DSI DATA SCIENCE AND INFORMETRICS
VOL.3, NO.3, AUG. 2023

RESEARCH ARTICLE

# Identification of ChatGPT answers and physician answers in the online medical community

Shengli Deng, Haowei Wang*

School of Information Management, Wuhan University, Wuhan, China

**ABSTRACT**

ChatGPT changes the way of knowledge production and information space structure of human society. In the healthcare industry, ChatGPT's powerful question-and-answer capability will drive its application in automated question answering in online healthcare communities. However, because ChatGPT answers are limited by factors such as the quality of data sets, their authority and accuracy cannot be guaranteed, and they are prone to misdiagnosis and damage to life and health. Therefore, the identification of ChatGPT answers in online medical communities with physician answers is crucial. In this paper, we collected medical question-answering data generated by the Haodafu platform and ChatGPT, respectively, constructed feature vectors from semantic features, syntactic features, and the fusion of both, and combined different feature vectors with XGBoost models to construct BERT-XGBoost, POS-XGBoost and Merge- XGBoost models for identifying ChatGPT answers and physician answers in online medical communities. The three models achieved accuracy rates of 0.960, 0.968, and 0.986, respectively. The difference in performance between the three models reflects the degrees of variation in different features of ChatGPT answers versus physician answers. The results indicate that the differences between ChatGPT and physicians in syntactic features, i.e., linguistic expression habits, are greater than their differences in semantic features, i.e., specific content suggestions.

**KEYWORDS**

ChatGPT; Online medical community; Text classification

## 1 Introduction

Since OpenAI announced ChatGPT, it has attracted much attention by demonstrating its powerful linguistic abilities in chatting, quizzing, writing, etc. ChatGPT has been initially applied in many language-related fields, such as scientific writing (Curtis & ChatG PT, 2023), novel creation (Thorp, 2023), online quizzing (Budler et al., 2003), and abstract summarization (Else, 2023). In the medical field, some scholars have studied the performance of ChatGPT on medical qualification tests (Gilson et al., 2023), and the results showed that the performance of ChatGPT is close to that of professional medical students. Some scholars have studied the application of ChatGPT in medical consultations, such as nurse training (Scerri & Morin, 2023) and clinical consultations (Rao et al., 2023; Sabry Abdel-Messih & Kamel Boulos, 2023), with good results.

* Corresponding Author: 2018301040137@whu.edu.cn

However, the powerful capability of ChatGPT also brings some risks, as ChatGPT's answers are affected by the quality of the dataset, so its authority and accuracy cannot be guaranteed, for example, one of the dangers posed by ChatGPT is its tendency to be used as a "weapon of mass deception" (WMD) and an enabler of criminal activities involving deception (Sison et al., 2023). The lack of authority and accuracy of ChatGPT answers in online medical question answering can increase patients' distrust of online medical question answering. In addition, ChatGPT answers may produce misdiagnosis, and thus affect the life and health of users. Therefore, identifying whether the answers in online medical communities come from ChatGPT or physicians can help reduce the potential harm when ChatGPT is applied in online medical communities.

The identification of ChatGPT answers versus physician answers is essentially a text classification problem. However, unlike judging the quality of answers in the past (Qiu et al., 2022), ChatGPT and physician answers come from different information sources, and their information sources change from human-human to human-AI, and the difference between the two does not only lie in the good or bad quality of the answers but there may also be differences in their expressions in the answers. Therefore, this paper uses semantic features, syntactic features, and fusion features to build recognition models, and compares the performance differences of different models to analyze the degree of difference between ChatGPT answers and physician answers in different features.

## 2 Related work

Research related to question answering in online medical communities is divided into two main directions: answer quality analysis and automated question answering. Answer quality analysis focuses on predicting the quality of physicians' answers (Qiu et al., 2022) and recommending answers in the community (Lin et al., 2021), and most studies use machine learning and deep learning methods to evaluate the answer quality or further design recommendation algorithms to promote the overall answer quality in the medical community. Research in automated question answering, on the other hand, has focused on methodological improvements, such as the use of knowledge graphs, neural networks, and automatic reasoning to improve the quality and accuracy of automated question answering.

