Fake News Detection using RoBERTa

Tejal Shanbhag
Department of Computer Science
Rochester Institute of Technology
Rochester, NY 14623
ts8583@rit.edu

Arjun Ravikumar
Department of Computer Science
Rochester Institute of Technology
Rochester, NY 14623
ar4038@rit.edu

Saloni Shah
Department of Computer Science
Rochester Institute of Technology
Rochester, NY 14623
ss7170@rit.edu

Abstract

Recently a lot of miss information and fake "verified" news has been spreading through different Social Media platforms. Most of them do not have a verified author or a verifies source. These fake news are created in such so that they appear as real as real news. These Fake news spread through Social Networking sites have caused mass chaos and economic loss to many. This raises the case for the need to judge the authenticity of the news before taking any action on the news article. This paper proposes a method to train a model to detect English Fake News on state-of-the-art RoBERTa and deploy the model on a website which will give the user the ability to test the authenticity of the news article and a novel method of giving the user an idea about why the model predicted so. We hope that instead of giving the user a simple Real or Fake the extra information will help the user determine if the prediction of the model is accurate. The final model created got a best 1 validation accuracy of 99.35% for one of the trails.

1 Introduction

In the day and age where social media has become the medium to share the thoughts and opinions of individuals the spread of fake news through this platform also has increased exponentially. Main factors leading to this is the lack of accountability of users for what they share. While most of the individuals share and spread these fake news without them knowing that they are fake, the creators of the news use these platforms to spread the fake news for malicious activity. The existence of fake news is not something which is relatively new but the amount of fake news being generated and the wide reach it gets is scary. The task of detecting fake news from real is not easy for an individual who is not an expert in the field, this is one of the main reasons why these social media platforms are exploited for spread of Fake News. Not very recently before the popularity of social media the main source of news was from reputed news media platforms, since these news came from a well established organisations they would have a team of experts who would validate the news before they are being send out. As any news if it turns out to be fake would affect the credibility of the organisation. In this paper we replace the team of experts using RoBERTa and developed a Fake News Detector website. The proposed method is trained and tested on multiple data sets consisting of more than 50,000 Fake and Real news in total. After which the trained model is used for testing current news and predicting if the news is Fake or Real. This is done by creating a website where individuals can enter the Title and Contents of the news and check the authenticity OF the news

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

provided. The website will also mention the most important group of words which contributed to the result.

The rest of the paper would be conducted as follows: Section 2 provides the motivation for choosing this topic. Section 3 provides information about the related work on the topic. Section 4 describes the Problem Statement. Section 5 will explain the dataset. Section 6 explains the methodology and architecture of the project. Section 7 discusses the results obtained. Section 8 describes the Lessons Learned throughout the project. Section 9 includes the future work which can be done on this project. Section 10 concludes the paper.

2 Motivation

Fake News has been an issue forever and is not something new. But recently the spread of fake news through social media has been one of the biggest issues discussed. Due to the nature of the current internet and the popularity of Social Media it has become more easy that ever to publish and spread Fake News. During the time of COVID-19 the fake news of how the virus originated, how the virus can be cured using household remedies and especially fake news regrading why it is not recommended to wear masks during the pandemic created chaos on a already vulnerable situation. These news had affected societies and people in general as they looked and seemed genuine. This is our main motivation for doing this project. To create a website where anyone can check the authenticity of the news.

3 Related Work

Research on fake news has being going on for the past many years most researchers in the past converted the data into tokens and these were then passed through some variation of Recurrent Neural Networks mostly LSTM[6]. However ever since the discovery of transformers and attention models the use of transformer based models for NLP tasks have significantly improved the accuracy of the prediction. Here we have studied some of the previous work done on this topic.

3.1 Fake News Detection

A previous research on the same topic was conducted in the paper "Natural Language Contents Evaluation System for Detecting Fake News using Deep Learning" [1]. This paper works with Korean News dataset this paper is very similar to our approach but uses BERT. The accuracy given in this paper using BERT was 96.25. But like most other papers this implementation also only gives out if the given news is Real or Fake and not what influenced the model to predict what it did. This leaves the user with no information. We plan to change this is and give the user more information as to why the model predicted what it did and what was the most important part in the news which influenced the model.

