Role of Natural Language Processing in Document Understanding and Semantic Analysis: A Chinese Perspective

Yun Liu

Recommended citation:

Liu, Yun (2024). "Role of Natural Language Processing in Document Understanding and Semantic Analysis: A Chinese Perspective". *Profesional de la información*, v. 33, n. 3, e330324.

https://doi.org/10.3145/epi.2025.ene.0324

Manuscript received on 08th June 2023
Accepted on 12th November 2023

Yun Liu
https://orcid.org/0000-0003-4885-0357
School of Humanities and Law
Henan University of Animal Husbandry and Economy
Zhengzhou, Henan 450000, China
80814@hnuahe.edu.cn

Abstract

With the recent advancements in Natural Language Processing (NLP), there is a growing need to enhance the effectiveness and accuracy of document understanding and sentiment analysis particularly for Chinese text as it presents unique challenges of linguistics. To conduct this study, a hybrid approach was implemented which combined CNN and BiLSTM models with an ensemble voting mechanism for sentiment analysis on Chinese text. This method attained an accuracy of 97% and outperformed other techniques such as Text_CNN and AttentionBiLSTM with significant improvements in F1 score and recall. Results demonstrated a superior performance achieving 97% accuracy, along with a 95% F1 score and 97% recall. The present study extends the growing body of literature by underscoring the effectiveness of integrating BiLSTM and CNN models in sentiment analysis within the context of Chinese linguistics. It showcased enhanced document comprehension and capabilities of semantic analysis. Practically, this study provides a robust framework for leveraging BiLSTM and CNN models in the sentiment analysis of real-world. It offers significant boost in adequacy and reliability for processing Chinese text. The research limitations and future research indications have also been addressed in the study.

Keywords

Natural Language Processing, Semantic Analysis, Bidirectional Long short-term Memory, Convolutional Neural Network.

1. Introduction

Natural language processing (NLP) has emerged as a transformative force concerning document understanding and semantic analysis. It enables the machines to interpret and process human language in those ways that were considered unimaginable once. Natural language processing is considered as an integral area of computer science in which computational linguistics and machine learning are widely used. However, this field is mainly concerned with making the human and computer interaction easy but efficient (**Fanni *et al.***, 2023). In recent years, the application of NLP has grown exponentially which have been driven by advances in artificial intelligence, big data analytics and machine learning. These technologies have paved the way for more sophisticated methods of extracting the contexts, meaning and insights from a vast amount of unstructured textual information (**Stenmark**, 2022). NLP has spread its application in different fields such as email spam detection, extraction of information, machine translation, medical, questioning and answering and summarization. In this regard, NLP is a tract of linguistics and AI which is devoted to make computers understand the statements or words which have been written in human languages (**Khurana *et al.***, 2023; **Tunca *et al.***, 2023). As the organizations or individuals increasingly relies on digital communication (**Gawer**, 2022; **Baptista *et al.***, 2020), therefore the ability to analyze and understand documents have become important for a wide range of applications from retrieval of information and categorization of content to the sentiment analysis and decision-making.

In China NLP has taken on a particularly significant role due to the unique characteristics of Chinese language and the rapid digital transformation of the country (**Baptista *et al.***, 2020; **Liu *et al.***, 2023; **Liu *et al.***, 2024). Chinese language with its complex scripts, diverse dialects and rich cultural contexts presents distinct challenges and opportunities for NLP research and development (**Shao *et al.***, 2024; **Cui *et al.***, 2020). However, document understanding in China not only involves the adequate interpretation of characters and words but also the understanding of tone, contexts and cultural aspects. Furthermore, the rise of Chinese social media platforms such as Weibo or WeChat has generated an immense amount of user-generated content. It further underscores the need for the effective NLP solutions tailored to the linguistic environment of China.

Despite rapid advancements in NLP (**Tyagi; Bhushan**, 2023), unique complexities of Chinese language pose significant challenges for effective document understanding and semantic analysis. In this regard, the existing models of NLP which are often developed with a focus on Western languages, struggle to adequately interpret and assess the Chinese text can possibly result in issues such as retrieval of information, sentiment analysis and categorization of content. This gap underscores the need for tailored approaches of NLP that can address the specific linguistic and cultural aspects of Chinese. It ensures the adequate and meaningful procedures of Chinese-language documents in different applications. Although, NLP has gained an increased scholarly focus in recent years (**Kamineni *et al.***, 2024; **Dai *et al.***, 2024; **Shen *et al.***, 2024; **Shakhnoza; Nargiza**, 2024), but to the best of the researcher's knowledge there are almost a little or no studies that explored the Role of Natural Language Processing in Document Understanding and Semantic Analysis from a Chinese Perspective. To bridge this gap, this study aims to underscores the significance of tailored frameworks and approaches that can enhance the effectiveness of NLP in processing Chinese language documents. In this way, it contributes to the broader fields of computational linguistics and retrieval of information (**Estok**, 2022).

2. Literature Review

2.1. NLP Technology and Techniques

NLP has advanced alongside recent developments in artificial intelligence (AI) and computing technologies shaping and being shaped by these innovations (**Lauriola *et al.***, 2022; **Koroteev**, 2021; **Chowdhary**, 2020). This evolution has resulted in new applications and novel ways for machines to interact with human languages. **Omar *et al.*** (2022) discussed that the key benefits of NLP lie at the heart of teaching computers regarding the way through which large amounts of textual data is analyzed. Although, it may seem as a new technology with the emergence of recent successful applications, the roots of NLP go back to the early 1950s when NLP was utilized for the first time for the machine translation. Modern NLP focuses on the assessment of language patterns and extracting meaningful insights (**Shaik *et al.***, 2022; **Olivetti *et al.***, 2020; **López-Úbeda *et al.***, 2022). Therefore, it is garnering a significant attention in both industry and academia. It also presents remarkable opportunities across distinct applications of AI such as question answering, retrieval of information, sentiment analysis and recommender systems and supporting critical tasks such as machine translation and reading comprehension with a persistent improvement in real-world performance (**Chen *et al.***, 2022).