ChatGPT-related research has focused on four main areas: application, impact, improvement, and answer. About applications, scholars have explored human collaboration with ChatGPT (Bockting et al., 2023), research on knowledge production based on ChatGPT (Dwivedi et al., 2023), and value assessment based on ChatGPT (Alshater, 2022), as well as the prospects of its applications in healthcare (Thorp, 2023), finance (Balakrishnan et al., 2022; Northey et al., 2022), education (Haque et al., 2022), and scientific research (Dwivedi et al., 2023). About impact, scholars analyze the ethical, legal, employment, technological dependence, and interpretability issues brought by ChatGPT (Dwivedi et al., 2023) and explore the positive impacts such as productivity improvements (Kshetri, 2023). Regarding improvement, studies have focused on the potential problems of ChatGPT, examining the improvement of real-time ChatGPT answer results, the diversity and innovation of generated content, the interpretability of the answer process, and the moral and ethical regulation of it (Kshetri, 2023). In answer, scholars have focused on the quality assessment of ChatGPT-generated content to reduce the problem of disinformation and fraud brought by ChatGPT (Dwivedi et al., 2023). Some scholars have also tried to identify ChatGPT-generated content and UGC

(user-generated content) using deep learning methods such as Roberta (Guo et al., 2023).

ChatGPT is a neural network model, but unlike previous neural network models in automated medical question answering, it has the advantages of a larger-scale corpus, a larger parameter scale of billions, and the incorporation of human preference mechanisms, and has made certain breakthroughs (Rao et al., 2023; Sabry Abdel-Messih & Kamel Boulos, 2023), and may be widely used in online medical community question-answering in the future. However, for the potential problems of ChatGPT application, the current research mainly focuses on the adverse effects of ChatGPT, and there are no effective measures that can solve the problems. Meanwhile, due to the lack of authority over ChatGPT-generated content, its application in automated question answering in online medical communities may generate life and health hazards, but the current research on similar automated question answering in online medical communities is mostly limited to algorithm improvement, without in-depth study of its possible risks and hazards. To solve the potential problems brought by automated question answering in online medical communities in the context of artificial intelligence-generated content (AIGC), as well as to maintain the community order and create a good ecology, the identification of ChatGPT answers and physicians' answers in online medical communities is crucial. However, in the current research on ChatGPT recognition, only simple applications of methods such as deep learning have been attempted, without considering the effects of different features on recognition effects or thinking about solutions from the perspective of the differences in behavioral features between ChatGPT and humans. Therefore, this study attempts to apply syntactic and semantic features to the recognition of ChatGPT answers and physicians' answers in online medical communities. This study can be used to suggest risk information in online medical communities, thus reducing the harm caused by misdiagnosis when ChatGPT is applied in online medical communities. On the other hand, this study uses semantic features and syntactic features to build recognition models to help understand the degree of difference between ChatGPT answers and physicians' answers about different features.

# 3 Method

## 3.1 Recognition methods of ChatGPT–generated content and user–generated content

In the recognition methods of ChatGPT-generated content and user-generated content (UGC), simple text classification is currently performed mainly using neural network models without considering the feature differences between the two. The differences between ChatGPT-generated content and user-generated content are reflected both in syntactic features (lexicality, syntax) and in deep semantic features (Guo et al., 2023). Past studies have shown the success of different feature fusion approaches in several tasks, such as question-answer matching (Zhang et al., 2017), and text classification (Xu, 2023). Therefore, in this study, we construct recognition models by syntactic features and semantic features and analyze the feature differences between ChatGPT and humans in the context of online medical questioning by model performance.

The recognition of ChatGPT answers and physician answers is essentially a classification of text. In the field of text classification, integrated algorithms tend to outperform single learning methods by building multiple learners to accomplish the task. The boosting algorithm is commonly used as an effective integration method and belongs to the iterative algorithm. It

serially constructs a stronger learner by continuously using a weak learner to make up for the "deficiencies" of the previous weak learner, and this strong learner can make the objective function value small enough. A representative one is the gradient boosting decision tree (GBDT) (Friedman, 2001), which is an additive model based on the idea of boosting integration, where a forward distribution algorithm is used for greedy learning during training, and each iteration learns a CART tree to fit the prediction of the previous t-1 trees with the residuals of the true values of the training samples.

However, the GBDT model has the disadvantages of strong dependence, difficult parallelism, and low efficiency. The extreme gradient boosting (XGBoost) (Chen et al., 2016) is an improvement of GBDT to address these disadvantages by optimizing its loss function and using the second-order Taylor formula expansion to improve the computational accuracy. At the same time, XGBoost uses regular terms to simplify the model, which improves the training efficiency on the one hand and avoids the overfitting phenomenon on the other hand. To address the drawback that GBDT cannot operate in parallel, XGBoost adopts a Blocks storage structure to realize parallel operation.

Therefore, this study uses the XGBoost model, combined with semantic features and syntactic features, to identify ChatGPT answers and physician answers in online medical communities.

