University at Buffalo
The State University
of New York

Final Report
for
CSE 555: Introduction to Pattern Recognition

Submitted to:
Sreyasee Das Bhattacharjee
Department of Computer Science
University at Buffalo

Multimodal Fake News Classification

by:
Group 15
Mahek Sangwan
Nishant Kapoor

December 10th 2020

Table of Contents

1. Introduction

Fake news or disinformation is one of the greatest challenges in today's digitized world. With a rapid spread of fake news having critical impacts like swaying elections, influencing people's bias and increasing political divide, classification and detection of fake news content is important. However, a lack of useful datasets has been posing a major problem for its research and detection model development. Nuances of machine learning and deep learning are being utilized to contest against this progressively expanding modern-problem. Existing fake news datasets do not provide multimodal text and image data, metadata, comment data, and fine-grained fake news categorization at a significant scale. The researchers at the University of California, Santa Barbara created Fakeddit, a multimodal dataset of over 1 million samples from multiple categories of fake news. Their approach processes information with the help of deep learning to process the information through several stages of review and labels the samples according to 2-way, 3-way and 6-way classification categories using hybrid text+image models. Fakeddit uses a 6-way classification to distinguish the information.

2. Related Work

S.no	Title	Critique	Future Scope	Reference
1	CSI: A Hybrid Deep Model for Fake News Detection	In this work, a model is proposed combining the three models for a better prediction. The performance of users who propagate fake news and articles, is measured. A model called CSI: Capture, Score, and Integrate is suggested.	A possible research direction would be building models incorporating machine learning concepts like reinforcement learning.	[1]
2	dEFEND: Explainable Fake News Detection	A method to make use of the content of news as well as user comments to know the explicable user comments is presented. Experiments are done on different datasets to show the accuracy of dEFEND for fake news detection.	The learning process can be improved by using fact-checking contents from expert journalists. The trustworthiness of the users who post different comments can be used to make fake news detection more efficient.	[2]
3	Multi-Source Multi-Class Fake News Detection	Approaches to merge data from various sources and to differentiate between numerous degrees of fakeness are introduced. A Multi-source Multi-class Fake news Detection framework is proposed for the same.	An all-encompassing set of features for fake news is not available despite the success of other work. This poses a greater threat for short statements because they are not significant enough for fake news detection.	[3]

3. Experiments & Results

Fake news classification is a method for preventing the set of rumours from the society. Initially, many methods have been previously implemented to improve the accuracy of the classifier. Classification is a task that requires the use of machine learning algorithms that learn how to assign a class label to examples from the problem domain.

- First, we implemented One-vs-rest model (logistic regression) for fake news classification. One-vs-rest (OvR) is an empirical method for using binary classification algorithms for multi-class classification. It involves splitting the multi-class dataset into multiple binary classification problems. A binary classifier is then trained on each binary classification problem and predictions are made using the model that is the most confident.
- We tried getting results from this binary classifier but the model didn't fit well and it resulted in low precision and accuracy.
- So, we used decision tree, random forest classifier and extra tree classifier models to obtain significant results. It works better because our data contains categorical variables and it is a multiclass classification.
- Natural Language Processing helps us to get more knowledge about the dataset. NLP processes include stopword removal, tokenization, stemming and lemmatization. To make the data more concise, stopword removal operation was performed that helped us get rid of the common words that are used in our day-to-day lives.
- Tokenization is a process used to break the sentences into tokens. Here, we had a dataset that contained column named "clean_title".
- We performed NLP operation on it for pre-processing of data. We applied word2vec model to our classification process. The purpose and usefulness of Word2vec is to group the vectors of similar words together in vector space. Word2vec creates vectors that are distributed numerical representations of word features, features such as the context of individual words.
- To reduce the dimensionality of data and pick the best features among all features, we made use of PCA. PCA stands for Principal component Analysis, it is an unsupervised, non-parametric statistical technique also used for filtering out noisy datasets.
- In this project we used to pick top 10 best features from vgg16 features of images. The problem faced in baseline was to make use of combination of text and image for the classification process and improve its accuracy.
- For Image, we extracted image url & saved in memory arrays in the form of pickle file. Whenever we need image, we deserialize the pickle file. We saved the images in pickle file because it takes too much time for extraction otherwise. We converted these images into vgg features by inserting each image into vgg16 pre-trained model and store these features into vgg_features.pkl.
- Combining all these features helped us get a good accuracy.

