Project Proposal

COMS 6156 Topics in SE - UNI: vk2497

Introduction

I would like to work 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.

Motivation

There are two primary reasons for the surge in fake news -

1. Misinformation - when bogus information is unintentionally spread across people without any intention to harm someone. Usually, people circulating such information are oblivious that the data is untrue.
2. Disinformation is when people intentionally try to cause harm to the readers by deliberately spreading false news with a more cynical intent.

Political fake news articles have tremendously 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.

Methods and Dataset

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.

I first plan to test the performance of the trained models only on the ISOT dataset and then again by adding the synthetic data (generated by GPT-J)

First, I plan on using only the ISOT dataset, splitting it 80-10-10 / 80-20 for training, k-fold validation, and testing. The dataset is valuable as they have multiple editors to fact-check the news. Therefore, I would first evaluate the performance of fake news detection using ML/NLP approaches on the ISOT 'test' dataset.

To evaluate robustness of the above models, we would need to test performance on unseen examples of fakes-news outside the dataset.

I was thinking of using a large-language model like GPT-J (an open-source version of ChatGPT) to generate fake news articles from real-world news examples and measure the performance of the machine learning and NLP models. (I hypothesize that this would be a form of adversarial attack that should reduce the performance of the ML detection models). On performing initial literature reviews in NLP research, the use of LLMs for adversarial attacks (generating fake synthetic data) is also very recent and novel; Existing approaches are more manual and rules-driven to perturb text examples.

Furthermore, To better explain my experiment on synthetic data, I experimented and realized that ChatGPT could not generate fake news directly. So, I plan to use the Large-Language Models like GPT-J / ChatGPT - as a 'next' sentence prediction task. So, the LLMs will generate fake new sentences from a given input news prompt/title. 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.
(2) If time permits, Perform adversarial perturbations on real-news data such as (a) fact perturbation (changing numerical information), (b) subject-object exchange, (c) cause-confounding, etc.

I will then retrain the machine learning and deep learning models on the augmented dataset to measure performance and see the impact of adversarial attacks on machine learning and deep learning models to detect fake news. If time permits, I will also summarize real news examples with real-world news articles to generate additional data for training the ML/DL 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/