Koonce *et al.* (2024) highlight the rapid advancements and accessibility of artificial intelligence techniques, offering biomedical informatics team a significant opportunity to explore and derive insights from both external and internal data on a large scale. It ultimately enhances health science research and clinical care. Their research indicated a novel aspect of implementing NLP which is the application of NLP to extract valuable knowledge for the extensive volumes of structured and unstructured clinical data that are captured daily through the electronic health record (EHR). **Widdows *et al.*** (2024) also indicated that NLP is a promising area of application through quantum computers. Research by **Jiang and Lu** (2020) NLP is an essential part of artificial intelligence technology and is rooted in multiple disciplines such as computer science, mathematics and linguistics. They mentioned that the rapid advancements in NLP provides a strong support for machine translation research. **Baclic *et al.*** (2020) also explained that NLP is routinely utilized in the virtual assistants such as "Alexa", "Siri" or in google translation and searches. Their research indicated beneficial implication of NLP as it provides the ability to analyze and extract information from the unstructured sources, automate question answering and conducting text summarization and sentiment analysis.

2.2. Historical Development in NLP

The development of NLP is rooted in the intersection of computer science, linguistics and artificial intelligence (AI). The origin of NLP can be traced back to 1950s during the early stage of computing (**Bahja**, 2020; **Church; Liberman**, 2021; **Feng**, 2023; **Omar *et al.***, 2022). At that time the researchers initiated to understand the way machines can understand and process human language. One of the earliest milestones was the famous question of Alan Turing i.e., "Can machine think?" (**Berrar *et al.***, 2012; **Gonçalves**, 2023; **Saha *et al.***, 2023) and his subsequent proposal of the Turing test as a measure of machine's ability to exhibit intelligent behavior. That behavior was expected to be equivalent to or indistinguishable from that of a human.

Johri *et al.* (2021) cites how Alan Turing, in 1950, had proposed the Turing test which is also known as the imitation game to investigate whether a computer could exhibit human-like thinking. The test involved three participants including a man (player 1), a woman (player 2) and an interrogator (player 3). The task of interrogator was to identify the gender of player 1 and player 2 based solely on written communication. However, the twist in their game is that player 2 (the woman) tried to help the interrogator to reach the adequate conclusion. However, player 1 (the man) attempted to device the interrogator. However, Turing suggested replacing player 1 with a machine. If the interrogator successfully identifies the gender of both participants, the machine fails the test. However, if the interrogator cannot distinguish between the human and the machine, the machine is considered to have passed (**Turing**, 1980). It demonstrates its ability to think in such a way that is indistinguishable from that of a human. The test was not about solving a specific problem but rather about evaluating whether a machine could perform tasks in a way that is indistinguishable from the human thinking. Nowadays, NLP is used in many real-time applications such as smart homes, smart offices like Alexa, Siri, Cortana and Google Assistant. In short, the history of NLP initiated in 1950s and has come a long way from then and improved to a great extent (**Zhang**, 2023).

2.3. Challenges of NLP

Apart from its effective and rapid advancements, there are also certain challenges of NLP which have been reported by different researchers. Research by **Abro *et al.*** (2023) also mentioned that natural language understanding (NLU) in NLP exhibit significant challenges due to its critical role. It involved generating natural language responses and understanding the context. However, current algorithms of NLU also struggles to fully grasp the natural language as a similar meaning can be expressed in numerous ways and language usually deviates from the grammatic norms. While fundamental principles for NLU exists, capturing the subtleties of human language remains difficult. Unlike machines, humans can infer meaning from grammatically incorrect or incomplete sentences through experience and context which makes acquisition of language more intuitive for people.

Another challenge in NLP involves training models on limited data such as with few short learning and achieving out of distribution generalization as explained in the research of **Kharrat *et al.*** (2024). Sequence to sequence models which have been commonly used in NLP also struggles with systematic generalization which makes it difficult for them to grasp general principles and reasons regarding the critical concepts of knowledge. Furthermore, working with the historical documents also presents unique difficulties as these texts usually contain noise and missing or incomplete characters which can further complicate the tasks for NLP model. **Omar *et al.*** (2022) also indicated that there are numerous real-world NLP projects which have failed after deployment due to different factors. It involves a lack of robustness, exploring robustness as a multi-dimensional concept that needs the development of new techniques holds great importance. Their research also indicated that the newly developed techniques should address the challenges of spurious correlation and achieve high out of distribution adequacy so that sufficient accuracy to uncertainties can be ensured. According to the research by **Krasadakis *et al.*** (2024), the legal domain presents many challenges for NLP. Some major challenges involve disambiguating the tiles, resolving nested entities and resolving the coreferences.

3. Methodology

The structure of the proposed model used in this study is depicted in Figure.1. As it is evident, three models were used namely BILSTM, and a CNN incorporating with their ensemble using voting mechanism. BiLSTM or Bidirectional LSTM was used for great purposes; it is a term that utilized for the succession model. It consists of 2 LSTMs; one is for input processing in forward direction while other is for processing in backward direction. It is specially utilized in natural language processing related tasks. The idea of using this method is processing of data in a pair of directions (**Brahma**, 2018). CNNs have advantages over other models in NLP due to its hierarchical features learning step, it captures in input data global and local relationships, and achieve well and correctly on NLP tasks using fewer computation and parameters (**Rhanoui *et al.***, 2019).

Figure 1: Flowchart of the Proposed System.

3.1. Dataset Collection

The dataset was collected from (**Hu *et al.***, 2019) THUCNews setup by the NLP (Natural language processing) Tsinghua University Laboratory. This data is established from Sina News historical data. In this dataset 740000 article news are available and are separated in 10 types. The text length is from 20 to 30 word categories and every category contains 20,000 parts of text data. The THUCNews dataset are shown in Table 1 and Figure 2.

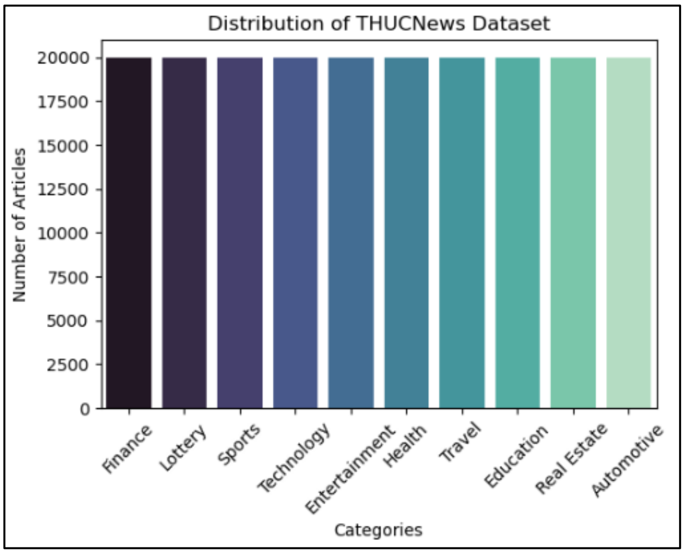

Figure 2: Details of THUCNews Dataset.

