# Two Important Models in Closed Domain Question Answering Chatbot

David Tong
April 2023

## Abstract

This article discusses two text classification models for question answering and emotion classification. Both of them are based on pre-trained BERT models and different approaches are used in the fine tuning stage, to improve their performance measured by popular metrics.

We discuss three methods to improve performances in the fine tuning stage. At first, different BERT models are being attempted, by their different network structure or training goals, to improve the testing result. Secondly, regard to the SQuAD 2.0 dataset, a two-stage model is proposed, to judge if the question is answerable or not, to achieve a better score. Finally, two two-stage classification models, based on classic psychologist emotion classification model and data clustering result, are proposed to see if they can improve the result or not.

**Keywords: text classification model, question answering model, emotion classification model**

## Introduction

The rapid development and widespread use of the Internet make people deeply attached to it. One of the highly demanding requirements  is to interact with websites in a conversational mode. Compared to searching for information as we use Search engines, people prefer to have a conversation with customer support and it becomes particularly true when it comes to some customized requirement communication.

Differ from open domain chatbots are able to chat about any topic, while closed domain chatbots are designed to chat about a specific topic and will admit they are able to provide an answer when the question is out of their scope. Advantages of closed domain chatbots include that they can be more accurate since they only need to know about one topic, and they can be more efficient since they can extract answers from predefined answer templates.

We aim to develop a closed domain chatbot that can provide accurate and efficient answers to user queries on a specific topic. To achieve this goal, we will leverage recent advances in NLP, specifically BERT and its derivative models, which have shown remarkable performance in NLP question-answering tasks.

However, accurately answering questions is not the only factor that determines the success of a chatbot. We also need to consider the emotional state of the user, as this can have a significant impact on their satisfaction with the conversation. Therefore, we will also incorporate solutions in NLP emotion detection tasks to help us better understand the user's emotional state during the conversation process.

To train and evaluate our closed domain chatbot, we will use both open datasets and popular metrics for question answering and emotion detection tasks. We will also explore both discriminative and generative models to improve the chatbot's performance.

Overall, the development of a closed domain chatbot that can accurately answer user questions and understand their emotional state could have significant implications for customer support and other conversational applications. We hope that our research will contribute to the continued improvement of chatbot technology and its applications in various industries.

# Background

## Question Answering Model

The progress in pre-trained large language models makes a remarkable contribution in NLP question answering tasks.

In recent years, BERT (Bidirectional Encoder Representations from Transformers), a pre-trained deep learning model based on transformers, has emerged as a powerful tool for natural language processing tasks. Many researchers have leveraged BERT for question answering on the SQuAD 2.0 dataset and achieved state-of-the-art results.

However, there is still room for improvement in question answering models. One area of improvement is the use of other pre-trained models that have better structures than BERT, such as SpanBERT, ALBERT, and XLNet. These models can be fine-tuned on the SQuAD 2.0 dataset and their performance can be compared to that of BERT-based models. SpanBERT, for example, has shown promise in capturing more accurate span representations for the answer.

Another area of improvement is the use of ensemble learning and two-stage prediction to improve the performance of question answering models. One approach is to use a Retrospective Reader model, which first classifies whether a question is answerable and then tries to answer answerable questions. This approach can help reduce the number of errors caused by the model attempting to answer unanswerable questions.

## Emotion Detection Model

In terms of emotion classification models, the GoEmotions dataset has been proposed as a fine-grained emotion classification dataset containing 27 emotion categories plus neutral. The baseline model for this dataset is a one-stage classification model using BERT.

# Approach

## Question Answering Model

Our first task is to develop a question-answering model using the SQuAD 2.0 dataset. We will use both the EM (exact match) and F1 scores as metrics to evaluate the model's performance.

We will start with the baseline model, DistilBERT, which is a pre-trained model with relatively less trainable parameters and simple structure. Do fine-tuning on the baseline model with the SQuAD 2.0 dataset will give us a baseline used to compare with further improvements.

To improve the model's performance, we will conduct two types of improvements. For Type I improvement, we will explore the use of other pretrained models, such as ALBERT and SpanBERT, which have more trainable parameters or complex architecture than DistilBERT. We will fine-tune these models with the SQuAD 2.0 dataset and compare their metrics to the baseline model. We will discuss why the increase in trainable parameters or changes in structure of these pretrained models may lead to improved performance on the SQuAD 2.0 dataset.

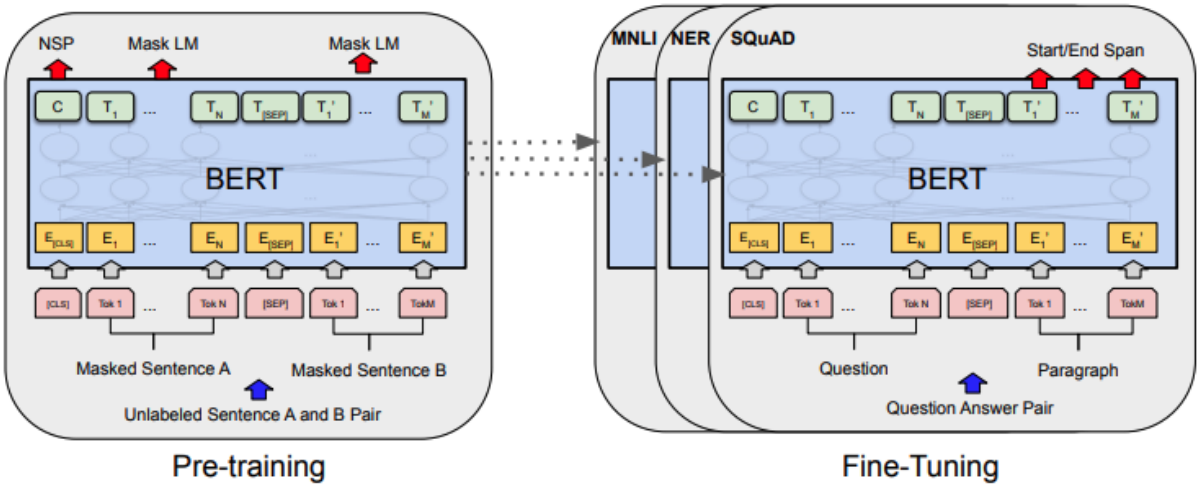


Fig. 1: Use BERT classification to predict start and end span of answer

For Type 2 improvement, we will leverage ensemble learning and apply a two-stage prediction process to further enhance the performance of the model on the SQuAD 2.0 dataset. The first model, called the skim model, will use logistic regression to predict whether the question is

answerable or not. The second model, called the read model, will then attempt to answer the answerable questions. This approach may improve the model's performance by reducing the noise in the dataset and focusing on the most relevant questions.
In summary, our approach involves fine-tuning the baseline models, exploring other pretrained models, and leveraging ensemble learning to improve the performance of our question-answering model on the SQuAD 2.0 dataset.

## Emotion Detection Model

For the second task, we will use the GoEmotions dataset, which contains 27 distinct emotions and a neutral label. Our baseline model is a one-stage classification model that uses the bert-base-uncased model maintained by Hugging Face to classify 28 emotion categories. We will use F1 scores as the metrics to evaluate the model's performance.

To improve the model's performance, we will explore two types of improvements.

As for Type 1 improvement, we will leverage a famous categorical emotion model proposed by psychologists to develop a multiple-stage classification model. By doing so, we aim to increase the accuracy of our emotion detection model. Specifically, we will use the parrot model. The parrot model involves two stages: First, classify the text into one of the six basic emotions (ambiguous, sadness, fear, joy, love, anger) and neutral. Second, refine the classification by mapping it to the GoEmotions categories. We will compare the performance of the multiple-stage models with the baseline model.

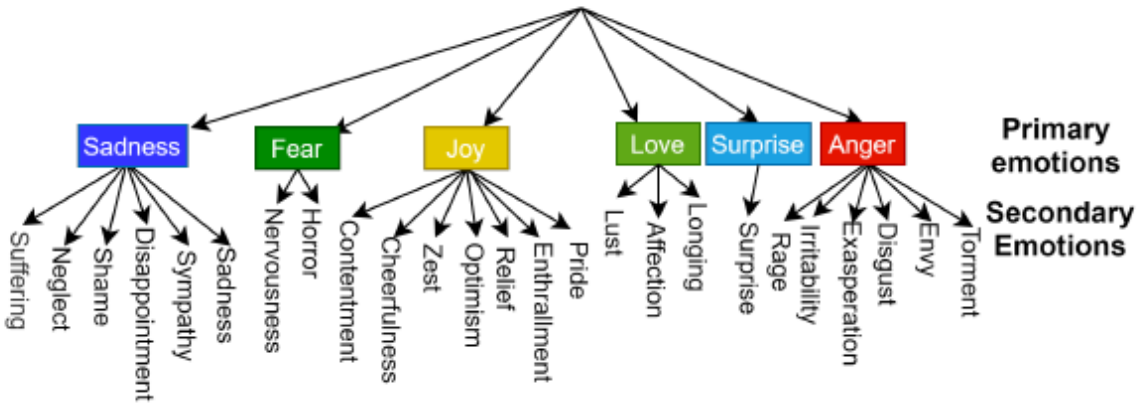


Fig. 2: Parrot's emotions model

```
parrot_mapping = {
    'neutral': ['neutral'],
    'ambiguous': ['realization', 'confusion', 'curiosity', 'surprise'],
    'sadness': ['sadness', 'disappointment', 'embarrassment', 'grief', 'remorse'],
    'fear': ['fear', 'nervousness'],
    'joy': ['joy', 'relief', 'amusement', 'optimism', 'pride', 'excitement', 'gratitude'],
    'love': ['love', 'admiration', 'desire', 'caring', 'approval'],
    'anger': ['anger', 'annoyance', 'disgust', 'disapproval']
}
```

Fig.3 Primary and secondary emotions mapping in parrot's model

The Type 2 improvement has a simple underlying logic. It involves utilizing texts that have been labeled with multiple emotions, and assuming that when multiple emotions are labeled for the same text, there is a close relationship among them. To construct 28 emotion vectors with 58009 features, we transpose the matrix of the GoEmotion dataset, which contains 58009 valid samples with 28 emotions. We then use correlation to gauge the similarity between any two emotions and perform clustering based on the correlation calculation results. We group them into major categories by setting the distance to 1.05.

Next, we create a multiple-stage classification model that first classifies the text into a major category and then maps it to the corresponding GoEmotions categories. To evaluate the performance of this model, we compare it with the baseline and the Type 1 improvements using F1 scores as the metrics.

