Fake news detection

Introduction:

In the past few years, “Fake news” has come at the center of the political and social discussion, and many proposed solutions emerged to detect them. Before describing the project, this paper starts off by briefly defining what we mean by “fake news” and why it is problematic. The next part overviews various methods used in detecting fake news. The implemented project relies on text analysis of the content of the news. The neural network used is BERT and the model is a classifier that predicts if a given text is fake or not. I would also say that the main purpose for me to use this framework and idea was to learn how to perform a classification problem using BERT. Before discussing the implementation part and BERT’s framework, I include a section for the data. Generally, and more specifically for this task, data is very important for the performance of the model on real world tasks. For this kind of task it is even important because of the political nature of the topic, and how “fake news” are viewed by different people. I chose an already collected dataset that I thought of being trustworthy. Following the data section I talk about BERT, from tokenization, text encoding and training on classification. I finally end the report with a conclusion where I view the results of the classifier.

Fake news as concept and why to detect it:

Written and electronic media is a pillar of a healthy democracy. As societies have been steadily moving towards an online life, social media and online platforms have been playing a major role in shaping trends and information across multiple domains in societies. Research conducted in the USA and Europe concluded that 60% percent of people totally depend on social media to get any kind of news and even consider it to be totally reliable and trustworthy (Allcott, H., & Gentzkow, M. (2017)). Along this shift of human behaviors in terms of news consumption, there has been a huge eruption of fake news that misleads people who consume them. Many of these news items are neither detected by the general public nor a sophisticated media.(Bhushan, Agrawal,Yadav.2020). This has led to a trend of unreliability and chaotic flow of information that is difficult to manage and trust. Fake news in recent years, not only influenced elections but also altered the choice and mood of many people across the globe. Some of the popular examples were the opinion spamming in the U.S election, and the intentionally provoking campaigns that targeted the public’s view and mindset towards the EU during the Brexit referendum (Bovet, A., & Makse, H. A. (2019) . Many websites and agencies behind these

organized campaigns employed bots and relied on virtual accounts to reach and manipulate their targets(Flammini, A., & Menczer, F. (2017). Before moving to the various methods used in recognizing fake news, a small elaboration on what "Fake news" means more specifically.

- Two types of wrong information :
    1. Misinformation: "Wrong information, Misleading data, information about anything, Not verified data"
    2. Disinformation: "Wrong news (Intentionally delivered or published) to fool people"

There is a thin line to distinguishing misinformation from disinformation. Misinformation cases are when the news includes information that hasn't been yet verified. Meanwhile, disinformation is to intentionally deliver information that is known to be falsehood, such as conspiracy theories that contradicts scientific facts or proven knowledge. Satire news, even if it is not based on prior investigation or proper knowledge, falls into the category of misinformation, rather than deceptive fake news. Especially, if it includes humorous elements which would make it seem as an opinionated expression of political view. Therefore, the definition of fake news, here, includes only the news that is meant to directly target and manipulate(disinformation), rather than the news that indirectly misleads as misinformation. (Bhushan, Agrawal,Yadav.2020)

Methods:

There are different methods for detecting fake news on the internet. The methods can fall into 3 main categories(Nordberg, Nohlberg, Kvärstad. 2020):

1. Text classification: Classification that is based on the articles' content
2. Network detection: a method that relies on tracing how a cluster of news is spread.
3. Human-machine hybrid: automatic detection with manul human aid.

The authors (Nordberg, Nohlberg, Kvärstad. 2020) point out that text classification that relies on machine learning gives promising results with descent accuracy. They refer to multiple machine learning techniques and assert that SVM models are most effective, as they state that SVM models tend "to not only distinguish fake news from real news, but also from opinion-pieces and propaganda." (Nordberg, Nohlberg, Kvärstad. 172). The paper also refers to other approaches in this category that relies on classifying the relation between an article's headline and its body, but they argue that "such a method fails to exploit the article structure, and enhanced text classification is presented". (172). Other presented techniques, such as (A. Giachanou, P. Rosso, F. Crestani, 2019),

incorporate the use of sentiment analysis, and particularly measure “emotion entensity” by detecting the amount of sensitive words used which fake news would rely on to trigger high emotional responses to achieve a higher spread online . These methods use CNN that analyses semantic and syntactic information, and including text analysis. The evaluation of the CNN syntactic,semantic analyses gives them a good credit given the simplicity of the classifiers and the used model.

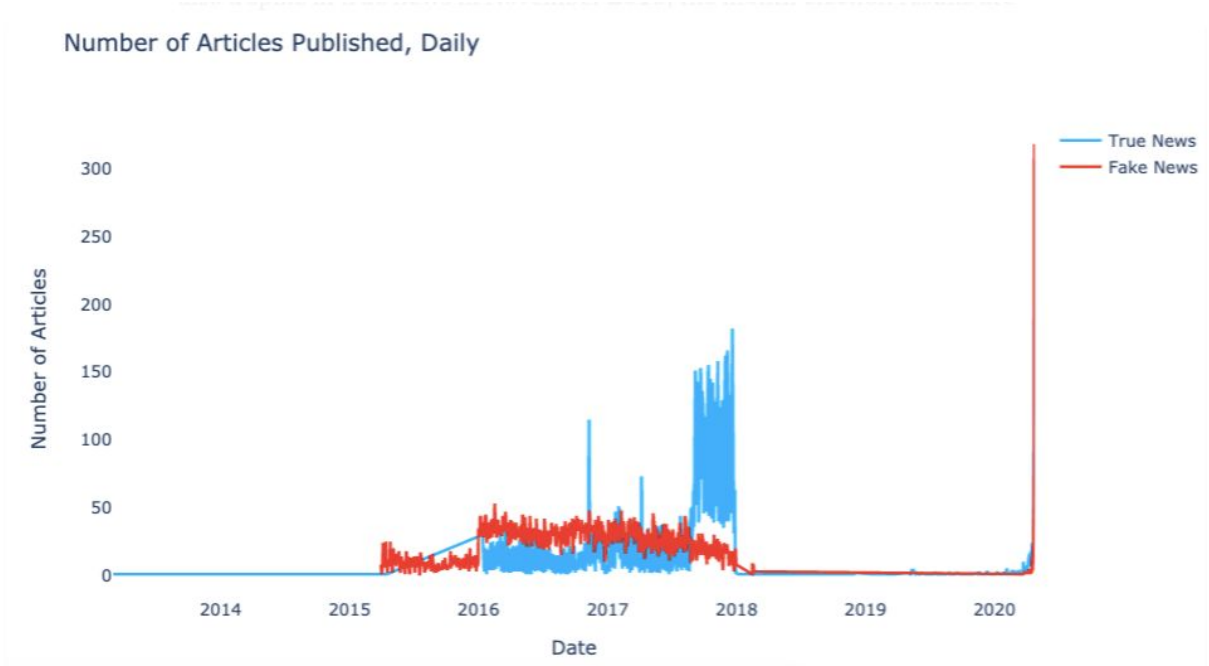
Fake news detectors that rely on monitoring the network are very effective in terms of the speed in detecting news threads online. This approach considers the data beyond the ones within the article, meaning that circulating data that connects or relates to a certain degree to an article, by assessing the environment where certain news are flowing, such as users in social media connected by their regenerating of certain threads. The employability of this technique is characterized by its operation in detecting fake news with limited information. It also makes a good alternative when the information available on fake news is too sprace to be analysed in text classification. The review asserts that detecting fake news based on text analysis has “a limitation in the form of not being able to detect fake news early, when the information required for verification or debunking is unavailable due to the early stage of news propagation”(172). Knowledge based approaches also go under this category. knowledge graphs can detect the relation between two entities. This can be achieved by gathering background information and relating the target to it in terms of falsehood and truthhood. This enables such models to detect deception, such as fact-checkers, even with short statements. Models as (N. K. Conroy, V. L. Rubin, Y. Chen, 2015) divide a statement, such as subject-predicate, into nodes and measure the likelihood of the pair to be false or true. Another paper from Chalmers University “ Investigating Content based fake news detection using knowledge graphs”(Germishuys,2019) uses a similar approach of “knowledge networks” using a GPT-2 language model, and proves a decent result with a small set of news articles.

Finally, Human machine-hybrid relies on both, automatic machine detection and on manual input from humans. These language models are most fit for advanced language usage, such as satirical news about politics where machines would fail to analyse properly as a human with a linguistic approach. In such a model, the human input depends on the confidence of the machine in detecting. It is natural that it achieves better results than only machine systems (Nordberg, Nohlberg, Kvärstad. 2020).

Data :

The dataset was already collected and downloaded from the internet. (The source at the end of the section). The dataset’s publisher collected it from two main sources. The first source is Kaggle

Dataset. The second is a collection of articles from multiple news outlets: Real news: CNN, BBC, The Guardian, Fox News, NBC News, and Washington Post ; Fake news: BreitBart, The Onion, and Infor words. The number of Fake News articles is 24,194, and the Real news articles 22,506. The articles from Kaggle are collected between (2015,2018), and the ones from the news outlets are collected up until 2020. The dataset builder pointed out that there was a spike of fake news at the start of the US presidential election year in January 2016 and increase of real news at the time of the election in November 2016. She demonstrates the flow of the published news articles throughout the years in this chart :



Source: At the end of the section

Hence there is a massive increase in the number of published news in 2020, either fake or real. This is due to the beginning of Covid crises all around the world.

An analysis included with the data shows a certain pattern between the article's length and its reliability. It shows that the fake news articles tend to have more words which demonstrates that the article length plays a role in determining whether the article is real or fake.

	Number of Words	Standard Deviation
True	390.16	338.08
Fake	420.55	363.51

The source of the dataset, and more analysis on the data can be found in:
https://github.com/riag123/FakeNewsDeepLearning/blob/master/EDA_%2B_Pre_Processing.ipynb

BERT and implementation :

BERT(Bidirectional Encoder Representation from Transformers) learns a deep contextualized representation of words. Even though it's called b-directional network, unlike directional networks,the transformer encoding in BERT learns the context of a word at once, rather than passing it from right to left then to left to right. In a normal sequential model, the systems tend to define the embedding as word prediction problems where the sequence " She ran in the ....." is used as the context for the next word embedding. However, BERT tackles this issue by relying on unique strategies:

- Masked LM (MLM) : BERT masks target words for the purpose of learning word embeddings depending on the context words, the un-masked words in the sequence of the target word (Devlin,Chang,Lee,and Toutanova,2018). The prediction of the target words happen through different technical stages:

    1) A classification layer following the encoding output.
    2) The sequence is transformed to a vocabulary dimension by multiplying the output vectors with an embedding matrix.
    3) Finally, softmax probability is computed for each single word.
    4) the loss function is only considers the predictions of the masked words

- Next sentence prediction: On a sentence level, sentences are paired. 50% of the sentences are paired with other sentences from the same document and the other 50% is randomly selected. They are randomly selected based on the assumption that random sentences would tend to disconnect from the first sentence. To differentiate between the sentences BERT relies on a labeling strategy for processing them before entering the model:

    1. [ACE] is inserted at the beginning of the first sentence (the sentence from the document) to indicate the start of the sentence, and [SEP] is inserted at the end of each sentence.
    2. Label A or B is inserted into each token as a reference to what sentence the token belongs. (sentence embedding)
    3. Thirdly, a vector of indices for the position of each token in the sentence. (Transformer positional embedding)

Since BERT provides pre-trained word-vectors following the method described briefly in this section, it was a good alternative for the two major tasks in the implementation. Extracting features from the data through BERT tokenizer and encoder; then fine tuning the existent BERT model for sequence classification. The extension uses BERT Base; 12 layers with 12 attention heads, and 110 million parameters; "BertFor Sequence Classification pre-trained model", this is the main reason to further consider BERT as its bottom layers and weights are already trained, (Devlin,Chang,Lee,and Toutanova,2018). Otherwise, building and training a neural network from scratch would require larger data and time in order for it to produce promising results.
The interface in the implementation is pytorch and transformers as I am used to working with pytorch and because BERT is set up of transformers. The maximum length of a document in BERT is 512 tokens. I tested from the highest possible length to lower, high as 512 causes runtime error due to the large space it takes to run at this size. I ended up choosing 70 tokens as the length of each article based on the performance as the standard.The size of the batches 16. For the loss function, I use Adam optimizer. To optimize the training, I use a learning rate scheduler to change the hyperparameters and reduce the learning rate as epochs increase. For the exploding gradient problem, where a large error gradient accumulates leading to an unstable model that is unable to learn, I use utils.clip_grad_norm from pytorch.

Conclusion and results:

Fake news as a concept is necessary to be defined in order to build the right model for the right purpose. This means the type of data and information collected is very important for the problem for fake news detection. Understanding the patterns behind the spread of fake news helps in finding appropriate solutions to detect them. The proposed model is a content based classifier. Such a task requires a strong language model of well-trained neural networks. BERT is the option chosen for the network. The results show a very good classification to a degree that even raises suspicion. Before showing the results, the validity of the models need to be tested on other datasets and performed on recent news. The accuracy of the model is as follow:

epoch	train_loss	val_Loss	val_Accur
0	0.034	0.012	0.998
1	0.020	0.012	0.998
2	0.017	0.012	0.998
3	0.017	0.012	0.998

In the training part we could see that the model's performance stabilized already after the third epoch, which indicates how BERT could train and learn with only a few epochs. The loss and accuracy on the validation part has been static.

References:

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211–236.

Bhushan, D., Agrawal, C., & Yadav, H. (2019, December). Fake News Detection: Tools, Techniques, and Methodologies. In International Conference on Information Management & Machine Intelligence (pp. 347-357). Springer, Singapore

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F. (2017). The spread of fake news by social bots. arXiv preprint.

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. Nature Communications.

Nordberg, P., Kävrestad, J., & Nohlberg, M. (2020). Automatic Detection of Fake News. In 6th International Workshop on Socio-Technical Perspective in IS Development, virtual conference in Grenoble, France, June 8-9, 2020 (pp. 168-179). CEUR-WS.

Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), 1-4.

Germishuys, J. (2019). Investigating Content-based Fake News Detection using Knowledge Graphs.

Giachanou, A., Rosso, P., & Crestani, F. (2019, July). Leveraging emotional signals for credibility detection. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 877-880)

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.