Table 1: Sample Dataset.

Category	Chinese text	English translation	class
Lottery	经济预计将增长，股市上涨，投资者信心恢复	The economy is expected to grow, stock markets rise, and investor confidence returns.	positive
Sports	全场观众为比赛中表现出色的选手报以热烈的掌声。	The audience gave a warm applause to the players who performed well in the match.	positive
Finance	新税法将给中小企业带来不成比例的负担，这让许多企业主感到担忧。	The new tax law will place a disproportionate burden on small and medium-sized businesses, which worries many business owners.	negative

3.2. Preprocessing

Three categories from downloaded dataset news were extracted. The starting line of each document was read and then inserted in a new file .txt, where every text data point consisted of a headline of news and body content. However, text title first line plus tab spacer plus type mapping with matching number was the format. 10% of data was selected for test and verification while the 90% was reserved for training from three categories. When data was completely extracted the training set was inserted in training.txt, testing set was inserted in testing.txt as well as the verification data was written is ver.txt files. At the end the meaningless or unimportant data was removed.

4. Results

4.1. Applying CNN and BILSTM models

This section explains the techniques that were utilized to build a system for user's sentiments classification on Chinese text. The CNN and BILSTM models. there is a famous option for NLP applications like machine translation, sentiment analysis and text classification. We trained two models on our prepared dataset. The framework of our applied methods is shown in Figure.3.

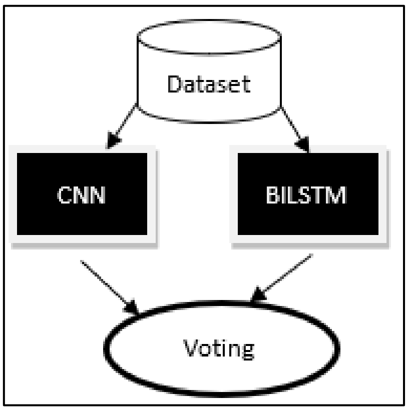

Figure 3: Proposed Models Framework.

4.2. CNN (Convolutional Neural Network) Model

CNN is often utilized in computer vision. It takes local patterns from the data especially helpful in recognizing local features or n grams in text and identifying common word or phrases related to precise sentiments. CNN can compute parallel, commanding quick training times. This can be adapted to many formats and text lengths, enabling it perfect for sentiment analysis. Following components were utilized using CNN

Embedding layer: This layer gives mapping function with highly dimensional place to input sequences, frequently utilized for indicating categorical variables or words. Embedding layer working formula is given below.

$$embedding\ layer = tf.keras.layer.embedding(L, d) \qquad (1)$$

Here L is an input sentence with a length fixed and d is the embedding dimension

Convolutional Layer: This layer is applied for Convolutional process to extract the features. In this we take kernels usually small numbers matrix and give beyond paragraph matrix, and then transformed on the basis of filter values. The result from convolutional layer on applying kernel (k) to the input can be calculated using eq 2.

$$(M - k + 1)\ \times 1\ \times filter\ numbers \qquad (2)$$

Max Pooling Layer: it is an operation of pooling that teks the largest element of the feature map area coated through the filter. Therefore, after this layer the output will be feature map containing very prominent features in the early feature map. Let applying max pooling with size of window w × 1 the result will be calculated using eq 3.

$$\frac{(M-k+1)}{w} \times 1\ \times filter\ numbers \qquad (3)$$

Flatten and Dense layer: After the embedding layer, the flat layer flattens the 2D output into a one-dimensional matrix suitable for input to the dense layer, which sorts the output values from the flat layer. See eq 4.

$$\frac{(M-k+1)}{w} \times filter\ numbers \qquad (4)$$

softmax activation: last output in dense layer with softmax for classification gives a normalized and clear probability distribution over possible sentiment classes i.e. positive or negative. Table 2 shows the CNN model training results.

Table 2: CNN Model Training.

Train on 270007 samples, validate on 300 samples
Epoch 1/2
- 28s - loss: 0.5473 - acc: 0.6672 - val_loss: 0.2780 - val_acc: 0.8484
Epoch 2/2
- 27s loss: 0.3094 - acc: 0.8188 - val_loss: 0.2625 - val_acc: 0.6564
Accuracy: 88.64%

4.3. BILSTM Model

BiLSTM or Bidirectional LSTM is used for great purposes; it is a term that utilized for the succession model. It consists of 2 LSTMs; one is for input processing in forward direction while other is for processing in backward direction. It is specially utilized in natural language processing related tasks. The idea of using this method is processing of data in a pair of directions. The model is able to better realize the sequences (for example, understanding which words follow and precede them in a sentence) (**Hu *et al.***, 2019). The following components were used in this model:

- ***Embedding layer:*** Embeddings are the representations of dense vector which is being transformed by input sequence. Embedding takes data points for semantic meaning and gives very meaningful and compact subsequent layers representation. Suppose n tokens of a sentence were given $D = [n_1, n_2, n_3, n_4, \ldots. n_m]$. every word n_i became vector consisting with meaning $n_i € V^u$. The word embedding technique we utilized with keras that creates every sentence of feature matrix. $M € V^{u \times m}$ indicated this matrix in which m shows the word count in sentence length and M is the embedded semantics that utilized as an input for the bilstm layer
- ***Bidirectional LSTM layer:*** One-way LSTM retains only previous information because it processes only previous inputs. By masking previously processed inputs, it conserves information. Bi-LSTM (BiLSTM), which can preserve past and future information, was created to reduce this problem. BiLSTM first establishes a connection between input and output while maintaining data order. The hidden state of BiLSTM can record past as well as future data using 2 layers, the forward layer and the backward layer. The collected feature maps are fed into the BiLSTM layer. In Forward LSTM present input *"u_s"* and the past input *"w_{s-1}"* are combined to examine the order of left to right. In backward LSTM the order is managed from right to left by combining 2 inputs present and the past i.e. *"u_s"* and *"w_{s+1}"*. Eq 5 indicated the both LSTMs combination

$$\overleftrightarrow{\boldsymbol{w}} = \overrightarrow{\boldsymbol{w}} \blacksquare \overleftarrow{\boldsymbol{w}} \qquad (5)$$

When lastly pooled map features were received by BILSTM layer, the forward and backward calculation of LSTMS will be found using below equations.

Forward LSTM

$$\boldsymbol{a_s} = \sigma(\boldsymbol{Y_a U_s} + \boldsymbol{Z_a w_{s-1}} + \boldsymbol{O_s}) \quad (6)$$

$$\boldsymbol{k_s} = \sigma(\boldsymbol{Y_k U_s} + \boldsymbol{Z_k w_{s-1}} + \boldsymbol{O_k}) \quad (7)$$

$$\boldsymbol{v_s} = \sigma(\boldsymbol{Y_v U_s} + \boldsymbol{Z_v w_{s-1}} + \boldsymbol{O_v}) \quad (8)$$

$$\boldsymbol{i_{\to s}} = \boldsymbol{\mathcal{A}}(\boldsymbol{Y_i U_s} + \boldsymbol{Z_i w_{s-1}} + \boldsymbol{O_i}) \quad (9)$$

$$\boldsymbol{i_s} = \boldsymbol{a_s} \blacksquare \boldsymbol{i_{s-1}} \oplus \boldsymbol{k_s} \bowtie \boldsymbol{i}\sim_{\boldsymbol{s}} \quad (10)$$

$$\boldsymbol{w_s} = \boldsymbol{v_s} \blacksquare \boldsymbol{\mathcal{A}}(\boldsymbol{i_s}) \quad (11)$$

Backward LSTM

$$\boldsymbol{a_s} = \sigma(\boldsymbol{Y_a U_s} + \boldsymbol{Z_a w_{s+1}} + \boldsymbol{O_s}) \quad (12$$

$$\boldsymbol{k_s} = \sigma(\boldsymbol{Y_k U_s} + \boldsymbol{Z_k w_{s+1}} + \boldsymbol{O_k}) \quad (13)$$

$$\boldsymbol{v_s} = \sigma(\boldsymbol{Y_v U_s} + \boldsymbol{Z_v w_{s+1}} + \boldsymbol{O_v}) \quad (14$$

$$\boldsymbol{i_{\to s}} = \boldsymbol{\mathcal{A}}(\boldsymbol{Y_i U_s} + \boldsymbol{Z_i w_{s+1}} + \boldsymbol{O_i}) \quad (15)$$

$$\boldsymbol{i_s} = \boldsymbol{a_s} \blacksquare \boldsymbol{i_{s+1}} \oplus \boldsymbol{k_s} \bowtie \boldsymbol{i}\sim_{\boldsymbol{s}} \quad (16)$$

$$\boldsymbol{w_s} = \boldsymbol{v_s} \blacksquare \boldsymbol{\mathcal{A}}(\boldsymbol{i_s}) \quad (17)$$

- ***Dropout Layer:*** The purpose of the dropout layer is to reduce the possibility of overfitting. In our study, we set the dropout layer value to 0.5, which operates ranges from 0 to 1. This layer follows the embedding layer and randomly sets the neurons in the embedding layer to 0 to systematically suppress their activation. This prevents the model from overly relying on specific neurons during training, thus improving the model's ability to generalize.
- ***Classification Layer:*** Final output layer classification layer. The first step in determining the segregation probabilities (i.e., whether they are positive or negative) is to calculate the cumulative inputs using eq 18

$$\boldsymbol{A_i} = \sum_{\boldsymbol{i}}^{\boldsymbol{k}} \boldsymbol{u_i d} + \boldsymbol{c_i} \quad (18)$$

The sigmoid function is utilized to examine y. see eq 19

$$\boldsymbol{y} \to \boldsymbol{1}/_{\boldsymbol{1} + \boldsymbol{\mathcal{A}^{-A}}} \quad (19)$$

4.4. Incorporating Ensemble Technique

By combining individual model performance, we enhance the robustness and performance of our proposed system in ensemble. Ensemble learning is a process in which we gather different model results and generate a powerful result. Ensemble learning can improve accuracy and reduce over fitting (**Saleena**, 2018). In our proposed system BILSTM and CNN were individually trained (See Table 3), BILSTMS has ability to take long term reliance and can improve local extraction of features in CNN. In SA (sentiment analysis) jobs, BiLSTM can specialize in extracting general sentiment from long reviews, while CNN can sharply recognize powerfully indicate sentiment key phrases.

Table 3: BILSTM MODEL TRAINING.

```
Epoch 001:
    Training Loss: 0.642 | Accuracy: 0.852.
    Validation Loss: 0.341 | Accuracy: 0.816. Improvement!
Epoch 002:
    Training Loss: 0.381 | Accuracy: 0.882.
    Validation Loss: 0.490 | Accuracy: 0.841. Improvement!
Epoch 007:
    Training Loss: 0.233 | Accuracy: 0.906.
    Validation Loss: 0.342 | Accuracy: 0.930. Improvement!
```

4.5. Voting Mechanism

Majority voting is an ensemble technique that can assist by ensuring the robustness of final prediction through consistency between different models consideration. Our trained model CNN and BILSTM combination utilizing majority voting mechanism process has following steps.

- First we train both models (CNN and BILSTM) individually
- Then we generate results from every model using test data
- Then in set of test data for every instance, the last prediction is the portion of all model predictions that received the majority of votes.

Majority voting working is done using the following eq 20, 21, and 22

$$yBILSTM = ModelB(a) \text{€} P \qquad (20)$$

$$yCNN = ModelC(a) \text{€} P \qquad (21)$$

$$z = mode(yBILSTM, yCNN) \qquad (22)$$

Where $yBILSTM$ and $yCNN$ are the final predictions of both models for a input and P is the class

Finally, z is the last prediction by returning most common class.

Example Case: suppose our model predictions have been achieved like this

$$yBILSTM = positive$$

$$yCNN = positive$$

Final prediction by using majority voting will be

$$z = mode(positve, positve) = positive$$

4.6. Comparison of Techniques

In this section, the performance of our proposed work is compared with different machine learning and deep learning models. First, we trained CNN and BILSTM individually and got 88% accuracy for CNN and 93% for BILSTM on the same dataset. Therefore, we combined both models' predictions using majority voting and achieved accuracy=97%, and 95% and 97% of f1 score and recall were obtained. However, we compared our results with (**Wu *et al.***, 2021; **Sun *et al.***, 2024; **Huq *et al.***, 2017) for sentiment analysis and achieved better results. Table 4 shows the performance of proposed study as compared to different techniques

Table 4: Comparison Table with Different Techniques.

