List of notebooks:

1. Preprocessing for testset of stage 1 classifier
2. Preparing positive class training set
3. Preparing negative class training set
4. Stage 1 classifier models
5. Stage 2 classifier models
6. Topic modelling
7. NER Casualty

List of text files:

1. violent_train.txt
2. not_violent_train.txt
3. 2019.tar.gz

1. Preprocessing for testset of stage 1 classifier:

This is a template for preprocessing the data of the test set being considered, on which the classifier models will be used in future. Only the input file needs to be changed in general (but may not work if articles from other news sources are in a different format)

- Lots of empty strings, taken care of them
- Somehow, some duplicate items were also present. Removed these redundancies.
- There were some words & phrases which were naturally present in all docs but would be of no use for the classification purpose. Removed them.
- A number of documents contained a promotional message for the newspaper and similar things which are not useful for our task. Removed them.

The end result, saved to .txt or .csv format, can now be used directly as the test set for performing the classification.

The next 2 notebooks are part of a 2-stage classification technique

2. Preparing positive class training set:

- Used violent_train.txt
- Filtered out from there the subset of the documents that contained:

a. IPC sections related to violence
b. High-precision keywords denoting positive class (attack, lynch, murder, etc)

- Added positive incidents identified from positive_incidents.xlsx
- All these combined form our new training set for positive incidents.

- This is aimed at building a high-precision training set. (with size about 70% of initial size)
- Idea is that training using this will lead to results where those predicted as positive are almost certain to be actually positive incidents as well

3. Preparing negative class training set:

This notebook tries to take care of some of the drawbacks of the 1st stage model and also achieve a higher degree of precision

- Preparation of negative class training set:
- Now aim is to make the negative training examples more precise
- This will help us deal with one of the drawbacks of the first stage model (negative examples which contain the violence related keywords and IPC sections)

- Used not_ violent_train.txt
- Filtered out from there the subset of the documents that contained:
    a. IPC sections
    b. Violence related keywords as discussed above
- Added negative incidents from positive_incidents.xlsx
- Made use of the results from the previous model
- Added the documents that had been predicted as negative instances
- All these together form our negative training set

- For the model building phase, we now use both these manually modified training sets(for both positive and negative classes)
- The size of negative training set is > positive training set. This is in keeping with the fact that the number of positive incidents in the test set is expected to be substantially less than that of others.

4. Stage 1 classifier models:

This notebook is related to building the stage 1 classifiers from the manually modified positive class training set (as stated above)

OBSERVATIONS:

- This stage 1 model was used this to predict on randomly sampled docs from test set
- A number of classifiers had been built for the same
- Best classifier was Random Forest on count_vectorised data

DRAWBACKS:

- Highly imbalanced dataset. Only about 5% of articles belong to the positive class. Very difficult to classify so accurately.
- There is a possibility of some incorrectly classified negative class as well

5. Stage 2 classifier models:

- Test set is again randomly sampled
- Best classifier in this case is Logistic Regression on tf-idf_ngrams at character level

DRAWBACKS:

- Similar problems of imbalanced dataset as discussed above

6. Topic modelling:

- Used LDA (latent dirichlet analysis) to come up with top few keywords of underlying topics in the data
- Highest coherence score was observed for around 30 topics
- Intuitively assigned topic label based on keywords
- Found out frequency of each topic using prevalence score
- Filtered out those topics which are unlikely to have any cases of violent incidents (sports, weather updates, movies, etc).
- This will help to decrease the size of data and also reduce the degree of imbalance

COMPARATIVE ANALYSIS OF RESULTS:

Model Details	Accuracy	Precision	F1 Score
2 months data - stage 1	0.93	0.31	0.43
2 months data - stage 2	0.93	0.42	0.49
full year - stage 1	0.95	0.4	0.35
full year - stage 2	0.96	0.4	0.44
filtered_full - stage 1	0.89	0.33	0.41
filtered_full - stage 2	0.94	0.4	0.48

- Accuracy isn't a very tangible metric here due to the high imbalance
- F1 score is the most reliable metric
- In all cases, stage 2 performs better than stage 1
- Filtered data (after topic modelling) does slightly better
- Other subsamples may perform even better

THE FOLLOWING ARE PART OF FURTHER SCOPE OF RESEARCH:

7. NER Casualty:

- NER (Named entity recognition) is a technique to identify the different type of entities present in a piece of text
- We are concerned with Cardinality
- Further, this will help to identify no of casualties in a violent incident
- CARDINALITY gives presence of numeric entities expressed in both integer and written forms
- (7 and seven)

Problems Encountered:

- Not the only presence of cardinality in the sentences, can denote other numbers as well
- Not necessarily in close proximity to words like murdered, killed, assaulted, etc
- RegEx can identify only in int64 format

COMPARISON OF TOPIC FREQUENCY WITH TIME: