Proposal for final project

Fake news specifically refers to news reports that are untrue or exaggerated. These reports may be deliberately created to mislead the public or promote a specific agenda, through traditional media channels like print, and television as well as non-traditional media channels like social media The wide and fast spread of fake news online has posed real-world threats in critical domains like politics, economy, and public health.

There are organizations, like the House of Commons and the Crosscheck project, trying to deal with issues as confirming authors are accountable. However, their scope is so limited because they depend on human manual detection, in a globe with millions of articles either removed or being published every minute, this cannot be accountable or feasible manually. Automatic fake news detection, which aims at distinguishing inaccurate and intentionally misleading news items from others automatically, has been a promising solution in practice. Though much progress has been made, understanding and characterizing fake news is still challenging for current models. This is caused by the complexity of the news-faking process: Fake news creators might manipulate any part of the news, using diverse writing strategies and being driven by inscrutable underlying aims.

Textual content-based fake news detection methods are mainly dependent on the features extracted from the text that the classifier relies on in identifying fake news, such as linguistic features and syntactic features, sentiment features, or features based on the style and quality of the writing. Therefore, textual content-based method is promising for detection of fake news. Large language models (LLMs, which are usually trained on the larger-scale corpus and aligned with human preferences, have shown impressive emergent abilities on various tasks and are considered promising as general task solvers. It is interesting to build a LMM model to detect fake news and compare its performance comparing to machine learning classification models such as logistic regression, random forest and LSTM, etc.

In this project, I propose to build a fake news detection model by fine tuning Llama 2 with Lora (Low rank adaption). The Llama2 model was proposed in LLaMA: Open Foundation and Fine-Tuned Chat Models, which is a collection of pretrained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters. Llama 2, like the original Llama model, is based on the Google transformer architecture, with improvements. Llama's improvements include RMSNorm pre-normalization, inspired by GPT-3; a SwiGLU activation function, inspired by Google's PaLM; multi-query attention instead of multi-head attention; and rotary positional embeddings (RoPE), inspired by GPT Neo. Llama training used the AdamW optimizer. Llama 2's primary differences from Llama are increased context length (4096 vs. 2048 tokens) and grouped-query attention (GQA) instead of multi-query attention (MQA) in the two larger models. LoRA (Low-Rank Adaptation of Large Language Models) is a popular and lightweight training technique that significantly reduces the number of trainable parameters. It works by inserting a smaller number of new weights into the model and only these are trained. This makes training with LoRA much faster, memory-efficient, and produces smaller model weights (a few hundred MBs), which are easier to store and share.

I will first explore the fake news data set to check the data characteristics. I will ran basic preprocess to filter out low quality features and records. After that I will check the difference of distribution of these characteristics between true and fake news. If some feature/characteristics are different between fake and true news, I will consider to include in the model. After that I will ran semantic analysis to further explore semantic features. The first step of semantic analysis is text preprocessing to filter out number, rare words, punctuation, stop words etc. After that I will run words counts visualization analysis such as word cloud. Thirdly, I will normalize the data based on tf-idf and further explore its corresponding characteristics.

I will build fake news classification model by fine tuning Llama2 model by lora. After that I will compare its performance with machine learning classification model such as logistic regression model, SVM and random forest based on metrics such as accuracy, F1 score and AUC, etc.