**Capturing the steps to build model.- Model 1**

1. The pre-processing was done using stopwords, lower case conversion, word net lemmatizer and snowball stemming.
2. The input file was read for the columns cause, title case, summary case and the data was cleaned by removing the rows with no data
3. The pipeline library was used to combine multiple steps into one.
4. The data was split into training and test in the ratio of 1:3(though this step is not needed I guess)
5. DTM was created using the word occurance and the dtm was imported into a file dtm.csv
6. The svn classifier was built and the accuracy of the model was determined. The accuracy achieved was 57.69%
7. Gridsearch was employed but we got the following warning message: “Warning: The least populated class in y has only 2 members, which is too few. The minimum number of groups for any class cannot be less than n\_splits=3.

% (min\_groups, self.n\_splits)), Warning)”. So, we need to re-look into the classifications.

**Model 1.1**

1. Custom stopwords were added to remove specific words

custom\_stopwords = ["victim", "year", "morning", "afternoon","00","00pm","10","107","11","12","13","13th","14","15","16","18","20","200","2011","22","24th","25","26","26th","28","2km","30","32","34","35","360","39","3rd","40","43","45","47","50th","52","55","6th","75","80"]