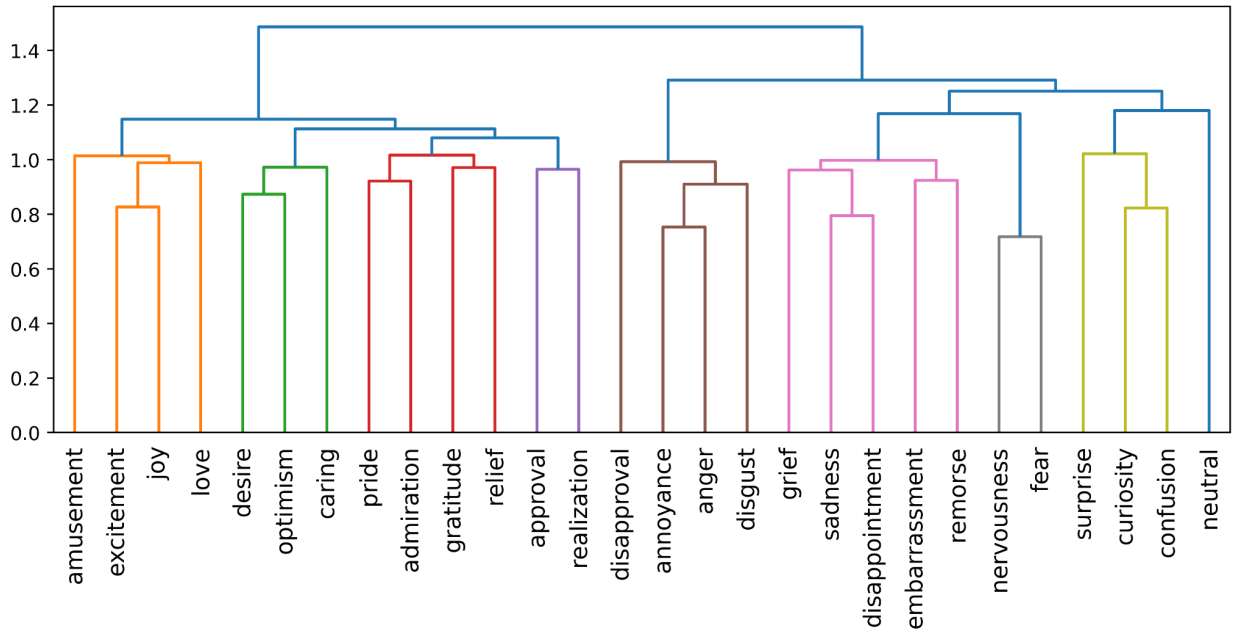


Fig. 4: Data clustering

In summary, our approach for the second task involves exploring two types of improvements: leveraging a famous categorical emotion model proposed by psychologists to develop a multiple-stage classification model and using data clustering to group similar emotions together and develop a new categorical emotion classification model. We will compare the performance of the models using F1 scores as the metrics.

