Team 19: Identification of fake news in an online news media

Tathagata Raha, 2018114017
Aayush Upadhyaya, 2019202010
Jeevesh Kataria, 2019201058
Pramud Bommakanti, 2020900011

Abstract

This document describes the project outline and the work done for the term project as part of the IRE course. In this work, we implement various models to detect fake news.

1. Introduction

With the advent of the internet, social media and online news media, there is an increasing number of news, posts that we get to see on a daily basis. However, sometimes some news, posts or articles do not clearly represent the factual information, but represent some form of biasing, factual incorrectness or hate speech, which clearly differs way from the actual information. Since the digital data is growing now at ever so fast rate, it calls for a automated digital mechanism to identify the veracity of such news articles or social media posts. Here we aim at experimenting with various models and coming up with the best approach that can help detect fake news in online news or social media posts.

2. Related Work

Several studies have been done to find out the whether AI can be used to effectively detect fake news.One such study finds that AI can effectively be used for such purpose [4], with accuracy over 90%. The most used method for automatic fake news detection is not just one classical machine learning technique, but instead a amalgamation of classic techniques coordinated by a neural network [3]. Recent studies have shown that fake and real news spread differently on social media, forming propagation patterns that could be harnessed for the automatic fake news detection. Propagation-based approaches have multiple advantages compared to their content-based counterparts, among which is language independence and better resilience to adversarial attacks [14].

1. **BERT**: BERT stands for Bidirectional Encoder Representations from Transformers.BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers.BERT adopts to wide range of tasks, such as question answering and language inference [7].

2. **XLNet**: The researchers from Carnegie Mellon University and Google have developed a new model, XLNet, for natural language processing (NLP) tasks such as reading comprehension, text classification, sentiment analysis, and others. XLNet is a generalized autoregressive pretraining method that leverages the best of both autoregressive language modeling (e.g., Transformer-XL) and autoencoding (e.g., BERT) while avoiding their limitations. [13]

3. **RoBERTa**: We find that BERT was significantly undertrained and propose an improved recipe for training BERT models, which we call RoBERTa, that can match or exceed the performance of all of the post-BERT methods [12].

Vlachos and Riedel(2014) [11], constructed a dataset for claim verification consisting of 106 claims, selecting data from fact-checking websites such as PolitiFact.

In the Fake-news challenge [1], the authors pose the problem of fake news identification as as stance detection problem: given a claim and an article, predict whether the article supports, refutes, observes (neutrally states the claim) or is irrelevant to the claim.

A differently motivated but closely related dataset is the one developed by Angeli and Manning (2014) [1] to evaluate natural logic inference for common sense reasoning.

3. Dataset

We have selected a few datasets for our fake news detection task. They are as follows:

Nela-GT 2019(NELA) This dataset is a improvement in the Nela-GT 2018 dataset which is a multi-labelled dataset

[1] `http://www.fakenewschallenge.org/`

which contained 713k articles from 194 news and media outlets including mainstream, hyper-partisan, and conspiracy sources.[5].

Covid-19 Fake News DataSet(COVIDFN) This is manually annotated dataset of 10,700 social media posts and articles of real and fake news on COVID19. [9]. This dataset is a part of an ongoing shared task where given a social media post, the objective of the shared task is to classify it into either fake or real news.

BanFake News Dataset(BNFK) This dataset consists of manually annotated datasets consisting of various fake and real news articles from different Bengali news sources. It consists of 3964 authentic news articles and 3238 fake news articles. [6]

4. Preprocessing

In the NELA dataset, we had three kinds of reliability labels: High, Mixed and Low. For a binary classification task, we have discarded the Mixed class. After removing the Mixed data points, there were 446251 data points with High reliability and 126009 data points with low reliability resulting in a High-Low ratio of 0.22. Thats why, we chose 50000 data points with high reliability and 50000 data points with low reliability to make a balanced dataset. We have used the standard methods of preprocessing text for both the NELA and COVIDFN dataset. It involves:

- Lowercasing the words
- Replacing irrevelant symbols with spaces
- Removing stopwords

For the BNFK dataset, we used only the first two steps because stopwords list for Bengali was not available.

5. Metrics for Evaluation

For all the three datasets and different approaches, we have reported macro F1-score and accuracy.

6. Baselines

We have implemented different simple baseline models for both NELA and COVIDFN datasets.

Word embeddings: We have chosed two different word embeddings for getting vector representations for our posts and sentences: **Word2Vec** and **tf-idf**. For the Word2Vec, we find embeddings for each word and take the mean of embeddings of each to get a 300-dimension vector representation for a text.

Models: After getting the word embeddings, we made a train-valid-test split of 0.7-0.1-0.2. Then, we train trained the following models:

- Naive Bayes
- Logistic regression
- Support Vector Machines
- Bagging models (Random Forests)
- Boosting models (XGBoost)

Results for NELA dataset: In table 1, we can see the results of Word2vec and tf-idf models for NELA-GT 2019 dataset.

Model	tf-idf	Word2vec
Naive Bayes Model	0.790	-
Linear Classifier	0.862	0.76
Bagging Model	0.825	0.80
Boosting Model	0.846	0.79
SVM Model	**0.881**	**0.82**

Table 1. F1-score for the baseline models on NELA-GT dataset

Results for COVIDFN dataset: In table 2, we can see the results of Word2vec and tf-idf models for COVIDFN dataset.

Model	tf-idf	Word2vec
Naive Bayes Model	0.90	-
Linear Classifier	0.91	0.88
Bagging Model	0.90	**0.91**
Boosting Model	0.91	**0.91**
SVM Model	**0.92**	0.90

Table 2. F1-score for the baseline models on COVIDFN dataset

7. Experiments

For our more advanced models we explored different transformer models. For the COVIDFN and NELA-GT dataset, we used the following pre-trained transformer models from HuggingFace repository and fine-tuned it to our classification task:

- bert-base-uncased
- distilbert-base-uncased
- textattack/roberta-base-SST-2
- google/electra-base
- xlnet-base-cased

For the NELA-GT dataset, the average number of tokens in each article was around 540 after removing stopwords with more than 1500 articles having more than 2000 tokens. We know that transformer models can take at max 512 tokens for training. To resolve this issue, we took two approaches:

- At first, we considered only the first 512 tokens for each article. We discarded the rest of the article. We consider this on the basis of the assumption that crux of a report or article is present at the top of the report.
- Then we implemented a model known as cascading transformers. This model is similar to as mentioned here [8]. Here we break down each article into smaller parts, say 200 tokens where the last 50 tokens of a chunk overlap with the first 50 tokens of next chunk for a particular article. This method helps in maintaining a continuity in between the articles.
- Apart from that, we also implemented transformer models on the titles of the articles.

For the COVIDFN dataset, we implemented the all the above transformer models.

For the BNFK dataset, we cannot use the above transformer models because they are built for English. Thats why, we have trained two multilingual models: multilingualBERT and xlm-roberta and one mono lingual model: bangla-bert-base.

8. Results

Table 3 show the Macro Averaged F1-score of different transformer models on the NELA-GT dataset on the first 512 tokens, cascading and title settings respectively. We have experimented with various hyperparameters for the transformers model. In our case, the hyperparameters we experimented with are the learning rate and overlapping tokens and sequence length in case of cascading transformers.

Model	512 tokens	Cascading	Titles
bert-base-uncased	0.702	0.881	0.856
distilbert-base-uncased	-	0.859	0.848
roberta-base	**0.744**	**0.905**	**0.866**
electra-base	0.721	0.873	0.861
xlnet-base-cased	-	0.865	0.852

Table 3. F1-score for the transformer models on NELA-GT dataset on titles, first 512 tokens and cascading tokens

Table 4 show the Macro Averaged F1-score of different transformer models on the posts of CovidFN dataset. Like the previous models, we have experimented with the learning rate of the models.

