The University of Hong Kong

COMP4801 Final Year Project

Deep Learning on Social Media Discussion

Project Plan

Prepared by
Lee Long Hin 3035475216
Tsui Wai Kin 3035436741

Supervised by
Dr. K.P. Chow

October 3, 2020

Table of Contents

1. Introduction

1.1 Background

In the last few years, the use of social media has become a necessary daily activity. It changes our ways to organize and communicate. This project aims to build a model that incorporate deep learning techniques together with natural language processing (NLP) to analyze social media discussions with the purpose to identify topics automatically and determine the trend of the discussion.

Social media plays an essential role in communication nowadays. Interactive platforms are comprised of creating and publishing content, and it allows users to share and exchange ideas, which facilitate social interaction. It not only aggregates opinions and feelings of diverse groups of people but also influences consumer's purchase decisions through marketing and advertising. People around the world can be easily connected with little limitation using these platforms. In particular, social media significantly improves and manipulates to some extent, our ability to communicate, spread and receive information.

One of the most popular social media websites, Facebook, has demonstrated its influence among society. According to the statistic from Alexa Internet, Inc. [1], Facebook is in the top 10 most-visited websites in the world. Facebook, there are more than 2 billion daily active users among the world and 4 petabytes (10^{15} bytes) of data are created daily [2]. Different social media forums specialize in specific areas, for example, sports, politics and games. As there are a lot of opinions and experiences flowing through social media, a wide range of viewpoints are aggregated.

1.2 LIHKG forum

In the project, data will be extracted from a well-known forum in Hong Kong, called "LIHKG" as known as "連登" in Chinese. Users of the forum discuss mostly in Cantonese. There are different categories for users to post in [3]. Registered users can share their opinion with others. A Hong Kong internet service provider (ISP) email or a University email is mandatory to sign up as a registered user. Regular users can read the forum content instantly.

1.3 Bidirectional encoder representation from transformer (BERT)

This project will mainly apply the Bidirectional Encoder Representations from Transformers (BERT) model to achieve the goals. BERT is the first finetuning based model to perform sentence-level tasks [4] and a strategy based on pattern recognition to complete different tasks, including but not limited to sentiment analysis and language inference. It makes use an attention mechanism, known as the Transformer, to learn the subject in a text. BERT model is different from a directional model that reads input unidirectionally (in either side). The transformer encoder in BERT reads the whole sequence of text at that same time [5]. It is considered as a bidirectional model.

1.3.1 BERT model architecture

This section will introduce the details of the BERT model architecture.

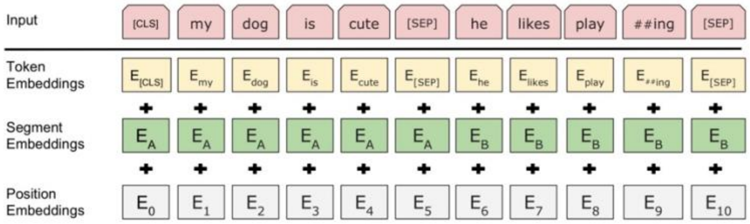
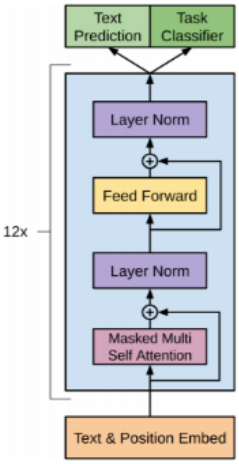

Input
[CLS]
my
dog
is
cute
[SEP]
he
likes
play
##ing
[SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{dog} E_{is} E_{cute} $E_{[SEP]}$ E_{he} E_{likes} E_{play} $E_{\#\#ing}$ $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

Figure 1: Architecture of Word Embedding in BERT [6]

