Kaggle Fake News predictor challenge solution overview

Dataset

- Balanced dataset containing total around 20k examples and three features which are title, author and total text of news article.
- Task is to predict whether a news article is fake or real given these three features.

EDA

Checked the relationship of each feature with the class and also analysed individual features.

Author vs label

- There are 4202 unique authors only and hence different authors have written different number of articles
- By looking at the density plot of the number of articles written by each author, it was observed that a large number of authors have written very low numbers of articles.
- Hence not every author is of equal importance.
- To find out important authors, 95% percentile of the number of articles written by each author is considered as the threshold which was 17 articles.
- Out of these important authors, by looking at the percentage of fake articles written by each author, it was found that many of them always wrote only fake articles.
- List of such authors were generated and a model only with this feature could also give good accuracy.

News article title vs label

- Null values were replaced by 'Not Available' and all titles were cleaned. Anything apart from alphabets were replaced with blank space.
- Length of news article titles were calculated and the distribution plot of it showed that length of news title for each class is having different mean.
- Though the distribution plots are overlapping, this looks as an important feature as mean and variance values are quite different.
- Word clouds were generated for all, real and fake news titles.
- From word cloud it was understood that this dataset primararly covers political news from the USA (most probably 2016 US election!).
- In fake news titles, there are many words with large weight shows that most of the fake news are the same or at least use the same words repeatedly.
- Whereas words from real news titles have equal usages and New York Times is the most trusted word in many real news.

News text vs label

- Length of news article text was calculated and the distribution plot of it for both classes are similar.
- Distribution plots are overlapping and mean and variance values are quite similar.
- Word clouds were generated for all, real and fake news text. It showed the same pattern as seen with titles.

Model training

Below chart shows summary of models trained for prediction. Precision, Recall and Accuracy values are for randomly generated validation dataset.

Model	Class	Precision	Recall	Accuracy	Comments
Logistic Regression	0	0.88	0.99	0.93	Only “title” feature
	1	0.99	0.89		
Logistic Regression	0	0.96	0.96	0.96	All features
	1	0.96	0.96		
Logistic Regression	0	0.97	0.97	0.97	All features, Hyperparameter tuning
	1	0.97	0.97		
Random Forest	0	0.97	0.98	0.98	All features, Default parameter values
	1	0.98	0.98		
Distil-BERT	0	0.95	0.95	0.95	Only “title” feature
	1	0.95	0.95		

RandomForest model with default hyperparameter values was used. RandomizedSearchCV followed by GridSearchCV could be used to further boost accuracy.

Model prediction

RandomForest model trained with all features was used for prediction as it was giving highest accuracy among the other models.