We have also reported the macro-averaged f1-score of different multilingual and monolingual models on the BanFake dataset in Table 5

Model	F1-score
bert-base-uncased	0.962
distilbert-base-uncased	0.957
roberta-base	**0.975**
electra-base	**0.972**
xlnet-base-cased	0.948

Table 4. F1-score for the transformer models on COVIDFN dataset

Model	F1-score
bert-base-multilingual-cased	0.59
xlm-roberta-base	0.64
sagorsarker/bangla-bert-base	**0.66**

Table 5. F1-score for the transformer models on BanFake dataset

9. Analysis

We observe that the TFIDF model, although simple was able to give a strong baseline. The Word2Vec model has performed slightly better than the baseline model.

We can observe that the cascading models performed a lot better than the first 512 tokens model because it stores the whole information of the article. roberta-base gave the best results with an f1-score of 0.905.

On the Covid Fake News dataset, again the roberta-base model outperform the other models with an f1-score of 0.975. Similar to the last dataset, electra-base also gave really good results.

For the BanFake dataset, the multilingual models outperform the multilingual models by a large margin. This is probably because the multilingual models like xlm-roberta and mBERT were trained on a smaller Bengali corpus as compared to other languages.

10. Streamlit App

We used streamlit to showcase our data app for models we built on covid and nelagt data sets. We have some of the best performed models available in the drop down. To use streamlit we saved all the final models, all the intensive preprocessing pipeline including text vectorization and all the different reliability classification models. We stored all the scikit learn models as pickle files and models from tensorflow as tf directory and torch model as *".pt"* files. In our app we provided user with two options to predict, in one user can manually input text and predict that and in the other one user can choose to predict for n random samples. Where n random samples are picked from either test or training data, which was used to train the model. In models we have Some of them are trained on content of fake/reliable news and some are trained on titles of fake/reliable news.

11. Code and Models

The code is made public at https://bit.ly/2IRgLNP. The models(as pretrained pickle files) are made available at this location https://bit.ly/3kLHs3Q

12. Summary and Future Work

In this work, we have tried and implemented the standard machine learning as baselines as well as the advanced transformer architecture like the roberta-base, bert-base for our fake news detection challenge.

In the future, we would like to implement Hierarchical Transformers for Long Document Classification. which contains a similar model like the cascading transformers that we implemented with a sequential network.[10]. We can also try to implement the Longformer architecture which can be used to encode long texts.[2] Also we would try to include both the title and article information through multi layer perceptron.

References

[1] Gabor Angeli and Christopher D Manning. Naturalli: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 534–545, 2014.

[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

[3] Rafael Garcia Ana Cristina Cardoso Durier da Silva, Fernando Vieira. Can machines learn to detect fake news? a survey focused on social media. 2019.

[4] Murat GoksuNadire Cavus. Fake news detection on social networks with artificial intelligence tools: Systematic literature review. 2019.

[5] Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. Nela-gt-2019: A large multi-labelled news dataset for the study of misinformation in news articles. 2020.

[6] Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar. BanFakeNews: A dataset for detecting fake news in Bangla. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2862–2871, Marseille, France, May 2020. European Language Resources Association.

[7] Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019.

[8] Raghavendra Pappagari, Piotr Żelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. Hierarchical transformers for long document classification, 2019.

[9] Srinivas PYKL Vineeth Guptha Gitanjali Kumari Md Shad Akhtar Asif Ekbal Amitava Das Tanmoy Chakraborty Parth Patwa, Shivam Sharma. Fighting an infodemic: Covid-19 fake news dataset. 2020.

[10] Jesús Villalba Yishay Carmiel Najim Dehak Raghavendra Pappagari, Piotr Żelasko. Hierarchical transformers for long document classification. 2019.

[11] Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA, June 2014. Association for Computational Linguistics.

[12] Naman Goyal Jingfei Du Mandar Joshi Danqi Chen Omer Levy Mike Lewis Luke Zettlemoyer Veselin Stoyanov Yinhan Liu, Myle Ott. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*, 2019.

[13] Yiming Yang Jaime Carbonell Ruslan Salakhutdinov Quoc V. Le Zhilin Yang, Zihang Dai. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv:1906.08237*, 2020.

[14] Meghana Moorthy Bhat Justin Hsu Zhixuan Zhou, Huankang Guan. Fake news detection via nlp is vulnerable to adversarial attacks. *arXiv:1901.09657*, 2019.