4. Challenges

There were numerous challenges that had to be overcome before reaching a good test accuracy. These were:

- identifying and removing unimportant features
- identifying and filling missing values
- converting text data into numerical features
- converting images into pre-trained vgg16 model features
- converting categorical features into numerical

- hyperparameter tuning of models

5. Conclusions

Using CNN and Long short term memory gives an accuracy of 30%. After we used a combination of random forest, decision tree and free multi-class prediction, the test accuracy improved significantly.
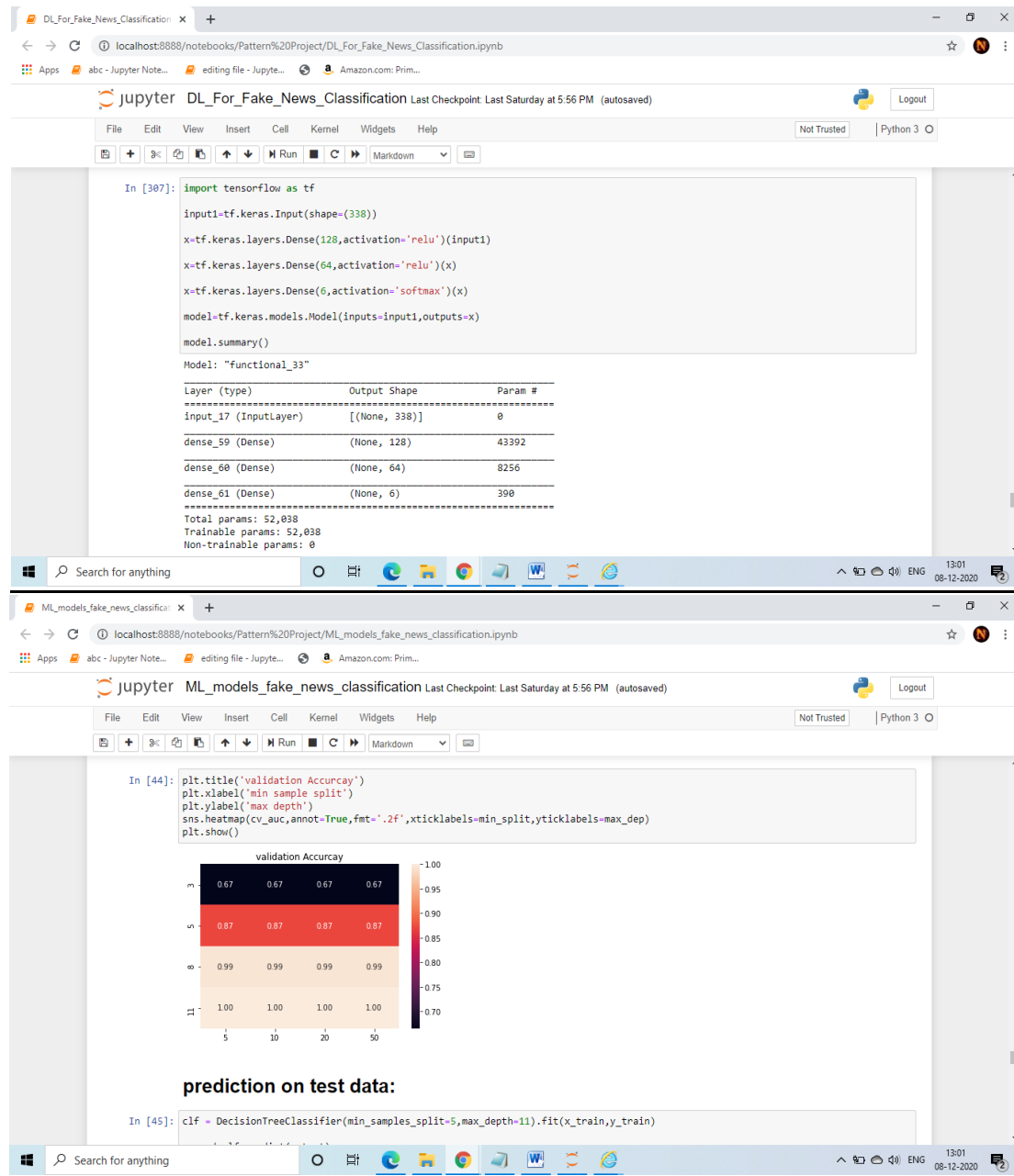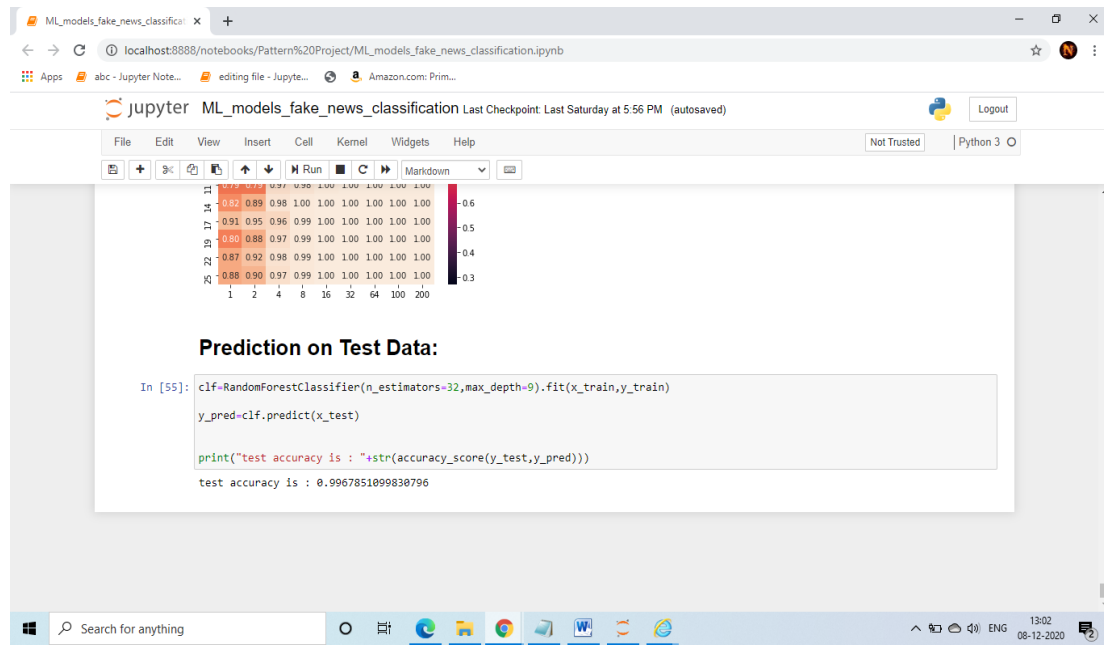
6. Contributions

Nishant Kapoor:
1. Model building using various ML techniques including Decision Trees, Random Forest.
2. Implemented vgg16 and word2vec model for feature extraction of text and images.
3. Hyper-parameter tuning and prediction on test data

Mahek Sangwan:
1. Pre-processing of data including NLP approaches such as stemming and stopwords removal.
2. Implemented one-hot encoder for converting categorical data into numerical data.
3. Hyper-parameter tuning and prediction on test data

References

[1] Ruchansky N., Seo S., Liu Y.,: CSI: A Hybrid Deep Model for Fake News Detection, 2017
[2] Shu K., Cui L., Wang S.,: dEFEND: Explainable Fake News Detection, 2019
[3] Karimi H., Tang J.,: Multi-Source Multi-Class Fake News Detection,2017