ANALYZING BERT-BASED HIGH RECALL INFORMATION RETRIEVAL MODELS WITH SOCIAL MEDIA DATA

by

Zhengyu Zhou

Submitted in partial fulfillment of the requirements
for the degree of Bachelor of Computer Science: Honours in Computer Science

at

Dalhousie University
Halifax, Nova Scotia
April 2022

Table of Contents

Abstract

The most common approaches to improve the performance of high recall information retrieval tasks are to label more data. However, these require much effort, so we would like to start from the perspective of model selection to improve the final performance of the task through a reasonable choice of models.

In this thesis, we performed two simulations of high recall information retrieval and compared several BERT-based models' performance. We slice our existing labeled data and feed it into the model in stages to simulate the multiple iterations performed in high recall information retrieval. We also do a more reductive one using RoBERTa-base and DeBERTa-base to label data instead of sliced data during iteration. We compared BERT-based models from three perspectives: cased/uncased, the number of parameters, and with/without domain-specific pre-training, and came up with three hypotheses from the results.

First, it is not always the larger model, the one that gives better performance. Some larger models can even perform worse than a smaller version of themselves, and larger models usually are more time-consuming.

The second is to choose a cased sensitive model since they can perform better in noisy social media data. Uppercase letters and words play a critical role in social media data. People often use capitalized words or phrases for emphasis or particular meaning, allowing case-sensitive models to understand text content better.

The final suggestion is domain-specific pre-training can improve the overall performance. Many domain-specific models that have already been pre-trained can serve a good purpose. We can use these domain-specific models to improve the overall performance if we do not have exceptionally high requirements for the task performance.

Many methods could improve the model's performance on high recall information retrieval tasks with social media data. Here, we gave the above three suggestions on achieving better performance with less time and effort.

Acknowledgements

I would first thank my thesis supervisor Dr. Evangelos E. Milios, of the Faculty of Computer Science at Dalhousie University. He consistently helped me and gave valuable comments to help me refine the thesis.

I would also thank Muthukumar Rajendran for his assistance with the data sets and experiment procedures.

Chapter 1

Introduction

BERT has been released for almost four years, and it is very commonly used in the NLP field. As time progressed, many variants of BERT were released, and each had different characteristics. Nevertheless, the number of examples and related studies using the BERT-based model in high recall information retrieval is minimal.

In high recall information retrieval tasks, we often only get a minimal number of training sets. We usually improve the model's performance by increasing the number of iterations and expanding the number of manually labeled data sets. However, these are extremely time and effort consuming. Therefore, we would like to improve the performance of BERT-based models in high recall information retrieval tasks by rational model selection from different perspectives. Motivated by the above intuition, this thesis conduct a case study on applying BERT-based models to high recall information retrieval task with social media data.

We simulate the process of building a high recall information retrieval classifier twice. The first approach is to slice the labeled data and stimulate the increase of the training set after iteration in the high recall information retrieval task. The second approach is a more reductive simulation. Compared to the real high-recall information retrieval classifier training, the only difference is that we label the data by combining the results of other models instead of manually labeling them.

Retrieving information from social media data is challenging because social media data contains noisy data, and some useful data is expressed using emotes or abbreviations. The BERT-based models have shown their potential on information retrieval tasks. According to the experiment results, we will give some suggestions on selecting a BERT-based model for high recall information retrieval tasks.

In this thesis, we tested three different perspectives to investigate their impact on the final performance of the model. After evaluating the recall rate on our test set, we come up with three recommendations on selecting a model for high recall information

retrieval tasks with social media data.

The first recommendation is to not always choose the largest model since the largest model does not always mean better performance. The performance varies from task to task, and we should select the model based on our requirements. Training a larger model will always be more time-consuming and computationally power-consuming. According to our experiment, the larger model can sometimes perform even worse than a smaller version of itself with a longer training time. For example, if we only have limited time, choosing a smaller model will make us achieve a similar or better score than throwing all the data to the large model and waiting for the result.

The other advantage of choosing a smaller model is tuning a smaller model is always easier than tuning a giant model. We can run the model with different hyper parameters multiple times to find the best result during training. While we are fine-tuning the small model, the large model may not have even finished its first training.

The second recommendation is to choose cased sensitive model since they can perform better in noisy social media data. It is mainly because capital letters are used very often in social media. When capital letters are used in emails, text messages, and social media, the intent is to show emphasis - like a person speaking in a confident voice or shouting to express emotions such as anger or frustration. In this way, a case-sensitive analysis can gain an advantage in understanding the social media data. In this thesis, we focused on a specific high recall information retrieval task: to retrieve as much mental health and psychology-related data from the Reddit data. Those capital letters and words with emotion will help us find the related data better.

Our last suggestion is to find a domain-specific pre-trained model rather than a generalized model. Domain-specific pre-training can improve the overall performance if we focus on a particular domain. As we mentioned before, the task we focused on in this thesis is to find the information related to the mental health and psychology domain. We introduced two domain-specific pre-trained models, one is pre-trained with scientific text data, and the other is a mental health and psychology enhanced version of the first one. The result shows domain-specific pre-training can achieve better performance with similar training time.

To conclude, if someone wants to apply the BERT-based model in the high recall information retrieval field with social media data, we suggest choosing the model

which is case sensitive and has been pre-trained with domain-specific data and choosing the smaller version of the model if there is not enough time and computational power. After doing simple tuning to the smaller version, if the performance still does not meet the requirement, try to use some larger versions of the model to see whether there is an improvement or not.

Chapter 2

Related work

2.1 Background

2.1.1 BERT-based models

BERT, known as Bidirectional Encoder Representation from Transformers, is a pre-trained language representation model. It emphasizes using the new masked language model (MLM) instead of the traditional unidirectional language model or the shallow splicing of two unidirectional language models for pre-training, which can generate deep bidirectional language representations. The BERT paper was published with significant better results in 11 natural language processing tasks.[6]

After BERT was introduced and open-sourced, many scholars in the field of natural language processing have developed various models based on BERT's architecture, which we will refer to below as BERT-based models. These models have different characteristics, ranging from lite version to those with secondary pre-training for specific domains. However, there is no doubt that these models are constantly setting new records for NLP tasks.

At the same time, some scholars have started to apply the powerful BERT-based models to the field of information retrieval. The most common approach is to combine the BERT-based models with other information retrieval systems. However, there is very little research on the impact of different types of BERT-based models on results. In particular, there is minimal research on the application of BERT models to high recall information retrieval tasks.

2.1.2 Social media data

With the rapid development of technology, social media plays a more critical role in our daily lives. Social media data also has contributed great to the arrival of the big data era.

Social media has given us endless opportunities. Twitter, Reddit, and other platforms allow us to showcase our skills and areas of competence. It makes communication much more reachable. We can express our ideas very quickly and straightforwardly, which was almost impossible a few years ago.

It is easy to get your point across on social media. If you are behind a nickname, everyone can feel more comfortable expressing their true thoughts - after all, no one will know who you are that way. Of course, many people also take advantage of anonymity to lie freely or be violent online. In this way, we should carefully handle social media data when doing data mining and data analysis.

Despite the massive amount of data, which is full of false and useless data, the real data is still very valuable.

