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

The Temperature of Feedback Chatbots: an Experiment in Humor Detection

Yang Cao[1], Tianle Chen[2, *] and Qian Wu[3]

[1]Department of Communication Engineering, School of Zhejiang Gongshang University, Hangzhou, China

[2]Department of Management, School of Huazhong Agricultural University, Wuhan, China

[3]Department of Computer Science and Software Engineering, School of Hebei University of Technology, China

*Corresponding author: a18186659614@163.com

Abstract. Humor is one of the measures of human intelligence, and similarly, humor is a major indicator of chatbot intelligence. To make chatbot answers more lively and interesting, thus increasing the interaction and entertainment between users and bots and providing a more interesting user experience. This paper collects a large text dataset and use models such as humor detection and humor scoring to detect the humor ability of chatbots, and improve the future work based on the results obtained from the experiments. It is shown that the CNN model's ability to recognize the humorous elements in the text and score the humorous information in the chat content gradually improves with multiple detections. This experiment implements the detection of humor ability of chatbots. A rating mechanism can assist in enhancing the humor and interactivity of robots, improve the existing computer humor model, enhance computer understanding of natural language, and promote the further development of artificial intelligence.

Keywords: Humor Detection; Humor Rating; Chatbot Interactivity.

1. Introduction

Humor is a subjective experience that involves a skillful combination of words, sentences, linguistic structures and contexts. Among humans, humor can trigger laughter and pleasure, and humor plays an important role in the social and emotional regulation in human interactions. Similarly, the ability of robots to humor is important for improving the interaction experience and emotional connection between humans and machines. Therefore, as the application of robots in human life continues to increase, so does the demand for the humor ability of robots.

However, most chatbots today can only generate humorous responses using pre-defined humor templates and sentences. These templates are usually pre-written, and the bots can select appropriate humor responses based on the triggering context or user input. This approach can produce simple humor effects in certain contexts, but it lacks the ability to generate and personalize them in real time. Furthermore, with advances in natural language processing and machine learning techniques, researchers have been able to use these techniques to analyze and evaluate the humor capabilities of machines.

Therefore, This paper will identify and score the humor in chat conversations between users and chatbots, and improve the robot's adaptability and refine the chatbot's humor ability based on the data obtained from the experiments.This paper will deal with humor detection from conversations. This paper mainly focuses on resource-rich training, allowing machines to deeply learn humor detection from big data, and the experiments are aimed at improving the chatbot's ability to judge and process humorous language so that the chatbot has a feedback "temperature".In addition, this experiment also works on the robot's scoring of humorous language, giving a grade to the humorous language, which in turn is more helpful for the chatbot to respond in different forms and degrees. In response to the different psychological needs of users, the chatbot can respond to humorous and non-humorous

language in different forms and degrees by detecting and rating humor in the user's language.This paper collects and designs a series of conversation scenarios, covering different conversation situations and user intentions, to train the robot in humorous language. Through deep learning, the robot improves its ability to judge and process humorous and non-humorous inputs in different contexts.In Task 1, a constant amount of the humor category in the code database is used to determine whether or not a joke is a joke. In Task 2, humor ratings are used to predict the value of the funny score in a conversation. This paper uses these two subprojects to categorize many conversations as humorous or non-humorous, using humor_detection_rating and deriving the funny score value. This paper will deal with humor detection from conversations, mainly focusing on resource-rich training to allow machines to learn humor detection in depth from big data, and experiments proceeding to improve the ability of chatbots to judge and process humorous language.

2. Dataset and experimental setup

2.1. Data Set

This paper collects conversation datasets from conversations and comments on social networks, humor shows, and joke books. These datasets can be used for humor detection. This study partitions the datasets into training, validation and test sets for model training, tuning and evaluation. Then, the data is filtered and labeled manually or automatically to distinguish humorous and non-humorous conversations. And the collected humorous dialogues dataset is classified in terms of humor types, such as sarcasm, puns, absurdity, and so on. For example, humor and sarcasm are two closely related sentiments - while humor is often expressed using exaggeration and irony, sarcasm mostly generates from incongruity for detecting humor from everyday sitcom dialogue [1].

2.2. Experimental Setup

For data processing and model construction, this paper imports libraries such as re, spacy, pickle, numpy, pandas, tensorflow, matplotlib, sklearn, keras, imblearn, textblob, urllib, collections, etc. to support text processing , machine learning, data processing and visualization. To control the training speed, effect and generalization ability of the model, this paper defines some key parameters, such as maximum sentence length (maxlen), vocabulary size (nb_words), embedding dimensions (embedding_dim) and so on.

