Project Number: P7

Project Title: Explainable Fake New Detection on Twitter

Project Clients: Dr Jiaojiao Jiang, School of Computer Science and Engineering

Project Specializations: Software development, Deep Learning, Data Science, and HCI;

Background:

Detecting fake news on social media is the process of analysing the news contents to determine the truthfulness of the news. The news could contain information in various formats such as text, video, image.

Previous fake news detection systems typically only give a final decision as to whether the news under investigation is fake or not, with little explanatory information revealed about why the decision is made. Such a coarse dichotomy does not account for the nature of fake news being a combination of mis/disinformation mixed with factual information and ignoring rich and highly contextualised psycho-social and cognitive factors.

Requirements and Scope:

The goal of this project is to explore interpretable machine-learning techniques and generate appropriate user-interpretable explanations. We will adopt the Linguistic Inquiry and Word Count's (LIWC) high-level language-bound psychological feature sets to extract the most explainable sentences out of a given news article and extract the most relevant psychological features to generate appropriate user-interpretable explanations. We will particularly focus on the social platform, Twitter, where we can crawl data with the open-source Twitter API. We aim to develop a tool for the automatic detection of fake news spreading on Twitter.

Required Knowledge and skills:

- Theoretical understanding of deep learning models.
- Hands on experience with deep learning models (e.g., Python libraries)
- Familiarity with UI design and development
- Interested students should have complete (or learned the topics covered in) and COMP9444 (or COMP3511) (or equivalent).

Expected outcomes/deliverables:

- Source code of developed models and its evaluation
- A tool for automatic fake news detection
- Project documentation in the form of written report