In the high recall information retrieval task that this thesis focuses on, we want to find as much information related to mental health and psychology problems as possible from the data we extracted from Reddit. During the COVID-19 pandemic, it became more difficult than usual to go out and socialize. The opportunities for people to use social media to communicate and voice their opinions became extraordinarily high. We hope to retrieve as much mental health-related information as possible from the Canadian Reddit data to better prepare for the analysis.

2.2 The potential of BERT in high recall information retrieval

The field of information retrieval, like natural language processing, has witnessed the growing advantages of neural network-based approaches. This advantage has become even more apparent with the release of deep language models led by BERT. A common way to apply the successful BERT-based model to information retrieval is to use a traditional information retrieval system for initial retrieving and a deep language model for re-ranking. The final result is the result of the ranking.

The results show that combining BERT with traditional information retrieval systems yields better performance and results. [2] This provides a sound basis for applying more BERT-based models to information retrieval. This is one of the reasons we compare the performance of BERT-based models in high recall information retrieval. Although the field of information retrieval has been developed for quite a long time, there are still too few such studies for applying deep language models to

the field of information retrieval.

2.3 Domain-specific pre-training

Recent research on domain-specific BERT models has shown that pre-training the models on in-domain data can improve the efficiency of downstream tasks. Through extensive experiments, studies show the importance of selecting in-domain source data to pre-train the BERT-based model.[5]

However, these existing studies mainly focus on some classical tasks in natural language processing, such as text classification and named entity recognition. There are not many studies on the performance of BERT models in the field of high recall information retrieval after domain-specific pre-training.

These studies in natural language processing made us curious whether such secondary domain-specific pre-training could also be effective in good improvement in high recall information retrieval. So this thesis conducts experiments to simulate the effect of domain pre-training on the BERT model.

2.4 Impact of number of parameters and model size

Increasing the size of a model before training on a natural language representation usually improves the performance of downstream tasks. However, at some point, further increasing the model becomes more difficult due to GPU/TPU memory limitations and longer training times.[8]

To solve the above problem, many models have been released in different versions for the number of parameters and other metrics to solve the problem of training resource consumption for large models. One of the most commonly used metrics to distinguish model size is the number of parameters, and in this paper, we will use this metric to compare model sizes. In our comparison, we will refer to models with a relatively more significant number of parameters as large models and vice versa. Although large models can achieve excellent results with sufficient training resources on their own merits in most cases, it is also shown that models with a small number of parameters can achieve similar results with limited time and computing power.[8] Choosing a smaller model version and applying some fine-tuning could be a better

solution for users with limited computing power and time.

However, as stated in the introductory section of this paper, the current research on the performance difference between large and small models also focuses on some classical tasks of natural language processing. In contrast, this paper will simulate the performance of large and small models in high recall information retrieval in experiments.

2.5 Previous high recall information retrieval classifier

Our work is an extension of the high recall information retrieval classifier built in another article.[10] We will afterwards refer to the content of this article as previous work.

This classifier is based on the RoBERTa-base model and results from 7 iterations. After each iteration, the classifier builder sifts through the top 200 posts given by the classifier to the domain experts for manual annotation and feeds the annotated ground truth back into the model to improve the model's overall performance. We have the complete annotated data set from the authors and a separate validation set. We also inherit the authors' data set from the original data, and we use the above data set in our experiments to simulate high recall information retrieval and analyze the experimental results.

The biggest problem in creating this high recall information retrieval classifier is the long training time and labor required for manual labeling. In the previous work, the authors invited domain experts to label the posts, but in the end, only 1002 labeled samples were obtained after months of effort. Even after obtaining 1002 samples, the classifier still performs with a recall of 0.3 on the validation set. This leads us to think about how we can improve this process and increase the final recall. This paper will focus on the model selection phase to test the impact of different BERT-based models on the performance of building a high recall information retrieval classifier.

Chapter 3

Methodology

3.1 Data sets

We mainly used three data sets in our experiments, which were created in previous work when building the high recall information retrieval classifier.

We first present the first data set, which is a purely manually annotated data set, and in our experiments, we consider it as ground truth. The whole data set consists of 1002 manually labeled Reddit posts. Among the data set there are 550 negative posts and 452 positive posts. (See Table 3.1) It is a collection of data manually annotated by experts in the domain during each iteration of the high-recall information retrieval classifier. In the previous work iteration, the authors gave the first 200 posts filtered by each classifier to domain experts for labeling. They performed a total of 7 iterations, and the data set has only 1002 data due to the limited performance of the classifier in the last few predictions.

The second data set is a collection of relevant posts collected by keyword search on a subreddit outside subreddit Canada. It contains 988 posts related to mental health/psychology field. Multiple keywords has been used in building this data set. All of the data in this data set not only has keywords related to international students and Canada but also has specific mental health/Psychology keyword such as mental illness, stressful, etc.. Since this data set are built with more detailed and complex keyword search, they are also considered ground truth and will serve as the validation

	Total Posts	Positive Posts	Negative Posts	Obtain Method
Labeled data set	1002	550	452	Manually labeled
988 data set	988	988	0	Complex Keyword Search
subset of the original data set	5806	N/A	N/A	Extract with Reddit API

Table 3.1: The statics of all three data set and obtain method, the subset of the original data set is unlabeled data, so we use N/A for both positive and negative posts.

set in the experiment. We will refer to this later as the 988 data set.

The third data set is a subset of the original data set obtained from Reddit. The original data set contains all the data collected on subreddit Canada from January 2020 to June 2021. The subset we use is the data set related to international students obtained from the original data set using keyword searches. This data set has 5806 information and we want to retrieve related data from this data set.

The specific steps for constructing the data set can be found in Appendix A.

3.2 Perspectives

In a high recall information retrieval task, we typically start with a model to classify the data. We then manually labeled some or all of the classified data. We will add the human-labeled data to the training set and continue training. This is called one iteration. Usually, multiple iterations will be performed in high recall information retrieval tasks to improve the performance. The model selection is an essential feature in this whole process. This thesis will focus on the model selection part and apply different BERT-based models to the iteration process to know different models' features' impact on the performance.

We mainly compare the performance of BERT-based models in high recall information retrieval tasks from three different perspectives. We will then expand to explain each of these three perspectives.

All the set up and parameters of models can be found in Appendix B, we will just list the model names in this section.

3.2.1 Cased/uncased

We focus on cased and uncased models' performance of high recall information retrieval classifiers in the first perspective. The main difference between the cased and uncased models is that in the uncased case, the text will all be turned into lowercased, while in the cased case, the text is the same as the input text. In other words, cased models have separate vocabulary entries for differently-cased words.

In general, the performance of cased and uncased models depends on our specific task and cased, and uncased versions of the same model will generally be consistent

in other parameters. If the task requires more information from the training text to improve performance, then the cased model may be a better choice because we can make the model understand the text better by adding cased information. Our data source is social media data on Reddit in our specific experiments. People often use plain capitalization to spell words frequently on social media to express their emotions better. We believe that using the cased model can better help us capture the emotions behind these words. Since we need to retrieve mental health and psychology posts, capturing these subtle emotions is particularly important. That is the reason we chose to compare the performance of the cased/uncased model, and we conjectured before the experiment that The cased model would have better performance. We will verify this idea through a series of experiments in the follow-up.

In this thesis experiment, we selected 2 different groups of models to compare the differences in performance between the cased/uncased versions of the same model, and they are:

1. BERT-base and BERT-base-cased
2. BERT-large and BERT-large-cased

3.2.2 Model size

Deep language models in natural language processing have evolved rapidly driven by large-scale computing, large data sets, and advanced algorithms and software for training these models. Language models with many parameters, more data, and more training time can yield more decadent and detailed language understanding.As a result, they generalize well to effective zero-shot or few-shot learners with high accuracy on many natural language processing tasks and data sets.

The BERT-large model was the largest model at its release, with a whopping 340 million parameters.[6] Moreover, when 2022 comes, the Megatron-Turing NLG model released by Microsoft in collaboration with NVIDIA already has a staggering model size of up to 530 Billion. It has demonstrated top accuracy in various natural language tasks, including reading comprehension, common sense reasoning, and natural language inference.[11]

It seems to be a trend to have large models with many parameters, but for most

training tasks, the time cost, computing power, and other resources required for training large models are also very impressive. When the ALBERT model was released, the authors released several versions with a different number of parameters to solve this problem.[8] Therefore, we would like to compare the differences in performance when different sized versions of the same model are used for the same training task from the perspective of the number of model parameters. We want to see whether the larger model will also has a consistent better performance in the high recall information retrieval classifier. We will also record the time required to train these models and give an analysis.

In the experiments of this thesis, we selected five different groups of models to compare the differences in performance between different parametric (different size) versions of the same model, and they are:

1. RoBERTa-base and RoBERTa-large
2. BERT-base and BERT-large
3. ALBERT-l and ALBERT-xxl
4. DeBERTa-base and DeBERTa-large
5. BERT-base-cased and BERT-large-cased

3.2.3 Domain-specific pre-training

Research on pre-trained domain BERT models based on in-domain data has shown that domain-specific pre-training can improve target task performance. [3] With the development of BERT-based models, many models have been pre-trained twice for different specific domains. These models can often be as easy to use as the native BERT, most of the work has been done by the authors in the pre-training phase, and these models tend to perform better in handling tasks within the domain.

We will compare whether our selected models that have been pre-trained with in-domain data will obtain better performance in our simulated high recall information retrieval task as in the traditional natural language processing task.

Because our task focuses on retrieving posts related to mental health and psychology, we selected two models with different levels of secondary training. The first one

is SciBERT[4], and the other one is psych-search, which is an extension of SciBERT trained with Psychology and Psychiatry PubMed research data. We will compare their performance with BERT-base to verify whether domain-specific pre-training is also advantageous in high-recall information retrieval tasks.

3.3 Simulating high recall information retrieval

Due to time and computational power, we will simulate the training process of a high recall information retrieval classifier in our experiments by inputting data in batches and labeling the data with the results of multiple models. Our experiments are divided into two phases. The first phase we call the first view. We input our existing labeled data into the model in batches as a training set, and we evaluate the model's performance by two different validation sets. The first validation set is the 988 data set we mentioned before. However, the problem with this data set is that all the data are positive examples, so to better evaluate the model's performance, we slice the labeled data and use train test split method to set aside 201 posts of it to serve as our second validation set (we will call it the TSS data set).

For the first simulation experiment, we took the top 200, top 400, top 600, top 800, and the complete data set from the labeled data to form 5 training sets to mimic the situation that the data set grows gradually with iterations in high recall information retrieval, and let the models trained in different training sets predict the 998 data sets to compare their recall rates. The work flow of the first simulation training is in Fig.3.1.

For our TSS validation set, we took the top 200, top 400, top 600, and the complete data set to form four training sets and let them predict the TSS data set to compare the performance of the models. The detail of the training set is in Fig.3.2.

To make the obtained experimental data more credible, we performed additional training for the case where the whole training set and 988 validation set were used, and for the case where the full TSS training set and validation set were used, respectively.

We repeat the additional training for four times and record the recall rate, precision rate and training time because we noticed that the prediction varies for each training. We counted the results of the additional training and calculated their arithmetic means and standard deviations. We expressed the final performance of the

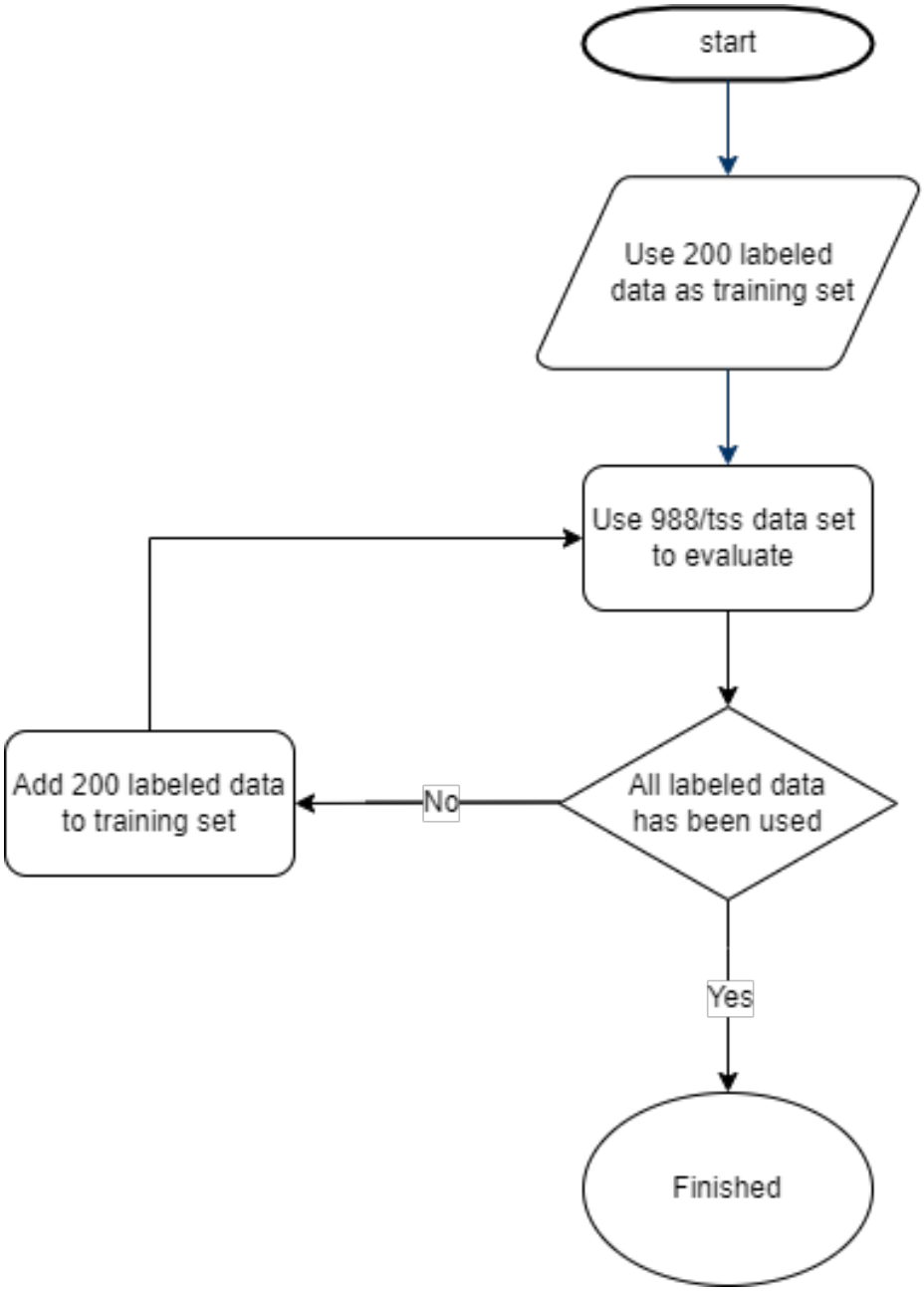


Figure 3.1: Experiment procedure for the first simulation, the parameter used in this simulation is the same from previous work, we trained model for 3 epochs in each iteration, the parameters and set up could be found in Appendix B.