Techniques	Results		
	Accuracy	F1 Score	Recall
Text_CNN	82%	82%	82%
AttentionBiLSTM	83%	83%	82%
ASN-CSD	91%	91%	94%
KNN	80%	79%	75%
Proposed (voting of CNN and BILSTM)	94%	95%	97%

Figure 4 shows the performance of our model as compared to other models.

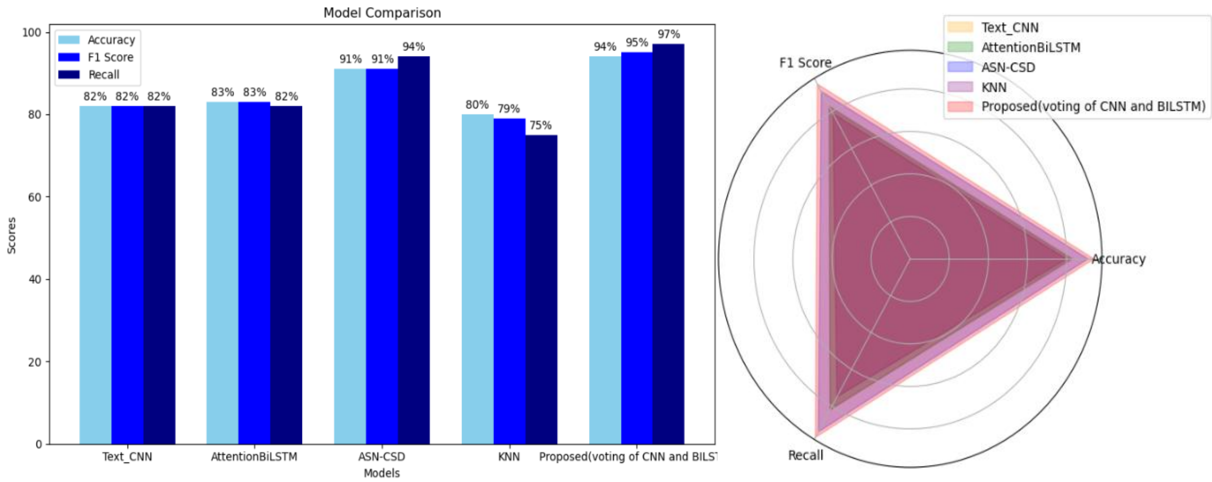

Figure 4: Performance of Our Model as Compared to other Models.

5. Discussion

The present research aimed to understand the role of Natural Language Processing (NLP) in document understanding and semantic analysis from a Chinese perspective. Findings of this study demonstrated the effectiveness of ensemble model combining CNN and BiLSTM for the sentiment analysis (**Tan *et al.***, 2022) on Chinese text data. The proposed model significantly outperformed the individual models and other comparative techniques. These findings correspond with the research conducted by **Luo *et al.*** (2021) which shows that the integrated model significantly improves the adequacy of text sentiment analysis and it can effectively predict the sentiment polarity of the text. Individually CNN attained 88% of accuracy and BiLSTM 93%. When combined through a majority voting mechanism, the ensemble model attained an impressive 97% of accuracy with F1 scores and recall rates of 95 and 97 percent respectively.

These results indicate that the ensemble approach effectively leverages the strengths of both models. CNN is capable of capturing the local patterns and the capability of BiLSTM to understand the sequential dependencies in text. **Jang *et al.*** (2020) also mentioned that BiLSTM model enhances accuracy in text classification. This synergy not only improves the adequacy but also ensures reliability and robustness in the sentiment classification. Also underscored that CNN-BiLSTM is considered as an effective technique as compared to others such as Naïve Bayes, SVM and CNN. They were also of the view that by combining these two models a hybrid model is generated that uses the strengths of CNN and BiLSTM.

The success of this study underscores the potential of ensemble learning in the complex tasks of NLP (**Jaradat *et al.***, 2024; **Zhang; Shafiq**, 2024) where integration of different models can mitigate the limitations of individual approaches. Furthermore, utilization of carefully selected datasets and rigorous pre-processing further contributed towards the superior performance of model. However, these findings resonate with the broader understanding that combining multiple models can often results in more adequate and generalizable results in machine learning (**Moein *et al.***, 2023) particularly within the context of natural language processing. This approach can also be extended to other tasks of NLP. It suggests that the ensemble methods should be considered as a vital strategy for attainment of high performance in text analysis.

The findings of this study underscore the effectiveness of utilizing ensemble learning techniques to enhance the performance of sentiment analysis models (**Qabasiyu *et al.***, 2023; **Alsayat**, 2022; **Chen; Pang**, 2023). It is particularly relevant while dealing with complex linguistic structures such as Chinese text. Therefore, by combining the strengths of convolutional neural networks (CNN) and Bidirectional long-short term memory networks (BiLSTM) (**Méndez *et al.***, 2023; **Jihado; Girsang**, 2024; **Farid *et al.***, 2021) through the voting mechanism, the study has achieved significant improvements in adequacy, F1 scores and recall as compared to utilizing these models individually. Therefore, through the combination of CNN and BiLSTM models, this study not only improved sentiment analysis performance but also offers a framework that can be adapted to other applications of NLP. It also paves the way for more accurate and reliable text analysis tools.

6. Research Implications

6.1. Theoretical Implications

The integration of NLP in document understanding and semantic analysis particularly from the Chinese perspective presents significant theoretical advancements. Theoretical implications of this integration involve the development of models that better understand the complexities of Chinese language. This research contributes to the body of knowledge through demonstrating the way NLP techniques such as ensemble learning can be effectively adapted to handle these unique features of linguistics. Findings also highlights that by combining CNN and BiLSTM the accuracy and depth of sentiment analysis can be enhanced.

6.2. Practical Implications

From the practical viewpoint, this study also provides effective practical implications particularly for industries and sectors that rely heavily on the document understanding and sentiment analysis. The demonstrated effectiveness of ensemble learning models in assessing Chinese texts can be implemented to distinct practical applications. It includes automating the processing of legal documents, improving the feedback analysis of customers and enhancing the systems of content recommendations. Businesses and government agencies in China can also leverage these insights for the development of more accurate and culturally relevant NLP systems that better understand the intent of users, contexts and sentiments.

7. Limitations and Future Research Indications

Despite of the significant research contributions, this study also involves a few limitations that suggest avenues for future research. One significant limitation is the focus on combining only CNN and BiLSTM models. Although, these techniques are effective but may not fully capture all the aspects of Chinese language. Future researchers can explore the inclusion of more diverse NLP models. It can involve transformer-based architectures so that semantic understanding can be enhanced further. Furthermore, the scope of this research can also be expanded to include multilingual datasets that can provide insights regarding the way these techniques can be adapted for other languages.