### 3.2 Semantic feature extraction and sentence vector generation methods

In natural language processing, the generation of sentence vectors has been an important research direction in pre-training. Vector semantics is an important issue in sentence vector generation, involving various aspects such as vector similarity, relevance, semantic frames, semantic roles, and implied meaning. Indicators such as sentence vector similarity in natural language processing reflect the semantic features of sentences.

Currently, there are two main ways to generate sentence vectors. One is to transform the generation of sentence vectors into the generation of word vectors by summing, meaning, or weighted meaning of the individual word vectors in a sentence to obtain the sentence vectors. The other one is to obtain the sentence vectors directly. Since the BERT model was proposed, it has been widely used for sentence vector generation. The BERT pre-training model (Devlin et al., 2019) contains two tasks: one is the masked language model (MLM), in which 15% of the word elements are randomly selected for masking to form a prediction task, and the model is trained by predicting the masked word elements; the other task is next sentence prediction (NSP), in which BERT incorporates a binary classification model into the pre-training, by extracting the original continuous sentence pairs from the corpus as the "true" corpus and connecting the originally discontinuous sentences to generate the "false" corpus, and training the model by making it predict whether the next sentence is continuous or not. BERT generates a fixed-length vector for each word in a sentence and then transforms multiple word vectors in a sentence into sentence vectors through pooling operations such as cumulative pooling, average pooling, maximum pooling, and concatenation. In this study, we adopt the method of average pooling of word vectors to obtain sentence vectors, which can prevent the sentence vectors from being too long to affect the training effect on the one hand, and average pooling can also prevent the sentence vectors from being affected by the number of word elements.

RoBERTa-wwm-ext (Cui et al., 2021) is a BERT model trained using a training method that combines the advantages of RoBERTa and BERT-www. In the pre-training phase, the model uses the whole word masking (WWM) strategy for masking. The roBERTa-wwm-ext model has made significant progress in many downstream tasks, such as simple Chinese reading comprehension, natural language inference, sentiment analysis, sentence pair classification, etc. The model's outstanding ability in semantic representation has been demonstrated in downstream tasks. The outstanding ability of the model in semantic representation is reflected in the downstream tasks. Therefore, in this study, we use RoBERTa-wwm-ext as a pre-trained model to generate sentence vectors to construct semantic features.

### 3.3 Syntax feature extraction and Part–Of–Speech tagging methods

Syntactic structure refers to the pattern of linguistic units such as sentences, phrases, and words about their syntactic structure and syntactic meaning. The features of lexical distribution in a sentence are an important part of the syntactic features of a sentence. In medical question-answering scenarios, lexical distribution is one of the important perspectives for studying the differences in syntactic features between ChatGPT answers and physicians' answers. The frequency distributions of different lexical properties in sentences are important features for identifying ChatGPT answers and physicians' answers in online medical communities.

In natural language processing, lexical annotation algorithms can be divided into dictionary-finding algorithms based on string matching, statistical-based algorithms, and algorithms based on neural network models Fasthan (Geng et al., 2021) is a BERT-based neural network model trained on 13 corpora by supervised learning, which can be used for tasks such as Chinese word separation, lexical annotation, and dependency analysis, and with fewer models, good results were achieved with fewer total parameters of the model. In this study, a base version of the Fasthan model is used for the lexical annotation of sentences. For the format of lexical annotation, the output of the Fasthan model defaults to the Chinese Treebank 9.0 (CTB9) tagging format (Xue et al., 2016), which is a tree-structured annotation format with annotation data sources from news, radio, and conversation, microblogs, etc. Its hierarchical annotation format can more clearly represent the relationship between individual lexical properties.

Therefore, this study uses the Fasthan model to classify sentences and annotate each word according to the CTB9 tagging format and then constructs Syntactic features based on the frequency of each lexical property in each sentence.

## 4 Experiment

The flow of this experiment is shown in Figure 1. Firstly, we collected medical question-answering data from the Haodafu platform and ChatGPT to construct the dataset. Then, we used the RoBERTa-wwm-ext model to extract semantic features and constructed syntactic features based on the lexical distribution of sentences, as well as fused these two features to construct three models: BERT-XGBoost, POS-XGBoost, and Multi-XGBoost. The experimental results show that the Multi-XGBoost model with multi-features perform the best compared with the other two models and other machine learning methods, with the highest accuracy and AUC values. Meanwhile, the accuracy of the POS-XGBoost model based on syntactic features is higher than that of the BERT-XGBoost model based on semantic features.

