Can AI help us in identifying misinformation and conspiracies?

Misinformation is considered to be the top global challenge. It is ubiquitous and spreads quickly, due to the popularity and limited regulatory measures of social media. Misinformation manipulates facts, which in turn triggers emotions and sentiments in citizens that can convince them to falsely believe in a topic. As a result, trust is eroded, causing harm to society. Taking advantage of the fact that fake news eliciting moral outrage is likely to generate many reshares, rumourmongers can ensure that appropriately worded false information diffuses widely within a very short space of time. However, identifying rumours and fake news among the huge volumes of information circulating on social media is highly challenging.

In this project, we have investigated the development of Artificial Intelligence (AI) approaches based on Natural Language Processing (NLP) and Large Language Models (LLMs) to distinguish fake information from genuine facts, with a particular focus on a specific type of misinformation, i.e., conspiracy theories, whose false content is intended to cause harm.

Large Language Models

Our approach to detecting misinformation uses LLMs, which have revolutionised the field of AI and NLP. A widely known example of an LLM is ChatGPT. LLMs gain their "intelligence" through training on vast amounts of textual data, which allows them to build up highly detailed knowledge about how to recognise, interpret and generate human language.

Armed with these capabilities, LLMs can be used to answer many types of questions and to perform various types of tasks. In the NLP field, LLMs have been successfully used to detect a wide range of different types of information in text. In this project, we have developed a novel LLM for accurate misinformation detection, focusing on a specific type of misinformation, i.e., conspiracies, which are a deliberate misinformation act.

The sophistication of LLMs means that they can often achieve good results when applied to specific NLP tasks, without any task-specific adaption or training. Nevertheless, the success with which LLMs are able to carry out particular tasks can often be improved through *fine-tuning*. A commonly used technique for fine-tuning is *instruction tuning*, which involves providing the LLM with a set of tasks/instructions, and a corresponding set of desired outputs. Such instruction datasets provide the means for LLMs to learn how to carry out specific tasks more accurately.

As part of our work on this project, we have developed a number of instruction tuning datasets for conspiracy detection, which we have subsequently used to fine-tune LLMs to support misinformation detection.

Importance of sentiment and emotion in detecting misinformation and conspiracies

Text that conveys a particular type of information often exhibits specific characteristics or *features*. The automatic recognition and use of these features can contribute towards the accurate detection of the information of interest (in our case, misinformation).

Several studies have shown that sentiment and emotions (which we collectively refer to as *affective information*) constitute important textual features that are inextricably intertwined with

misinformation. Sentiment analysis (SA) and emotion detection (ED) are two types of NLP techniques for analysing human expressions that can help us to understand people's feelings towards specific topics. SA aims to capture the overall emotional tone conveyed by a data source (usually positive, negative, or neutral), along with the strength of this tone. ED is the process of classifying data at a finer-grained level, according to the emotions that it conveys. Compared to sentiment, the term *emotion* refers to more specific and stronger feelings. For example, positive sentiment encompasses a range of different emotions, such as happiness and joy, while negative sentiment includes the emotions of sadness and anger, among others.

Previous studies have shown that misinformation is generally associated with a significant level of high-arousal emotions, such as anger, sadness, anxiety, surprise, and fear. Rumours conveying anger, sadness, anxiety, and fear are likely to generate a large number of shares or re-tweets, and to be long-lived and viral, while emotional appeals (like anger and disgust) can increase users' engagement with fake posts. Fake news also expresses higher overall emotion, negative sentiment and lower positive sentiment than genuine news.

During this project, we have conducted our own affective analysis of tweets related to conspiracy theories, which revealed that such tweets predominantly convey negative sentiments and emotions (e.g., anger, fear and disgust), and are generally long-lived. In contrast, tweets that are unrelated to conspiracy theories are more likely to express positive sentiments and emotions (e.g., joy, love and optimism).

The findings above clearly demonstrate the strong links between various types of misinformation and affective information, and illustrate that a wide range of different emotions can be expressed in text that conveys misinformation. Accordingly, automated methods capable of carrying out a comprehensive, fine-grained affective analysis of text are needed to fully support accurate misinformation detection.

Although a range of previously proposed NLP-based approaches to automated misinformation detection have made use of affective information, we are not aware of any existing LLM-based approaches to misinformation detection that employ affective information.