References

Abro, Abdul Ahad; Talpur, Mir Sajjad Hussain; Jumani, Awais Khan. (2023). "Natural language processing challenges and issues: A literature review". *Gazi University Journal of Science,* v. 36, n. 4, pp. 1522-1536. *https://doi.org/10.35378/gujs.1032517*

Alsayat, Ahmed. (2022). "Improving Sentiment Analysis for Social Media Applications Using an Ensemble Deep Learning Language Model". *Arabian Journal for Science and Engineering,* v. 47, n. 2, pp. 2499-2511. *https://doi.org/10.1007/s13369-021-06227-w*

Baclic, Oliver; Tunis, Matthew; Young, Kelsey; Doan, Coraline; Swerdfeger, Howard; Schonfeld, Justin. (2020). "Artificial intelligence in public health: Challenges and opportunities for public health made possible by advances in natural language processing". *Canada Communicable Disease Report,* v. 46, n. 6, pp. 161-168. *https://doi.org/10.14745/ccdr.v46i06a02*

Bahja, Mohammed. (2020). "Natural Language Processing Applications in Business." In: *E-Business-Higher Education and Intelligence Applications. https://doi.org/10.5772/intechopen.92203*

Baptista, João; Stein, Mari-Klara; Klein, Stefan; Watson-Manheim, Mary Beth; Lee, Jungwoo. (2020). "Digital Work and Organisational Transformation: Emergent Digital/Human Work Configurations in Modern Organisations". *The Journal of Strategic Information Systems,* v. 29, n. 2, pp. 101618. *https://doi.org/10.1016/j.jsis.2020.101618*

Berrar, Daniel; Konagaya, Akihiko; Schuster, Alfons. (2012). "Can machines make us think? In memory of Alan Turing (1912–1954)." In: *The 26th Annual Conference of the Japanese Society for Artificial Intelligence.* pp. 3P2-IOS-2b-2. The Japanese Society for Artificial Intelligence, Tokyo, Japan. *https://doi.org/10.11517/pjsai.JSAI2012.0_3P2IOS2b2*

Brahma, Siddhartha. (2018). "Improved Sentence Modeling using Suffix Bidirectional LSTM". *arXiv preprint arXiv:1805.07340*. *https://doi.org/10.48550/arXiv.1805.07340*

Chen, Jiansong; Pang, Hongjing. (2023). "Analyzing Factors Influencing Student Achievement: A Financial and Agricultural Perspective Using SPSS Statistical Analysis Software". *Journal of Commercial Biotechnology,* v. 28, n. 1, pp. 304-316. *https://doi.org/10.5912/jcb1118*

Chen, Xieling; Xie, Haoran; Tao, Xiaohui. (2022). "Vision, status, and research topics of Natural Language Processing". *Natural Language Processing Journal,* v. 1, pp. 100001. *https://doi.org/10.1016/j.nlp.2022.100001*

Chowdhary, K R. (2020). "Natural Language Processing." In: *Fundamentals of Artificial Intelligence.* pp. 603-649. Springer, New Delhi. *https://doi.org/10.1007/978-81-322-3972-7_19*

Church, Kenneth; Liberman, Mark. (2021). "The future of computational linguistics: On beyond alchemy". *Frontiers in Artificial Intelligence,* v. 4, pp. 625341. *https://doi.org/10.3389/frai.2021.625341*

Cui, Yiming; Che, Wanxiang; Liu, Ting; Qin, Bing; Wang, Shijin; Hu, Guoping. (2020). "Revisiting Pre-Trained Models for Chinese Natural Language Processing". *arXiv preprint arXiv:2004.13922*. *https://doi.org/10.48550/arXiv.2004.13922*

Dai, Shuying; Li, Keqin; Luo, Zhuolun; Zhao, Peng; Hong, Bo; Zhu, Armando; Liu, Jiabei. (2024). "AI-based NLP section discusses the application and effect of bag-of-words models and TF-IDF in NLP tasks". *Journal of Artificial Intelligence General science (JAIGS),* v. 5, n. 1, pp. 13-21. *https://doi.org/10.60087/jaigs.v5i1.149*

Estok, Simon C. (2022). "Anthropocene becomes the world: Indra Sinha's Animal's People, Rohinton Mistry's A Fine Balance, and Paulo Bacigalupi's The Windup Girl as world literature". *Cultura,* v. 19, n. 2, pp. 43-55. *https://doi.org/10.3726/CUL022022.0003*

Fanni, Salvatore Claudio; Febi, Maria; Aghakhanyan, Gayane; Neri, Emanuele. (2023). "Natural Language Processing." In: *Introduction to Artificial Intelligence.* Klontzas, M.E.; Fanni, S.C.; Neri, E. (Eds.), pp. 87-99. Springer. *https://doi.org/10.1007/978-3-031-25928-9_5*

Farid, Ahmed Bahaa; Fathy, Enas Mohamed; Eldin, Ahmed Sharaf; Abd-Elmegid, Laila A. (2021). "Software defect prediction using hybrid model (CBIL) of convolutional neural network (CNN) and bidirectional long short-term memory (Bi-LSTM)". *PeerJ Computer Science,* v. 7, pp. e739. *https://doi.org/10.7717/peerj-cs.739*

Feng, Zhiwei. (2023). "Past and Present of Natural Language Processing." In: *Formal Analysis for Natural Language Processing: A Handbook.* pp. 3-48. Springer. *https://doi.org/10.1007/978-981-16-5172-4_1*

Gawer, Annabelle. (2022). "Digital platforms and ecosystems: remarks on the dominant organizational forms of the digital age". *Innovation,* v. 24, n. 1, pp. 110-124. *https://doi.org/10.1080/14479338.2021.1965888*

Gonçalves, Bernardo. (2023). "Irony with a Point: Alan Turing and His Intelligent Machine Utopia". *Philosophy & Technology,* v. 36, n. 3, pp. 50. *https://doi.org/10.1007/s13347-023-00650-7*

Hu, Weixiong; Gu, Zhaoquan; Xie, Yushun; Wang, Le; Tang, Keke. (2019). "Chinese Text Classification Based on Neural Networks and Word2vec." In: *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC).* pp. 284-291. IEEE. *https://doi.org/10.1109/DSC.2019.00050*

Huq, Mohammad Rezwanul; Ahmad, Ali; Rahman, Anika. (2017). "Sentiment analysis on Twitter data using KNN and SVM". *International Journal of Advanced Computer Science and Applications,* v. 8, n. 6, pp. 19-25. *https://doi.org/10.14569/IJACSA.2017.080603*

Jang, Beakcheol; Kim, Myeonghwi; Harerimana, Gaspard; Kang, Sang-ug; Kim, Jong Wook. (2020). "Bi-LSTM Model to Increase Accuracy in Text Classification: Combining Word2vec CNN and Attention Mechanism". *Applied Sciences,* v. 10, n. 17, pp. 5841. *https://doi.org/10.3390/app10175841*

Jaradat, Shadi; Nayak, Richi; Paz, Alexander; Elhenawy, Mohammed. (2024). "Ensemble Learning with Pre-Trained Transformers for Crash Severity Classification: A Deep NLP Approach". *Algorithms,* v. 17, n. 7, pp. 284. *https://doi.org/10.3390/a17070284*

Jiang, Kai; Lu, Xi. (2020). "Natural language processing and its applications in machine translation: A diachronic review." In: *2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI).* pp. 210-214. IEEE. *https://doi.org/10.1109/IICSPI51290.2020.9332458*

Jihado, Anindra Ageng; Girsang, Abba Suganda. (2024). "Hybrid Deep Learning Network Intrusion Detection System Based on Convolutional Neural Network and Bidirectional Long Short-Term Memory". *Journal of Advances in Information Technology,* v. 15, n. 2, pp. 219-232. *https://www.jait.us/uploadfile/2024/JAIT-V15N2-219.pdf*

Johri, Prashant; Khatri, Sunil K; Al-Taani, Ahmad T; Sabharwal, Munish; Suvanov, Shakhzod; Kumar, Avneesh. (2021). "Natural Language Processing: History, Evolution, Application, and Future Work." In: *Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020.* Abraham, A.; Castillo, O.; Virmani, D. (Eds.), pp. 365-375. Springer. *https://doi.org/10.1007/978-981-15-9712-1_31*

Kamineni, Shasank; Tummala, Meghana; Kandimalla, Sai Yasheswini; Koduru, Tejodbhav; Manikandan, V M. (2024). "Advancements and challenges of using natural language processing in the healthcare sector." In: *Digital Transformation in Healthcare 5.0: Volume 2: Metaverse, Nanorobots and Machine Learning.* Malviya, Rishabha; Sundram, Sonali; Dhanaraj, Rajesh Kumar; Kadry, Seifedine (Eds.), pp. 317-342. De Gruyter. *https://doi.org/10.1515/9783111398549-013*

Kharrat, Asma; Drira, Fadoua; Lebourgeois, Franck; Kerautret, Bertrand. (2024). "Advancements and Challenges in Continual Learning for Natural Language Processing: Insights and Future Prospects." In: *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3.* pp. 1255-1262. SCITEPRESS – Science and Technology Publications. *https://doi.org/10.5220/0012462400003636*

Khurana, Diksha; Koli, Aditya; Khatter, Kiran; Singh, Sukhdev. (2023). "Natural Language Processing: State of the Art, Current Trends and Challenges". *Multimedia Tools and Applications,* v. 82, n. 3, pp. 3713-3744. *https://doi.org/10.1007/s11042-022-13428-4*

Koonce, Taneya Y; Giuse, Dario A; Williams, Annette M; Blasingame, Mallory N; Krump, Poppy A; Su, Jing; Giuse, Nunzia B. (2024). "Using a Natural Language Processing Approach to Support Rapid Knowledge Acquisition". *JMIR Medical Informatics,* v. 12, pp. e53516. *https://doi.org/10.2196/53516*

Koroteev, Mikhail V. (2021). "BERT: A Review of Applications in Natural Language Processing and Understanding". *arXiv preprint arXiv:2103.11943*. *https://doi.org/10.48550/arXiv.2103.11943*

Krasadakis, Panteleimon; Sakkopoulos, Evangelos; Verykios, Vassilios S. (2024). "A Survey on Challenges and Advances in Natural Language Processing with a Focus on Legal Informatics and Low-Resource Languages". *Electronics,* v. 13, n. 3, pp. 648. *https://doi.org/10.3390/electronics13030648*

Lauriola, Ivano; Lavelli, Alberto; Aiolli, Fabio. (2022). "An introduction to deep learning in natural language processing: Models, techniques, and tools". *Neurocomputing,* v. 470, pp. 443-456. *https://doi.org/10.1016/j.neucom.2021.05.103*

Liu, Meiyu; Li, Chengyou; Wang, Shuo; Li, Qinghai. (2023). "Digital transformation, risk-taking, and innovation: Evidence from data on listed enterprises in China". *Journal of Innovation & Knowledge,* v. 8, n. 1, pp. 100332. *https://doi.org/10.1016/j.jik.2023.100332*

Liu, Wei; Wang, Zhengbin; Shi, Qiwei; Bao, Siqintana. (2024). "Impact of the digital transformation of Chinese new energy vehicle enterprises on innovation performance". *Humanities and Social Sciences Communications,* v. 11, n. 1, pp. 592. *https://doi.org/10.1057/s41599-024-03109-y*

López-Úbeda, Pilar; Martín-Noguerol, Teodoro; Aneiros-Fernández, José; Luna, Antonio. (2022). "Natural Language Processing in Pathology: Current Trends and Future Insights". *The American Journal of Pathology,* v. 192, n. 11, pp. 1486-1495. *https://doi.org/10.1016/j.ajpath.2022.07.012*

Luo, Siyin; Gu, Youjian; Yao, Xingxing; Fan, Wei. (2021). "Research on Text Sentiment Analysis Based on Neural Network and Ensemble Learning". *Revue d'Intelligence Artificielle,* v. 35, n. 1, pp. 63-70. *https://doi.org/10.18280/ria.350107*

Méndez, Manuel; Merayo, Mercedes G; Núñez, Manuel. (2023). "Long-term traffic flow forecasting using a hybrid CNN-BiLSTM model". *Engineering Applications of Artificial Intelligence,* v. 121, pp. 106041. *https://doi.org/10.1016/j.engappai.2023.106041*