The first layer of BERT is word embedding which consist of 3 main steps. They are token embeddings, segment embeddings and position embeddings (Figure 1). Each of them captures certain information of the input sequence. First, BERT uses WordPiece to tokenize each input word, such as ‘playing’, split into ‘play’ and ‘##ing’, each token is indexed according to the BERT corpus, and each pair of sentences is splited with a separator token [SEP] [7]. There is a special token [CLS] embedded at the beginning of each input. It will aggregate information from the entire sequence for classification tasks. Then, in segment embeddings, each word of the first sentence will be marked as ‘A’ and second sentence as ‘B’ and so on (if there is more segments). The position embedding take cares of the position of each word in the sentence. The final embeddings will be the sum of these 3 parts [8]. It translates the input sequence into number representations, known as vectors, for further analysis. These embeddings are useful for semantic search and information retrieval.



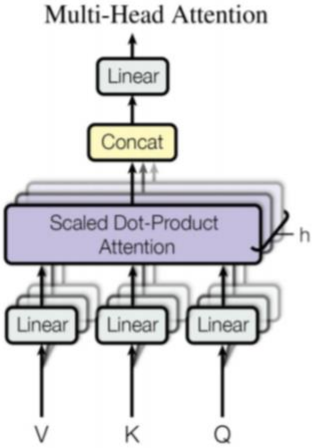


Figure 2: BERT model Transformer architecture [9]

Figure 3: Multi-Head Attention layer detailed view [10]

BERT makes use of the transformer to achieve different tasks (Figure 3Figure 2). There are N encoder blocks tie together to generate the output. Each block is responsible to not only find the relationship between the input representations but also encode them to the output.

The Multi-Head Attention mechanism is one of the crucial features implemented in the Transformer of BERT [11]. It computes the attention of the vectors several times with different weight. It concatenates the results so that the model can address information from various input at different positions (Figure 3Figure 3). It results in increasing parallelism and reducing training time as the whole sequence can be process at the same time.

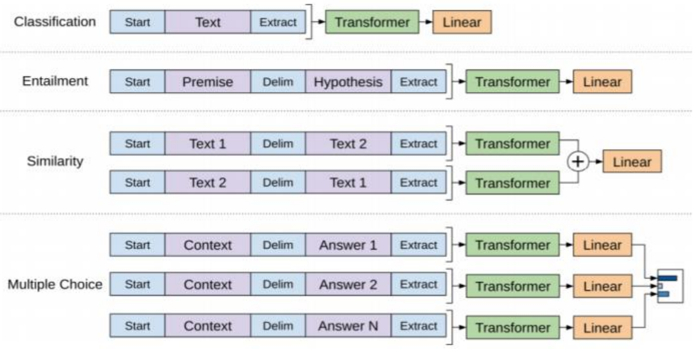


Figure 4: fine-tuning on different tasks. [9]

The model is then fine-tuned to achieve specific task by adding layers to the model (Figure 4Figure 4). For example, for a Question Answering task, we can fine-tune with adding a softmax layer to normalize the result from BERT and produce a distribution of reasonable answer.

1.3.2 BERT pre-training process

It is common to use pretrained BERT model as it is expensive to train one. Plenty of pretrained models in various languages are available online.

BERT uses millions of words to perform the masked language modelling and "next sentence prediction" in the pre-training procedure [12]. It is pre-trained by considering the surrounding context of unlabelled texts. The masked language model pre-training of BERT masks some tokens from the input randomly and predict the original word based on its surrounding context (Figure 5). Next Sentence Prediction is another important pre-training technique in BERT (Figure 6). It is because lots of NLP tasks rely on the relationship between sentences such as text inference. The input of the pre-training process involves a pair of sentences which separated by a special token [SEP] and appended with [CLS] [13]. The model is trained to predict whether the second sentence is possible to be the subsequent sentence of the first one.

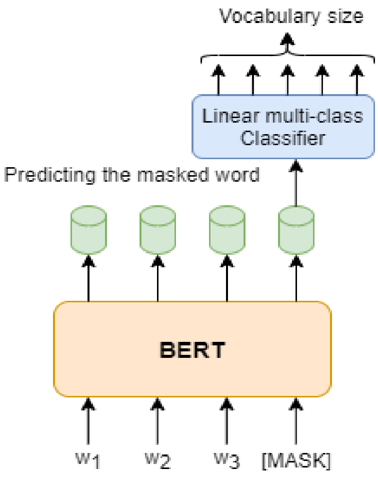


Figure 5: Masked Language Model

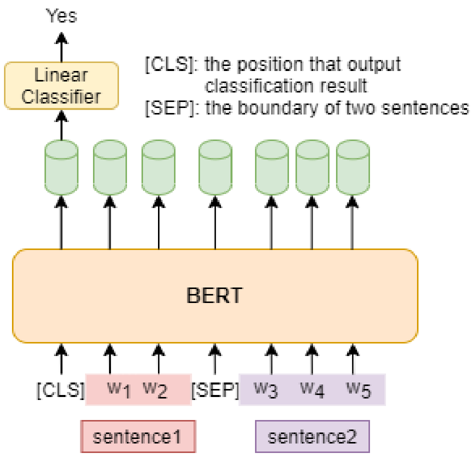

Yes
Linear Classifier
[CLS]: the position that output classification result
[SEP]: the boundary of two sentences
BERT
[CLS]
w_1
w_2
[SEP]
w_3
w_4
w_5
sentence1
sentence2

Figure 6: Next Sentence Prediction

1.4 Research on existing solutions

There are several studies using deep learning to predict public attention in social media discussion. Guandan Chen et al. [14] use data from Twitter to analyze the discussion. They proposed a neural model for predicting the popularity of social media content, which incorporate time-series information and time embedding enhanced RNN. They also reduce the noise of the text by using an attention mechanism.

Y. Liu and M. Lapata [15] proposed a modified encoder based on BERT that able to obtain the sentences representations and figure out the semantics of a text. They show the way to summarize the document from abstractive and extractive settings by using the modified encoder. Abstractive summarization setting processes sequence to sequence task. It takes a series of inputs. Then, the encoder maps inputs to a set of representations, and the decoder produces a target summary from the representations. Extractive summarization setting recognizes the critical sentences in a document. The encoder generates the classifier and representations for those sentences to predict which sentence to be the representative of the document summary.

1.5 Motivation

Social media platforms have increasingly grown beyond personal use. Recently, the data on social media can lead to useful predictions in many aspects, such as finance, marketing and politics. It is critical for businesses to recognize new market prospects by predicting the trend automatically. Analyzing the content on social media allow us to discover social structure characteristics.

There are good reasons for identifying topics automatically. Firstly, using a machine to perform automatic prediction require much lower cost than human. Secondly, events with extremely small or high probabilities are poorly predicted by a human. Thirdly, human usually makes decisions based on their interests and benefit, not purely by objective reasoning. Lastly, automatic prediction can provide a quicker response and process more massive amounts of data efficiently.

The remaining of the project plan proceed as follows. First, we will illustrate the objective of the project. Then, the methodologies of the implementation will be discussed. Subsequently, the project plan will end with a tentative schedule of the project, which shows how future work is going to be conducted.

2. Objective

This project aims to set up a model to identify the trend of the discussion of a day of current affair channel. Level of popularity of the content can be measured by quantitative representations, such as number of upvotes. By analyzing the trend, we can understand the public interest behind user interactions. If time allows, we will also predict the change of trend of a week and month. On top of that, we will also classify human emotion through sentiment analysis to identify social sentiment towards the topic. It may provide more insight about the discussion.

3. Methodology

This section explains the general procedures and technologies that will be included in this project. The project will be divided into two parts: data preparation, model implementation. The second part will be iterated continuously in order to get better results and more functionalities. The flow of the project is shown in Figure 7. The following paragraphs will illustrate each part in detail. Please note they are prone to change since there are uncertainties in each procedure.

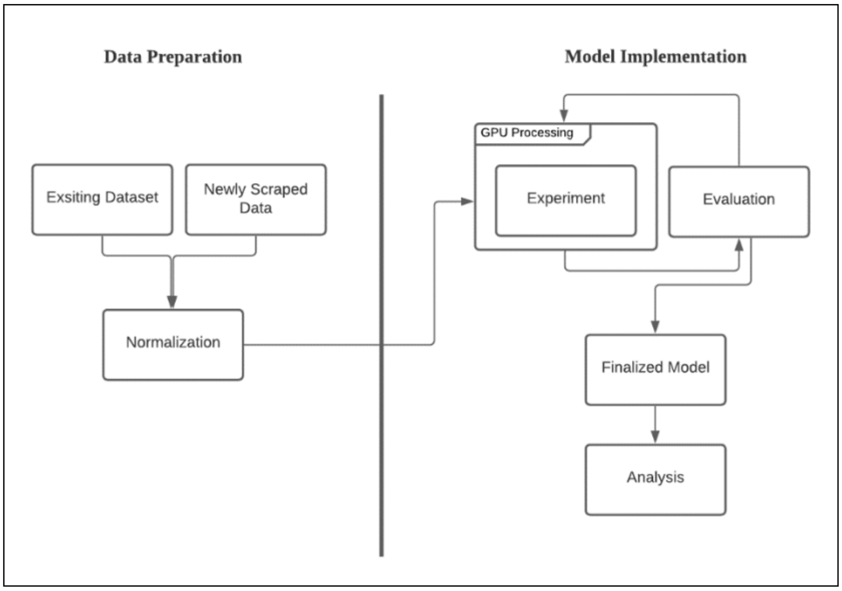


Figure 7: The Flow of the Project

3.1 Data Preparation

Posts and comments on LIHKG will be scrapped and processed in this phrase. At the moment, an existing dataset is available. It consists of 7992 comments and their corresponding timestamp. Although the size of the dataset is sufficient for the tasks, important information is missing in the data, such as the number of likes and dislikes and the topic of the post. Scrapping might be done to get a new set of data. It is proposed that only the Current Affair channel will be scrapped as the size will be huge if all the channels on the forum are included.

3.2 Model Implementation

This part includes both model experimentation and evaluation. These two processes will be iterated continuously. Different models and methods related to BERT will be tried in order to get the best result. Since the project is still in its exploration stage, areas with potentials will be listed in the following paragraphs.

For all methods, if applicable, 75% of the data will be used as training data while the rest 25% will be used as testing data.

3.2.1 Finetuning BERT

Pretrained BERT models will be used, and transfer learning might be performed due to the expensive cost of training a model from scratch. Although a majority of the BERT application is performing tasks in English, models trained on Chinese corpus are available. In particular, “bert-base-Chinese” [16] and “bert-base-Cantonese” [17] will be tested. “bert-base-Chinese” is trained on cased Chinese Simplified and Traditional text while “bert-base-Cantonese” has limited documentation. There exists a significant difference in the training corpus and the content in the LIHKG forum, as people discuss in spoken Cantonese there. Thus, transfer training might be performed on both models in order to allow the model to adapt to the content of the discussion. If the performance is not ideal and time permits, “Chinese-BERT-wwm” (whole word mask) [18] and some multilingual models might also be tested.

Then, the model will be fine-tuned to suit the tasks. As for identifying the trends, a few methods of fine-tuning will be researched and experimented on. They are named entity recognition, text summarization and keyword extraction. On the other hand, sentiment analysis of the topic will be done by fine-tuning BERT with classification.

3.2.2 Features of LIHKG

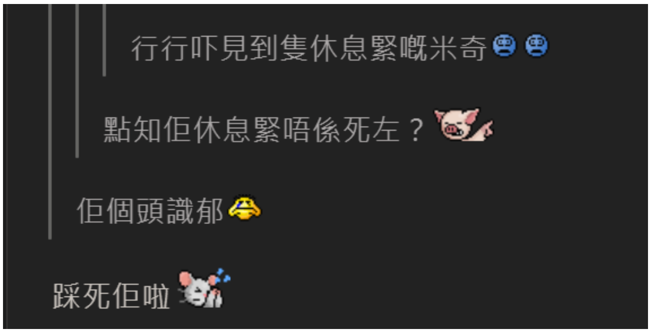
The features of LIHKG will also be considered in the process of model implementation. There is possible information to be extracted. For example, LIHKG allows user to like and dislike a comment. The difference in likes and dislikes might be included as an input for the classification.

3.2.3 Evaluation

Models will be repeatedly evaluated during model implementation. As for trends identification, basic natural learning process techniques might be used to generate testing data since the feasibility of doing so manually is questionable. Techniques might include word frequencies, finding n-gram word and removing stop word etc.

3.2.4 Possible Obstacles / Limitations

The BERT model might fail to capture some of the features in LIHKG discussions. Due to Hong Kong culture, it is common to see comments on the forum contain both Chinese and English. For example, “我 log in 唔到 moodle” (I cannot log in moodle). There are difficulties in tuning the model to understand both languages. It may, otherwise, require additional work to filter out the words. Besides, the comments in LIHKG are mostly written in spoken Cantonese, such as “小妹啱啱入咗城大” (I just got into CityU). Its meaning is different from verbal Chinese. Moreover, there are lots of slang in LIHKG discussion, such as “lm”, “ching”. In the forum, they might be used in a sentence normally as if any ordinary word. As they never appear in written Chinese and so the training corpus, the model will fail to understand their meaning.



行行吓見到隻休息緊嘅米奇

點知佢休息緊唔係死左？

佢個頭識郁

踩死佢啦

Figure 8: An example of conversation with emoji in LIHKG [19]

In addition, non-textual contents cannot be analyzed. Emoji are some commonly used symbols in a sentence (Figure 8). People usually use emoji to represent their subjective preference. It implies critical information on how people react to that topic, and it may also change the context of the sentence. Images, memes and GIFs are also widely used in the discussion. People can demonstrate feelings and emotions by using these tools. It might add additional content to the discussion. Because of its complexity, the complete content of the discussion might not be extracted.

3.3 Hardware Setup

Apple MacBook Pro 2017 is our primary tool to run the model. Its base configuration comes with 2.5 GHz Dual-Core Intel Core i7, 8 GB 2133 MHz LPDDR3, Intel Iris Plus Graphics 640 1536 MB.

Considering the demanding processing power, one of the two GPU farms offered by the Department of Computer Science of HKU will be used to train the model [20]. The computer cluster consists with 100 pieces of NVIDIA GeForce GTX 1080 Ti GPU cards/ NIVIDA GeForce RTX 2080 Ti GPU cards installed in 25 nodes, 35.5 Tera-FLOPS FP64/ 44 Tera-FLOPS FP64 processing power and 6 file servers providing 576 TB usable disk storage in total. It dramatically accelerates the training of neural networks.

4. Schedule and Milestones

The tentative schedule and milestones are shown in the table below and they are subject to change with different situations in the future:

Time	Task & Milestones
September 2020	• Ideation and literature review • Project Website creation • Detailed project plan
October 2020 – December 2020	• Research on datasets • Data Cleaning, normalization • Python package practice • 1st model Prototype
January 2021 – February 2021	• Modification of the model • First Presentation • Detailed interim report
March 2021	• Finalizing the model • Evaluation of the model
April 2021 – May 2021	• Final report • Final presentation

5. References

[1] Alexa Internet Inc, "Alexa Top 500 Global Sites," Alexa Internet Inc, [Online]. Available: http://www.alexa.com/topsites. [Accessed 19 September 2020].

[2] M. Osman, "Wild and Interesting Facebook Statistics and Facts (2020)," [Online]. Available: https://kinsta.com/blog/facebook-statistics/. [Accessed 19 September 2020].

