Project Progress Report

COMS 6156 Topics in SE - UNI: vk2497

Overview

I am working on the topic of **Detecting fake news using NLP, Machine learning/Deep learning techniques**.

A recent analysis by Statista [4] reveals that people worldwide find it increasingly challenging to distinguish fake news from real news and are losing their trust in the information they are exposed to. These statistics have made it critical to address and develop effective techniques to detect fake news. Machine Learning (ML) and Natural Language Processing (NLP) techniques are gaining popularity in fake news detection as it is challenging and time-consuming for humans to go through a large number of news articles manually.

I will be using the ISOT Fake News dataset made publicly available for research by the University of Victoria [1], [2]. The ISOT Fake News dataset [1], [2] is a compilation of several thousands of fake news and truthful articles (It contains more than 20,000 articles of Fake news and Real news each), obtained from different legitimate news sites and sites flagged as unreliable by Politifact.com.

I will be implementing machine learning (ML) algorithms along with natural language processing techniques (NLP) for detecting fake news on this dataset. Along with the traditional NLP techniques such as TF-IDF, word2vec, etc., as a novel contribution, I plan to also use state-of-the-art transformer models such as BERT, S-BERT, RoBerta, etc. to generate text embeddings as input to my machine learning models and compare performance.

Although there are a few other publicly available datasets for fake news detection, with the growing popularity of LLMs such as ChatGPT for generation of fake data [3], I plan to create synthetic example of fakes news data from real news data and fake input prompts to LLMs such as GPT-J / GPT-3 :-

(1) I plan to leverage GPT-J (a publicly available smaller version of GPT-3) to augment the dataset - I will generate fake news examples with fake input prompts.

Based on the feedback given in the revised proposal, I am exploring the use of the temperature [5] hyper-parameter which might significantly perturb the commentary generated by these (Large-Language Models). If the results are not significantly perturbed, I would follow the original feedback of dividing the ISOT fake news dataset into 80/10/10 split to measure performance.

Research Questions

1) Can we use Machine Learning algorithms with NLP embeddings to detect fake news articles successfully?
2) What percentage of real-news articles are tagged as 'fake' by these ML/DL algorithms?
3) If I am able to successfully use LLMs to generate fakes-news from real-news - Can adversarial attacks (generating fake news articles from real news articles) reduce the performance of Machine Learning models for fake news detection ?

Value to the User Community

Political fake news articles have manipulated the public by trying to influence and change their beliefs about their choice of leaders. This could create a long-lasting negative impact on the people and their countries altogether.

Furthermore, in more critical domains like healthcare, there has been a surge in cases where people have followed fake news articles/blogs on self-treatment for various medical conditions, with the COVID-19 pandemic being the most recent example. As a result, they have made decisions significantly harmful to their health and put their lives at risk.

We can measure the performance (accuracy, F-1 score, precision, recall), to address Research Question (1). In addition, we can compare different machine learning / deep learning models (using different NLP embeddings) to determine the best-performing model for fake news detection. The performance metrics will help address RQ (2) to replicate the findings and utilize these models in real time for fake news detection. With the popularity of 'ChatGPT' and other Large-Language Models (LLMs), it's essential to evaluate the performance of these models when the input real news prompt is perturbed (adversarial attacks). This will demonstrate the robustness and identify future-research opportunities to improve ML models for fake-news detection.

Fake news detection is crucial today as the fake news generation has greatly influenced decisions on political and socio-economic issues in every society and country.

Demo

As part of the elevator pitch, I would highlight the importance of using machine learning models for fake news detection. With the popularity of 'ChatGPT' and similar LLMs, generating fake news and spreading misinformation & disinformation through social media platforms has become more accessible.

As part of the demo, I would show PowerPoint slides to demonstrate the workflow and performance of the different machine learning models to detect fake news (Running the full code for training / testing will take a lot of time). I also plan to show how the model performs on a real news example and illustrate the impact of adversarial attacks (by perturbing the example)

Delivery

I plan to upload the code to GitHub, making it publicly accessible for anyone to reproduce the experimental results. Hence, anyone in the research community can use the machine learning model for fake news detection. In addition, I will comment on the code and include a ReadMe file so that everyone can run the code with the required packages and models.

References

[1] Ahmed H, Traore I, Saad S. "Detecting opinion spams and fake news using text classification", Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018. https://onlinelibrary.wiley.com/doi/epdf/10.1002/spy2.9

[2] Ahmed H, Traore I, Saad S. (2017) "Detection of Online Fake News Using NGram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127 138).
https://link.springer.com/chapter/10.1007/978-3-319-69155-8_9

[3] Abstracts written by ChatGPT fool scientists
https://www.nature.com/articles/d41586-023-00056-7

[4] Level of difficulty differentiating between true and false information online among consumers in selected countries worldwide as of June 2020
https://www.statista.com/statistics/1227193/identifying-misinformation-difficulty-worldwide/

[5] Harshit Sharma, July 15, 2022. “Softmax Temperature”

https://medium.com/mlearning-ai/softmax-temperature-5492e4007f71#:~:text=Temperature%20is%20a%20hyperparameter%20of%20LSTMs%20(and%20neural%20networks%20generally,utilize%20the%20Softmax%20decision%20layer.