Detecting fake news with Python

by

Fake News Detectives

Kristofer DeYoung
Dzianis Makarevich
Martin Liang
Morten Lønning
Gard Lindås Nordli

Submitted for assessment in

UC1ST1104– Studio 1 Final Assessment

at

Noroff University College

Contents

1. Introduction

Nowadays internet, especially social media, is the main tool for news and communication (Zhang & Ghorbani, 2020). However, the news is not always based on facts and accurate information. News is being spread everywhere on the internet, through online newspapers, social media, e-mail, and various other channels (Pérez-Rosas et al., 2017). It is quite easy for someone to spread fake news and reach a huge group of people within a short time, thereby manipulating people into thinking this is accurate information. Reducing the spread of misinformation, especially through social media is a major challenge (Pennycook & Rand, 2019). The misinformation being spread concerns a diverse range of subject matters, one subject that brought attention to fake news, was the US presidential election in 2016 (Pennycook et al., 2018). During the election, the sharing of known false news was estimated to be as high as 30 million shares for false news favouring Donald Trump, and 8 million shares for false news favouring Hillary Clinton. Further, it is believed that an article shared on Facebook on average generates 14 page visits per share, and with 30 million shares, it is understandable that some people think it had a significant impact on the result, favouring Trump during the campaign and election. When a person has an idea about a specific subject, it is easier for that person to believe similar fake news, due to psychological factors. Therefore, it is easy for producers of fake news to feed more fake news about the same subject (Allcott & Gentzkow, 2017). Other recent examples of subjects where fake news was, and still are being produced include the covid-19 pandemic and the ongoing war in Ukraine.

To overcome the challenge of determining what news is real using automation, it is especially important to utilize good historical datasets (Wang, 2017). The demand for such an application is increasing every day, as the amount of data being produced and published is growing every day (Zhou et al., 2019). Over the past years, a lot of research has been done concerning fake news. This research shows that fake news comes in many different forms, such as propaganda, misinformation, disinformation, rumours, hoaxes, parodies, satire, and clickbait (Hangloo & Arora, 2021).

Detecting fake news is not an easy task for a human being. Using Natural Language Processing and historical incidents of fake news, this study aims to address fake news, how it is being spread, and develop an application prototype for fake news detection, and then get an output based on statistics, that tells if the news articles are fake or not.

2. Literature Review

2.1. The prevailing threat of fake news

The concept of misleading and misinforming through publishing false information is not new; misinformation has always been a problem, showing itself in various shapes and forms, such as for example rumours or conspiracy. In relation to a survey on the amount of Americans believing in conspiracy theories, a poll was made in 2013 (Bowman & Rugg, 2013), marking the 50th anniversary of John F. Kennedy's assassination. The results showed that 59% of the respondents believe the assassination to be the work of more than one person, where the most popular theories vary from it being a plot by the CIA, Cubans, or the Soviet Union. However, there are many other examples of similar events of misinformation.

To some extent there has previously been a form of control over the News outlets, where the institutions responsible maintain quality control through validating facts and accuracy to uphold their integrity and reputation. During the 2016 US presidential election, the emergence of the term "fake news" grew rapidly due to social media becoming a new way to easily spread news and information (Allcott & Gentzkow, 2017). A post-election poll done by the Pew research centre estimated that 62% of American adults get news from social media, whereas 18% of respondents reported doing so often. Users reported that they were primarily receiving news from platforms such as Reddit (70%), Facebook (66%), and Twitter (59%) (Gottfried & Shearer, 2016). During the election, a study, (Allcott & Gentzkow, 2017) found a total of 41 pro-Clinton fake news stories and 115 pro-

Trump fake news stories totalling 38 million shares. Multiple sources estimate each share to have generated 14 page visits, which leads to a total of 760 million.

Information and disinformation spread on social media can sway public and political opinion and it has even been suggested in some news articles (Read, 2016) that it managed to influence the outcome of the 2016 U.S. presidential election. In contrast, in a post-election survey (Allcott & Gentzkow, 2017), 1200 respondents were asked to look through a dataset of both fake and real news to recall whether they believed the stories; 15% of the respondents recall seeing the fake news headline while only 8% recalled believing them. The study continues to analyse how important fake news is in changing a voter's bias and concludes that if the impact of fake news is as effective as a tv advert or ad campaign, it would only affect poll results by one-hundredth of a percentage point; this would likely not be enough to affect the end outcome of the election. It is unfortunately hard to measure the true impact of spreading fake news online, as there are too many fluctuating variables. due to multiple variables.

