Detecting Context Misinformation in Text and Images

Spring 2022 project - Final Report

Author
Deepika Sivasankaran
Mohd Aamir

Supervisors
Raphaël Troncy
Paolo Papotti
Youri Peskine

Spring semester 2022

Contents

Detecting Context Misinformation in text and images

Deepika Sivasankaran, Mohd Aamir
EURECOM
Antibes, France

ABSTRACT

The objective of this project is to prevent the spread of fake news in the society. Fake news has affected our society both politically and culturally. It affects us everywhere from social media to a day-to-day conversation amongst groups. With the help of machine learning classification in a fitting way we plan to avoid this spread of fake news. With the help of the Fakeddit dataset, which is such a complete dataset including multimodal text and image data, comment data and fine-grained fake news categorization consisting over 1 million samples out from different categories of fakes news. The samples are labeled according to 2-way, 3-way and 6-way classification. We have modeled a hybrid text + image model to carry out different experiments for different ways of classification.

Code Repository: https://github.com/deepika2502/fake_news_EURECOM

1 INTRODUCTION

Fake news brings us panic and misunderstanding against the truth, especially under some unusual circumstances. It's crucial to detect fake news on social media early to avoid further propagation. Today, the problem of fake news has obtained significant consideration among researchers due to its harmful nature to deceive the people of the society. Fake news has an adverse impact on society as it may manipulate public opinions. Thus, it is essential to investigate the credibility of news articles shared on social media. Fake news detection has recently garnered much attention from researchers and developers alike**.** To build a fake news detection model, one must obtain sizeable and diversity training. Within this area of research, there are several existing published datasets. However, they have many constraints like limited size, modality etc. We don't have to face these issues as we are using Fakeddit which is large scale multimodal news dataset consisting of over 1 million samples containing text, image, metadata, and comments data from a highly diverse source. Each data sample consists of multiple labels, allowing users to utilize the dataset for 2-way, 3-way, and 6-way classification. This enables both high-level and fine grained fake classification. Finally, we evaluate our data through text, image, and text+image models with neural network architectures that integrate both the image and the text data. We run several different models which helps get a vast view of the classification results.

2 RELATED WORK

A variety of fake news dataset have been available recently. We are going to be using a small part of the Fakeddit dataset which is a multimodal dataset for fine grained fake news detection.

3 DATASET

3.1 Fakeddit

We sourced our dataset from Fakeddit, which consists of over 1 million submissions from different subreddits. It contains title and image, comments made by users who submitted the data along with subreddit score, number of comments and up-vote to down-vote ratio. We have three labels for each sample, allowing us to train 2-way, 3-way, and 6-way classification. The 2-way classification determines whether sample is fake or true. The 3-way classification determines whether a sample is completely true, the sample is fake and contains text that is true, the sample is fake and contains text that is true or the sample is fake with false text. The 6-way contains categories of different types of fake news than just a simple 2-way or 3-way classification. For the 6-way classification, the first label is true and the other five are defined as other five types of news.

True: True content is accurate in accordance with fact.

Satire/Parody: This category consists of content that spins true contemporary content with a satirical tone or information that makes it false.

Misleading Content: This category consists of information that is intentionally manipulated to fool the audience.

Imposter Content: This category contains two subreddits, which contain bot-generated content and are trained on a large number of other subreddits.

False Connection: Images in this category do not accurately support their text descriptions.

Manipulated Content: Content that has been purposely manipulated through manual photo editing or other forms of alteration.

2-WAY CLASSIFICATION	
TRUE	FALSE
6951	10671

Table 1

3-WAY		
TRUE	FALSE	PARTIALLY TRUE/FALSE
10671	910	6034

Table 2

6-way					
True	Satire	Misleading	Manipulated	False Connection	Imposter
10671	851	3447	590	872	1191

Table 3

We can see from Tables 1,2 and 3 above mentioning the number of labels for each way of classification that the data is not at all well distributed as we move from 2-way to 3-way to 6-way classification which affects the training of the classification models which we can clearly notice in the experiments and the results. The image dataset is also noisy since it contains some images that are inherently of no meaning.

3.2 Experiment Settings

We used different methods for text and image feature extraction. SBERT was used to generate text embeddings and ResNet50 was used to extract the features of the images. We used SBERT since it performs very well as a sentence embedding generator and provides great results on array classification tasks.

For the images, we used ResNet50, VGG16 models. Using the pretrained weights, a transfer learning-based approach was employed to extract features. Here the last layer was removed and the model was put in eval mode since there is no learning involved and the features were extracted without any gradients.

To be able to combine both the embeddings for pooling operations, the smallest embedding ie, the text embedding

was padded with 0s to match the length of the image embeddings.

4. Experiments

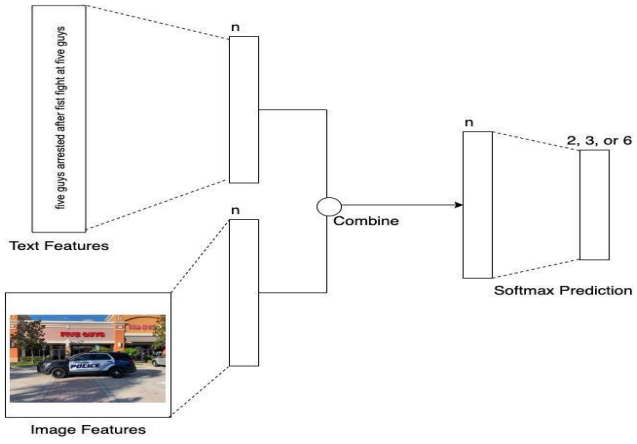
4.1 Text & Images Only

For **text only,** we used different versions of S-BERT on the sentences for 2-way, 3-way and 6-way classification.

For **images only**, we used different versions of ResNet i.e ResNet18, ResNet50 and ResNet101. Apart from this we also ran VGG16 and EfficientNet for the images only model.

4.2 Multimodal (Text + Image)

When combining the features in multimodal classification, we merged the text and image embeddings through four different methods: concatenate, maximum, minimum and average. These features were then passed through a Multi-Layer Perceptron. For all experiments, we tuned the hyperparameters to find optimal hyperparameters for the hidden layer. The MLP contains 4 Linear layers with Kaiming uniformity for the non-linearity and ReLU for activation. The number of nodes in each layer reduced from 100 to 30 to 6. The MLP was trained for 100 epochs for each type of labelling.



TYPE	MODEL	ACCURACY		
		2-way	3-way	6-way
TEXT	S-BERT(paraphrase-MiniLM-L6-v2)	0.80	0.79	0.73
	S-BERT(all-mpnet-base-v2)	0.84	0.82	0.75
IMAGE	ResNet 18	0.75	0.76	0.76
	ResNet 50	0.80	0.80	0.81
	ResNet 101	0.77	0.61	0.60
	VGG-16	0.78	-	-
	EfficientNet	0.76	0.80	0.78
TEXT + IMAGE	SBERT+ResNet by concat	0.82	0.79	0.61
	SBERT+ResNet by maxpool	0.80	0.79	0.62
	SBERT+ResNet by minpool	0.82	0.82	0.76
	SBERT+ResNet by avgpool	0.80	0.80	0.61
	SBERT+ResNet by minpool on SVM	0.84	0.83	0.81

Table 4

4.3 SVM Approach

Apart from the experiments performed in the Fakeddit paper, we conducted our own experiment in order to try finding a better approach towards finding a better result.

We used SVM with RBF kernel because in an MLP approach the model learns based on the parameter we give to the model. But with an SVM it fine tunes the parameters on its own which makes it less complicated yet effective.

Also, the boundaries of the SVM help provide a more refined classification between classes which we regarded as an important factor.

5. Results

The results are shown in Table obtained from the text, image and multimodal experiments. We can see that SBERT (all-mpnet-base-v2) perform better than the other versions. For image ResNet50 for feature extraction performed better than the other models. Multimodal features performed the best where text and image embeddings were combined to give more information to understand the extent of fake-ness. The ResNet50+SBERT embeddings trained on SVM yielded the highest accuracy. Overall, the multimodal model that combined the embeddings by the elementwise minimum method performed most optimally. Finally, we can see the SVM performs better than the MLP since the SVM adjusts its hyperplane on its own to obtain a clear boundary between the various types of labelling. Thus, the SVM is better than the MLP in our experiments.

Although the results seem fairly good, the results are not reliable since the dataset is biased due to the imbalanced number of instances in each class.

6. CONCLUSION

In this paper, we used subset of the novel dataset used in the original paper for fake news detection provided by Fakeddit. It contained a large number of multimodal samples with multiple labels for various level of fine-grained classification. We conducted several experiments with multiple baseline models and compared the accuracy of the different models.

Since the results are biased, to fine tune the model for more reliability on its predictions, we must train

with more and more data.

For **future research**, we can also try to look into tacking a user's credibility through different comment data provided. Implicit fact-checking could also be done. Also, since Image captioning involves the use of RNN and LSTMs to generate caption for a given image. The inconsistencies that are obtained with the image and the captions can be analyzed in the same. Furthermore, the project can be extended to work with videos too, where the inaccuracy obtained while captioning the videos could be eliminated.

7. Acknowledgement

We would like to thank professor Raphael Troncy and Paolo Papotti along with Youri Peskine for supervising and helping us throughout the project.

8. References:

[1]. r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection; Kai Nakamura, Sharon Levy, William Yang Wang; LREC 2020

[2]. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding; J. Devlin, M.-W. Chang, K. Lee, and K.Toutanova; 2019.

[3]. Very Deep Convolutional Networks for Large-Scale Image Recognition; K. Simonvan and A. Zisserman; 2015.

[4]. MLP references: https://pytorch.org/vision/master/generated/torchvision.ops.MLP.html

[5]. SVM reference: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC