivers outstanding performance, it is also noted for its high computational demands and potential challenges in handling very long sequences.

On the other hand, Word2Vec, developed by Mikolov et al., provides a fast and resource-efficient solution for mapping words into vector spaces but struggles with static embeddings and fixed contextual windows, making it less suitable for nuanced sentiment analysis tasks where context is key[6].

In evaluating the performance of these models for sentiment analysis of Chinese black humor, BERT stands out for its robust semantic understanding, enabling it to better interpret complex and subtle emotional nuances. While Transformer and Word2Vec each have their merits, BERT's bidirectional encoding uniquely equips it to deeply analyze complex contexts within texts, leading to more accurate sentiment identification, particularly in recognizing intricate emotions within black humor.

9. Conclusion

This study conducts a comparative analysis of the Transformer, Word2Vec, and BERT models in the task of Chinese black humor text classification, providing an in-depth examination of their respective strengths and weaknesses in handling complex semantic relationships. The results demonstrate that the BERT model outperforms both Transformer and Word2Vec models in terms of classification accuracy and generalization ability. The bidirectional encoding mechanism of BERT, combined with pre-training on a large-scale corpus, enables it to thoroughly understand the context of the text, effectively capturing subtle emotions such as sarcasm and reversal inherent in black humor. As a result, BERT achieves superior performance in classification tasks. In contrast, while the Transformer model excels in parallel processing and capturing long-range dependencies, its unidirectional encoding limits its ability to fully grasp contextual information when processing Chinese black humor texts. On the other hand, the Word2Vec model, due to its lack of contextual awareness and dynamic adaptation mechanisms, performs less effectively in complex sentiment classification tasks. Therefore, for the task of classifying Chinese black humor texts, the BERT model emerges as the most suitable choice, highlighting its superiority in this domain. This study provides a novel perspective on the classification of Chinese black humor texts and contributes to enhancing the accuracy of information retrieval, sentiment analysis, and content recommendation. Future research could further explore optimization strategies for the models and feature fusion techniques to improve both classification performance and generalization capability.

References

[1] Turing, A. M. (2009). Computing machinery and intelligence. Springer Netherlands. pp. 23-65

[2] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arxiv preprint arxiv:1810.04805.

[3] Ahuja, V., Bali, T., & Singh, N. (2018). What makes us laugh? Investigations into automatic humor classification. In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. pp. 1-9.

[4] Chen, P. Y., & Soo, V. W. (2018). Humor recognition using deep learning. In Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, 2, 113-117.

[5] Samson, M., & Gifu, D. (2021). FII FUNNY at SemEval-2021 Task 7: HaHackathon: Detecting and rating Humor and Offense. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021). pp. 1226-1231.

[6] Mikolov, T. (2013). Efficient estimation of word representations in vector space. arxiv preprint arxiv:1301.3781, 3781.

[7] Lilleberg, J., Zhu, Y., & Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. In 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC). pp. 136-140.

[8] Sidan, M., & Dongsu, L. (2019). Text classification method based on weighted Word2Vec. Information Science, 11, 38–42.

[9] Li, S. (2020). Subword-level Chinese text classification method based on BERT. Computational Science and Applications, 10, 12677.

[10] Yu, Q., Wang, Z., & Jiang, K. (2021). Research on text classification based on bert-bigru model. In Journal of Physics: Conference Series. 1746(1), 012019.

[11] Thomsen, D. B., de la Broıse, J. B., & Mielonen, E. Recognizing Humor and Predicting Humor Ratings in Short Texts.

[12] Li, Y., Zhou, Q., Luo, Y., Ma, S., Li, Y., Zheng, H. T., ... & Yu, P. S. (2024). When llms meet cunning questions: A fallacy understanding benchmark for large language models. arxiv preprint arxiv:2402.11100.

[13] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Proceedings of the Neural Information Processing Systems (NeurIPS), 1–11.

[14] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.

[15] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1-67.

[16] Dai, Z. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. arxiv preprint arxiv:1901.02860.

[17] Lan, Z. (2019). Albert: A lite bert for self-supervised learning of language representations. arxiv preprint arxiv:1909.11942.

[18] Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10). pp. 807-814.