Fake News Detection
By - Akshit Khanna (2017A7PS0023P)

Two Approaches were broadly used to calculate test accuracy which are Classical Machine Learning techniques and Deep Learning Techniques.

Classical Machine Learning Techniques -

- **Binary Classification -** Logistic Regression , Naive Bayes and SVM all three techniques were tried on TF-IDF Vectorized text as input to the models. The Train and Validation datasets were used as the corpus to train the models. The TF-IDF vectorization worked better than Bag-of-Words and even Word2Vec Embedding for all the models. Surprisingly using both statement and justification led to lower accuracy in all cases. Word Embedding had 59.53 % accuracy on Logistic regression and 57.93 % on SVM on just statement. All functions were used from sklearn and gensim.

Binary Classification	Train,Val - Statement	Train,Val - Statement and Justification
Logistic Regression	**62.59 %**	59.51 %
Naive Bayes	62.19 %	57.74 %
SVM	62.12 %	59.67 %

- **6-Way Classification -** Logistic Regression , Naive Bayes and SVM all three techniques were again tried on TF-IDF Vectorized text as input to the models.The Train and Validation datasets were used as the corpus to train the models. Justification was missing for some rows and was replaced with the statement itself. All functions were used from sklearn and gensim.

6-Way Classification	Train,Val- Statement	Train,Val - Statement and Justification
Logistic Regression	25.56 %	23.99 %
Naive Bayes	26.12 %	22.10 %
SVM	25.89 %	23.28 %

Deep Learning Techniques -

- **Binary Classification -** A LSTM model is used for language modeling and then classifier is used using Fastai library which uses transfer learning to enhance the accuracy on small datasets. The Test dataset has an accuracy of 62.38 % .

- **6-Way Classification -** Similar models were used as in binary classification but the models were trained on both statement and justification to try to find a better accuracy. The model accuracy on just statement is **26.71 %** and on just justification it is 21.11% . The model accuracy on combination of statement and justification bi-LSTM is 26.12 % . These models are also made in Fastai library.

Instructions -

- There are 2 Ipython notebooks and they have both run google colabatory. All the commands for libraries and downloads are written in the notebook. It is recommended to run them on Google Colab for fast and easy execution.

Resources Used -

- Multiple online blogs were used for references for the task . Fastai Docs were also used a reference material for the deep learning part.

Libraries Used -

- Sklearn
- Fastai
- Gensim
- Numpy
- Pandas

IPython Notebook Links -

- https://colab.research.google.com/drive/1vNNMTRy5fvkMe92OmDBp9Tcru3h12sfe
- https://colab.research.google.com/drive/1y8j_D3A5KD_rm6_oSazdbXhadltwGYT8