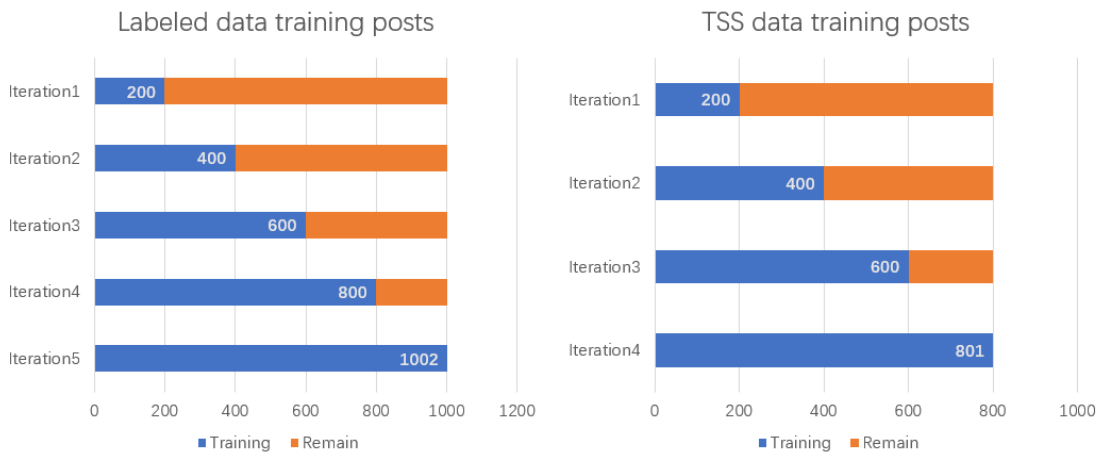


Figure 3.2: Training data set used for the first simulation, we sliced the data set to simulate the growing training set in high recall information retrieval

model as the mean of the recall, the mean of precision and their standard deviations.

In the first phase of the experiment, the parameters used are the same as those used in the previous work. The specific parameter settings for the experiment can be found in the appendix after the text.

The second simulation is a more reductive experiment of the training process of the high recall information retrieval classifier. We will combine the experimental data obtained in the first stimulation to filter the appropriate models and train them for iterations. The only difference from the actual training process is that we do not bring in domain experts for manual labeling; we combine the prediction results of the remaining models to label the results and re-input the top-ranked results into the model as the training set.

Based on the first simulation results, we will select four models to simulate with a higher degree of reduction, and we will use the first 400 data of labeled data as the base training set to train the models.

After completing the training, we will let the models retrieve data from the original data set and record the recall of the models for the 988 data sets at this simulation. After retrieving the original data set, we select the top 200 data with the highest probability of prediction based on descending order and call two models to evaluate the data independently of the four models we have selected. We will train these two models in advance using the complete labeled data, and we will combine the

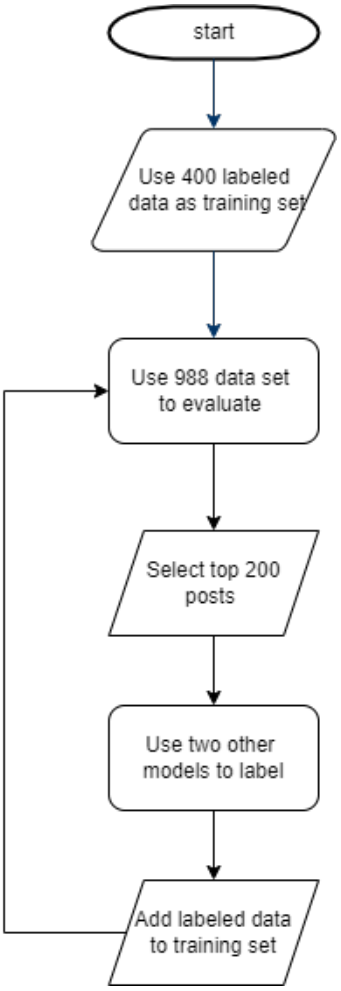


Figure 3.3: Experiment procedure for the second simulation, the parameter used in this simulation is different from previous work, we trained model with larger batch size in each iteration, the parameters and set up could be found in Appendix B.

prediction results of these two models to label the data and stitch it with the base training set as the training set for the next iteration.The work flow of the second simulation training is in Fig.3.3.

Because we do not use manual labeling, and the model performance shows that they tend to predict positive data as negative, we raise the threshold of the positive label to in order to make the machine labeling more accurate. We will label data as positive only when the prediction probability of our two labeled models sums up to more than 0.9. We will train the selected models for three iterations to compare the performance of each other.

Chapter 4

Results

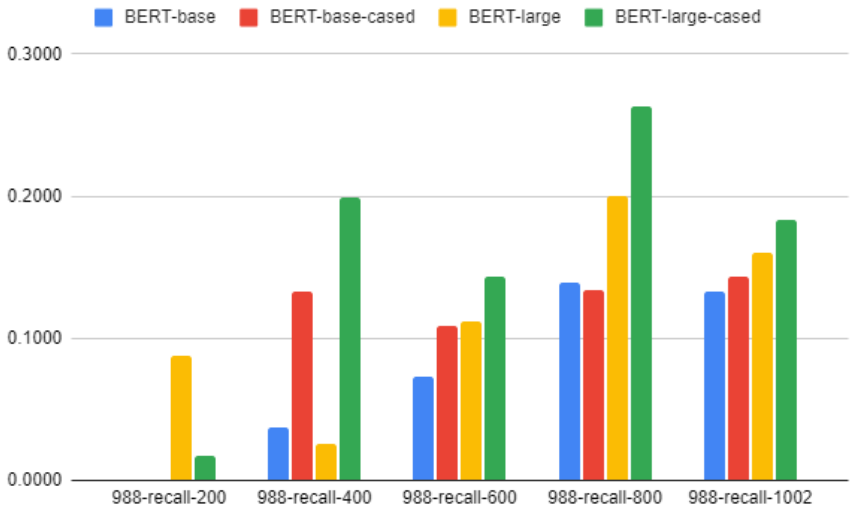
4.1 First Simulation

We followed the experiment steps we mentioned earlier and obtained the following results. Our experiments mainly use the recall rate as our metric, and we keep four decimal places in our results. There may be cases where there is no data, or the data is equal to 0 in the Figure and the table. We have conducted additional experiments for these cases and attribute them to the fact that there is too little information for the model to understand the knowledge well enough to make predictions from them, and we keep the values of these cases as 0.