Moein, Mohammad Mohtasham; Saradar, Ashkan; Rahmati, Komeil; Mousavinejad, Seyed Hosein Ghasemzadeh; Bristow, James; Aramali, Vartenie; Karakouzian, Moses. (2023). "Predictive models for concrete properties using machine learning and deep learning approaches: A review". *Journal of Building Engineering,* v. 63, pp. 105444. *https://doi.org/10.1016/j.jobe.2022.105444*

Olivetti, Elsa A; Cole, Jacqueline M; Kim, Edward; Kononova, Olga; Ceder, Gerbrand; Han, Thomas Yong-Jin; Hiszpanski, Anna M. (2020). "Data-driven materials research enabled by natural language processing and information extraction". *Applied Physics Reviews,* v. 7, n. 4, pp. 041317. *https://doi.org/10.1063/5.0021106*

Omar, Marwan; Choi, Soohyeon; Nyang, DaeHun; Mohaisen, David. (2022). "Robust Natural Language Processing: Recent Advances, Challenges, and Future Directions". *IEEE Access,* v. 10, pp. 86038-86056. *https://doi.org/10.1109/ACCESS.2022.3197769*

Qabasiyu, Muhammad Garzali; Zayyad, Musa Ahmed; Abdullahi, Shamsu. (2023). "An Ensembled Based Machine Learning Technique of Sentiment Analysis". *Journal of Telecommunication, Electronic and Computer Engineering (JTEC),* v. 15, n. 1, pp. 23-28. *https://doi.org/10.54554/jtec.2023.15.01.004*

Rhanoui, Maryem; Mikram, Mounia; Yousfi, Siham; Barzali, Soukaina. (2019). "A CNN-BiLSTM Model for Document-Level Sentiment Analysis". *Machine Learning and Knowledge Extraction,* v. 1, n. 3, pp. 832-847. *https://doi.org/10.3390/make1030048*

Saha, Dipanjan; Brooker, Phillip; Mair, Michael; Reeves, Stuart. (2023). "Thinking Like a Machine: Alan Turing, Computation and the Praxeological Foundations of AI". *Science & Technology Studies,* v. 37, n. 2, pp. 66-88. *https://doi.org/10.23987/sts.122892*

Saleena, Nabizath. (2018). "An Ensemble Classification System for Twitter Sentiment Analysis". *Procedia Computer Science,* v. 132, pp. 937-946. *https://doi.org/10.1016/j.procs.2018.05.109*

Shaik, Thanveer; Tao, Xiaohui; Li, Yan; Dann, Christopher; McDonald, Jacquie; Redmond, Petrea; Galligan, Linda. (2022). "A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis". *Ieee Access,* v. 10, pp. 56720-56739. *https://doi.org/10.1109/ACCESS.2022.3177752*

Shakhnoza, Otajonova; Nargiza, Inogamova. (2024). "Natural Language Processing (NLP) Applications in Language Teaching". *«contemporary Technologies of Computational Linguistics»,* v. 2, n. 22.04, pp. 278-283. *https://www.myscience.uz/index.php/linguistics/article/view/65*

Shao, Yunfan; Geng, Zhichao; Liu, Yitao; Dai, Junqi; Yan, Hang; Yang, Fei; Li, Zhe; Bao, Hujun; Qiu, Xipeng. (2024). "Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation". *Science China Information Sciences,* v. 67, n. 5, pp. 152102. *https://doi.org/10.1007/s11432-021-3536-5*

Shen, Zepeng; Pan, Yiming; Tan, Kai; Ji, Huan; Chu, SH. (2024). "Educational Innovation in the Digital Age: the Role and Impact of NLP Technology." In: *Old and New Technologies of Learning Development in Modern Conditions.* pp. 281-291. International Science Group. *https://doi.org/10.46299/ISG.2024.1.6*

Stenmark, Mikael. (2022). "Secular Worldviews: Scientism and Secular Humanism". *European Journal for Philosophy of Religion,* v. 14, n. 4, pp. 237-264. *https://doi.org/10.24204/ejpr.2022.3640*

Sun, Xinjie; Liu, Zhifang; Li, Hui; Ying, Feng; Tao, Yu. (2024). "Chinese text dual attention network for aspect-level sentiment classification". *PloS One,* v. 19, n. 3, pp. e0295331. *https://doi.org/10.1371/journal.pone.0295331*

Tan, Kian Long; Lee, Chin Poo; Lim, Kian Ming; Anbananthen, Kalaiarasi Sonai Muthu. (2022). "Sentiment Analysis With Ensemble Hybrid Deep Learning Model". *IEEE Access,* v. 10, pp. 103694-103704. *https://doi.org/10.1109/ACCESS.2022.3210182*

Tunca, Sezai; Sezen, Bulent; Wilk, Violetta. (2023). "An exploratory content and sentiment analysis of the guardian metaverse articles using leximancer and natural language processing". *Journal of Big Data,* v. 10, n. 1, pp. 82. *https://doi.org/10.1186/s40537-023-00773-w*

Turing, Alan M. (1980). "Computing Machinery and Intelligence". *Creative Computing,* v. 6, n. 1, pp. 44-53. *https://eric.ed.gov/?id=EJ216711*

Tyagi, Nemika; Bhushan, Bharat. (2023). "Demystifying the Role of Natural Language Processing (NLP) in Smart City Applications: Background, Motivation, Recent Advances, and Future Research Directions". *Wireless Personal Communications,* v. 130, n. 2, pp. 857-908. *https://doi.org/10.1007/s11277-023-10312-8*

Widdows, Dominic; Alexander, Aaranya; Zhu, Daiwei; Zimmerman, Chase; Majumder, Arunava. (2024). "Near-term advances in quantum natural language processing". *Annals of Mathematics and Artificial Intelligence,* pp. 1-24. *https://doi.org/10.1007/s10472-024-09940-y*

Wu, Xiaohan; Wu, Zejun; Feng, Yuqi. (2021). "A Text Category Detection and Information Extraction Algorithm with Deep Learning". *Journal of Physics: Conference Series,* v. 1982, n. 1, pp. 012047. *https://doi.org/10.1088/1742-6596/1982/1/012047*

Zhang, Hong. (2023). "Analysis of College Basketball Injuries: Implications for Healthcare and Patient Well-being". *Journal of Commercial Biotechnology,* v. 28, n. 2. *https://doi.org/10.5912/jcb1363*

Zhang, Hongzhi; Shafiq, M Omair. (2024). "Survey of transformers and towards ensemble learning using transformers for natural language processing". *Journal of Big Data,* v. 11, n. 1, pp. 25. *https://doi.org/10.1186/s40537-023-00842-0*