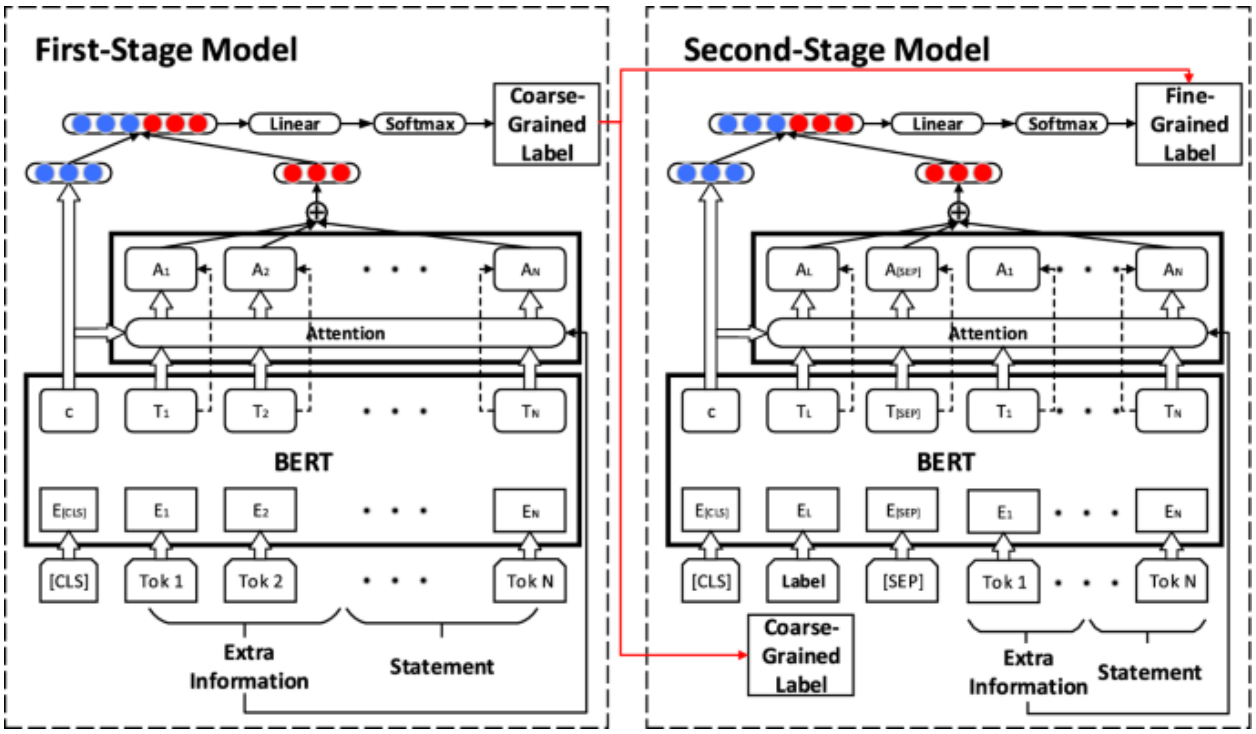


Fig. 5 Two stage classification model

# Result

## Question Answering Model

| Model | Model Type | Improvement Type | EM | F1 |
|---|---|---|---|---|
| DISTILBERT distilbert-base-uncased | Discriminative Model | Baseline | 64.9457 | 68.3097 |
| ALBERT albert-base-v2 | Discriminative Model | Improved Type 1 | 78.2111 | 81.2822 |

| | | | | |
|---|---|---|---|---|
| SpanBERT<br>SpanBERT/spanbert-large-cased | Discriminative Model | Improved Type 1 | 83.1382 | 86.3032 |
| T5<br>t5-small | Generative Model | Improved Type 1 | N/A | N/A |
| Skim-Read Model<br>David-Tong/squad2-skim-read-predictor<br>Skim Model Logistic Regression | Ensemble Model | Improved Type 2 | 83.381 | 86.519 |
| Skim-Read Model<br>David-Tong/squad2-skim-read-predictor<br>Elbow Method | Ensemble Model | Improved Type 2 | 84.0984 | 87.0521 |

The results indicates the following points,

1. A pre-trained model with more trainable parameters will have better results, by comparing results of ALBERT to DISTILBERT
2. A pre-trained model with a suitable training target will have better results, by comparing results of SpanBERT to DISTILBERT
3. An ensemble model with a prediction to if the question is answerable will have slight improvement in result, by comparing Skim-Read Model using SpanBERT to SpanBERT.

## Emotion Detection Model

| Model | | F1 | | |
|---|---|---|---|---|
| | | accuracy | macro avg | weighted avg |
| One-Stage Classification<br>bert-base-cased | Baseline | 0.60 | 0.49 | 0.59 |
| Two-Stage Classification<br>(Parrott's Emotions model) | Improved Type 1 | 0.57 | 0.48 | 0.57 |
| Two-Stage Classification<br>(Data Clustering Modell) | Improved Type 2 | 0.58 | 0.45 | 0.57 |

The results indicates the following points,

1. Two-Stage classification doesn't bring improvement in performance but even a little downgrade in performance.
2. Data based clustering has a different model from psychological emotion classification model.

# Conclusion

In conclusion, we discuss two important models in close domain chatbot and achieve performance can be compared to SOTA counterparts by leveraging pre-trained LLM and referring to papers of Retrospective Reader for Machine Reading Comprehension and GoEmotions: A Dataset of Fine-Grained Emotions.

Looking forward, further discussion of using generative language models like GPT or T5 for SQuaD 2.0 dataset and question answering model will be an interesting and promising topic. Special attention will be given to how to prevent generative language models from giving trumped-up answers.

In parallel, we will make further attempts to remove the neutral label from the GoEmotion dataset so that it can focus on emotion classification and see if the 2 stage models will have improvement in performance then. Also more clustering models will be tried to cluster GoEmotion or other public emotion dataset and try to map emotions to a 2D plane and further discussion relationships among them. In contrast to question answering tasks, emotion detection tasks are naturally easily influenced by bias caused by cultural background and it gives us another interesting topic to research.