Fake News Detection: An Unsupervised Approach

Dr. Yashvardhan Sharma, Ayush Kumar Tarun*1, Shreyas Samir Kolte#2

Department of Computer Science and Information Systems, BITS Pilani,
BITS Pilani, Vidyavihar Campus, Pilani – 333031, India

*1f20180258@pilani.bits-pilani.ac.in

#2f20180376@pilani.bits-pilani.ac.in

***Abstract*— This Report introduces a Novel Pipeline of Pre-trained State-of-the-art NLP Models along with Live web-scraping modules called FakeDetector. FakeDetector tries to mimic Human behavior in verifying whether a claim is real or fake. We have tried the model on benchmark datasets like LIAR and achieved results that are far better than any supervised or unsupervised model (an accuracy of *94.55%*). Thus we also prove our hypothesis that live web scraping as well as Unsupervised models (like our pipeline) are the future for Fake News Detection.**

***Keywords*— Fake News Detection, Semantic Similarity, Live Web-Scraping, Pipeline of Pre-trained Models, Human-like evaluation of claims**

1. Introduction

1.1. Background and motivation

Social media are a platform for quick transfer and access of information all over the world. Nowadays, however, misinformation travels just as fast as information, and one of the major sources of misinformation is fake news. Identification of fake news has become much more important a problem since the start of the COVID-19 pandemic in 2020. Moreover, the form or the context of the fake news can vary largely, i.e., fake news is not focused on specific domains of information such as political, socio-economical, healthcare, etc. or in any specific format (text-only, image only, etc.). It can be from any domain and come to us in any form. Thus, an approach for detecting fake news must be able to tackle data from any domain and possibly any format. Existing work in this field has been mainly done using Supervised Approaches. The results of these models are limited to a particular dataset from a particular time frame. Fake news detection is a more dynamic process and should not be limited to a particular time frame or domain. The model should not rely on being trained/tested on data being from a fixed time or domain. This is a major problem that we want to tackle.

1.2. Objective

Our final goal was to build a multi-modal fake news detector which works on Text-, Image- and Text+Image-based data using a pipeline of pretrained state-of-the art natural language processing and image processing models, along with live web-scraping modules. As of the date of submission of this report, the part of our model that handles text-based data has been completed and is completely unsupervised in nature. However, there are intricacies like correlation between the article and image, an edited image and others which are important for detecting fake news and can be best detected only by a supervised approach and hence, we wish to try an ensemble of our model and a state-of-the-art supervised model. We have only tackled Text-based data at the current

stage of the Pipeline, and yet it performs better than Multi-modal models as well.

2. Related Work

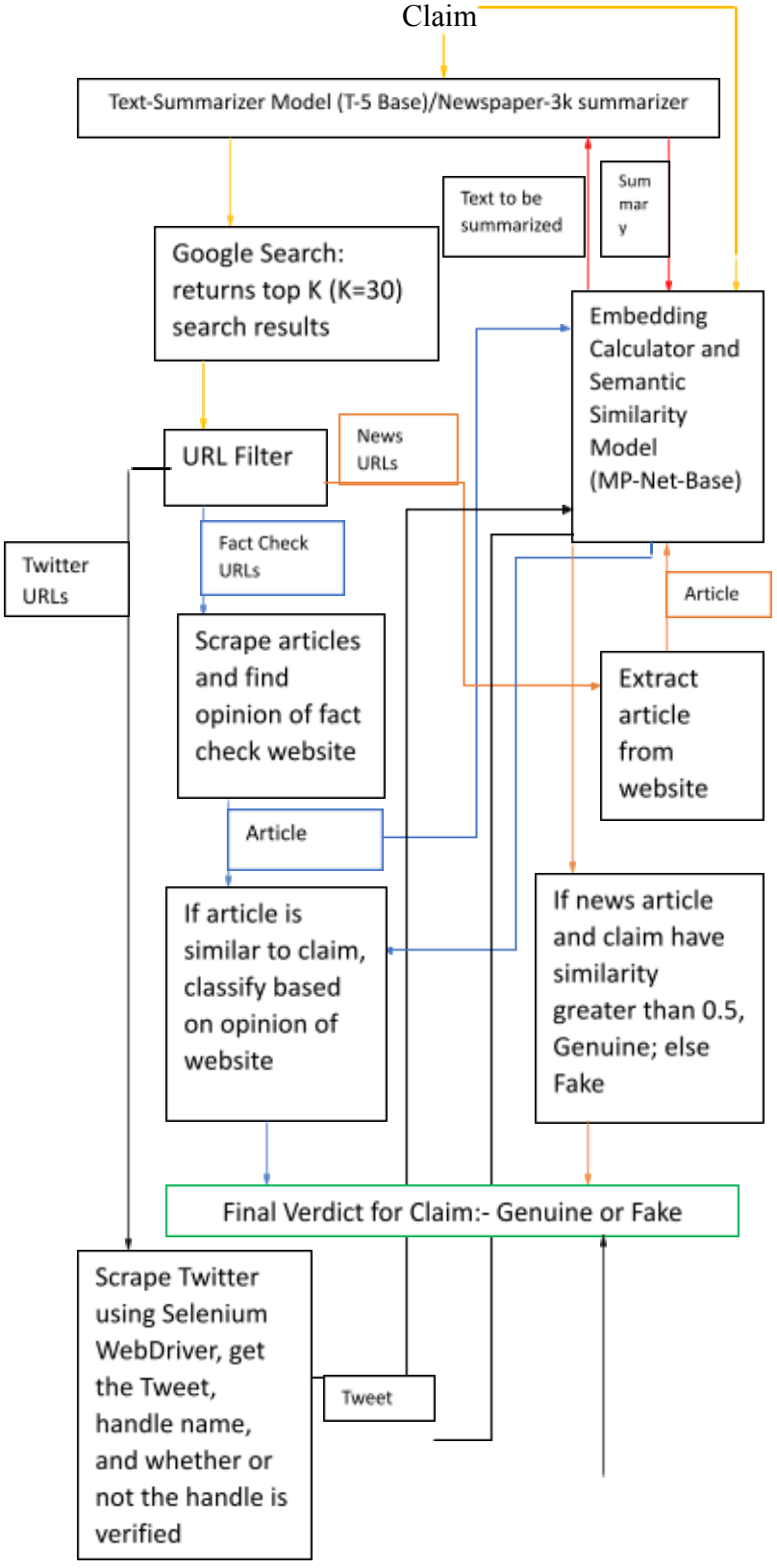
State-of-the-art supervised models for fake news detection include MVAE [1], EANN [2], and SpotFake [3]. These models have been tried and tested on benchmark datasets and have achieved significantly good results on the datasets. These show that Multi-modal models can perform considerably better than unimodal models by combining information from both the text and image. EANN and MVAE both extract embeddings of text and images through state-of-the-art models and concatenate them. EANN then feeds this vector to two fully connected neural network classifiers, one for event discriminator and another for fake news classification. MVAE feeds this vector into a decoder for reconstructing the original samples. The same latent vectors are also used for fake news detection. These models can have problems while generalizing as here the main classifier is always trained in tandem with a secondary task, and there may be a lack of data for the same. SpotFake and SpotFake+ improve over these by classifying without the help of any other sub-task.