3.2 RoBERTa

RoBERTa [3] is an optimisation on BERT [4] architecture. The BERT is a multi-layer bidirectional Transformer encoder, which uses Next Sentence Prediction and masking techniques to train the attention part of the model. RoBERTa [3] builds on this model by training the model more and removing the Next Sentence Prediction. The authors of RoBERTa [3] suggested BERT was severely under trained hence trained the model with 10 times more data than the original BERT model. RoBERTa also modifies the masking technique by changing it from the static masking of BERT to dynamic masking.

4 Problem Statement

We plan to train our RoBERTa model with a classifier head on top using the datasets [2,5] and validate it with the same using an 95-5 approach. The validation step was done so that we know the model doesn't over fitting on the existing data. The accuracy and loss metrics of validation even though impressive was not the major concern for us. We plan to use a pre-trained model for the same as this will help the model learn the language faster. We then intent implement the trained model on a website which would give any individual the access to check the autenticity of any English language news also would inform the user about what caused the model to predict the way it did. Giving the user more information instead of a Real or Fake news we hope the extra information will help the individual understand the news and the model better.

5 Dataset

We used two datasets for training the model. Each of them were taken from Kaggle[6]. The datasets [2,5] were both in csv format. The dataset [2] has 6299 records. Out of these 6299 records about 50% of them are real news and the rest of them are fake news. The dataset [2] has 4 columns : ID, title,text and label.

The dataset[5] was is two separate csv files having fake and real news separately. This dataset[5] has 5 columns: title, text, subject, date and label.The dataset[5] has in all 44898 records.

Each of these datasets[2,5] have news from all categories with topics ranging from politics to world news. The content and titles are of varying length containing special characters, empty spaces and upper and lower case words. We handled all the possible situations to clean and prospect the data in the cleaning stage before giving the data for training. The frequency of lengths of the title and text after preprocessing them are are Figure 1 and Figure 2. There were 51,197 news records which was used for training and validation. We decided to give the model a high amount of training data so as it can consider all situations while learning.

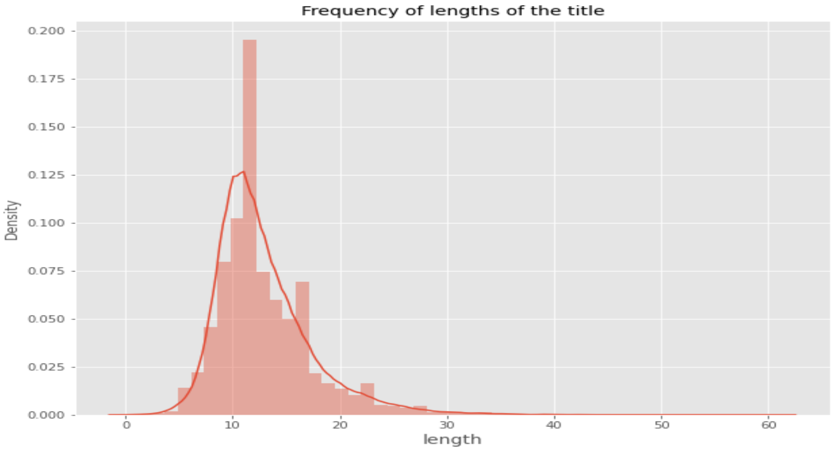


Figure 1: Length of title in the news dataset [2][5]

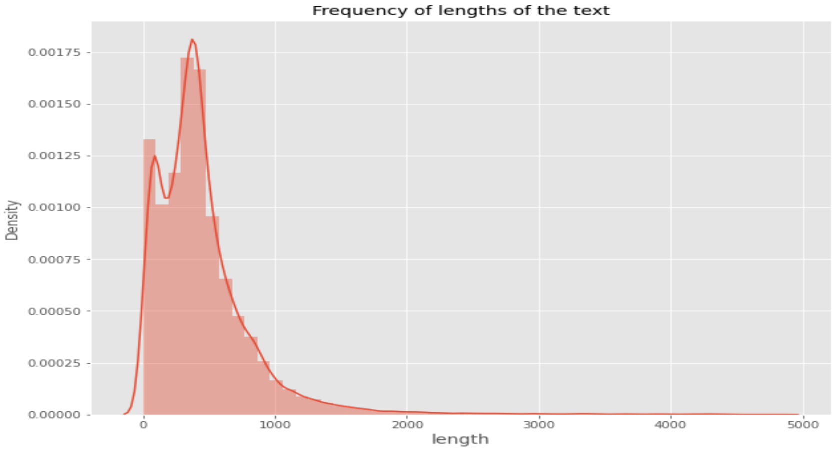


Figure 2: Length of text in the news datasets [2][5]

