Fake News Detector and Data Bias

Zhige Wang

May 2, 2024

Abstract

This is the age of information where people encounter countless news article that they themselves would have difficulties to distinguish whether it's true or false. There have been many fake data detectors research and many models proposed, yet most of them failed to recognize that there exists bias in training data, that we can expect all news report written in the West has some bias towards the so called "US Adversaries". Should we expect the training data to contain any truthful news targeted towards countries such North Korea? The goal of this project is to display such bias and try to provide a solution for it. This project demonstrates the effect of trying to remove specific country names from the training data and yet that proves to be difficult, and further word-representation models are needed to better perform such a task.

Contents

---

1 Introduction

When developing language models, one problem that all developers needs to take care of are data bias, that the percentage the data with label 0 and label 1 might be very different and therefore cause problems. We have seen problems like this with sex and gender for resume reading models. And yet in the world of fake news detection, there seems to be a missing component that people failed to recognize the West itself have general bias towards the global south, or even countries known as "US Adversaries". Could we expect any news agency to depict truthful information about countries such as China, Russia, or North Korea, and can we expect these news agency to not lie about countries such as Israel. Therefore it is very important to recognize these issues and try to solve these biases.

2 Related Works

Many research has been done on the topic of fake news detection[1, 2, 3], yet only one seems to have mentioned the topic of data bias of this sort which is entity bias [4]. Their proposed solution is to generate a new model to deal with such bias with an increased accuracy

on average of 2%. So it suffices to say that the problem has yet to be solved as the real solution would be to provide a new dataset that is less biased, or introduce a way to remove the effect of specific countries/regions name as the easy way out. Introducing such a dataset would prove to be difficult as machine learning model generated fake news are much more likely to be identified than fake news written by humans[5]. And so the current solution seems to lie in the preprocessing stage of removing influence of all countries to get rid of the bias.

3 Method

3.1 Data Preparation

The dataset was composed of two separate dataframes, fake and true, representing fake and true news articles respectively. Each dataframe was assigned a label, with '0' for fake and '1' for true news, and combined into a single dataframe. This combined dataset was then randomized and its index reset to ensure an unbiased sample distribution. The dataset was then divided into the training dataset and the testing dataset with the testing dataset containing 20% of the data.

3.2 Text Processing and Tokenization

The text data underwent several preprocessing steps to facilitate model training. Initially, all occurrences of specific country names, derived from the 'pycountry' database, were replaced with the generic placeholder "country" to prevent the model from learning biased associations based on specific countries. Additionally, non-breaking spaces were replaced with standard spaces.

The preprocessed text was then tokenized using the spaCy natural language processing library with the en_core_web_sm model, configured to disable the parser and named entity recognizer to focus on tokenization and part-of-speech tagging. During tokenization, each document was further processed to remove stopwords and to lemmatize the remaining tokens, resulting in a list of significant words from each document.

A batch processing function was implemented to process the texts in chunks, optimizing the handling of large volumes of data. The resulting tokens from the training and test datasets were then rejoined into continuous strings for each document.

For vector representation, the texts were tokenized using the Tokenizer class from the ten-

sorflow.keras.preprocessing.text module, which was fitted on the training data. This tokenizer converts the texts into sequences of integers, where each integer represents a unique token. To standardize the length of input sequences, padding was applied to achieve a uniform length of 150 tokens using the pad_sequences function from keras.utils.

3.3 Model Architecture

The neural network architecture for classifying the news articles was a recurrent neural network (RNN) built using the Keras library with TensorFlow backend. The model featured an embedding layer with a dimensionality of 128 to transform the integer sequences into dense vectors. This was followed by a GRU (Gated Recurrent Unit) layer with 96 units, equipped with dropout rate of 0.4 to prevent overfitting and to allow the model to learn long-range dependencies of the sequence.

The GRU output was subjected to global max pooling to reduce the dimensionality and to capture the most significant features. Several dense layers followed, interspersed with dropout layers to enhance regularization. The dense layers utilized 'elu' (exponential linear unit) activation functions, except for the final output layer, which used a 'sigmoid' activation function for binary classification.

3.4 Compilation and Training

The model was compiled using the Adam optimizer and binary crossentropy loss function, which is appropriate for binary classification tasks. The training process was conducted over 4 epochs with validation on a separate test set to monitor the model's performance and to mitigate overfitting.

3.5 Evaluation

The model's performance was evaluated on the test set, measuring loss and accuracy. These metrics provided insights into the model's generalization capability and its effectiveness in distinguishing between fake and true news articles.

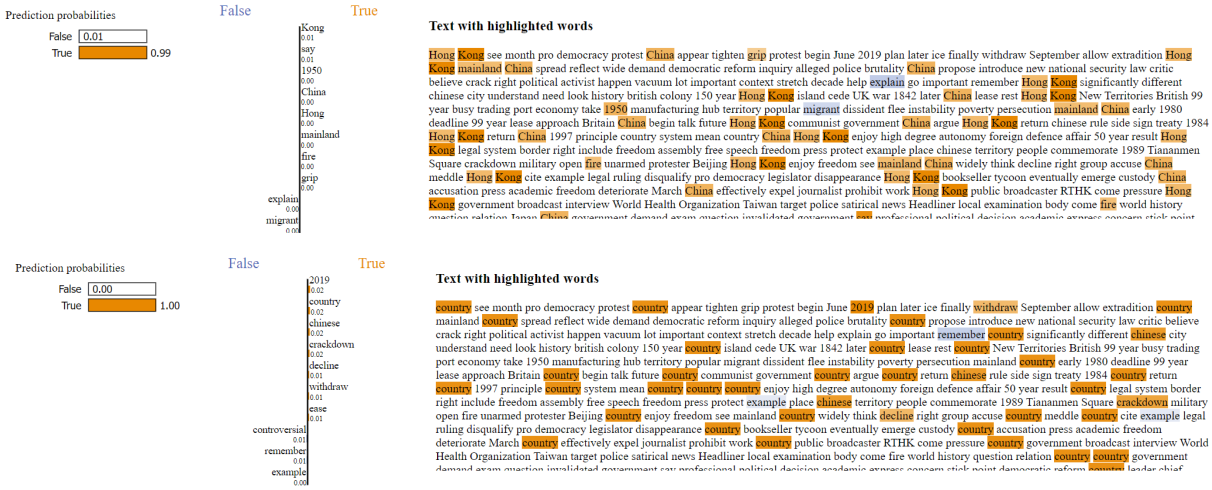

Figure 1: The above 2 images illustrates the LIME explanation result for both models. With the baseline model on the top and the improved model at the bottom. Both models inaccurately predicts the news article to be true with the main contributors to the baseline model being "Hong Kong" and "China", whereas the improved model's main contributors to be "country" and "Chinese"

4 Result

Two distinct versions of the models were trained to assess the impact of specific preprocessing steps on model performance. One version included a preprocessing step where all mentions of country names, as identified by the pycountry library, were replaced with the word "country." This version was designed to evaluate whether generalizing geographical identifiers affects the model's ability to classify news accurately. The second model, which served as the baseline, did not implement this replacement step. The baseline model was utilized for initial parameter tuning and served as a reference point for evaluating improvements in model performance resulting from the preprocessing adjustments involving country names.

Following meticulous parameter optimization, both models demonstrated robust performance, achieving an accuracy of at least 98% on the testing dataset, with a margin of error approximately around 1% due to inherent randomness. However, when subjected to a practical test using a contrived fake news article manufactured by the New York Times and analyzed using the LIME (Local Interpretable Model-agnostic Explanations) tool, the results were notably poor (Figure 1). This discrepancy highlights potential limitations in the models' ability to detect such type of data bias and replacing specific country names with filler words is not going to help.

5 Discussion

As highlighted in the introduction, the suboptimal performance of the baseline model was anticipated and served as the impetus for initiating this project. However, the disappointing results from the improved model, which attempted to mitigate bias by removing references to specific countries, also indicate that the mere mention of any country tends to be associated with true news. This suggests several potential issues, such as the dataset not being representative of the true distribution of fake and true news or the possibility that global news articles, which often mention other countries, are seldom labeled as fake in the dataset.

A further challenge was identified upon a detailed examination of the LIME results for the improved model. Although specific country names were removed, terms like "mainland," commonly referring to mainland China in the context of comparisons with Hong Kong or Taiwan, were not replaced. This oversight suggests that "mainland" is effectively a synonym for "China" in certain contexts and should also be substituted with a neutral term like "country."

For future research, it is crucial to conduct a thorough analysis of the dataset to determine why the placeholder "country" correlates strongly with genuine news. Additionally, the development of a machine learning model capable of recognizing and adjusting terms contextually similar to specific geographical identifiers, such as "mainland," will be essential. This approach will help in refining the model's ability to generalize across different types of news content without inheriting biases based on geographic references.

References

[1] Zhang, Xichen, and Ali A. Ghorbani. "An overview of online fake news: Characterization, detection, and discussion." Information Processing & Management 57.2 (2020): 102025.

[2] Jain, Akshay, and Amey Kasbe. "Fake news detection." 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). IEEE, 2018.

[3] Reis, Julio CS, et al. "Supervised learning for fake news detection." IEEE Intelligent Systems 34.2 (2019): 76-81.

[4] Zhu, Yongchun, et al. "Generalizing to the future: Mitigating entity bias in fake news detection." Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022.

[5] Su, Jinyan, et al. "Fake news detectors are biased against texts generated by large language models." arXiv preprint arXiv:2309.08674 (2023).