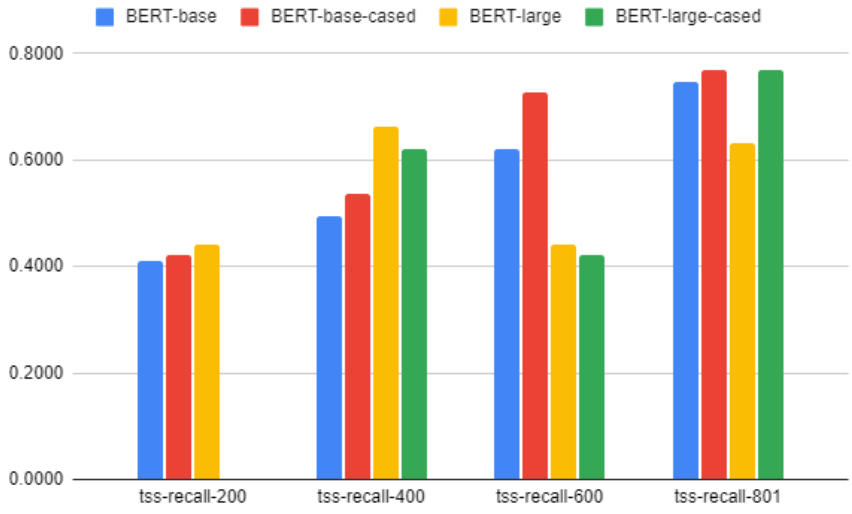
In both 988 data set and TSS data set, cased models always have a higher recall rate than uncased ones when we use all the labeled data (for 988 data set the training set is 1002 and for TSS data set it is 801) as training set. (See Fig. 4.1) For the example against our hypotheses, we will recognize them as counter examples. We choose to ignore the 200 size training set because it contains limited information for the model to learn. We can observe three other counter examples which are 988 with 800 training data, and TSS with 400 and 600 training data. In these cases, the cased model perform worse than uncased models.

Given the small number of such counterexamples, the difference between the performance of the cased and uncased models in the counterexamples is not significant, and the cased model still performs better overall with the complete training set. We believe that our theory is correct and that cased models tend to outperform uncased models.

We got a super solid result shows that more domain-specific pre-training better performance. (See Fig.4.2) The 988 with 200 input training data case is ignored since all three models cannot get enough information with it. The TSS with 801 training data case is kind of a counter example to our hypotheses, since SciBERT has better performance than Psych-search, but all three models have a similar result and

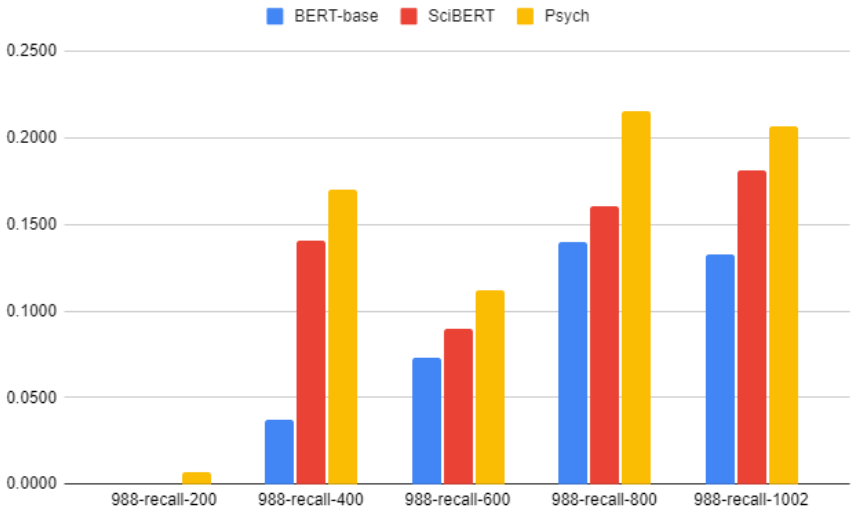


(a) 988-data set

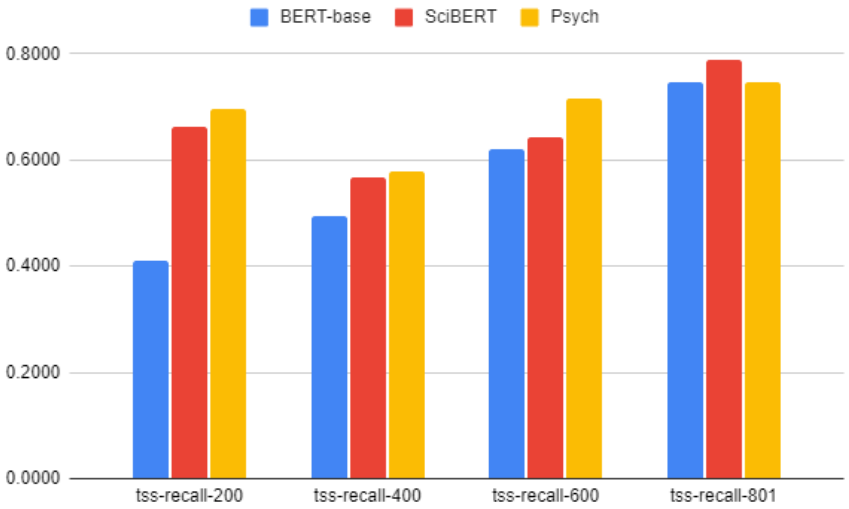


(b) TSS-data set

Figure 4.1: The performance between cased and uncased model in the first simulation, the bar chart shows the recall rate of each model in this simulation. The 200 size training data set has very poor performance, especially for the 988 validation set. Most of the cases fit our hypotheses, but 988 data set with 800 training data and TSS data set with 400 and 600 training data show counter cases against our hypotheses.



(a) 988-data set



(b) TSS-data set

Figure 4.2: The performance between model with or without domain-specific pre-training in the first simulation, the bar chart shows the recall rate of each model in this simulation. The 200 size training data set with the 988 validation set cannot make good prediction because lack of information. Most of the cases fit our hypotheses, TSS data set with 801 training data cases shows SciBERT has better performance than Psych-search, which is against our hypotheses.

	988-recall-200	988-recall-400	988-recall-600	988-recall-800	988-recall-1002
RoBERTa-base	0.0000	**0.1043**	0.1184	0.1538	**0.2298**
RoBERTa-large	0.0000	**0.0455**	0.1245	0.1842	**0.1761**
BERT-base	0.0000	**0.0374**	0.0729	0.1397	0.1326
BERT-large	0.0881	**0.0253**	0.1113	0.2004	0.1599
ALBERT-l	0.0000	0.0789	**0.1397**	0.1184	**0.1893**
ALBERT-xxl	0.0101	0.1377	**0.1275**	0.1872	**0.0789**
DeBERTa-base	0.0000	0.0850	0.0972	0.1164	**0.1680**
DeBERTa-large	0.0334	0.2196	0.1852	0.2115	**0.1528**
BERT-base-c	0.0000	0.1326	0.1083	0.1336	0.1437
BERT-large-c	0.0172	0.1994	0.1437	0.2632	0.1832

Table 4.1: Different Size of Models' Performance Comparison-988-simulation-1, we have 5 of 25 sets that show larger models perform worse than smaller models.

	TSS-recall-200	TSS-recall-400	TSS-recall-600	TSS-recall-801
RoBERTa-base	0.0000	0.3895	0.5158	**0.8421**
RoBERTa-large	0.1684	0.4526	0.7895	**0.7895**
BERT-base	0.4105	0.4947	**0.6211**	**0.7474**
BERT-large	0.4421	0.6632	**0.4421**	**0.6316**
ALBERT-l	0.5684	**0.6316**	**0.3263**	0.5895
ALBERT-xxl	0.6316	**0.4000**	**0.2316**	0.6000
DeBERTa-base	0.0000	0.0000	0.7368	0.7474
DeBERTa-large	0.0000	0.7158	0.7368	0.8421
BERT-base-c	**0.4211**	0.5368	**0.7263**	0.7684
BERT-large-c	**0.0000**	0.6211	**0.4211**	0.7684

Table 4.2: Different Size of Models' Performance Comparison-TSS-simulation-1, we have 7 of 20 sets that show larger models perform worse than smaller models.

the BERT-base model did not outperform the others, so we still think our theory is correct.

To better demonstrate the impact of model sizes, we created two tables 4.1 and 4.2 to explain the relationship between model size (number of parameters) and the overall performance.

In the tabular data, all numbers bold in italics mean that the large model does not perform as well as the small model. The large model does not perform as well as the small model in nearly one-third of our experiments' 45 sets of comparisons. Hence, we conclude that the large model does not always perform better than the small model in high recall information retrieval tasks under the same conditions.

4.2 Second Simulation

Before implementing the second simulation of the simulation process, we performed additional training with the complete labeled data. We combined these training results with the first simulation results to select the model for the second simulation. We trained each model with complete labeled data and six epochs to record more results and make our suggestion more convincible.

We record the recall rate and training time for each of the model and display them in Fig.4.3 and Fig.4.4. We have conducted the experiment for four times so the result here are the mean of those four experiment results.

From those Figures, we can extract two import pieces of information, the first one is that a larger model always means longer training times, but large models do not always mean better performance. As we can see in the Figures, ALBERT-xxl model has surprisingly long training time compared to other model. But in the TSS validation set it has a very poor performance.

The second important aspect is about the ALBERT-xxl model, which can achieve a super high result with 988 validation set, in our record, it can even sometimes outperform our previous work, which is a classifier under seven iterations' training. But it can also perform poorly in the TSS data set, we will discuss this interesting result later in the discussion section.

We also recorded the precision of each model during this experiment. The result shows that the model did not perform bad precision when we try to increase the recall rate. We can also noticed that not all the larger model has a better precision rate than the smaller model in Fig.4.5.

In the end, we choose BERT-base, BERT-base-cased, BERT-large and psych as examples to enable us to compare the result from three different perspectives because they all have a relatively good performance. We will use RoBERTa-base and DeBERTa-base to label the top 200 data from each iteration because they have the best performance if we do not take psych into account.

The experiments reaffirm our conjecture. We performed three iterations and observed that in each iteration the cased model performed better than the uncased model, the model with domain-specific pre-training performed better, and the large model did not always perform better than the small model.

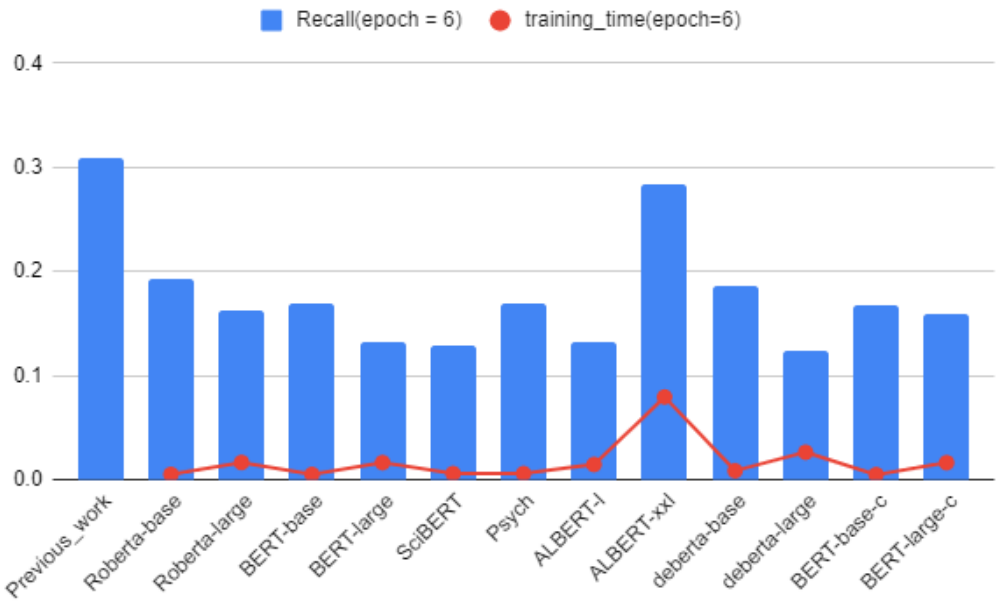


Figure 4.3: Result of Extra test on 988 data set, models such as RoBERTa-base has better performance than its larger version. But model such as ALBERT-xxl has better performance than its smaller version, it has the best performance across all models and even outperform previous work in our experiment result.

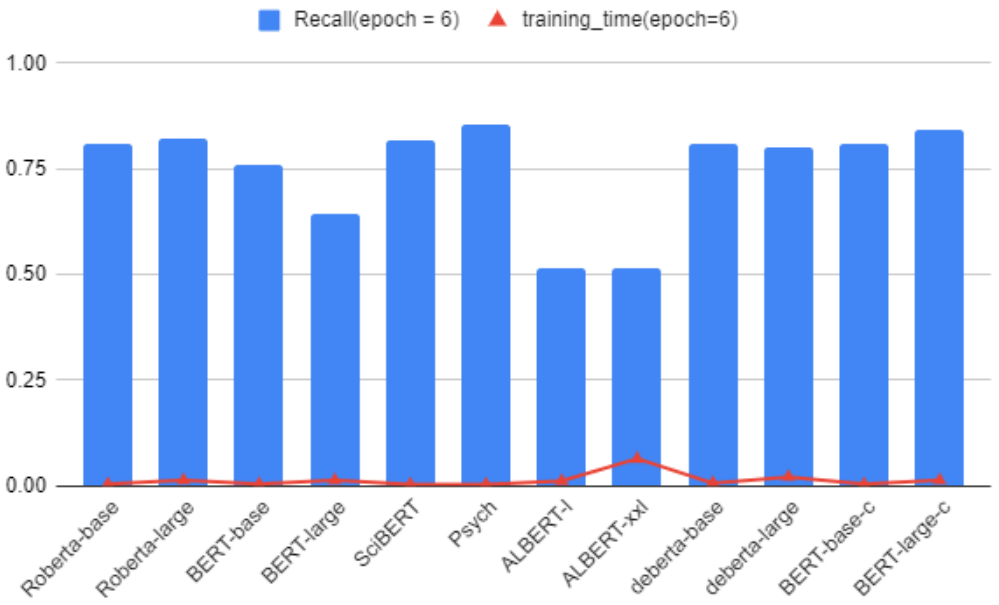


Figure 4.4: Result of Extra test on TSS data set, the performance of TSS validation data set is very different from the 988 validation set. But in both validation set, ALBERT-xxl both has long training time.

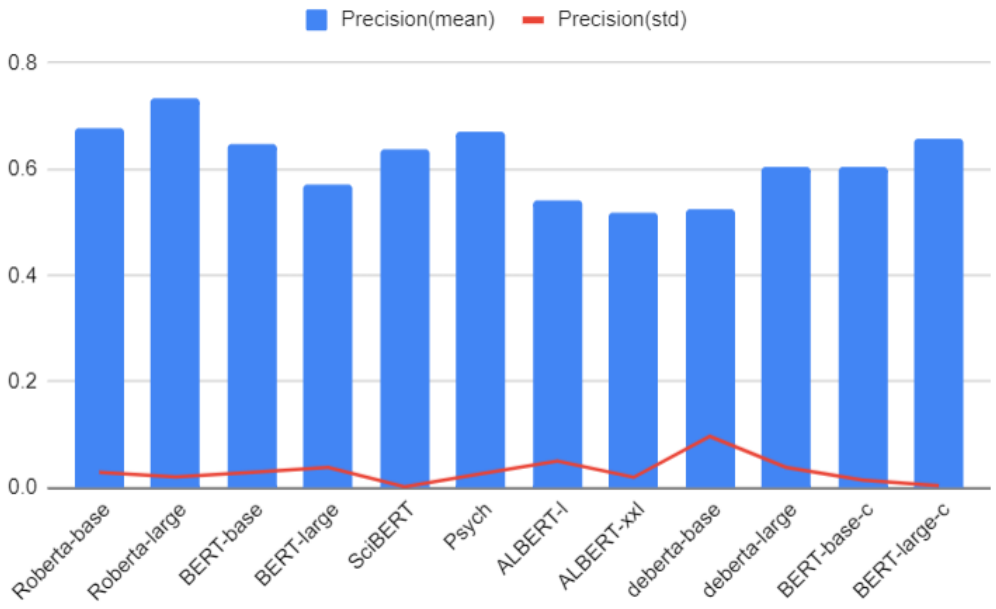


Figure 4.5: Result of Extra test on TSS data set-Precision, the models all have acceptable precision rate while we try to increase the recall rate and the standard deviation shows all the model have stable performance on precision rate.

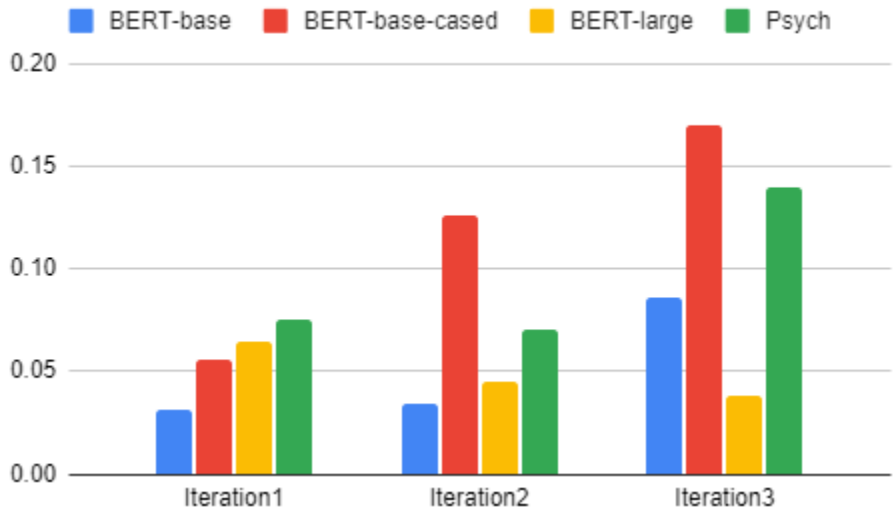


Figure 4.6: Result of second simulation of high recall information retrieval, the results fit our three hypotheses very well.

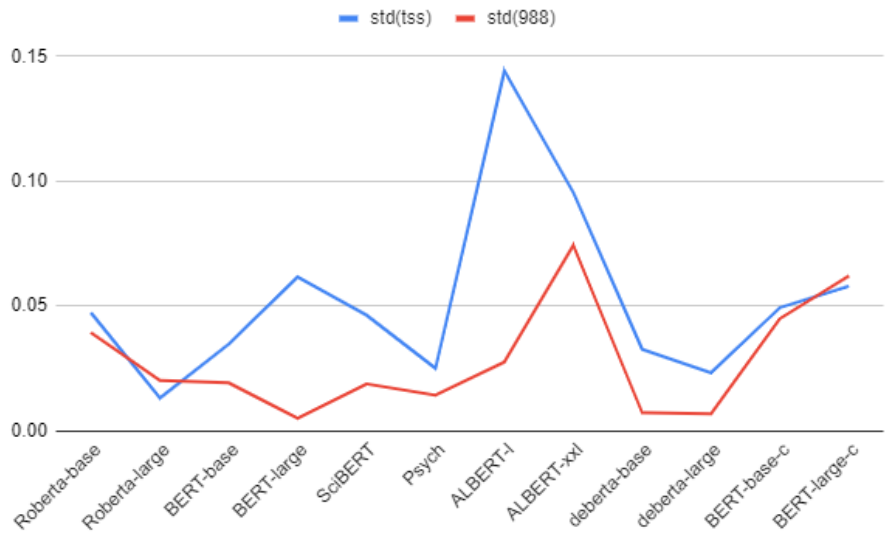


Figure 4.7: The Standard deviation of Extra Test before second simulation, ALBERT-l has highest standard deviation among all models with TSS validation data set, ALBERT-xxl has highest standard deviation among all models with 988 validation data set.

4.3 Discussion

In the first simulation of our experiments, we observed the results of our experiments. We found an exceptional case: the ALBERT-xxl model, whose performance is volatile, when increasing the number of epochs and keep other parameters as previous work for training, even if we do not iterate exhaustively but use all labeled data as the training set, we can sometimes get a very high recall rate. The recall rate can be even higher than the previous work, which is manually labeled for seven iterations. The other models cannot achieve a recall rate over previous work even with more epochs. However, we also observe that the standard deviation of the performance of the ALBERT model is very large in Fig.4.7. In other words, its performance is not stable.

We attribute this performance to both the lack of labeled data and the structure of the model itself. However, we have not yet argued this point, and it is only one possibility. If we need to verify this conclusion, we need to conduct more experiments and analyses on the ALBERT model.

Before we entered the second simulation, we found that the prediction result varies

at each training, even with the same data set. This may be because we reach the local optimum because of the small number of epochs. Our results show that most models have very similar or equal results during four training, but the ALBERT series models' results vary a lot in the experiment.

The overall performance of the rest of our models in the experiments is not very high. This is because we only simulated the training process of a high recall information retrieval classifier. The accuracy of manual labeling would be more reliable than relying on the performance of other models. However, because we aim to verify the differences in the performance of different models, our conclusions are still reliable under such experimental conditions.

There are huge difference between our two validation sets. We addressed this to the complex keyword search methods used in the 988 validation set. The model may classify the posts only according to the keywords instead of understanding whether it is a related post or not. In our TSS data set, since it is manually labeled by domain experts, it contain posts related to the mental health/psychology domain even without related keywords, this will help model to understand the knowledge behind the words.

And for our three suggestions we make the following explanatory analysis.

The cased models generally perform better than the uncased models. We attribute mainly to the fact that the data sets at our disposal are often very limited in high recall information retrieval tasks. In this case, we need to obtain as much information as possible to allow the model to understand the knowledge better. In social media data, uppercase characters often contain more information, such as the author's sentiment, so incorporating these word vectors with uppercase characters into the model can improve the recall of the model.

The large model does not consistently achieve higher recall than the small model. We speculate that this phenomenon is the limited data set we use, but this is a widespread phenomenon in constructing high recall information retrieval classifiers, where more training sets mean more time and labor. Larger models tend to learn deeper knowledge, which also leads to a decline in prediction recall rate due to overfitting when training with a small data set.

One of the advantages of large models is their ability to generalize, which is an

advantage of many parameters. At the same time, in our experiments, we focus on only one task, which may be one of the reasons why large models are not as powerful as expected.

For large models, we think more experiments can be conducted to explore, and we can try larger data sets. We can also try to compare the performance of models in different domains. If there are no domain-specific models already developed for some certain domains, and if it is too expensive to pre-train data in the domain by ourselves, maybe we can choose large models for training

Domain-specific secondary pre-training can similarly improve the recall of the model. We are not surprised by this performance; it has been verified for a long time that domain-specific pre-training improves model performance, and we use more experimental data here to verify this idea. The performance improvement is mainly the input of more pre-training data, especially domain-related information, allowing the model to understand a particular domain better to make more reasonable predictions.

Chapter 5

Conclusion

To summarize, this thesis makes three hypotheses for the problem of BERT-based model selection in high recall information retrieval based on social media data. The first is that choosing the cased version of the model helps achieve better results in high recall information retrieval based on social media data because of the more information brought by the capitalized characters.

The second is not to blindly choose a large model because, in many cases, we do not have a large training set for high recall information retrieval tasks, and large models tend to overfit, resulting in even worse performance than small models.

The third is to try to choose a model that has been pre-trained with data in the specific domain for the task. Using a model that has been pre-trained by others does not cost more resources than an ordinary BERT model. The second pre-training can bring immediate performance improvement.

5.1 Future Work

In future work, we will invite experts in the field to annotate the data of each iteration, based on which we will verify the reliability of our theory through the performance of the model. With the involvement of domain-experts, we can implement the actual high recall information retrieval building process and compare our results with the previous work to see how much these suggestions can help us to improve the performance.

Another essential point to note is that we only performed experimental analysis for a specific task in this thesis. In future work, we will try to find other areas of high recall information retrieval tasks based on social media data, where more data will also help demonstrate our suggestions' reliability.

Bibliography

[1] Nlp4good/psych-search · hugging face.

[2] Zeynep Akkalyoncu Yilmaz, Shengjin Wang, Wei Yang, Haotian Zhang, and Jimmy Lin. Applying BERT to document retrieval with birch. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 19–24, Hong Kong, China, November 2019. Association for Computational Linguistics.

[3] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *EMNLP*. Association for Computational Linguistics, 2019.

[5] Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. Cost-effective selection of pretraining data: A case study of pretraining BERT on social media. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1675–1681, Online, November 2020. Association for Computational Linguistics.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[8] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[10] Muthukumar Rajendran. Topvis: Visual text analytics for deep topic modeling of reddit data. 2021.

[11] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.

[12] ThilinaRajapakse. Thilinarajapakse/simpletransformers: Transformers for classification, ner, qa, language modelling, language generation, t5, multi-modal, and conversational ai.

[13] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

Appendix A

Data set details

labeled data set The labeled data contains all the manually labeled data during the build process of previous work's high recall information retrieval classifier. It contains 1002 posts which are manually labeled relevant or non-relevant.

988 data set The 988 data set is extracted from outside the Canada subreddit but with the keywords as follows:

1. 'Depression & International student & canada',
2. 'Depressed & International student & canada',
3. 'Depressing & International student & canada'
4. 'Stress & International student canada',
5. 'Stressful & International student canada',
6. 'Stressed & International student canada',
7. 'Stressing & International student canada'
8. 'Anxiety & International student canada',
9. 'Anxious & International student canada',
10. 'Disorder & International student canada',
11. 'Bipolar & International student canada',
12. 'OCD & International student canada',
13. 'Psychiatry & International student canada',
14. 'Psychiatrist & International student canada',

15. 'Psychological & International student canada',
16. 'Mentalhealth & International student canada',
17. 'mental illness & International student canada',
18. 'mental health & International student canada',
19. 'mental problem & International student canada',
20. 'mental awareness & International student canada',
21. 'mental condition & International student canada',
22. 'Anorexia & International student & canada',
23. 'distressful & International student & canada',

The data based on the above filtration sums to 988 posts which are related to international students and mental health in Canada.

TSS data set The TSS data set we are using as the validation set is made by train test split, we separate the labeled data with a test size = 0.2, which means we have a 801 posts training set and a test data set contains 201 posts.

original data set The original data set in our thesis is a data set which contains 5806 posts from the Canada subreddit. It is a subset of the original data set extract with Reddit API. We apply a simple keyword search to find the posts related to international student and build this data set. The keywords we are using is "International student" and "International students".

More information about the data set can be found in the appendix of Muthukumar's thesis.[10]

Appendix B

Experiment details

The code and data sets that can help reproduce the experiments can be found in the GitHub repository.

`https://github.com/ArcherZY80/BERT-based-models-HRIR`

The model we are using in this thesis are from Hugging face[13], we are using two different libraries to access those models, SimpleTransformer and Transformer, the details are as follows.

In the Table B.1, "1" stands for using the library and "0" stands for not using. We choose SimpleTransformer[12] for most of the model since it is easier to use, for the Psych and SciBERT, we cannot access them with SimpeTransformer so we decided to use Transformer library instead.

B.1 First simulation

We run all of our experiment code with Google colaboratory. The detailed model's hyper parameters are as follows :

For SimpleTransformer:

train_batch_size=2,

gradient_accumulation_steps=16,

learning_rate=3e-5,

num_train_epochs= 3,

max_seq_length= 512

For Transformer:

per_device_train_batch_size=2,

per_device_eval_batch_size=2,

num_train_epochs=3,

seed=0,

load_best_model_at_end=True,

	SimpleTransformer	Transformer
RoBERTa-base [9]	Yes	No
RoBERTa-large	Yes	No
BERT-base [6]	Yes	No
BERT-large	Yes	No
SciBERT [4]	No	Yes
Psych [1]	No	Yes
ALBERT-l [8]	Yes	No
ALBERT-xxl	Yes	No
DeBERTa-base [7]	Yes	No
DeBERTa-large	Yes	No
BERT-base-cased [6]	Yes	No
BERT-large-cased	Yes	No

Table B.1: The library used for different models

learning_rate=3e-5,

gradient_accumulation_steps=16

B.2 Second Simulation

We first conduct some extra experiments based on the first simulation experiment, the extra experiment parameters settings are:

For SimpleTransformer:

train_batch_size=2,

gradient_accumulation_steps=16,

learning_rate=3e-5,

num_train_epochs= 6,

max_seq_length= 512

For Transformer:

per_device_train_batch_size=2,

per_device_eval_batch_size=2,

num_train_epochs=6,

seed=0,

load_best_model_at_end=True,

learning_rate=3e-5,

gradient_accumulation_steps=16

We then did three iterations for the select models, the hyper parameters settings for the simulation are:

For SimpleTransformer:

train_batch_size=4,

gradient_accumulation_steps=16,

learning_rate=3e-5,

num_train_epochs= 3,

max_seq_length= 512

For Transformer:

per_device_train_batch_size=4,

per_device_eval_batch_size=4,

num_train_epochs=3,

seed=0,

load_best_model_at_end=True,

learning_rate=3e-5,

gradient_accumulation_steps=16