Despite this, in more recent events, such as the covid 19 pandemic, a report (Mitchell & Walker, 2021) showed that 48 % of Americans believed the government should be actively taking steps to prevent the spread of misinformation on social media platforms. Half of the respondents even believe in restricting people's rights to read or publish content, only to hinder the spread of disinformation. Further a total of 61% of the respondents believe the responsibility lies with social media providers to reduce the spread of misinformation. Despite the heavy criticism towards social media platforms after 2016 election, 48 % of U.S. adults still report receiving news often or sometimes from these platforms. While Facebook is the predominant source of information, with 31% of its users reporting regularly getting the news there, other sources such as YouTube at 22%, Twitter at 13%, and various sources such as Instagram, TikTok, and Snapchat are on the rise, especially for the younger demographics (Walker & Matsa, 2021).

2.2. Humans' ability to distinguish fact from fiction.

An alarming statistic is that 64% of respondents report; only getting news from one source before deciding whether it is believable, 26% from 2 sources, and only 10% from 3 or more sources. Not being source critical makes it harder to distinguish between what is fake and what is true (Gottfried & Shearer, 2016). The problem is further amplified as social hubs often result in a phenomenon known as "the echo chamber effect," creating "homogeneous and polarized communities"(Del Vicario et al., 2016). Many social networking circles share similar opinions, making them more likely to share information and disinformation that reconfirm their own bias. This effect is known as "confirmation bias" and appears where the spread of information usually coincides with community norms, not necessarily differentiating between what is accurate and false (Del Vicario et al., 2016). "Echo chambers" are a phenomenon in social media outlets, the resulting effect is that high exposure to identical or similar news increases the chance of believability. Even fake news headlines that are familiar can be perceived as more accurate There is little difference in the spread between fake news and real news; however, the psychological effect remains and can influence our choices (Pennycook et al., 2018). The same study also found that placing a warning label on questionable information reduced the amount of reshares of articles containing misinformation. The effect of repetition to change someone's opinion about the accuracy of a given piece is a phenomenon known as the "sleeper effect." A similar phenomenon is when a person engages with an article where the source is of low credibility, when given enough time, people tend to only recall the content of the article and not necessarily the credibility of the source. "When information volume is low, recipients tend to favour experts, but when information volume is high, recipients tend to favour information from other users." (Paul & Matthews, 2016). Another effect is where the reader initially believes something to be valid only to have information later retracted or proven false. The information often still manages to subconsciously influence opinions and memories of the given event.

Humans are naturally poor fake news detectors due to various psychological phenomena such as cognitive bias (Shu et al., 2017). Machines do not have the same disadvantage, although they can still adopt biases based on the data used to train them. This poses significant concerns over the safety of unsupervised data-driven decision making, where studies show that neural networks are vulnerable to noise and errors in the input data (Alshemali & Kalita, 2020). Researchers in the field now face a unique set of challenges in automated fake news detection in social media (Bondielli & Marcelloni, 2019).

2.3. Datasets

In a research paper published in 2020 (Zhang & Ghorbani, 2020), fake news characteristics are broken down into three main categories: Volume, Variety, and Velocity. The ability to disseminate fake news has never been more effortless, or to use a more popularized term, "viral." Various websites have now popped up whose sole purpose is to spread disinformation, such as denverguardian.com, wtoe5news.com, and ABCnews.com.co (Allcott & Gentzkow, 2017).

The variety of fake news presents problems because the distribution of fake news now occurs across multiple different platforms, in various forms; "such as rumours, satire news, fake reviews, misinformation, fake advertisements, conspiracy theories, false statements by politicians, etc." (Zhang & Ghorbani, 2020). Research (Zhou et al., 2019) shows that these different news outlets create distinct characteristics to look for in specific datasets, for example; author or publisher, the headline aimed at catching the reader's attention, main text body that usually makes a claim or an image/video that encapsulates the data needed to be visualized to convince the reader of its authenticity. These distinct characteristics require innovative approaches when it comes to processing data, and techniques suitable for specific datasets.

The various articles emerging on social media, like the ones before the 2016 U.S. presidential election, tend to be short-lived (Allcott & Gentzkow, 2017). Separate investigations made by BuzzFeed and the Guardian revealed that teenagers from a small town in Macedonia were responsible for over 100 sites posting fake news, where the primary motivation was advertising revenue when clicking on the link to a webpage (Hughes & Waismel-Manor, 2021). Though this is only one example it paints a picture of people who choose to engage in these activities and provides insight into the motivation behind such acts. Due to the short life span of sources, it further complicates analysing how many are affected by fake news.

A reliable dataset is crucial when developing accurate and automated fake news detecting tools. Datasets are usually split into multiple categories, and one of the more reliable ones is called export-oriented. This method is utilizing domain experts to determine the truth of articles; however, it is a comprehensive and resource-consuming task, and therefore expensive (Shu et al., 2017). Websites such as PolitFact.com, Factcheck.org, and snopes.com are examples of websites providing datasets of fake news statements that are labelled either true or false (Bondielli & Marcelloni, 2019). Another popular resource is crowdsourcing-oriented, for example Fishkit (Shu et al., 2017). Using this approach, you can utilize a larger community of people who discuss questionable content, which allows better scalability than previously discussed (Bondielli & Marcelloni, 2019). The latest edition in fake news detection is computational-oriented data sets, utilizing algorithms and crowdsourced data. One benefit of this is that the data set checks external links through a knowledge graph to compare factual claims, consistency, and frequency (Shu et al., 2017). After the Facebook and Cambridge Analytica data scandal surfaced in 2018, the landscape for gathering data changed; most social media companies now have policies restricting their users' data from being available to the public, making it increasingly hard to collect relevant data to train accurate models (Bondielli & Marcelloni, 2019).

However, progress is constantly being made, like in the paper by Wang (2017), where a data set called "Liar Liar" with 12836 short statements spanning over a decade worth of content from various contexts, with real-world ideas was made. The benefit of this compared to many previous datasets that rely on a large body of text which is more appropriate for deception detection; is that the shorter body of text conforms better with news from tv, speaker statements and social media as this is often in a more concise form.

2.4. Reliability and scalability using Natural language processing

Many Tech providers are constantly looking for solutions to detect and debunk the spread of false information. In the aftermath of the 2016 U.S. presidential election, Facebook (Meta), in collaboration with IFCN (International Fact-Checking Network) certified companies, started a third-party fact-checking program (Hangloo & Arora, 2021). The software relies on "independent fact-checkers to review and rate the accuracy of stories through original reporting" (Meta's third-party fact-checking program, 31 May, 22). The devised process follows a sequence of manual checks by its collaborative partners to determine the accuracy of the content to reduce the distribution when necessary.

Google News Initiative is one of Googles many answers to the spread of disinformation "One of our primary goals for the Google News Initiative is to support journalists with the tools, training, and resources to help them find, verify and tell engaging stories across the web" (Google News Initiative, 31 May,22). The program committed $9.5million in 2020 and 2021 to fight COVID-19 disinformation and now has over 7000 different partners.

One of the latest models in Natural Language Processing (NLP) was created by Google and is named BERT (Bidirectional Encoder Representations from Transformers). The model boasts increased accuracy on "eleven natural language processing tasks, including pushing the GLUE (General Language Understanding Evaluation) score to 80.5% (7.7% absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement)" (Devlin et al., 2019). BERT is currently used for most language search queries made through Google. In addition to excelling at next word prediction, and next sentence prediction, BERT has a wide range of applications in the field of NLP. One of these use cases, which will be expanded upon in this paper, is classification. During classification, the model assigns classes, tags, or categories to the input content it is analysing. The built-in method of classification has a wide range of applications and is currently the most used method used in the field of automated fake news detection. (Alshemali & Kalita, 2020).

In a survey of 150 Deep Learning networks, it has been stated that Deep learning models have surpassed machine learning models in various NPL tasks, specifically in News categorization (Minaee et al., 2022). The survey gives insight into these models' complexity, and one example is the BERT Large model which consists of 340million parameters and is trained on 3.3 billion words.

The latest from Microsoft is MT-DNN (Multi-Task Deep Neural Networks), and this program incorporates BERT and further improves upon the GLUE score with a new benchmark of 82.7%, which is a 2.2% absolute improvement. (Liu et al., 2019). In a subsequent study, other newer models that are outperforming BERT in NLP classifications were found (Minaee et al., 2022). For fake news detection in this literature review, it was found that the most relevant research incorporated the BERT model. BERT uses a Bidirectional approach making It optimal for processing sentence structures whereas previous models such as ELMO only can analyse text from right to left or left to right. BERT can process both on a sentence level and token level (Devlin et al., 2019). Furthermore, BERT exceeds at allowing the user to utilize both fine-tuning and feature-based approaches. The paper displays evidence of the value of a rich pre-trained dataset in NLP. Even low-resource tasks can benefit from a bidirectional architecture in their models.

3. Design and Implementation

At the beginning of the project, a "Bag of Words" (BoW) approach was considered. As the literature review progressed in parallel with the artefact production, BERT became a more appropriate choice for the project. Because BERT is a more recent NLP framework, the decision was made to proceed with that model. BERTs advantage over previous frameworks is that it can weigh sentences, not only single-word tokens. BERT's performance on a variety of natural language understanding tasks was state-of-the-art when it was first released. This improvement over previous methods stems from the fact that it can work through a corpus in both directions, allowing it to assess the context of entire phrases and texts. BERT enables transferable learning because it is pre-trained on a vast corpus. This makes it incredibly useful for generating models for smaller, more specialized tasks. The model may be fine-tuned on a small dataset and deployed quickly.

We used the BERT Base uncased model and the matching preprocessing model for our artefact. BERT is available in two sizes; BERT Base and BERT Large. The BERT Base model contains 12 layers of encoders and around 110 million trainable parameters, whereas the BERT Large model has 24 layers and about 340 million trainable parameters. Apart from adding one or more classification output layers, the BERT models are identical for all tasks.

BERT Large was tested during the project but produced inferior results than the Base model when used on the same dataset, with otherwise equal model fitting. The BERT pre-processor simplifies the data cleaning process, saving time when building an NLP model. It accepts raw text as input and utilizes the following traditional preprocessing steps:

- Lower-casing the corpus
- Removing punctuation
- Removing stop words
- Tokenizing the corpus
- Stemming and lemmatization
- Learning word embeddings using Count Vectorizer and TF-IDF

The WELFake dataset was chosen for training the model. It consists of several datasets combined, put together for a research paper published in IEEE Transactions on Computational Social Systems (Verma et al., 2021). It contains approximately 72 000 articles, labelled “1” for real news, and “0” for fake news. It was selected in large part because of its’ size, and the fact that it has been published in a highly ranked journal. Working with such a large corpus increased model training time and was initially done out of a belief that it would improve model accuracy. Further testing showed that after a certain threshold the amount of training features did not make a significant difference.

The Python libraries TensorFlow, Scikit-learn, and pandas were used to implement the model. The dataset, which was in a .csv format, was loaded in a pandas DataFrame to simplify the handling and manipulation of the data. First, the title and body were combined into one column to train BERT on both. After the data had been prepared, it was processed using Scikit-Learns’ train_test_split function. This function partitions the dataset into a training set and a testing set, allowing the model to be tested on data it has not been trained on to avoid false results. At this point, the data is ready for training.

TensorFlow is a collection of tools used to develop and train machine learning models. For this project, TensorFlow-Hub is used to download the BERT Base uncased model and the BERT preprocessing model. Then the Neural network layers are set up in Keras through TensorFlow, one 'dropout' layer and one 'dense' layer. The dropout layer randomly discards input data at a set rate to help prevent overfitting. The dense layer implements a sigmoid activation function, limiting the range of outputs from the neural network to a range of 0 to 1. The model is then compiled using the preconfigured Adam optimizer and a binary cross entropy loss function. The optimizer determines how the model learns while training, and the loss function quantifies the error between the output of the algorithm and the given target values. Attempts were made to tweak the learning rate of the Adam optimizer. The default learning rate for the version included in TensorFlow is 0.001. (Devlin et al., 2019) suggests a learning rate starting at 5e-5, or 0.00005. Through the training process these rates decay for each step until they reach zero by the end of the process. When these parameters have been configured the model is ready for training.

Using the TensorFlow's. fit function, the model can be fed training data. It also needs to know how many epochs it should run whereas one epoch is a full pass through the training dataset. In the BERT paper (Devlin et al., 2019), it is suggested that 2-4 epochs are enough for any BERT model training. The most significant increases in the model’s accuracy occurred in the first and second epoch, while further fitting of the model resulted in highly diminishing returns by the third and fourth epoch. The model developed for this project was fitted over 3 epochs.

4. Results

A BERT model complemented with a dropout layer and a dense neural network layer was used in this experiment. BERT Base was chosen for its high accuracy and the simplicity of implementing it, as it was shown to be faster to train than BERT Large. The BERT model trained during this project shows an accuracy of ~0.85 when training on the training set and ~0.87 when tested using the test partition of the same dataset. In the different training attempts made during artefact development, the “loss value” decreases rapidly until it reaches ~0.35-0.45 by the third epoch where it stagnates. Different approaches were tested in an attempt to achieve better results. The learning rate of the Adam optimizer was tweaked to 5e-5 (0.00005) and a gradual decay to 0 throughout the learning process. This is in line with what Devlin et al. (2019) states to be the optimal learning rate in their paper, however this gave subpar results compared to simply using the default rate of the optimizer packaged with TensorFlow, which starts at 0.001 before gradually decaying. This might be caused by faulty implementation. Returning to the standard Adam configuration, tests were run using the BERT Large model instead of the BERT Base model. This increased training time while giving slightly worse results, even though BERT Large performs better on most NLP tasks than the Base model. Changing batch size from the default value of 32 to values between 128-512 did also not improve accuracy scores. To improve the accuracy of the model, different variations of the dataset were tested. Training the model on either the article title or body did not yield better results.

Further tests reveal what seems to be a flaw in the models' pipeline. Testing on a test set partitioned from the same dataset as the training set yields expected results that correspond to the accuracy values shown during training. When using the trained model to predict articles from an unrelated dataset the result is completely opposite of expected values. The model wrongly predicts a large share of the datasets, and we can find no apparent cause for this. The corpus is seemingly in a similar state before being fed to the preprocessing model, and the binary labels correspond with the classification of real and fake articles. As shown in the confusion matrices below (fig. 1-4), the result when testing on unseen datasets are completely opposite of the results when testing on unseen data from the original dataset.

Figure 1 shows the trained model yielding in total an 87% accuracy on the test partition of the dataset used for training, while figure 2 shows the same model being used to classify articles from another dataset found on Kaggle (HASSAN, 2019). Several models were trained on other datasets, but the result remained the same. The models test well on unseen data partitioned from the same set as the training data but cannot reliably predict the class of data from other datasets, as seen in figure 3 and 4. It was theorized that this could be a result of overfitting to the training dataset. When the artefact development was first started the impression of the team was that training on a larger feature set would yield more accurate results. In the late stages of the development process this theory was subverted as trial runs using only 10% of the original training set returned equivalent results.

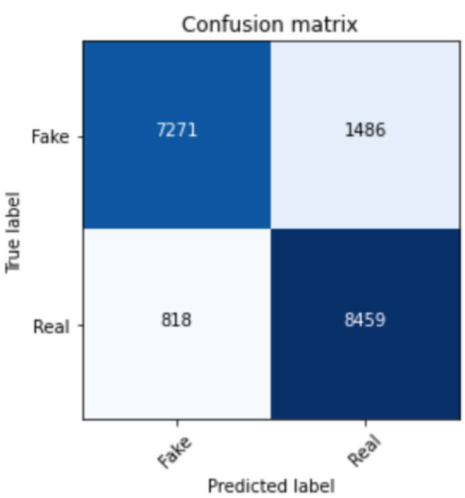


Figure 1. Model trained and tested on WELFake dataset.

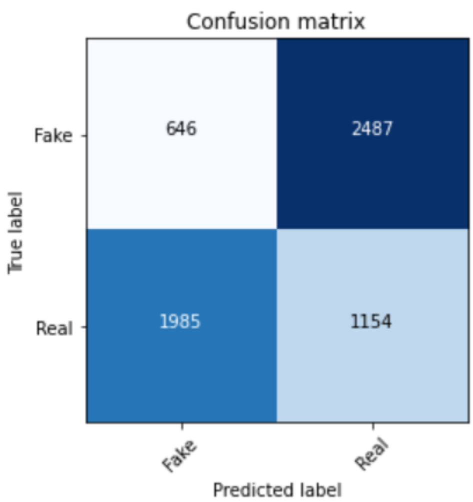
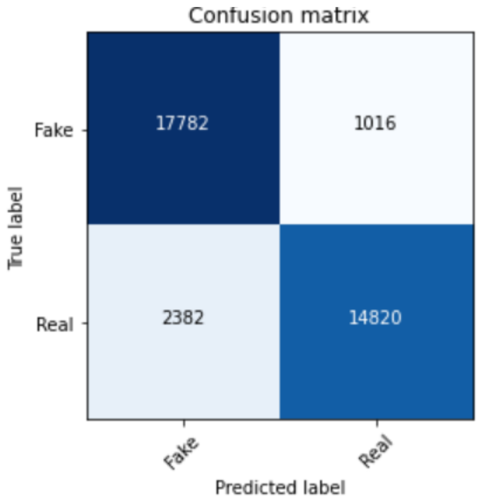


Figure 2. Model trained on WELFake dataset tested on different dataset



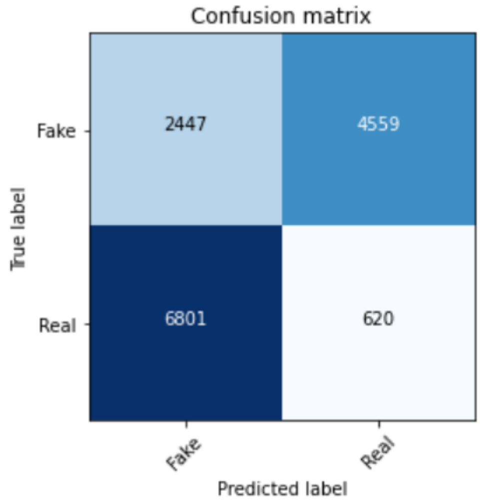


Figure 3. Model 2 trained and tested on dataset taken from Kaggle.

Figure 4. Model from fig. 3 tested on WELFake dataset

It is likely that using the BERT preprocessing model is a point of failure for this model. The first prototype that was produced before BERT became the focus of this project used Python libraries NLTK (Natural Language ToolKit) and RegEx (Regular Expressions) to clean the data. This did allow for more control and oversight of the process. When the choice was made to use BERT instead of more traditional Machine Learning techniques, using the BERT preprocessing model alongside felt like a more complete approach. Cleaning the data was now baked into the TensorFlow model for simpler training and testing with several datasets. This came at the cost of possibly missing important cues as to what was going wrong during training because the data was not examined closely enough. At the time of turning in this project, it is still uncertain exactly where the artefact failed. This might have been avoided, had a simpler approach been chosen, but learning about and working with all mentioned libraries and tools has proven a valuable experience.

5. Conclusion

Machine learning (ML) is used everywhere, for example when developing internet search engines and developing voice recognition applications. Another subset of Artificial Intelligence, Natural Language Processing, is vital to progress towards a data driven society, wherein the chosen topic of fake news is an important issue and an area where further NLP research is necessary. Through this study and analysis, the group has aimed to concretize the necessity of fake news detection, and the potential pitfalls of automatic detection.

Furthermore, we did an in-depth analysis of datasets and the toolkits available, we found BERT to be a state-of-the-art model, with a wide range of real-life use cases in fake new detection, making it ideal for this study. The BERT pre-processor makes data cleaning more manageable and saves time when developing an NLP model. The BERT model trained in this experiment gave satisfactory results on the test partition of the same dataset as it was trained. However, when the pre-trained model was used on unseen datasets, it resulted in a lower accuracy rate.

This project was challenging due to the complexity of machine learning models but at the same time a valuable experience in the field of data science. Working with datasets, researching, and learning about deep neural networks gave good insight into the field. There is a vast number of resources available in the chosen field of study but given time constraints and limited knowledge as first-year students, it is a complicated subject to understand. In conclusion, the artefact developed for this project did not deliver the results we were hoping for when we first started. Still, the group concludes that the subject is highly relevant in today's news landscape. In our opinion a model built on BERT could be a very useful method for classifying the legitimacy of news articles. Even though the artefact did not work as expected the project itself was an interesting endeavour that not only gave us insight into the strategies used in fighting misinformation but also allowed us to expand our knowledge into the world of data science through machine learning and artificial intelligence. We believe that further development in this project will create an attractive addition to a portfolio.

