Misinformation Challenge

Camila Barbagallo, Paula García, Rocío González Lantero

IE University

Professor: David López

People worldwide constantly receive and read news, whether online or on paper. Have you ever stopped and wondered whether the information you are reading is authentic or fake? Fake news is defined as false stories published in newspapers, articles, or other media platforms that mislead readers from the facts. Fake news may happen in different ways, such as misleading headlines, visualizations that lead to biased conclusions, omission of information that may intentionally lead to false interpretations or strictly fake news.

The rise of social media has also eased the increase of fake news, mainly due to information overload, a “situation when people are confronted with a massive amount of information created on social media which exceeds the capacity they can handle” (Eliyana et al., 2020). We see millions of posts using social media daily, mainly from our friends, who usually share the same ideologies. If they don’t, we are less likely to pay attention to their posts regarding some matters. Furthermore, cookies and targeted advertising have also eased the reach and diffusion of fake news. Through these technologies, you read what the algorithm thinks you may want to read, not exposing you to contra-arguments or other sources that may indicate that whatever you are reading is fake news.

Many politicians have used the rise of fake news to their advantage to “undermine legitimate opposition” (About the author Damian Tambini Posted In: Filtering and Censorship|Intermediaries|LSE Media Policy Project et al.), not necessarily creating it but claiming it is fake news. Donald Trump, former President of the United States of America, has used this on various occasions. Specifically, his win in the elections back in 2016 has been “blamed” on the use of fake news.

All of this has increased the demand for fake news detection and intervention. Therefore, we will be exploring a possible way to detect fake news in the following.

The Dataset

Our dataset is from the Hugging Face platform, named *mrm8488/fake-news*. This dataset contains only a training set composed of two features: text and label. The text shows the whole body of the article, and the label is a binary variable, which indicates whether the article is fake. We have 44,898 labelled articles, from which we will use 31,428 (70%) for training and 13,470 (30%) for validating. We have no further information about where they come from regarding the articles used. However, through visual inspection, we have concluded all articles are in English but come from different parts of the globe. Many come from Reuters, “a global information and news provider headquartered in London, England, that serves professionals in the financial, media, and corporate markets” (Hayes, 2021). The articles from Reuters indicate so in the text field, and all of them are labelled as not fake news. This could suppose a bias for our model, which will be considered in the discussion section. Lastly, we will assume all articles were originally written in English, as there is no indication of previous translations, which may alter the content.

The Model

We have used the roberta-fake-news model from the Hugging Face repository, uploaded by ghanashyamvtatti. This specific model was trained using a hyperparameter search with ten trials and using this dataset: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset. Liu et al. (2019) explain how Roberta is an improved version of the BERT model. Researchers found that BERT was significantly undertrained when carefully evaluating the effects of

hyperparameter tuning and training set size. The implemented modifications include: (1) training the model longer, with bigger batches, over more data; (2) removing the next sentence prediction objective; (3) training on longer sequences; and (4) dynamically changing the masking pattern applied to the training data.

When evaluating the performance of this model on our validation set, the accuracy resulted to be 99.87%, and the evaluation loss was 0.0074.

Model Improvement

We believe that the first thing that we could do to improve the model's performance is increasing the number of epochs we train it with. Our computing resources only allowed us to train one epoch, which took approximately 2 hours. The weights could be optimized through more training, and the model could act to its full potential.

Furthermore, we could include more features in the dataset. This could be, for example, publisher, writer, date, location, headline, etc. This will give more information to the model, and it won't base its predictions solely on the content but also on the context. Furthermore, we could also train an initial model, which extracts key facts from the article. These would then be compared with trusted sources and, if necessary, labelled as fake. Our model would also adopt the knowledge-based fake news detection technique. If we consider our model performs very well, it could also fact-check against its predictions or even learn some facts are true and some are not. However, the drawback is that there is no global truth, and it would be difficult for the model to accept news stories as true.

Moreover, we could adopt some techniques of style-based fake news detection. This consists in computing some statistics regarding different characteristics, such as quantity, complexity, uncertainty, etc. These characteristics

could improve the accuracy of our model; however, they should be further explored to prove their significance.

Additionally, it would also help to know which factors make our model decide whether an article is a fake news. In our current model, this could be some words or a group of words that trigger the algorithm to believe the article is fake. Once again, if we had more features, we could test our model's reliability by understanding how decisions are made.

Predicting a reliability score instead of a fixed classification could also make our results more understandable and could help us draw more profound insights. For example, we could establish the reliability of sources or journalists, which we could announce publicly so that the audience could be aware.

Through social media, the news is mainly spread through images. Therefore, our model should also accept image features. This is more than the preprocessing stage, as we would need to train an additional model, which detects letters from pictures, transcribe them, and detect different objects in the images. This could also help us to identify fake news.

Lastly, we believe it would be helpful to double-check how our training model was created, as we have no information. It would be interesting to know how the data creator has defined whether an article is a fake news. This will allow us to know if any potential biases could affect the offset performance of our model.

Processes to Create and/or Update Datasets

We would develop two kinds of processes to create/update datasets to keep fake-news classifiers updated and relevant to emerging disinformation. The first process would be to automate data collection, so information is constantly being

checked for misinformation. Our second process would be to re-check those articles that our model has already reviewed to see if these change from fake to non-fake.

To automatize the process of data collection, our idea would be to create datasets per newspaper or online website for article publication. For example, imagine you want the articles' headlines of all the articles published in The New York Times. To get these headlines, you would need to create a blank dataset, and each time a new piece is issued, the name would be automatically inputted into this dataset. Overall you would have datasets from each different online newspaper or website containing the headlines of all the articles published within them. With all these datasets, you would then need to merge all these datasets as a general dataset and use this dataset to declare the article's headlines as fake or non-fake. This would be beneficial to keep fake-news classifiers updated and relevant to emerging disinformation because all the information constantly being uploaded to the web would be checked automatically. You would be able to automatically detect whether the article recently published by The Washington Post is fake or non-fake.

Our second approach to keeping fake-news classifiers updated and relevant to emerging disinformation would be to re-run the model. For example, El País launches an article on Monday whose headline reads, "A Meteor Will Fall in Madrid Within the Upcoming Days." If the model were to get this headline, it would probably classify it as fake. However, El Periódico de España launched a new article on Wednesday, which headline reads, "Meteor Approaching Madrid, Expected Crash on Friday." Now imagine if El Mundo, La Razón, and El Diario all launch similar headlines during the same week. If the model were to get the original article from El País and re-run it through the model, this headline would most likely change from

being declared fake news to being a real article. Information is constantly evolving and deepening. We can't have a model declare a news article as fake news simply because there is no more information that proves the message stated. Models should be re-run as data flows. This way the accuracy of the prediction of fake news would be much higher than before.

Overall we believe that both approaches would work best together. Eventually, if you create a general dataset with all the information and this information is constantly being run through the model, fake news would be detected faster and more accurately. Readers deserve to get the latest information available and detected as soon as possible to share with whomever they would like. Having this dataset created with the permission of all the newspapers and blogs around the world would greatly benefit the detection of fake news. But also, having the models re-run after a certain specified amount of time would help correct those articles that could have been detected wrongfully beforehand.

Bibliography

About the author Damian Tambini Posted In: Filtering and Censorship | Intermediaries | LSE Media Policy Project, Tambini, D., Posted In: Filtering and Censorship | Intermediaries | LSE Media Policy Project, & *, N. (n.d.). *Who benefits from using the term 'fake news'?* MediaLSE Who benefits from using the term fake news Comments. Retrieved March 10, 2022, from https://blogs.lse.ac.uk/medialse/2017/04/07/who-benefits-from-using-the-term-fake-news/

Eliyana, A., Rohmatul Ajija, S., Rizki Sridadi, A., Setyawati, A., & Permana Emur, A. (2020). Information Overload and Communication Overload on Social Media Exhaustion and Job Performance . *Sys Rev Pharm 2020;11(8):344-351*.

Hayes, A. (2021, May 19). *Reuters*. Investopedia. Retrieved March 10, 2022, from https://www.investopedia.com/terms/r/reuters.asp#:~:text=Reuters%20is%20a%20global%20information,Thomson%20Financial%20Corporation%20in%202008

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.