Ankhbayar Delgerchuluun (12211529)

COVID-19 rumor detection and trustworthy AI

What is a rumor? According to the Oxford dictionary rumor is the uncertain or doubtful truth. It disseminates misinformation and disinformation. Misinformation is claims that can be directly or indirectly verified (Andrew M. Guess and Benjamin A. Lyons, 2020). The main difference between misinformation and disinformation is the intention. The misinformation is spread unintentionally whereas disinformation is intentionally prepared and a made-up fake story. Social media has become an indispensable tool in people's daily life and allows a plethora of information to be spread at a dizzying speed across countries and continents. The plethora of information also contains rumors which harm people and undermine their dignity, social institution, and justice. Unfortunately, the rumors spread faster than facts.

The COVID-19 rumors become a threat to public health and human and social well-being as well as hindering the response action against the pandemic. Also, the rumor detection methods became essential but faced unique challenges to detect the COVID-19 rumors. Despite challenges and time constraints, the COVID-19 rumor detection methods have to be based on trustworthy AI. The assignment consists of two parts. In the first part, I will try to give an overview of COVID-19 rumors and its damage to human rights and public health as well as new methods of detection. Then, the result of simple classification models which are intended to detect rumors among the tweets will be presented.

The pandemic, COVID-19, introduces an infodemic to humanity. The COVID-19 rumor began to circle around causes of the disease, wrong preventative measures, treatment, conspiracy theories, and anti-vaccination just after the first breakouts. Its large scale, newness, speed and lack of formal and scientifically proven information ease the spread of COVID-19 rumors among the social media. Since the pandemic was new for everyone, even including scientists and policy makers, every piece of information started to be categorized as rumor, undoubtful truth. Moreover, the pandemic affects everyone's life; therefore, personal stories about the pandemic, their losses, and headlines of the top media agencies flood social media.

A real damage, or threat of the rumor, came to light during the outbreak of the Covid-19 pandemic. The rumors COVID-19 even lead people to the use of ultraviolet lamps, bathing in hot water, drink methanol, ethanol, or bleach to cure the virus and discriminate people who are infected or certain race. Referring the fact that COVID-19 pandemic was first reported in Wuhan, China, anti-asian discrimination was emerged. Hahm et al. concluded that anti-asian discrimination might be more widespread than initial reports and showed psychological impact on both Asian, Asian American and non-Asian people's lives (Hahm, 2021). False preventative measures, conspiracy theories, and anti-vaccination campaigns threaten public health as well as hindering response to the pandemic. In March 2020, mostly conservative politicians in U.S. pushed the hashtag #FimYourHospital and encourage people to break quarantine rules and prove that the pandemic is an elaborate hoax by posting pictures of hospitals' empty parking space and waiting rooms in Twitter (Gruzd, A., & Mai, P., 2020). The hospitals banned visitors to prevent the spread of the pandemic and cancelled or delayed un-urgent appointments and surgeries to deal with shortage of medical staff. These preventative measures created empty parking space and waiting rooms. This rumor, conspiracy theory, undermined the medical institutions and people's trust towards them as well as hindering the preventative measures.

In response of the rumors, the rumor detection methods are implemented by the BigTechs and Governments and began to warn to people about the reliability of the information they are seeing in the social media. The effectiveness of these rumor detection methods is problematic due to certain unique features of the pandemic. The rumor detection methods are based on content of the rumor, a rumor source, and propagation path. Due to the time constraints in the outbreak of the COVID-19, the methods based on propagation is not applicative because it takes time to create it. The source-based methods are not common in social media because it mainly relies on a web address. Therefore, the detection methods for rumors COVID-19 mainly focus on content of it and NLP techniques.

The BERT model, Bidirectional Encoder Representations from Transformers, has been used as pretrained NLP model to detect the rumor among social media. Kim et al. used various BERT models to detect the tweet that contains misinformation about a garlic and the pandemic (Kim, M. G., Kim, M., Kim, J. H., & Kim, K., 2022). Since the start of the pandemic, the misinformation about various food supplements and their preventative effect against the COVID-19 are began to spread in the social media. The false preventative measures can rise the risk of the infection of that person and people around him/her. As a result of the study, $\text{BERT}_{\text{weet-large}}$ model showed the highest performance in classifying the tweets in the garlic specific dataset. In particular, F1 score was around 89.4%. Heidari et al. examined whether a bot detection can improve the performance of COVID rumor detection using the BERT model (Heidari, M., Zad, S., Hajibabaee, P., Malekzadeh, M., HekmatiAthar, S., Uzuner, O., & Jones, J. H., 2021). The result claimed that there is no significant difference spreading the misinformation between human and social bots. On other words, the bots are not main source of the COVID rumors. They added a new feature whether bot or not feature to the BERT model, and didn't find any improvement in accuracy and F1 scores. Another issue that governments and researchers are faced with is a language limitation. Due to the global scale of the pandemic, the demand of the COVID rumors detection was high across all countries and languages. Therefore, Kar et al. proposed a new method to detect the COVID rumors for Hindi and Bengali as well as creating dataset in these languages (Kar, D., Bhardwaj, M., Samanta, S., & Azad, A. P., 2021). They used the BERT model which is augmented with extra features from the tweets, and the proposed model achieved a performance of 79% and 81% F score in Hindi and Bengali languages, respectively.

Even though it has some unique challenges compared to a normal rumor detection method, the COVID-19 rumor detection should be built upon trustworthy AI. It shouldn't contain any bias about certain gender, race, or ethnicity, especially during the pandemic. Yang and Pan proposed the rumor detection (CR-LSTM-BE) which uses content of the rumor and other people's responses in social media. The experiment showed that the new detection method can significantly improve the COVID-19 rumor detection (Yang, J., & Pan, Y., 2021).

I am going to use the twitter dataset which is built by Cheng and Wang (Cheng, M., Wang, S., Yan, X., Yang, T., Wang, W., Huang, Z., ... & Bogdan, P., 2021). They collected tweets about the pandemic with key hashtags as well as reply/retweet comment, reply number. They manually labelled the stance, sentiment, veracity of each tweet by checking multiple sources. The veracity consists of true, false, and unverified categories. The procedure follows steps of checking authoritative sites, and common sense. If the tweet isn't able to be verified even after authoritative sites and common senses, it is labelled as the unverified. If the statement is referred to differently among various sources, the voting mechanism is used among them. The authors also conducted several revisions and verification in order to validate accuracy of the manual labeling. The datasets contain 2705 tweets.

With the dataset, I tried to classify tweets by true and false after excluding tweets that are classified as unverified. I trained the classification models using the dataset. I used 80% of my dataset as training while another 20% as a test. I used the scikit-learn package in python and used Logistic regression, Naïve Bayes, LinearSVM, Decision tree and XGBoost models. The table 1 one shows the accuracy of the models. LinearSVM and decision tree models are performing the best and classify the tweets as false and True.

	Type	Model	Precision	Recall	F1-score
1	TF_2	LinearSVM	69.0	69.0	69.0
2	TF_3	Decision Tree	68.7	68.7	68.7
3	CV_1	Decision Tree	68.7	68.7	68.7
4	TF_3	LinearSVM	68.7	68.7	68.7
5	TF_2	Decision Tree	68.7	68.7	68.7
6	CV_2	Decision Tree	68.7	68.7	68.7
7	CV_3	Decision Tree	68.7	68.7	68.7
8	TF_2	Logistic Regression	68.0	68.0	68.0
9	TF_2	XGBoost	67.4	67.4	67.4
10	CV_3	XGBoost	67.4	67.4	67.4

These simple classification model doesn’t take account of the content or linguistic features of the model. This is the main limitation of this exercise and reason of why we are seeing quite low performance as well many other limitations.

References

Andrew M. Guess and Benjamin A. Lyons. (2020). Misinformation, Disinformation, and Online Propaganda. In N. p. A.Tucker, *Social Media and Democracy.* Cambridge University Press.

Cheng, M., Wang, S., Yan, X., Yang, T., Wang, W., Huang, Z., ... & Bogdan, P. (2021). A COVID-19 rumor dataset. *Frontiers in Psycholog*, 12, 644801.

Gruzd, A., & Mai, P. (2020). Going viral: How a single tweet spawned a COVID-19 conspiracy theory on Twitter. *Big Data & Society*.

Hahm, H. C. (2021). Experiences of COVID-19-related anti-Asian discrimination and affective reactions in a multiple race sample of US young adults. *Public Health*.

Heidari, M., Zad, S., Hajibabaee, P., Malekzadeh, M., HekmatiAthar, S., Uzuner, O., & Jones, J. H. (2021). Bert model for fake news detection based on social bot activities in the covid-19 pandemic. *12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference*.

Kar, D., Bhardwaj, M., Samanta, S., & Azad, A. P. (2021). No rumours please! A multi-indic-lingual approach for COVID fake-tweet detection. *Grace Hopper Celebration India* .

Kim, M. G., Kim, M., Kim, J. H., & Kim, K. (2022). Fine-Tuning BERT Models to Classify Misinformation on Garlic and COVID-19 on Twitter. *International Journal of Environmental Research and Public Health.*

Yang, J., & Pan, Y. (2021). COVID-19 Rumor Detection on Social Networks Based on Content Information and User Response. *Frontiers in Physics*, 570.