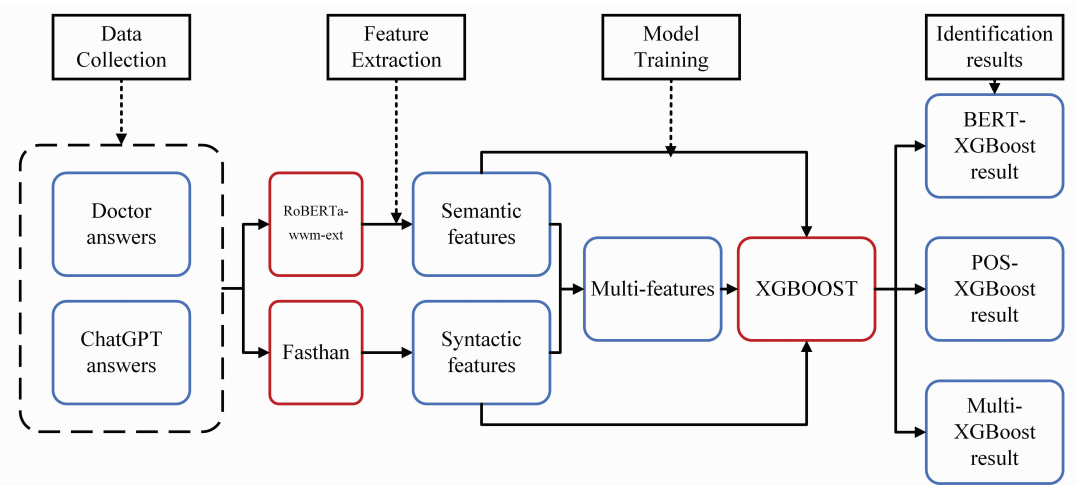


**Figure 1** Experimental flowchart

## 4.1 Data collection

This experiment collected 8820 classical question-and-answer data and corresponding expert advice from the Haodafu platform. 208,000 doctors are registered on the platform under their real names and use it personally to provide medical consultation, appointment booking, disease management, scientific knowledge, and other services directly to patients. Among these 208,000 active doctors, the proportion of doctors with the title of attending physician and above accounts for 70%, and they can give sufficient authoritative treatment advice to patients.

To collect the medical question answering data from ChatGPT, we entered the questions from the previously collected 8820 online medical community question answering datasets into ChatGPT's API interface "get-3.5-turbo" to generate bulk answers and collect the answer data from ChatGPT. It should be noted that ChatGPT may indicate that it is an artificial intelligence and cannot give medical advice when answering some medical-related questions. In the dataset of this study, 3293 ChatGPT samples prompted that they could not answer medical consultation questions, and 5364 ChatGPT answers that could be used for the experiments were finally retained after eliminating this part of the samples. As shown in Table 1, a total of 14,184 experimental data were collected, including 8,820 questions from patients, 8,820 answers from physicians, and 5,364 answers from ChatGPT. After that, it was divided into the train, test, and validation set according to the ratio of 8:1:1.

**Table 1** Data collection table

| | Advice | No advice | Total |
|---|---|---|---|
| Physician | 8820 | 0 | 8820 |
| ChatGPT | 5364 | 3293 | 8657 |
| Total | 14184 | 3293 | 16477 |

## 4.2 Feature Extraction and Model Training

### 4.2.1 Semantic features based on RoBERTa–wwm–ext with XGBoost: BERT–XGBoost

The semantic features of a sentence represent the intrinsic meaning of the sentence, and sentences with similar semantic features are closer in vector space distribution. Identifying ChatGPT and physicians' answers in the context of medical consultation by semantic features can help us study the differences between ChatGPT and physicians in the specific content of consultation answers.

First, we perform data preprocessing on the answer parts in the dataset to remove punctuation, special characters, and numbers from the sentences, intercept the excessively long parts, and use the processed text as the input for the RoBERTa-wwm-ext pre-training model.

Second, as shown in Figure 2, after inputting a sentence into the RoBERTa-wwm-ext pre-training model, we extract the last four Transformer (Trm) hidden layers in the model, each of which has 768 dimensions, and the vector formed by averaging the pooling operation of the four hidden layers is used as the word vector of each word element, thus forming a matrix of n*768 (n is the number of lexical elements in the sentence). Then, we access another averaging pooling layer to average the matrix along the row direction, which finally generates a 768-dimensional sentence vector for each sentence.

