Sheetal Report

The planned responsibilities for the code are to perform data-preprocessing, provide deep insights into the data using visual graphs and statistics and the develop a machine learning algorithm powered by Natural Language Processing that can identify reddit posts that belong in the science domain.

# Data Pre-processing

The data pre-processing section in the code handles all the common discrepancies in the dataset like duplicate and null instances in the dataset. It also reshapes the data to perfectly fit the needs of the code. This section of the code also provides analysis on the dataset using graphs and statistics, more specifically central tendency measures.

The features of the dataset are as follows:

Title: title of the reddit post

Score: reddit score for the post (upvotes)

Id: reddit post id

Subreddit: to which subreddit it belongs

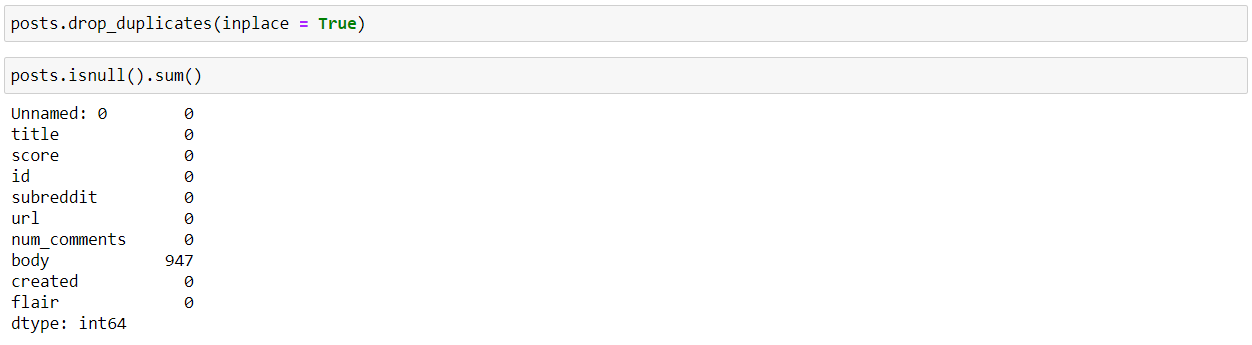
url: url of the reddit post

num\_comments: number of comments on the reddit post

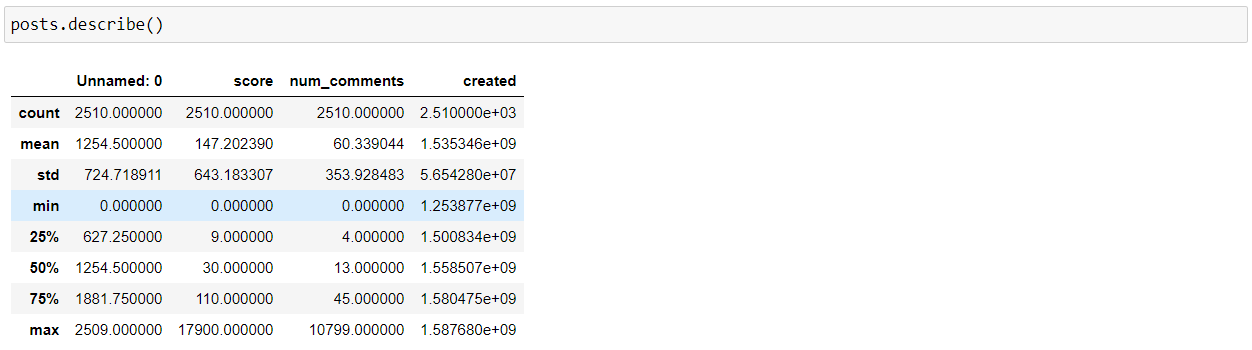
body: text content of the reddit post

created: date and time at which the post was created

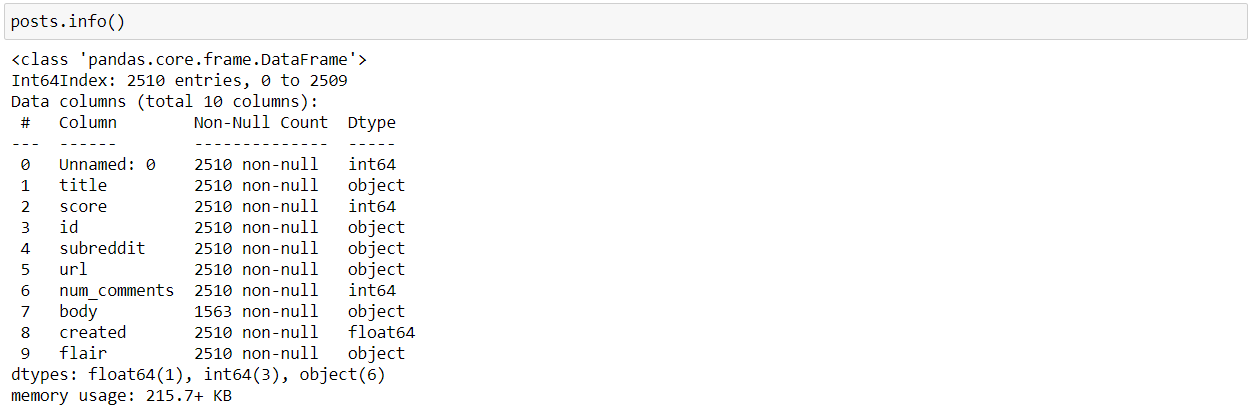
flair: the output class of the dataset denoting the niche of each reddit post instance



