Centre for Digital Trust and Security

Seedcorn Final Report 23/24

Project Title:

Emotion Detection and Misinformation Harms arising from Large Language Models

Project Investigators:

Sophia Ananiadou, Peter Knight, Stephen Hutchings, Zhiwei Liu, Paul Thompson, Boyang Liu, Kailai Yang

Project overview:

Misinformation is ubiquitous. It manipulates the emotions and sentiments of citizens and can convince people to falsely believe in a topic, potentially eroding trust and thus 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, as is the application of regulatory measures to reduce their diffusion.

Large Language Models (LLMs) possess sophisticated language understanding capabilities acquired through training on vast amounts of text. However, in common with many AI technologies, LLMs can be viewed as both a blessing and curse, especially with regard to misinformation. The intelligent characteristics of LLMs can be exploited maliciously to rapidly generate false information that appears highly convincing. Furthermore, since LLMs are trained on web text that includes a certain proportion of misinformation, they may unintentionally generate misinformation during tasks such as automatic summarisation or question answering. LLMs could thus substantially exacerbate the misinformation problem, making it even easier for false information to infiltrate society. Accordingly, there is an urgent need to develop robust automated approaches to distinguish fake from genuine information. Luckily, the advanced capabilities of LLMs mean that they can also contribute positively towards the fight against the spread of rumors and fake news, by detecting misinformation automatically.

We are investigating how to best exploit LLMs for the automated detection and analysis of misinformation, building upon previous approaches based on conventional machine learning and deep learning. Misinformation detection is a muti-factorial problem, reliant not only on establishing whether or not a piece of text is factual, but also on determining a variety of features concerning both the textual content and structure of social media posts, which could interact to signal that information is fake. Developing an enhanced understanding of these features is a key aspect of developing accurate automated methods. In collaboration with social science scholars working on misinformation, disinformation, conspiracy theories, argumentation and trust, we have analysed collections of social media posts in topics surrounding the "Great Replacement" deep state and global elite conspiracies to identify a range of semantic, lexical and stylistic features that are characteristic of misinformation. These features include emotions, sentiment and stance, along with structural and discourse-level information, such as dialogue acts and temporal dynamics.

We have evaluated how interactions among different combinations of features, and instruction-tuning based on affective features have improved the recognition of misinformation and conspiracies using mainstream models such as LLaMa2, ChatGPT and Vicuna.

Key findings:

1. Development of a multitask conspiracy detection instruction dataset for fine-tuning LLMs (ConDID) based on affective features
2. Development of a multitask conspiracy detection LLM (ConspEmoLLM)
3. Application of ConspEmoLLM to five tasks related to conspiracy detection
4. Comparison of ConspEmoLLM to different LLMs and different conspiracy datasets to test its generalisation
5. Development of RAEmoLLM, the first retrieval augmented (RAG) LLMs framework to address cross-domain misinformation detection using in-context learning based on affective information. The framework was applied to different datasets, including the COCO dataset of conspiracy-related tweets, and showed significant improvements over zero-shot methods
6. Development of a multi-lingual dataset on the subject of the Great Replacement

Outputs to date:

1. Published a review on emotion-based methods for misinformation detection, which has been published in the journal *Information Fusion*:

   Zhiwei Liu, Tianlin Zhang, Kailai Yang, Paul Thompson, Zeping Yu and Sophia Ananiadou (2024). Emotion detection for misinformation: A review. *Information Fusion* (107): 102300
   https://www.sciencedirect.com/science/article/pii/S1566253524000782

2. Written a paper on the development of emotion-enhanced LLMs, which has been accepted for presentation at the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 24), to be held in Barcelona, Spain, from 25th - 29th August 2024.:

   Zhiwei Liu, Kailai Yang, Qianqian Xie,Tianlin Zhang and Sophia Ananiadou (2024). EmoLLMs: A Series of Emotional Large Language Models and Annotation Tools for Comprehensive Affective Analysis. To appear in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 24)*
   https://arxiv.org/abs/2401.08508

3. Written a paper on the application of emotion-enhanced LLMs to conspiracy theory detection, which has been accepted for presentation at the 13th Conference on Prestigious Applications of Intelligent Systems (PAIS-2024), which will be held in Santiago de Compostela, Spain, from 19th – 24th October 2024:

   Zhiwei Liu, Boyang Liu, Paul Thompson, Kailai Yang and Sophia Ananiadou (2024). ConspEmoLLM: Conspiracy Theory Detection Using an Emotion-Based Large Language Model. To appear in *Proceedings of the 13th Conference on Prestigious Applications of Intelligent Systems (PAIS-2024)*
   https://arxiv.org/abs/2403.06765

4. Produced a paper on the application of emotion-enhanced LLMs to cross-domain misinformation detection:

   Zhiwei Liu, Kailai Yang, Qianqian Xie, Christine de Kock, Sophia Ananiadou and Eduard Hovy (2024). RAEmoLLM: Retrieval Augmented LLMs for Cross-Domain Misinformation Detection Using In-Context Learning based on Emotional Information. arXiv
   https://arxiv.org/abs/2406.11093

5. Invited talk by Sophia Ananiadou, entitled *Emotion Detection and Misinformation Harms from Large Language Models*, at the 8th Annual Women in Data Science Event, held at the American University of Beirut, held on 22nd April 2024
   https://www.aub.edu.lb/osb/wids/WiDS2024/Pages/talkabstracts.aspx

6. Keynote talk by Sophia Ananiadou, entitled *Emotion Detection and LLMs: Transforming Mental Health and Countering Misinformation on Social Media,* at the 2nd Symposium on NLP for Social Good (NSG), held at University of Liverpool, United Kingdom on 25th – 26th April 2024 https://nlp4social.github.io/nlp4socialgood/

7. Keynote talk by Sophia Ananiadou, entitled *Emotion detection for Misinformation and Conspiracy Detection,* at the 5th ACM Europe Summer School in Data Science, held in Athens on July 8th-12th , 2024.
https://europe.acm.org/seasonal-schools/data-science/2024/lecturers

Were all planned outcomes achieved?

If not, how did you mitigate non-achievement?

All planned outcomes were achieved.

Planned activities post-project:

We plan to extend our initial plan to compare our models with output produced by the most recent LLMs and to evaluate our models on additional misinformation datasets and conspiracy tasks. We wish to apply for further funding from UKRI to allow us to further develop our research into detecting misinformation and conspiracies, to improve our LLMs, to develop annotated datasets concerning conspiracy theories relating to the Great Replacement and those found on Russian "proxy" websites, and to construct additional instruction tuning data to cater for a wider range of types of misinformation. We also wish to investigate how the automated identification of social moral principles in text (e.g, Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Purity/Degradation) could feed into improving the detection of conspiratorial information in text.