In continuing the work on this artefact, the main issue is the accuracy of the model. Further testing and potentially a full rework of the underlying code may be necessary to ensure accurate results on any data the model is fed, not only the testing set. After successfully training a model, the aim would be a practical implementation of the model as a web application that accepts either a hyperlink to a website or a raw text input. The model would then provide an output of "Real" or "Fake" after performing a binary classification on the input.

6. References

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Alshemali, B., & Kalita, J. (2020). Improving the Reliability of Deep Neural Networks in NLP: A Review. *Knowledge-Based Systems*, *191*, 105210. https://doi.org/10.1016/j.knosys.2019.105210

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, *497*, 38–55. https://doi.org/10.1016/j.ins.2019.05.035

Bowman, K., & Rugg, A. (2013). *Public Opinion on Conspiracy Theories*.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559. https://doi.org/10.1073/pnas.1517441113

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805

Gottfried, J., & Shearer, E. (2016). *News Use Across Social Media Platforms 2016*.

Hangloo, S., & Arora, B. (2021). *FAKE NEWS DETECTION TOOLS AND METHODS – A REVIEW*. 10.

HASSAN, A. (2019). *Fake News Balanced dataset for fake news analysis*. https://www.kaggle.com/datasets/hassanamin/textdb3

Hughes, H. C., & Waismel-Manor, I. (2021). The Macedonian Fake News Industry and the 2016 US Election. *PS: Political Science & Politics*, *54*(1), 19–23. https://doi.org/10.1017/S1049096520000992

Liu, X., He, P., Chen, W., & Gao, J. (2019). *Multi-Task Deep Neural Networks for Natural Language Understanding* (arXiv:1901.11504). arXiv. http://arxiv.org/abs/1901.11504

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2022). Deep Learning--based Text Classification: A Comprehensive Review. *ACM Computing Surveys*, *54*(3), 1–40. https://doi.org/10.1145/3439726

Mitchell, A., & Walker, M. (2021). *More Americans now say government should take steps to restrict false information online than in 2018* (p. 7).

Patel, D. (2021, August 28). *Text Classification Using BERT & Tensorflow | Deep Learning Tutorial 47 (Tensorflow, Keras & Python)*. https://www.youtube.com/watch?v=hOCDJyZ6quA&ab_channel=codebasics

Paul, C., & Matthews, M. (2016). *The Russian 'Firehose of Falsehood' Propaganda Model: Why It Might Work and Options to Counter It*. RAND Corporation. https://doi.org/10.7249/PE198

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521–2526. https://doi.org/10.1073/pnas.1806781116

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2017). *Automatic Detection of Fake News* (arXiv:1708.07104). arXiv. http://arxiv.org/abs/1708.07104

Read, M. (2016, November 9). Donald Trump won because of Facebook. *New York Magazine*. https://nymag.com/intelligencer/2016/11/donald-trump-won-because-of-facebook.html

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 15. https://doi.org/10.1145/3137597.3137600

Verma, P. K., Agrawal, P., Amorim, I., & Prodan, R. (2021). WELFake: Word Embedding Over Linguistic Features for Fake News Detection. *IEEE Transactions on Computational Social Systems*, *8*(4), 881–893. https://doi.org/10.1109/TCSS.2021.3068519

Walker, M., & Matsa, K. E. (2021). *News Consumption Across Social Media in 2021* (p. 15).

Wang, W. Y. (2017). *'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection* (arXiv:1705.00648). arXiv. http://arxiv.org/abs/1705.00648

Wasike, B., & Willies, O. (2021, December 20). *How to Build a Text Classification Model using BERT and Tensorflow*. https://www.section.io/engineering-education/classification-model-using-bert-and-tensorflow/

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, *57*(2), 102025. https://doi.org/10.1016/j.ipm.2019.03.004

Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2019). Fake News: Fundamental Theories, Detection Strategies and Challenges. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 836–837. https://doi.org/10.1145/3289600.3291382