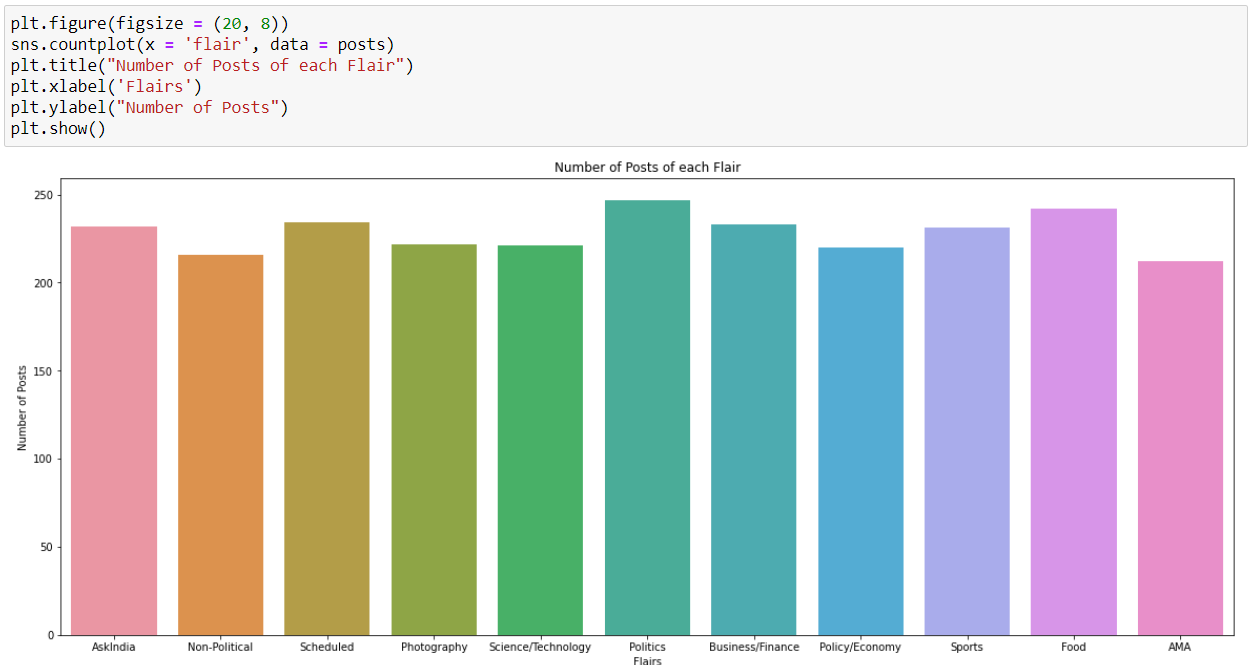
The figure above shows the code used to remove duplicate instances and view the number of null instances for each feature. The dataset is clean and ready for further use.



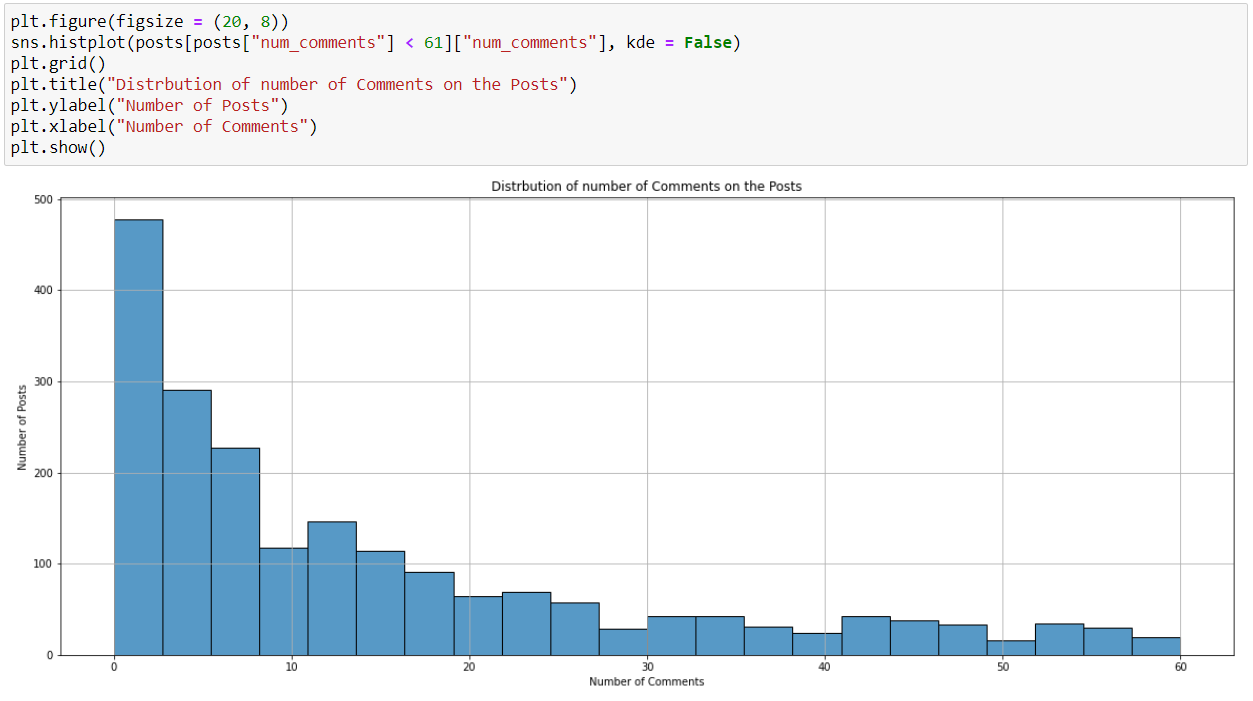
The figure above shows the code that provides a statistical analysis of each feature in the dataset using central tendency measures. These central tendency measures work best on numerical data only as the figure above only shows features with numerical data.