6 Methodology

The main problem we are trying to resolve with this project was to create an easy to use web application running on the trained model for individuals to test the authenticity of the news. For this we needed to train the model on as many data as we can and validate on very few. The choice of a very small validation set was due to the fact that every record we lose on validation was a record which we could train the model on. We also wanted to give the user enough data to show why the model predicted the way it did as with anything related to machine learning there a very small chance that the model would be wrong. Hence instead of the model just displaying Real or Fake we also will display what made the model decide the authenticity of the news. There are mainly four steps in the proposed approach. Which will be explained in the upcoming subsections.

6.1 Pre-Processing News

As the data in the dataset is in it's raw format there needed to be some text preprocessing before being send to the model for training. The aim of the preprocessing step was to remove any unique traits which either of the classes had and to convert all of the models to base text without loosing any information in the process. There are four textual preprocessing done on the dataset. While checking for the most common words in the Fake dataset and Real Dataset we found out that more than 90% of the Real data had the name of the news agency at the beginning of the data. Removing this was essential as it is not necessary that the name of the news organization need to be present and this would also make the model lazy by just looking for that specific feature in the given data. The next step was to remove any and all special characters as these characters do not provide any specific information toward the prediction. Then we wanted to remove any and all extra spaces which is present in the data thus removing all multiple spaces to single space. Finally we converted all the text to lowercase so that the model doesn't assign separate tokens just because the words are in a different case. We noticed that news agencies try to put some words in upper cases to convey the importance of the words. We didn't want the model to pickup on the case difference instead to pickup on the actual content and the style of writing the content to differentiate the classes.

We also made a few assumptions in the process about the data. We assumed as the data is being created by news organizations the will be spell checked and grammar checked. Hence we have not done spell correction on the data. We also have not done stemming and lemmatization as we didn't want to loose any information provided. We wanted to preserve as much information as possible in the training process. As grammatical mistakes and spelling mistakes are generally more likely to be found in fake news than in real news.

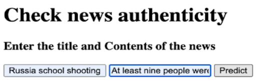
6.2 Training the Model

We started with a smaller dataset [2] and hoped to more epochs on this dataset. Even though this model converged and gave a really high validation accuracy in relatively quick time, the model worked terribly on current news data which we had tested. This made us use a bigger dataset [5] which we had mixed with the smaller dataset and this bigger dataset was split into validation and train sets.

We started by removed all the unwanted fields in the dataset and kept only the title, contents and label field. Then the title and the label were merged into a single field. We also tried to run the title and label separately and give weights to each of them but this did not give us any improvement in accuracy of the model. Each string of news was then clipped to a length of 256 words. After which we used a pre-trained RoBERTa model and implemented a classifier on top of the RoBERTa model. AdamW optimiser was user as the preferred optimiser because most papers which we referred as part of our ground work have which have used BERT and RoBERTa for classification preferred AdamW as the optimiser. This model was trained for 4 epochs on the dataset by splitting the data into batches of size 16. We trained the model on Google Colab on a GPU with 16 GB RAM. The total training and validation time is approximately 2 hours and 35 minutes. The epochs, batch size and the clipping length was decided due to limitations on Google Colab.

6.3 Website

We decided to use Flask to host the model along with the a HTML front end. Flask was selected due to it's simplicity to work with in python and because hosting a website in Google Colab will be easier with Flask. The website has two text boxes through which the user can input the title and the contents of the news. The website with the text and title passes the data to the trained RoBERTa model which will predict the authenticity of the news, mark the important word and redirect the user to a new page.



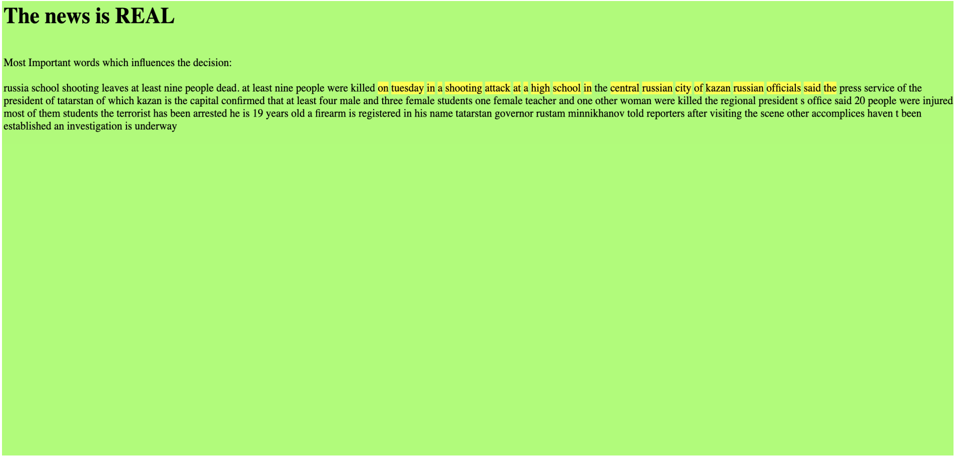
Check news authenticity

Enter the title and Contents of the news

Russia school shooting

At least nine people wen

Figure 3: Main Page

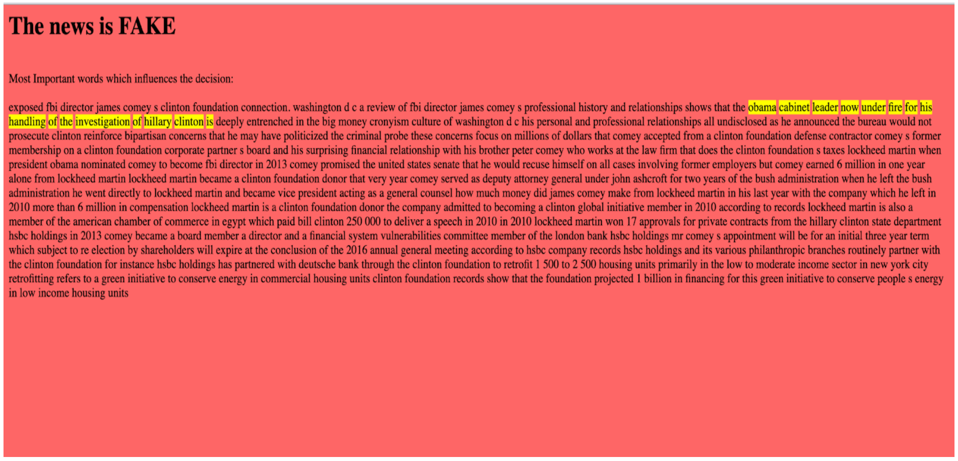


The news is REAL

Most Important words which influences the decision:

russia school shooting leaves at least nine people dead. at least nine people were killed on tuesday in a shooting attack at a high school in the central russian city of kazan russian officials said the press service of the president of tatarstan of which kazan is the capital confirmed that at least four male and three female students one female teacher and one other woman were killed the regional president s office said 20 people were injured most of them students the terrorist has been arrested he is 19 years old a firearm is registered in his name tatarstan governor rustam minnikhanov told reporters after visiting the scene other accomplices haven t been established an investigation is underway

Figure 4: Real News Detected



The news is FAKE

Most Important words which influences the decision:

exposed fbi director james comey s clinton foundation connection. washington d c a review of fbi director james comey s professional history and relationships shows that the obama cabinet leader now under fire for his handling of the investigation of hillary clinton is deeply entrenched in the big money cronyism culture of washington d c his personal and professional relationships all undisclosed as he announced the bureau would not prosecute clinton reinforce bipartisan concerns that he may have politicized the criminal probe these concerns focus on millions of dollars that comey accepted from a clinton foundation defense contractor comey s former membership on a clinton foundation corporate partner s board and his surprising financial relationship with his brother peter comey who works at the law firm that does the clinton foundation s taxes lockheed martin when president obama nominated comey to become fbi director in 2013 comey promised the united states senate that he would recuse himself on all cases involving former employers but comey earned 6 million in one year alone from lockheed martin lockheed martin became a clinton foundation donor that very year comey served as deputy attorney general under john ashcroft for two years of the bush administration when he left the bush administration he went directly to lockheed martin and became vice president acting as a general counsel how much money did james comey make from lockheed martin in his last year with the company which he left in 2010 more than 6 million in compensation lockheed martin is a clinton foundation donor the company admitted to becoming a clinton global initiative member in 2010 according to records lockheed martin is also a member of the american chamber of commerce in egypt which paid bill clinton 250 000 to deliver a speech in 2010 in 2010 lockheed martin won 17 approvals for private contracts from the hillary clinton state department hsbc holdings in 2013 comey became a board member a director and a financial system vulnerabilities committee member of the london bank hsbc holdings mr comey s appointment will be for an initial three year term which subject to re election by shareholders will expire at the conclusion of the 2016 annual general meeting according to hsbc company records hsbc holdings and its various philanthropic branches routinely partner with the clinton foundation for instance hsbc holdings has partnered with deutsche bank through the clinton foundation to retrofit 1 500 to 2 500 housing units primarily in the low to moderate income sector in new york city retrofitting refers to a green initiative to conserve energy in commercial housing units clinton foundation records show that the foundation projected 1 billion in financing for this green initiative to conserve people s energy in low income housing units

Figure 5: Fake News Detected

6.4 Inference of Fake News Detection

The title and contents received from the website is preprocessed and merged on to a single text. This text is initially passed on to test the authenticity by running the data through the RoBERTa model on eval mode. Once this is done the data is masked with all combindation of continuous words of length five to twenty. The dataset created with the masked data is batched into batches of 16 and passed onto the model for checking the most important masked sequence in the dataset. The logits receieved from the masked dataset is compared with the original dataset and the string with the maximum difference is chosen from them. This is the most important part of text in the whole sentence. We initially planned to mask words in all combinations continuous and non continuous from the original data randomly but soon had to change this plan as the number of dataset created with the removed words became exponential as the number of words increased. Also we felt that the words removed conveys more meaning when they are continuous.

7 Results

We conducted in total 14 trails and created 14 models in total. Since our main aim throughout the project was to create the model for predicting real and current news. The final results shown are being taken from the model with the least training and validation loss. Since the validation set is relatively smaller than what is usually practised for usual NLP training tasks we had a lot of fluctuation in validation accuracy. We received validation accuracy from the range of 74% - 99%. The Figure 6 shows the confusion matrix of the validation set after 4 epochs. Figure 7 shows the validation and accuracy loss of the model we can see that the validation loss almost became saturated at 0.065 after 3 iteration this pattern was consistent across multiple trials. We can also see that the validation accuracy also became stagnant after 3 epochs. This was one of the reasons for us choosing the number of epochs as 4. This helped us avoid over fitting the training data.

The second part of the results were the ones where we manually tested the website with actual daily news. We took most of our Real news from The Washington Post, The New York Times and The Reuters. And we took most of our Fake News from Valencia College Website[8]. We had in total tested 44 news which had 32 Real News and 12 Fake News. The model predicted 31 of the Real News as Real and all of the Fake news as Fake. This is really encouraging as these news were mostly just posted minutes before we took them hence there is no possible way the model had ever seen those data.

