**Problem solving approach**

As it’s a text classification problem determining the text as fake or real, I have divided the task into subtask as follows:

* Text Cleaning
* Modelling
* Results

Text Cleaning:

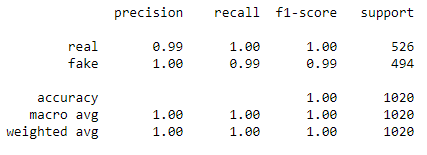
1. First, I have removed the empty text in the dataset.
2. Removed the duplicates text as it can bias the model over specific text.
3. Concatenated title, author, text and made a new column as full\_text.
4. Removed punctuation from text.
5. Converted all text to lower case because our model is case sensitive.
6. Used some contractions to make text similar and more promising.

Modelling: I have tried **bilstm** and **bert-base-cased** but bert model was giving better results so have finally gone with it.

1. BERT is a transformers model pretrained on a large corpus of English data in a self-supervised fashion.
2. It used MASK language modelling which allows the model to learn a bidirectional representation of the sentence.

Results: Using bert-base-cased modelling I got results as follows.

1. Precision and Recall for respective classes are as following



1. Almost all my labels were predicted accurately, confusion matrix of predicted label is as following.



1. Result can be found at ***fake\_submit.csv***