Project Details

Project Title

Multilingual COVID-19 Fake News Detection and Intervention

Project Abstract - 500 words

The emergence of the Coronavirus Disease 2019 (COVID-19) epidemic in late 2019 has produced massive information related to COVID-19. Information distribution platforms such as mass media and social media allow information to be spread widely. Unfortunately, not all of the information is accurate or trustworthy. Some of the information spreading around those platforms can be categorised as misinformation or even be identified as fake news. Particularly, different countries may have different situations and strategies to control the spread of COVID-19, which also leads to a huge amount of inappropriate news sharing. For example, “*The spread of COVID-19 is linked to 5G mobile networks*”, “*Place a halved onion in the corner of your room to catch the COVID-19 germs*”, “*Sunny weather protects you from COVID-19*”. These fake news stories and others like them spread rapidly on social media during the early stages of the pandemic. The wave of misinformation was so massive that the authorities have coined a word for it: “infodemic”. Meanwhile, a lot of fake news were produced in various languages so that they can spread more easily in particular ethnic groups. Clearly, the detection of multilingual COVID-19 fake news is essential for countries such as Australia and Indonesia which consist of hundreds of ethnic and linguistic groups. Unfortunately, existing fake news detection methods are too general to be effective in recognising COVID-19 related fake or inappropriate news in multiple languages. Thus, it is very challenging for the authoritative organisations to make timely responses to the spread of fake news. Failing to detect and intervene the spread of multilingual COVID-19 fake news can easily cause mistrust and wrong public responses, such as panic buying, non-compliance with social distancing, and even refuse to COVID tests and vaccinations.

Based on these emerging demands and to address the research challenges behind them, this project aims to (1) collect and analyse the data in Australia and Indonesia regarding individual online behaviour characteristics in COVID-19 news propagation and communication, (2) to conduct theoretical analysis about the existing fake news detection models, and design a multilingual COVID-19 fake news detection model using advanced machine learning techniques, and (3) to apply the proposed COVID-19 fake news detection method into the process of risk communication management to intervene the spread of fake news, and help authoritative organisations to generate their customised COVID-19 warning policies.

This project will use a computational-based detection, machine learning, and human factor engineering approach to design a decision support system that functions as a warning mechanism for misinformation and fake news related to COVID-19. The outcome of this project can be taken into practice by organisations and governments in determining the right risk communication and education strategy. By implementing an appropriate education strategy, public awareness of COVID-19 can increase so that the number of victims of fake news will decrease significantly.

Research Project Plan - 1500 words

Research Project Plan including the project scope, objectives, R&D activities, expected outcomes including proposed methods to publish and promote the research outcomes, and delivery milestones, including the ability to commence and complete research, or a distinct phase of the research, within a 12 month timeframe.

Project Scope

The emergence of the Coronavirus Disease 2019 (COVID-19) epidemic has produced massive information related to COVID-19. Information distribution platforms such as mass media and social media allow information to be spread widely. Unfortunately, not all of the information is accurate or trustworthy. Some of the information spreading around those platforms can be categorised as misinformation or even be fake news. Meanwhile, a lot of fake news were produced in various languages so that they can spread more easily in particular regions and linguistic groups. The detection of multilingual COVID-19 fake news is essential for countries such as Australia and Indonesia which consist of hundreds of ethnic and linguistic groups. However, the mainstream of existing fake news detection research mainly focuses on English based content which cannot recognise fake news spreading through multiple linguistic groups. Failing to detect and intervene the spread of multilingual COVID-19 fake news can easily cause mistrust and wrong public responses, such as panic buying, non-compliance with social distancing, and even refuse to COVID tests and vaccinations. This project aims to develop a complete set of strategies for multilingual COVID-19 fake news detection and intervention.

Objectives

This project aims to achieve the following objectives:

1. **To produce the datasets for multilingual COVID-19 fake news**: we plan to collect diverse types of users' information following and discussing COVID-19 outbreak in Australia and Indonesia based on users' privacy policies and agreement, and analyse their online behaviour characteristics in terms of COVID-19 news propagation, communication, and opinions. Users' sensitive information will be removed as their privacy is protected during dataset collection and use.
2. **To design and implement multilingual COVID-19 fake news detection model**: we plan to conduct theoretical analysis about the existing fake news detection models, and then develop a novel effective fake news detection model for society-related disaster events like COVID-19. To detect multilingual COVID-19 fake news, we also plan to develop a framework which supports advanced machine learning techniques such as transfer learning and federated learning techniques.
3. **To generate policy and guidance to intervene the spread of COVID-19 fake news**: to apply our research outcomes to help authoritative organisations and governments, we plan to apply our proposed multilingual fake news detection method into the process of risk communication management, and help to generate their customised COVID-19 warning policies to intervene the spread of COVID-19 fake news.

R&D activities

1. Dataset Generation and Preparation Stage

One of the key issues with detecting misinformation related to COVID-19 is that there is a lack of corpus to test methods for fake news detection. Therefore, at this stage, an online survey targeting 1500 Australian and Indonesian residents with different generation groups (including baby boomer, millenial and gen Z) from selected cities will be designed to classify individuals based on their characteristics in believing and processing COVID-19 outbreak-related information. This survey is also designed to identify and analyse their online behaviour characteristics in news propagation and communication, and help to determine which group is the most vulnerable to fake news. Accordingly, by identifying the most vulnerable group to fake news, the government will be able to generate policy and guidance strategies in risk communication during COVID-19 outbreak and similar situations in the future. This survey will be conducted via online questionnaire/survey platforms, following the user privacy protection policies.

Meanwhile, to validate the classification result from online survey, an experiment with 90 respondents (i.e. 30 respondents for each generation group) from some cities of each country with fairly high spread of COVID-19 will be conducted. In this experiment, a stimulus of information/news classified into true/factual (*signal*) and misinformation/fake (*noise*) related to COVID-19 will be given. Respondents will be asked to identify which of them to be *signal* or *noise*. Their response will be further analysed to determine both their ability to discern fake news and the most sensitive group to fake news.

2. Model Design and Implementation

2.1 Hoax characterisation

This stage is carried out to clearly understand the scope and varieties of hoaxes (intentionally misleading and deceptive fake news) over the time. In this stage, some important aspects to define the characteristics of a hoax are identified. The data to be retrieved includes location, time, language features, response features, user information, and so on. The raw data collected in the first stage is transformed into spatial-temporal data. Topic modelling, for example, Bidirectional Encoder Representations from Transformers (BERT) developed by Google, can be utilised to extract the main topic occurring at a specific time and location. Since it is very likely for local residents to generate hoax news for some particular topics at particular times, using spatial-temporal-topic data can help to mine the characteristics of hoax better than a single feature data.

2.2 Fake News detection and classification

With the support of the feature extraction and topic analysis in Task 2.1, we will further develop the effective deep learning models for detecting fake news. Here, we implement a recurrent neural network (RNN) with the embedded BERT-based spatial-temporal attention mechanism as the supervised fake news detection model. A new architecture of spatial-temporal BERT RNN will be proposed for the detection of COVID-19 fake news, specifically in Australia and Indonesia. Existing pre-trained models for fake news detection in other topics (such as politics and celebrity news) will also be applied as the benchmarks to further compare and evaluate the accuracy of the developed model.

2.3 Multilingual Fake News detection and classification Model Design

Since the variety and velocity of fake news keep changing, the existing detection methods and the region/language specific fake news detection methods in Task 2.2 may fail to detect misinformation in a multilingual context. To address such a problem, we plan to develop a macro-micro bi-level model that can be constructed and trained by the individual models generated in Task 2.2. The benefit of this design can maximise the users' privacy protection for each region where micro-level models are trained. Meanwhile, the macro-level model can be trained effectively by using the differential samples across multiple regions. To incorporate the multilingual transformation, we will advance the existing transfer learning and federated learning techniques by adding appropriate topic entity and relationship alignment attention into the architecture of spatial-temporal BERT RNN.

2.4 Experimental Evaluation, Framework Implementation and Optimisation

Comprehensive metrics will be used for evaluation of the proposed models, and correlation analysis will be performed to find out the influence of each characteristic and feature in determining hoax. To enable the use of research outcomes, we will develop a prototype system to encapsulate all the processes with a friendly interface. It will include data uploading module, data storage module, model training module, model parameter optimisation module, model sharing module, and suspicious fake news reporting module. Meanwhile, the system will be optimised to handle large-scale datasets using cloud servers and supercomputers.

3. Automatic Policy and Guidance Strategy Generation

The objective of this stage is to develop the policy, guidance and education strategy targeted for each generation group. This strategy will combine (a) an in-depth understanding of Australian and Indonesian residents' characteristics in processing information and discerning COVID-19 fake news, and (b) the application of fake news detection method into the process of risk communication management, and help organisations or governments to generate their customised COVID-19 warning policies to intervene the spread of fake news.

Expected Outcomes and Communication of Results

- A project website will be created and updated regularly to enable other researchers to obtain our up-to-date progress on this project in a timely manner, including the collected datasets and research publications.
- The techniques, methods and systems designed and developed in this project will be demonstrated at international and Australian/Indonesian conferences for both

researchers and practitioners. Research outcomes will be publish in top venues, e.g. IEEE TKDE, IEEE TSC, ICDE, ICWS and other CORE Ranked A*/A journals and conferences.

- Joint research workshops will be organised by Deakin and UGM with attendants invited from local industry and government organisations. The prototype system will be demonstrated to showcase how the detection method can be introduced into the process of risk communication management, and help to generate customised COVID-19 warning policies to intervene the spread of fake news.

<u>Research Timeline with Milestones</u>

This project will be conducted from October 2021 to September 2022.

- October 2021 to January 2022: online survey and dataset preparation (CI Liu, CI Hilya, with RA#1 and RA#3)
- October 2021 to December 2021: hoax characterisation (CI Li, CI Wijayanto with RA#2 and RA#4)
- January 2022 to April 2022: fake news detection and classification models (CI Li, CI Mulyani, CI RIfai, with RA#1 and RA#3)
- January 2022 to May 2022: multilingual fake news detection and classification models (CI Liu, CI Hilya, with RA#2 and RA#4)
- March 2022 to June 2022: system implementation and evaluation (CI Liu, CI Hilya, with RA#1 and RA#3)
- May 2022 to June 2022: automatic policy and guidance strategy generation (All CIs)
- July 2022 to September 2022: dissemination of project outcomes through publications, joint workshops, newspapers and social media (All CIs and RAs)