For more in-depth analysis and processing of text data, this paper uses spaCy to load the English language model in order to realize tokenization, linguistic annotation, entity recognition, etc. of text data. Subsequently, this paper defines a series of preprocessing functions, including the removal of stopwords, translation enhancement, synonym enhancement, etc. Stopwords are words with low discrimination power. If they carry no meaning, they can also affect efficiency, resulting in a large amount of unproductive processing [2]. Among them, the removal of stopwords aims to reduce the noise during model training and enhance the model's focus on keywords; translation enhancement and synonym enhancement are used to expand the training data and improve the model's generalization ability. Then, in order to help the model better understand the text data, the pre-trained words are loaded and embedded into a matrix (e.g., GloVe). In the data processing stage, this paper uses Tokenizer to segment the text data into words and convert them into numeric codes for processing by the neural network model. Finally, this paper loads and preprocesses the dataset for humor detection and scoring.

2.3. Humor Detection

In humor detection model, this paper uses CNN for feature extraction and classification. The CNN model is defined in terms of HumorDetection class, which detects whether the text data contains humor elements or not by learning the semantic features of the text data.

This paper performs model compilation in the _compile_model method and sets up the loss function, optimizer, and evaluation metrics. These settings will help the model optimize the parameters to more accurately determine the presence of humor in the text.

The main method performs model training while loading pre-trained models as needed. The model will continuously learn from the training data to improve accuracy and generalization.

In evaluate method, this paper carries out to evaluate the performance of the model on a test set to verify the model's ability to assess the degree of humor accurately.

The predict function scores the humor of a new sentence, determines the sentence's humor level, and gives the corresponding score.

The view_graphs function is used to visualize the loss and accuracy changes during the training process to help analyze the performance of the model during training and further optimize the model.

In the main program, the HumorRating class is instantiated, and the training, evaluation, and prediction processes are performed to ensure the complete construction and application of the humor rating model.

2.4. Humor Rating

This paper uses bi-directional LSTM (Long Short-Term Memory Network) to model the sequence information of textual data to rate the text's humor level. Meanwhile, this paper defines this bidirectional LSTM model in terms of HumorRating class.

The model is compiled using the _compile_model method, and the loss function, optimizer, and evaluation metrics are set. These settings help the model to evaluate the level of humor in the text data.

In the main method, the model is trained with the option to load a pre-trained model. The model will evaluate the humor level of the text by learning information about the text data sequence.

The performance of the model on the test set was evaluated using the evaluate method to verify the model's ability to accurately assess the degree of humor.

Use the predict function to score the humor of a new sentence, determine how humorous the sentence is, and give the appropriate score.

The view_graphs function is used to visualize the loss and accuracy changes during the training process to help analyze the performance of the model during training and further optimize the model.

In the main program, the HumorRating class is instantiated and the training, evaluation, and prediction processes are performed to ensure complete humor rating model construction and application.

3. Results and Discussion

The purpose of the research in this paper is to evaluate the effectiveness of a deep learning based model for humor detection and scoring in a bot chat application. Humor detection is performed using a convolutional neural network (CNN), and humor scoring is performed using a bidirectional long short-term memory network (bidirectional LSTM).

3.1. Loss Analysis

Both the training loss (loss) and the validation loss (val_loss) decrease with increasing training periods (Epochs) (as shown in Fig. 1), indicating that the model gradually improves its ability to fit the data during the learning process. The rate of decline of the training loss is faster initially and slows down as the period increases, which may be a sign that the model is starting to overfit, or that it is approaching an optimal solution. The validation loss is similar to the training loss for the first few

cycles, but begins to stabilize or increase slightly after a certain point. This may indicate that the model is overfitting the training data and cannot generalize on the unseen validation data. In the humor detection task, the training loss of the model decreases significantly with increasing training cycles, indicating that the model gradually improves its ability to recognize humorous content on the training data. However, the validation loss (val_loss) stabilizes later in the training cycle, which may mean that the model has achieved a better fit on the training set, but the risk of overfitting on the validation set still exists.

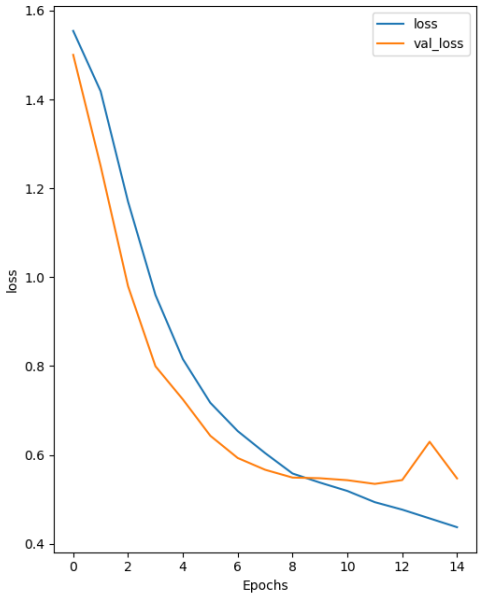

Fig. 1 Training loss and validation loss

val_loss:- val_loss is the loss that is applied to the testing set, whereas, the loss is applied to the training set. val_loss is the best to depend on to avoid the problem of overfitting. Overfitting is when the model fits too closely with the training data [3].