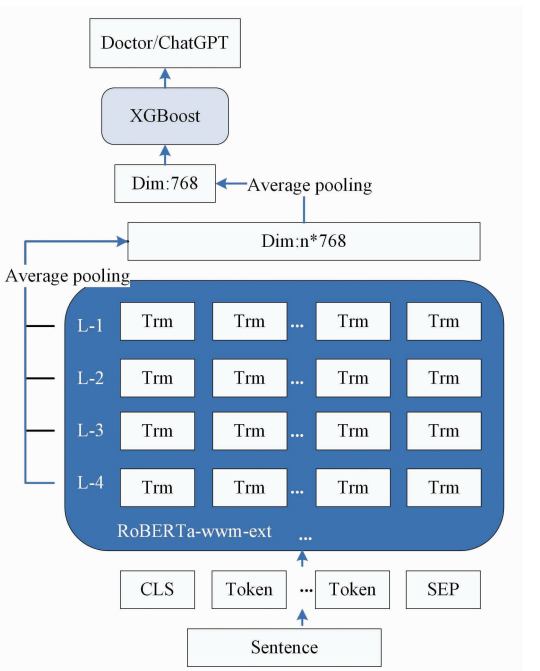


**Figure 2** BERT-XGBoost model

Finally, after obtaining the sentence vectors and the corresponding labels of the sentences, we use the XGBoost model for training and convert the output probability distribution of the XGBoost model into prediction results. The parameters of the XGBoost model are shown in Table 2.

**Table 2** XGBoost parameter setting

| parameter | BERT–XGBoost | POS–XGBoost | Multi–XGBoost |
|---|---|---|---|
| max_depth | 3 | 3 | 3 |
| learning rate | 0.1 | 0.1 | 0.1 |
| n_estimators | 200 | 200 | 200 |
| objective | ´binary: logistic´ | ´binary: logistic´ | ´binary: logistic´ |
| booster | ´gbtree´ | ´gbtree´ | ´gbtree´ |
| gamma | 0 | 0 | 0 |
| min_child_weight | 1 | 1 | 1 |
| max_delta_step | 0 | 0 | 0 |
| subsample | 1 | 1 | 1 |
| colsample_bytree | 1 | 1 | 1 |
| reg_alpha | 0 | 0 | 0 |
| reg_lambda | 1 | 1 | 1 |

#### 4.2.2 Syntactic features based on sentence lexical distribution with XGBoost: POS–XGboost

The lexical distribution of sentences in online medical consultation can reflect the Syntactic features and expression features of sentences. Therefore, by incorporating the differences in lexical distribution into the recognition model, we can discover the similarities and differences in the linguistic expression habits between ChatGPT and physicians.

As shown in Figure 3, first, we input the pre-processed answers into the Fasthan model for lexical annotation (POS) to generate word separation results with lexical labels. We used the CTB (Chinese Treebank 9.0) standard for lexical annotation.

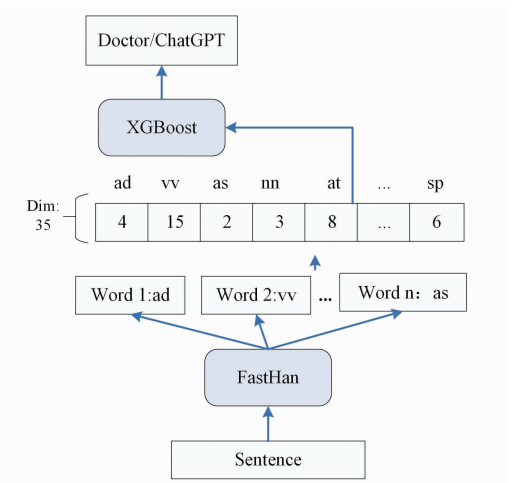


**Figure 3** POS-XGBoost model

Then, we counted the corresponding number of a total of 35 lexical properties in each sentence by summation to form a 35-dimensional lexical distribution vector. The number of lexical properties in each sentence may be different due to factors such as sentence length, so we normalize the lexical distribution vector to avoid the effect of sentence length on model training.

Finally, we also use the XGBoost model for training, and the output probability distribution

is transformed into prediction results to construct POS-XGBoost. The specific parameter settings are shown in Table 2.

#### 4.2.3 Fusion of semantic and syntactic features: Multi–XGBoost

By fusing semantic features and syntactic features of a sentence, we can more comprehensively identify the answers of ChatGPT in online medical communities with those of physicians, and focus on the differences in their contents and expression habits.

As shown in Figure 4, first, we connect the semantic feature vector extracted by RoBERTa-wwm-ext and the syntactic feature vector based on lexical distribution extracted by Fasthan to form an 803-dimensional hybrid feature vector.