[3] 出嚟食飯, "ELECTRA from Hong Kong Data," [Online]. Available: https://medium.com/@kyubi_fox/electra-from-hong-kong-data-ff268f55697. [Accessed 2020 September 27].

[4] S. Chi, Q. Xipeng, X. Yige and H. Xuanjing, "How to Fine-Tune BERT for Text Classification?," arXiv.org, 5 Feburary 2020. [Online]. Available: http://cips-cl.org/static/anthology/CCL-2019/CCL-19-141.pdf. [Accessed 28 September 2020].

[5] R. Khandelwal, "Intuitive Explanation of BERT- Bidirectional Transformers for NLP," April 2017. [Online]. Available: https://towardsdatascience.com/intuitive-explanation-of-bert-bidirectional-transformers-for-nlp-cdc1efc69c1e. [Accessed 28 September 2020].

[6] S. Kalra, "BERT Language Model," 27 Feburary 2019. [Online]. Available: https://medium.com/@shreyasikalra25/predict-movie-reviews-with-bert-88d8b79f5718. [Accessed 28 September 2020].

[7] Mohd Shukri H., "Why BERT has 3 Embedding Layers and Their Implementation Details," MC.AI, 19 Feburary 2019. [Online]. Available: https://mc.ai/why-bert-has-3-embedding-layers-and-their-implementation-details/. [Accessed 28 September 2020].

[8] C. McCormick, "BERT Word Embeddings Tutorial," [Online]. Available: https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/. [Accessed 29 September 2020].

[9] A. Radford, K. Narasimhan, T. Salimans and I. Sutskever, "Improving Language Understanding," [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. [Accessed 28 September 2020].

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need," arXiv.org, 6 December 2017. [Online]. Available: https://arxiv.org/abs/1706.03762. [Accessed 28 Septemeber 2020].

[11] A. Prakash, "BERT: Bidirectional Encoder Representations from Transformers," [Online]. Available: https://medium.com/swlh/bert-bidirectional-encoder-representations-from-transformers-c1ba3ef5e2f4. [Accessed 27 September 2020].

[12] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arxiv.org, 24 May 2019. [Online]. Available: https://arxiv.org/abs/1810.04805. [Accessed 30 September 2020].

[13] R. Horev, "BERT Explained: State of the art language model for NLP," Towards Data Science, [Online]. Available: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270. [Accessed 19 September 2020].

[14] C. Guandan, K. Q. chao, X. Nan and M. Wenji, "NPP: A neural popularity prediction model for social media content," ScienceDirect, 29 December 2018. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0925231218314942. [Accessed 19 September 2020].

[15] Y. Liu and M. Lapata, "Text Summarization with Pretrained Encoders," arxiv.org, 5 September 2019. [Online]. Available: https://arxiv.org/pdf/1908.08345.pdf. [Accessed 28 September 2020].

[16] Hugging Face, "Pretrained models - transformers 3.3.0 documentation," [Online]. Available: https://huggingface.co/transformers/pretrained_models.html. [Accessed 29 September 2020].

[17] Z. Lai, "Model: depa92/bert-base-cantonese," [Online]. Available: https://huggingface.co/denpa92/bert-base-cantonese. [Accessed 29 September 2020].

[18] Joint Laboratory of HIT and iFLYTEK Research (HFL), "Chinese-BERT-wwm," [Online]. Available: https://github/yumchi/Chinese-BERT-wwm/. [Accessed 29 September 2020].

[19] LIHKG, "[行路拎] 寶琳行去屯門," [Online]. Available: https://lihkg.com/thread/2226149/page/2. [Accessed 29 September 2020].

[20] Department of Computer Science, Faculty of Engineering, The University of Hong Kong, "HKU CS GPU Farm," [Online]. Available: https://www.cs.hku.hk/gpu-farm/home. [Accessed 29 September 2020].