Recent work in the field of fake news detection using unsupervised approaches includes [4] and [5]. The approach of [4] is based on Bayesian Modelling. The model is tested on a dataset from twitter. The exact approach is as follows: The classification of each of tweet is modelled using a Bernoulli distribution, and the parameter of the Bernoulli distribution is modelled using a Beta-Distribution. Along with this, the ‘opinions’ (of them thinking whether the claim in the tweet is “True” or “Fake” of ‘Reliable’ and normal twitter users are modelled using Gamma distributions. All of these become the prior distributions. Now, the tweet is searched for using a Twitter Search API, and the opinions of the reliable and unreliable users is gathered. Using this and the prior distributions, a score is generated on which the tweet is classified. The approach of [5] is similar to our approach, i.e., to find reliable news articles related to the claim (from a dataset of WhatsApp messages) and to find the semantic similarity between the news articles and the claim. Then, depending on whether the claim has similarity greater than a threshold, the claim is classified as real or fake. A major limitation of this model is that it checks for the similarity of articles with claims from a fixed dataset of articles. Our text-based model is derived from this approach to quite some extent, and we tackle their main limitation problem using live web-scraping. In [6] the method proposed is a tensor modeling of the problem, where we capture latent relations between articles and terms, as well as spatial/contextual relations between terms, towards unlocking the full potential of the content. Furthermore, they propose an ensemble method which judiciously combines and consolidates results form different tensor decompositions into clean, coherent, and high-accuracy groups of articles that belong to different categories of false news. We have drawn inspiration in our methodology from both [4] and [5]. From [4], we have used the idea of scraping twitter and using tweets from Verified handles for determining the

genuineness of a claim, and from [5] we have used the idea of semantic similarity between a verified genuine piece of news and the claim for obtaining a genuineness score.

3. Proposed Method

The following is a chart-like overview of our pipeline.



Given to the bo
We shall now e

ur Pipeline FakeDetector: The various color encodings how the various elements being worked on

We have used the Google T5-base summarizer for getting an abstractive summary of the input text claim.

2. Searching for relevant Articles: Using the summary of the input text claim, we have used google search for extracting the topmost results (30 in number) and getting their URLs. From these, we only store those URLs which come from reliable sources like Times of India, BBC, CNN, etc. **We built a custom search engine for extracting results from relevant fact-checking websites, news websites and from twitter.**
3. Scraping the articles: We have used the newspaper3k library for extracting the article's text from the URLs stored in the above step. We have also extracted the 'top image' of the article, but we are not using it as of now.
4. Article Summarization: We have used the function available in the newspaper3k library to find out an extractive summary of each of the articles. We did not use T-5 Base here because that model has an input limit of 512 words.
5. Fact-Checking Websites: We have chosen a set of reliable fact checking websites, noted the specific part of the article where the website mentions its opinion about the claim (e.g. title, end of article, image containing level of

truth, etc.). We decide on the basis of this.