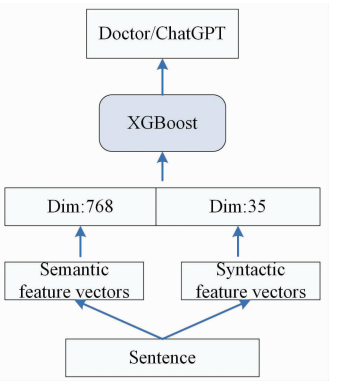


**Figure 4** Multi-XGBoost model

Then, we normalize them and input them into the XGBoost model for training and prediction.

Finally, we transform the probability distribution of the model output into prediction results. The specific XGBoost model parameters are shown in Table 2.

### 4.3 Experimental results

We used several metrics commonly used in the evaluation of dichotomous model results: accuracy, precision, recall, F1 value, and the area under the curve (AUC) value, and plotted the receiver operating characteristic (ROC) curves for the K Nearest Neighbors (KNN), Gaussian Naive Bayes (GaussianNB), and Decision Tree models under this task (Figure 5).

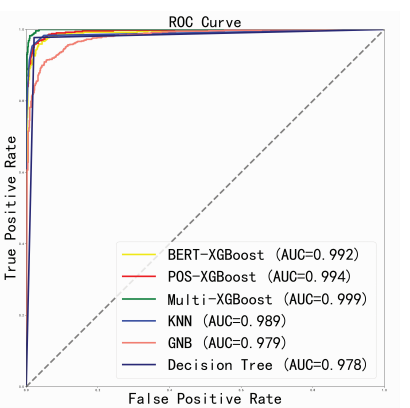


**Figure 5** ROC curves

Among the three models in this study and the comparison with other machine learning models, Multi-XGBoost based on multi-features has the highest AUC value (Figure 5) and the best result, indicating that both semantic and syntactic features play an important role in recognizing ChatGPT answers and physician answers in online medical communities and fusing the two can effectively improve the recognition effect of the model.

**Table 3** Experimental results

| Model | label | accuracy | AUC | precision | recall | F1 |
|---|---|---|---|---|---|---|
| BERT–XGBoost | physician | .959 | .992 | .963 | .973 | .968 |
| | ChatGPT | | | .952 | .934 | .943 |
| POS–XGBoost | physician | .968 | .994 | .971 | .978 | .975 |
| | ChatGPT | | | .961 | .950 | .955 |
| Multi–XGBoost | physician | **.986** | **.999** | .988 | .990 | .989 |
| | ChatGPT | | | .982 | .979 | .981 |
| KNN | physician | .954 | .989 | .926 | .991 | .958 |
| | ChatGPT | | | .990 | .914 | .950 |
| GaussianNB | physician | .909 | .979 | .963 | .858 | .908 |
| | ChatGPT | | | .862 | .964 | .910 |
| Decision Tree | physician | .976 | .978 | .978 | .976 | .977 |
| | ChatGPT | | | .974 | .976 | .975 |

## 4.4 Discussion

The experimental results (Table 3) show that most of the models are better at recognizing physicians' answers than at recognizing ChatGPT answers, whether based on semantic features, syntactic features, or multi-features, indicating that the models are more likely to recognize ChatGPT answers as physicians' answers, and it reflects the ability of ChatGPT to rival physicians' answers in medical questioning.

In similar text classification in the previous research, word vectors generated by pre-trained models similar to BERT as features outperformed feature vectors constructed in other ways. However, when identifying ChatGPT question answering with physicians, the model based on RoBERTa-ww-ext's 768-dimensional semantic feature vector instead performed lower than the 35-dimensional syntactic feature vector constructed based on lexical distribution. There are two possible reasons for this result: first, ChatGPT is similar to the physician in content, and ChatGPT can give similar measure suggestions as the physician when answering the patient's question. Therefore, it is more difficult to determine whether it is ChatGPT or the answer given by the physician by semantic features. On the contrary, there are large differences in language expressions between ChatGPT and physicians, as physicians may tend to use colloquial expressions while ChatGPT's expressions are more formal. This leads to a greater difference in their syntactic features, such as word distribution, and thus the word distribution enables a better distinction between ChatGPT and physicians. Another possible reason is that two averaging pooling operations are used in extracting sentence vectors using RoBERTa-wwm-ext, which can lose some features in the sentences and thus affect the recognition ability of the model.

# 5 Implication

## 5.1 Theoretical implication

With the application of AI in real-life and network environments, the recognition of AIGC will become an important part of social governance and network governance. In this study, the Merge-XGBoost model, which fuses semantic and syntactic features, achieves good results in recognizing ChatGPT with doctor's answers, which provides a research idea based on multi-feature fusion for AIGC recognition in the context of other kinds of tasks.