Fine-tuning LLMs for comprehensive affective analysis of text

A number of previous studies have fine-tuned LLMs to perform certain affective analysis tasks, e.g., detection of positive or negative sentiment and recognition of specific emotions, such as anger or joy. However, accurate misinformation detection is reliant on more fine-grained affective information, such as the *strength* with which a particular sentiment is expressed, or the *intensity* with which a given emotion is conveyed. Prior to this project, there was a lack of fine-tuned LLMs that were able to perform such comprehensive affective analyses. This was mainly due to a lack of datasets that could be used for instruction tuning and for evaluating the performance of the fine-tuned models.

In response, we have developed novel instruction tuning datasets, and used them to fine-tune a set of novel, open-source *emotional* LLMs (EmoLLMs) to carry out a range of fine-grained affective analysis tasks, e.g., to assign numerical scores denoting the strengths with which particular emotions and sentiments are expressed in text. We have also developed an evaluation benchmark dataset, which we used to compare the performance of different approaches to recognising this detailed affective information. Our results showed that the EmoLLMs perform exceptionally well in carrying out fine-grained affective analyses. Our demonstration that EmoLLMs can mostly outperform both open and closed source general purpose LLMs ina range

of affective analysis tasks illustrates the positive impact of fine-tuning LLMs for these tasks, as well as the utility of the novel datasets that we have developed.

Fine-tuning EmoLLMs for conspiracy detection

We have carried out further work to demonstrate how the comprehensive affective information recognised by the EmoLLMs introduced above can be usefully exploited to detect various types of information relating to conspiracy theories. To facilitate this, we developed a further novel instruction tuning dataset, called *ConDID*, to support the fine tuning and evaluation of novel LLMs that are specialised to carry out several different tasks related to conspiracy theory detection. The tasks concern determining whether or not a given tweet is related to a conspiracy theory, the type of conspiracy theory that is discussed in the tweet, and the level of relatedness of the tweet to a conspiracy theory (i.e., *closely related*, *broadly related* or *not related*).

We used ConDID to further fine-tune the best-performing EmoLLM discussed in the previous section, to create the first emotion-aware LLM that is specialised for conspiracy detection (*ConspEmoLLM*). Given the previously discussed strong relationship between conspiracy theories and affective information, it was hypothesised that the detailed affective knowledge encoded within ConspEmoLLM would act as an aid to effective conspiracy theory detection. To test this hypothesis, we also used the ConDID dataset to fine-tune an LLM that does not use affective information (*ConspLLM*), and compared the performance of ConspEmoLLM and ConspLLM.

Our experimental results firstly demonstrate the importance of fine-tuning LLMs to achieve optimal performance in detecting information relating to conspiracy theories, since both ConspLLM and ConspEmoLLM are able to outperform open and closed source LLMs that had not been fine-tuned for these tasks. Moreover, ConspEmoLLM surpasses the performance of ConspLLM on most of the conspiracy-related tasks evaluated. This reinforces the importance of exploiting affective features as an aid to detecting various types of information relating to conspiracy theories, and also provides strong evidence to support out hypothesis that fine-tunning an emotion-aware LLM is a better solution then fine-tuning a general purpose LLM for the task of conspiracy theory detection.

Main contributions

Our work on this project has been concerned with harnessing the power of LLMs to facilitate the accurate automated detection of misinformation (in particular, conspiracy theories) in text. The development of such methods is of crucial importance to help to curb the spread of false information across the internet.

We have developed a number of novel resources, which we have demonstrated to be useful for detecting misinformation, including the first LLM for conspiracy detection that is based on fine-grained emotions. We hope that our work will promote further research into the use of LLMs and affective information to fight against the many harms that misinformation can bring.

We have shown that our novel instruction tuning datasets can be successfully employed in fine-tuning LLMs to carry out accurate and comprehensive affective analysis of text. Furthermore, given the strong links between affective features and misinformation, emotion-aware LLMs can form an important basis in optimising LLMs to detect different types of misinformation. We have conducted experiments with well-known conspiracy datasets and demonstrated that our method outperforms other approaches.

Future Work

In order to build upon the work carried out in this project, we aim to collaborate with other teams working on misinformation and conspiracies in areas such as political discourse, online bullying, harassment. This will allow us to investigate further how we can mitigate the societal impact of conspiracy threats using AI.