COMP4211 Project Proposal

Topic: Natural Language Processing (NLP) in Text Classification

CHAN, Chun Hin (20853893)
LAW, Hui Nok (20860535)

Background Information

You may have already heard of different kinds of chatbots, such as ChatGPT, Llama, Gemini, ERNIE etc., as well as different voice assistants like Siri, Google Assistant, and Amazon Alexa, in your surroundings. You may wonder how they can communicate with us naturally, just like you are talking to your human friends at university. All of them use the technique of natural language processing.

Natural language processing can be dated back to the 1940s. At that time, there was a famous test called the Turing Test, which was aimed at testing whether a machine has the intelligence to behave equivalently as a human. In a nutshell, a human is masked and he receives two responses (one from the machine while the other is from another human). If the machine's reply can imitate a human's behaviour so well that a human cannot distinguish whether the responses are from a machine or a human, this machine passed the Turing Test.

Later in the 1960s, one of the world's first chatbots called "ELIZA" was born. It showcased the ability to talk with humans in a very natural way using natural language processing. Due to its excellence in pretending to have human-like psychological thinking, this has given rise to the development of natural language processing.

Formulation of the application as a machine learning problem

The application of natural language processing is very wide in machine learning problems. Here, we will list out the top 3 applications nowadays.

1. Sentiment Analysis

Usually, sentiment analysis involves analyzing the main idea of each opinion, commenting and reviewing different things. This often includes deciding the emotional tone of the opinion, based on the use of wordings, attributes and tones. It is quite common to see that this is widely used in analyzing the reviews and comments on the products. This can help firms understand how the customers comment about the product in general, thus using appropriate strategies to improve their products.

2. Text Classification

Based on the words input by the user, this can allow the model to classify whether the piece of text is positive, neutral or negative. In general, this is widely used in social media. It is because it is unavoidable that some of the comments under some posts contain inappropriate comments, such as terrorism, pornography, spam etc. If comments of negative attitude are detected, this can filter out these comments.

3. Chatbots

Based on what the user has input, the chatbot would need to digest the content and give the best response to the user. Currently, chatbots have been more important in business aspects because they are capable of understanding customers' needs and giving the corresponding solutions in a real-time manner.

Description of related applications of machine learning in the literature

1. Sentiment Analysis (Nasukawa & Yi, 2003)

The article describes the sentiment analysis in detecting the trends for specific subjects over time and conducting a qualitative analysis of the result. In the experience, it utilizes four camera brands' by comparing the favorability between the NLP analysis and real human counting on the webpage's review.

The result shows that the ratio of favourite sentiment in NLP is comparable to the human evaluation which concludes that it is possible to analyze the market trend through NLP sentiment analysis.

2. Text Classification (Alhogail & Alsabih, 2021)

Since the phishing email problem is an important cybercrime in the digital world, this article showcases the usage of NLP GCN with deep learning to classify phishing emails from an English corpus perspective. By utilizing their model which results in accuracy, precision, recall and f-measure are all higher than 98%. The proposed detection classifier detects the body text data which do not been seen before, so they assume that the model can have a good performance in detecting zero-day phishing attacks.

3. Chatbots (Adamopoulou & Moussiades, 2020)

The article states serval applications on chatbots by utilizing the NLP technique. The first application is in education, chatbots can provide the information that students missed by repeating the previous lecture and answering questions whenever students get confused. It can boost the comprehensive learning of students.

Moreover, chatbots can help in customer service by always interacting with clients which can reduce a lot of waiting time for getting a response.

Furthermore, chatbots can perform health diagnoses by asking about the patient's symptoms and providing immediate treatments or suggestions. Also, it can be deployed to provide useful information during the pandemic since it will not be infected.

Description of the dataset

The dataset is retrieved from this website: https://huggingface.co/datasets/ag_news.

For the training dataset, there are 120000 instances. While for the testing dataset, there are 7600 instances.

Each instance contains 2 feature columns. The first feature column is the title of the news article, which is of string data type. The second feature column is the class labels that specify the categories of the news title, which is of int data type. There are 4 class labels in total, where '0' stands for "World", '1' stands for "Sports", '2' stands for "Business" and '3' stands for "Sci/Tech" respectively.

Expected computing resources needed

We expect we would need to use Google Colaboratory (CPU runtime) as our main computing resource for our project. It is because in most cases, small projects related to natural language processing only require CPU but rarely use GPU.

In case we need to access GPU, we can switch to the GPU runtime in Google Colaboratory or use the GPU quota from Kaggle.

Google Colaboratory Services: CPU with 12.7 GB RAM
Backup GPU plan from Google Colaboratory Services: T4 GPU
Backup GPU plan from Kaggle Services: T4 GPU

Reference Material:

1. Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. Machine Learning with applications, 2, 100006.
2. Alhogail, A., & Alsabih, A. (2021). Applying machine learning and natural language processing to detect phishing email. *Computers & Security*, *110*, 102414-. https://doi.org/10.1016/j.cose.2021.102414
3. Nasukawa, T., & Yi, J. (2003). Sentiment analysis: capturing favorability using natural language processing. *International Conference On Knowledge Capture: Proceedings of the 2nd International Conference on Knowledge Capture; 23-25 Oct. 2003*, 70–77. https://doi.org/10.1145/945645.945658