3.2. Recall analysis

The recall metric measures the ability of the model to identify positive classes correctly. In the experiment, the recall value increased from 0.2 to 0.8 (as shown in Figure 2), showing a significant improvement in the model's ability to identify positive classes. The increase in recall indicates that the model is better able to capture the features of positive samples during training, which is especially important for unbalanced datasets. The recall metric (recall) of the humor scoring model continues to improve during training, showing the model's increased ability to identify humorous content. This suggests that the model is able to effectively score the humor elements in the chat, which is crucial for enhancing the fun and interactivity of bot chats.

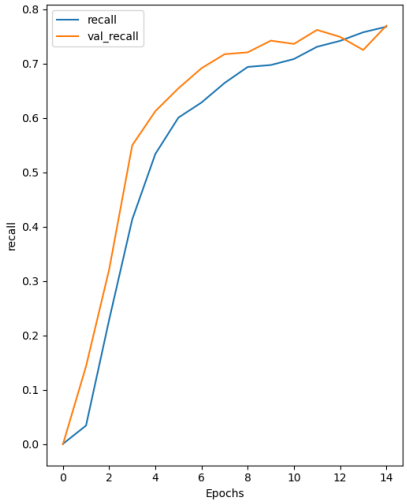

Fig. 2 Recall indicators

Overall, the model achieves the lowest training loss (as shown in Fig. 1) and the highest recall value (as shown in Fig. 2) around training cycle 14. This may indicate a better equilibrium in the model's performance around this cycle. However, it is important to note that an increase in the recall value does not always mean an increase in overall performance. In some cases, too high recall may lead to too many false positives. Therefore, a combination of other metrics (e.g., precision, F1 score) is needed to assess model performance fully.

4. Future Direction of Work

This experiment currently has limitations such as small dataset size, single language level and modeling construct. The group will make improvements in the direction of tuning, architecture, accuracy and naturalness. Future work directions focus on dataset extension and diversification, type optimization and hyperparameter tuning, contextual understanding and sentiment analysis.

Collecting and integrating a broader chat dataset, including conversations from different cultures, age groups, and social environments, to enhance the model's ability to understand and adapt to diverse humor styles. Multilingual datasets are also introduced to enable the model to handle humor expressions in different languages and improve the model's internationalization.

Hyper-parameter tuning, such as learning rate, batch size, embedding dimension, etc., is performed to find the optimal model configuration and improve the model's performance on different datasets. Different model architectures, such as deep CNN, Transformer model, etc., are also explored to determine which architecture performs best in phantom testing and scoring tasks.

Integration of the Context Understanding module allows the model to understand the context of conversations better, leading to more accurate detection and scoring of humor. It also incorporates sentiment analysis techniques to assess emotional tendencies in chats to improve the accuracy of humor scores and the naturalness of user interactions. Sentiment analysis (SA) is a process of computationally identifying and categorizing opinions expressed in a piece of textual content, particularly to decide the writer's attitude towards a particular topic, product or issue [4,5].

5. Conclusion

This paper builds and evaluates a deep learning-based humor detection and scoring model for a chatbot to train the bot's ability to judge and evaluate humorous language. This paper uses a convolutional neural network (CNN) to detect the humor elements in the text, and a bidirectional long

short-term memory network (bidirectional LSTM) to score the humor level. The research methodology includes collection and processing of large-scale text datasets, model training and optimization, and performance evaluation. Evaluating the model's fitting ability versus the robot's ability to recognize humorous content based on training loss and recall metrics. It was found that the humor detection model's loss on the training set decreased significantly with training, showing an increased ability to recognize humorous content. However, the validation loss stabilized at a later stage, hinting at a possible overfitting problem. Meanwhile, the recall metrics of the humor scoring model continued to improve, and the model could effectively assess the humorous elements in the chats.

This experiment currently suffers from limitations such as a small dataset size and a single language hierarchy and model construct. Future work will focus on the expansion and diversification of the dataset to enhance the model's understanding of humor styles across cultures and languages. Meanwhile, model optimization and hyperparameter tuning will be carried out in this paper to explore a more efficient model architecture. The significance of this study is to improve the robot's ability to detect and evaluate humor, improve the chat machine's fun and interactivity, give users a richer and more enjoyable communication experience, and enhance user satisfaction. It can also broaden the application prospects of chatbots in education, entertainment, mental health and other fields, laying the foundation for building a more intelligent and humanized AI system.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In Proceedings of the four- teenth conference on computational natural lan- guage learning, pages 107–116. Association for Computational Linguistics.

[2] Jashanjot Kaur, Preetpal Kaur Buttar. 2018. A Systematic Review on Stopword Removal Algorithms. In Proceedings of the Int. J. Future Revolut. Comput. Sci. Commun. Eng.2018, 4, 207–210.

[3] Shilpa Gite1, Hrituja Khatavkar1, Ketan Kotecha2, Shilpi Srivastava1, Priyam Maheshwari1 and Neerav Pandey1. 2021. Explainable stock prices prediction from financial news articles using sentiment analysis.

[4] Medhat W et al 2014 Ain Shams Eng. J. 5(4) 1093–1113

[5] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.