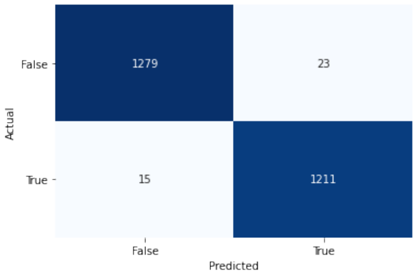


Figure 6: Confusion Matrix

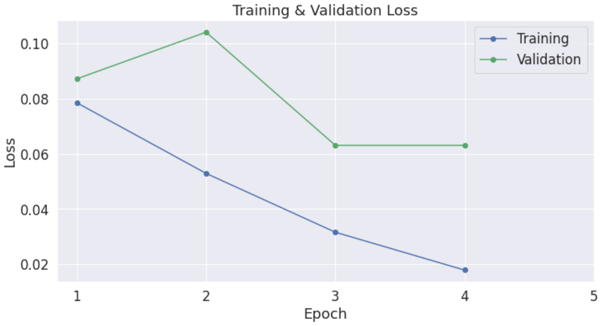


Figure 7: Training and Validation Loss

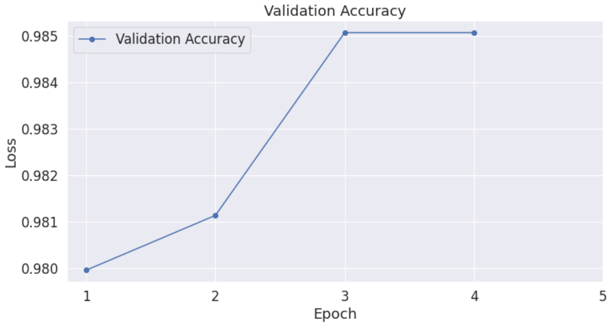


Figure 8: Validation Accuracy

8 Lessons Learned

While proposing the project we planned on using only one dataset [2] for training the model. However when we started implementing we learned that the model was performing with 43 % accuracy with only one dataset [2]. So we decided to use another dataset for training the model. This will help the model to learn all possible situations. After introducing the dataset [5] for training the model, the accuracy of the model increased as well. The model started to perform much better with current news data.

Along with this we also checked how our model was performing by changing only a few words at random places. Small changes in the news data by a word or date did not affect the over all prediction of the news. But however if the selected text to change was part of the set of words which influenced the prediction then after the change was executed the model chose a different set of words in the news as the most important part.

One of the major issue we faced while training was we couldn't train our model for higher epochs. The epochs and the word lengths were chosen because of the limitations on google colab. The higher epochs would enable us to try out a few more layers in the final classification layer and try and understand better as to why the model predicted the output.

One of the most influencing factors in the project was lack of actual Fake news. This was a major limitation in the project as even though we could attach the model to some famous news websites to make an ever training model. Due to the lack of Fake News this would make the dataset highly imbalanced.

9 Future Work

As a future work we would like to create the model more accessible and easier for users to test the news authenticity. We would want to create a scrapper which would scrape from a website the title and the contents when a user enters the link of the website. Another improvement for this was the need for scrapping of text from images. We planned to do a regression head on top of the RoBERTa model along with the current classification head for selecting the most important words instead of the finding the change in logits. But this requires a manual labelling of the dataset to identify the important words. We would also like to increase the clipping length of the words from 256 to 512 and use a dedicated GPU cluster to train the same. We would also like to improve the quality and feel of the webpage to make it look more appealing to the users.

10 Conclusion

Even though our implementation of Fake News detection has a very high accuracy and low average loss. This model has several limitation like language barrier, over dependence on text and not considering the source, type and date of the news. The practical implementation of the same is of more demand than ever however there are many major concerns to it. Since the model deals with very sensitive data any False prediction of Real or Fake can have catastrophic results. Thus any future improvement or real word implementation of this projects needs to done with at most care proper training. As the style of writing real news and fake news evolves every day the model also needs to evolve on a consistent basis. If the model if left stagnant and not trained on with newest content then the model accuracy will decrease drastically as the time increases. Overall we feel the novel idea presented in the paper of identifying the most important words in a given news and clearly marking them to the user will decrease the spread of fake news.

References

[1] Y. Ahn and C. Jeong, "Natural Language Contents Evaluation System for Detecting Fake News using Deep Learning," 2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE), Chonburi, Thailand, 2019, pp. 289-292, doi: 10.1109/JCSSE.2019.8864171.

[2] Real and Fake news dataset. https://www.kaggle.com/nopdev/real-and-fake-news-dataset

[3] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). "Roberta: A robustly optimized bert pretraining approach.", arXiv preprint arXiv:1907.11692.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805, 2018

[5] Fake and real news dataset. https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

[6] L. Yao and Y. Guan, "An Improved LSTM Structure for Natural Language Processing," 2018 IEEE International Conference of Safety Produce Informatization (IICSPI), 2018, pp. 565-569, doi: 10.1109/IICSPI.2018.8690387.

[7] Kaggle https://www.kaggle.com/

[8] Fake News Datset https://libguides.valenciacollege.edu/c.php?g=612299p=4251645