In this study, two classification models are constructed based on different features from the syntactic and semantic features of linguistics, and by analyzing the performance difference between the two on the task of recognizing ChatGPT and the doctor's answer, we study the difference between ChatGPT and human in specific content and language expression and provide a research idea to analyze the feature difference between AI and human based on the performance difference of the models, which provides a reference to the research on machine behavior and features.

## 5.2 Practical implication

ChatGPT's strong ability in the Q&A field brings new opportunities for online medical communities. However, with the application of ChatGPT in medical Q&A, patients may suspect that the Q&A suggestions come from robots rather than professional doctors, leading to an increase in patient distrust. The recognition model of ChatGPT with doctors' responses in online medical communities proposed in this study can be used to recognize and alert ChatGPT responses in online medical communities to reduce patients' uncertainty in online medical consultations.

Meanwhile, the medical field is directly related to human life and health, but ChatGPT responses are not completely accurate. Recognizing ChatGPT answers and reminding them during medical consultations can reduce the occurrence of medical accidents due to misdiagnosis and wrong diagnosis, avoid the risk of legal liability due to ChatGPT-generated Q&A, and contribute to the maintenance of the order of online medical communities in the context of the AIGC era.

In the past, studies of physician contribution in online medical communities were mainly comparisons between physicians, but the addition of ChatGPT may interfere with the evaluation of physician contribution. Therefore, identifying ChatGPT responses versus physician responses in online medical communities can help provide accuracy and credibility in the evaluation of physician contribution.

The identification of ChatGPT versus physician responses in online medical communities assists in solving potential problems in online medical communities in the AIGC era, which can help maintain community order and create a favorable community ecosystem.

# 6 Conclusion and future work

In this study, models are constructed by extracting semantic features, syntactic features, and fusion features from the perspective of syntactic features and linguistic features, respectively, for the recognition of ChatGPT answers and physicians' answers in online medical communities. The performance of each model is also compared to analyze the

degree of feature differences between ChatGPT and physicians. The experiments reflect that the lexical distribution of ChatGPT answers and physician answers may differ significantly in comparison with the semantic features, i.e., the specific content and that the fusion of syntactic features with semantic features can improve the recognition accuracy.

However, for model selection, we used a generic classification model from natural language processing, with relatively little optimization in the design of the model structure. In addition, the differences between ChatGPT answers and physicians' answers in online medical communities may not only be reflected in semantic and syntactic features, but also emotional and logical structure features. Therefore, we can try to incorporate more features to further improve the model performance. Meanwhile, UGC from different communities and different fields can be compared with AIGC in the future to extend the applicability of the model by enriching the dataset and to find more general conclusions on the differences between AI and human behavioral features.

## Authors details

Shengli Deng (ORCID:0000-0001-7489-4439), Ph.D.; Haowei Wang (ORCID:0000-0002-5085-6574), M.S.

## Acknowledgment

This research is supported in part by National Natural Science Foundation, PR China (Grant No. 72374158 )

## Declarations of interest

There are no financial conflicts of interest to disclose.

## Data available on request from the authors

The data that support the findings of this study are available from the corresponding author, [Wang H.W.], upon reasonable request.

## References

Alshater, M. M. (2022) . Exploring the role of artificial intelligence in enhancing academic performance: A case study of ChatGPT. *SSRN Electronic Journal*.

Balakrishnan, J., Abed, S. S., & Jones, P. (2022) . The role of meta–UTAUT factors, perceived anthropomorphism, perceived intelligence, and social self–efficacy in chatbot–based services?. *Technological Forecasting and Social Change, 180,* 121692. https://doi.org/10.1016/j.techfore.2022.121692

Bockting, C. L., van Dis, E. A. M., Bollen, J., van Rooij, R., & Zuidema, W. (2023) . ChatGPT: Five priorities for research. *Nature, 614* (7947) , 224–226. https://doi.org/10.1038/d41586–023–00288–7

Budler, L. C., Gosak, L., & Stiglic, G. (2023). Review of artificial intelligence–based question–answering systems in healthcare. *Wiley Interdisciplinary Reviews–Data Mining and Knowledge Discovery, 13* (2), e1487. https://doi.org/10.1002/widm.1487

Chen, T. Q., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In*Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining (Kdd´16)* . ACM Digital Library.

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021) . Pre–training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 3504–3514. https://doi.

org/10.1109/TASLP.2021.3124365

Curtis, N., & ChatGPT. (2023) . To ChatGPT or not to ChatGPT? The impact of artificial intelligence on academic publishing. *The Pediatric Infectious Disease Journal, 42* (4) , 275–275. https://doi.org/10.1097/inf.0000000000003852

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019) . BERT: Pre–training of deep bidirectional transformers for language understanding. In *Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human language Technologies (naacl hlt 2019) , vol. 1*, 4171–4186. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/N19–1423

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al–Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., . . . Wright, R. (2023) . "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 102642. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2023.102642

Else, H. (2023) . Abstracts written by ChatGPT fool scientists. *Nature, 613* (7944) , 423–423. <Go to ISI>://WOS:000928175300013

Friedman, J. H. (2001) . Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29* (5) , 1189–1232, 1144. https://doi.org/10.1214/aos/1013203451

Geng, Z. C., Yan, H., Qiu, X. P., & Huang, X. J. (2021) . fastHan: A BERT–based Multi–Task Toolkit for Chinese NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, 99– 106. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2021.acl–demo.12

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023) . How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR medical education, 9*, e45312 –e45312. https://doi.org/10.2196/45312

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023) . How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv preprint arXiv: 2301.07597*. http://export.arxiv.org/abs/2301.07597v1

Haque, M. U., Dharmadasa, I., Sworna, Z. T., Rajapakse, R. N., & Ahmad, H. (2022) . " I think this is the most disruptive technology" : Exploring sentiments of ChatGPT early adopters using Twitter data. *arXiv preprint arXiv:2212.05856*.

Kshetri, N. (2023) . Regulatory technology and supervisory technology: Current status, facilitators, and barriers. *Computer, 56* (1) , 64–75. https://doi.org/10.1109/MC.2022.3205780

Lin, C. Y., Wu, Y. H., & Chen, A. L. P. (2021) . Selecting the most helpful answers in online health question answering communities. *Journal of Intelligent Information Systems, 57* ( 2) , 271 –293. https://doi.org/10.1007/s10844–021–00640–1

Northey, G., Hunter, V., Mulcahy, R., Choong, K., & Mehmet, M. (2022) . Man vs machine: How artificial intelligence in banking influences consumer belief in financial advice. *International Journal of Bank Marketing, 40* (6) , 1182–1199. https://doi.org/10.1108/IJBM–09–2021–0439

Qiu, Y., Ding, S., Tian, D., Zhang, C. Y., & Zhou, D. (2022) . Predicting the quality of answers with less bias in online health question answering communities . *Information Processing & Management, 59* (6) , 19, Article 103112. https://doi.org/10.1016/j.ipm.2022.103112

Rao, A., Pang, M., Kim, J., Kamineni, M., Lie, W., Prasad, A. K., Landman, A., Dreyer, K. J., & Succi, M. D. (2023) . Assessing the utility of ChatGPT throughout the entire clinical workflow. *medRxiv : the preprint server for health sciences*. Retrieved 2023 Feb, from <Go to ISI>://MEDLINE:36865204

Sabry Abdel–Messih, M., & Kamel Boulos, M. N. (2023) . ChatGPT in Clinical Toxicology. *JMIR medical education, 9*, e46876–e46876. https://doi.org/10.2196/46876

Scerri, A., & Morin, K. H. (2023) . Using chatbots like ChatGPT to support nursing practice. *Journal of clinical nursing ,32* (15–16) ,4211–4213.

Sison, A. J. G., Daza, M. T., Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2023). ChatGPT: More than a "Weapon of Mass Deception" ethical challenges and responses from the human-centered artificial intelligence (HCAI) perspective. *International Journal of Human- Computer Interaction*, 1-20. https://doi.org/10.1080/10447318.2023.2225931

Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science (New York, N.Y.), 379* (6630), 313. https://doi.org/10.1126/science.adg7879

Xu, F. (2023). Automatic quantitative assessment of English writing proficiency based on multi-feature fusion. *International Journal of Continuing Engineering Education and Life-Long Learning, 33* (1), 114-127. https://doi.org/10.1504/ijceell.2023.127852

Xue, N., Zhang, X., Zixin Jiang, Palmer, M., Xia, F., Fu-Dong Chiou, & Chang, M. (2016). Chinese Treebank 9.0 https://doi.org/10.35111/gvd0-xk91

Zhang, S., Zhang, X., Wang, H., Cheng, J., Li, P., & Ding, Z. (2017). Chinese Medical Question Answer Matching Using End-to-End Character-Level Multi-Scale CNNs. *Applied Sciences-Basel, 7* (8), 767, Article 767. https://doi.org/10.3390/app7080767