6. Twitter Scraping: We have used the Selenium WebDriver module for simulating the Google Chrome browser for scraping Twitter and extracting the Tweets, Handles, and whether or not the handle is Verified. We then check whether there is semantic similarity between the tweet and our claim. Based on this, if the handle is verified and the tweet is similar, we classify the claim as Genuine.
7. Semantic Similarity Computation: We had earlier used Bert-Large for finding the Sentence Encodings and then used these encodings for finding the semantic similarity. We found that the results were not good and the model failed to differentiate between basic sentences like “Prince is dead” and “Prince is not dead”. **Hence, after doing a lot of research on the best models for semantic similarity, we arrived at MP-Net [7], which performs extremely well on our datasets and gives good results.**
8. Final Classification: We find the article with highest similarity. If this is greater than or equal to the threshold value 0.5, we classify the article as Genuine, else it is fake.
9. The lifetime of a claim through our pipeline is as follows. Claim->{Google Search}->{Fact-Checking Websites followed by Twitter handles followed by News Websites for finding whether claim is Genuine or Not}->Claim is classified as Genuine or Fake.

4. Datasets and Their Description

Since our method is a completely unsupervised method, we did not have to train any component of the model. However, for comparing how our model performed with some state-of-the-art models, we chose the Test sets of 2 datasets: the LIAR dataset and the COVID-19 dataset. Their descriptions are as follows: -

1. LIAR Dataset: It is a publicly available dataset for fake news detection. A decade-long of 12.8K manually labelled short statements were collected in various contexts from politifact.com, which provides detailed analysis report and links to source documents for each case. This dataset can be used for fact-checking research as well. Notably, this new dataset is an order of magnitude larger than previously largest public fake news datasets of similar type. The LIAR dataset4 includes 12.8K human labelled short statements from politifact.com’s API, and each statement is evaluated by a politifact.com’s editor for its truthfulness. We have used only the test set which consists of 1284 claims.
2. COVID-19 Dataset: A publicly available dataset with more than 10,000 tweets labelled “real” or “fake”. This was part of a competition from November, 2020. We used the test set consisting of 2100 tweets for testing our pipeline.

5. Experiments and Results

LIAR Dataset: -

As has been stated earlier, the labels available in the dataset has various magnitudes of truthfulness. Specifically, there are 6 labels: True, Mostly True, Half-true, Barely True, False, and Pants-Fire. Since we build a binary classification model, we chose to label the first 3 categories as 'Genuine' and the other 3 as 'Fake'. We then tested our pipeline and got an **Accuracy of 94.55%** on the entire dataset. **This is far better than any state-of-the-art Supervised or Unsupervised Model.** The claims where our model failed were incomplete sentences, opinion-poll like questions, etc., where not even a human can determine whether the claim is real or fake.

COVID-19 Dataset: -

Due to constraints on time and some issues with the web scraping modules, we have not been able to test our pipeline on the complete test set. However, we tested on the first 500 data items and got an **accuracy of 72%,** which is far better than that of any other unsupervised model.

6. Conclusion and Future Work

Thus, we present our very novel Pipeline consisting of Pre-trained, state-of-the-art NLP models and Live Web-scraping Modules. This pipeline is practically useful and can be developed into an Application for Fake News Detection.

Future Work: We first need to fine-tune our pipeline w.r.t. scraping twitter and some other fact checking-websites to achieve even higher accuracies. We also intend to include image data (specifically image data with text embedded into it) into our pipeline, to make it an even more powerful model.

7. References

1. Singhal, Shivangi and Shah, Rajiv and Chakraborty, Tanmoy and Kumaraguru, Ponnurangam and Shin'ichi Satoh, SpotFake: A Multi-modal Framework for Fake News Detection, IEEE BigMM, 2019, http://precog.iiitd.edu.in/pubs/SpotFake-IEEE_BigMM.pdf
2. D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in The World Wide Web Conference, ser. WWW '19. New York, NY, USA: ACM, 2019, pp. 2915–2921. [Online].
3. Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "Eann: Event adversarial neural networks for multi-modal fake news detection," in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ser. KDD '18. New York, NY, USA: ACM, 2018, pp. 849–857. [Online].
4. Unsupervised Fake News Detection on Social Media: A Generative Approach, Copyright c 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org).
5. Unsupervised WhatsApp Fake News Detection using Semantic Search Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020) IEEE Xplore Part Number: CFP20K74-ART; ISBN: 978-1-7281-4876-2
6. Unsupervised Content-Based Identification of Fake News Articles with Tensor Decomposition Ensembles, © 2018 Copyright held by the owner/author(s). 123-4567-24-567/08/06. . . $15.00 DOI: 10.475/123 4
7. MPNet: Masked and Permuted Pre-training for Language Understanding, Advances in Neural Information Processing Systems 33 (NeurIPS 2020)