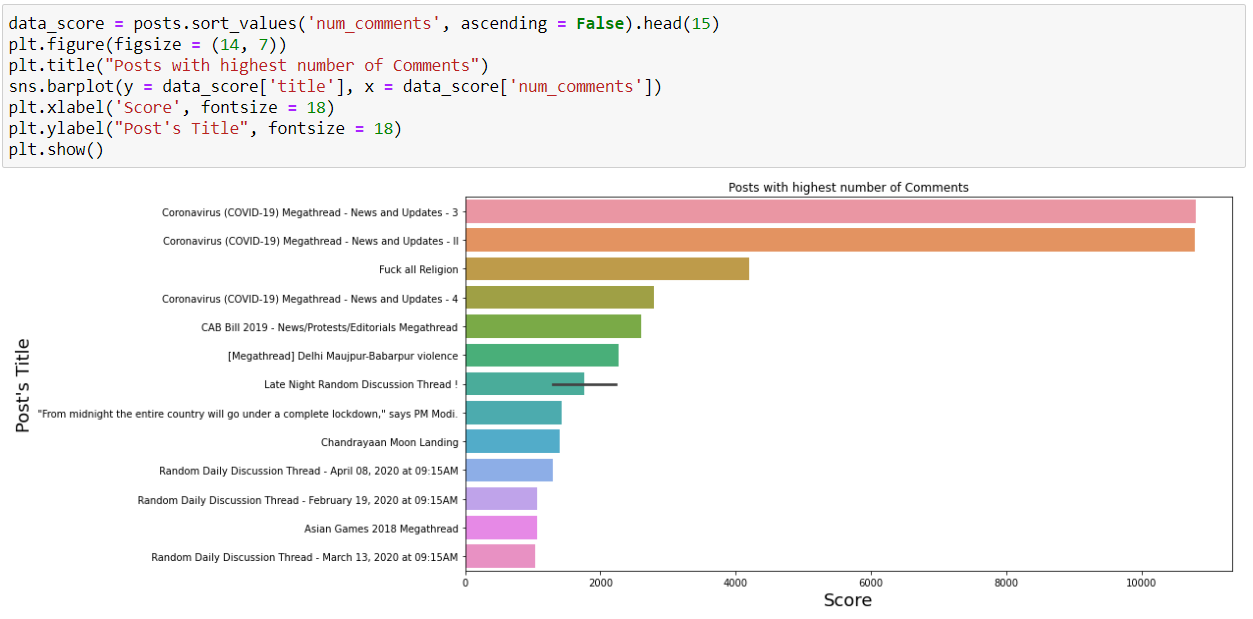
The code in the figure above shows the metadata of the dataset. The metadata contains the number of non-null instances for each column along with their names and data types. This is useful when you need to identify if any of the features in the dataset require any encoding or not since categorical features denoted by the “object” data type require encoding to be useful. The next is the visualization of data to study the trends in it.



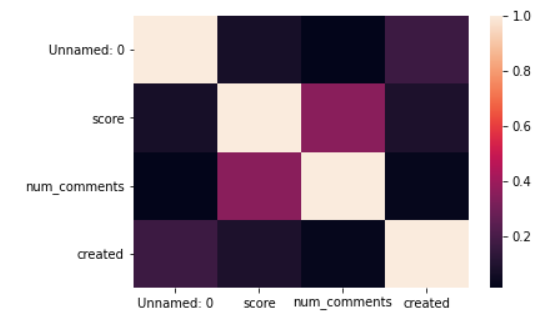
The graph above shows the number of reddit posts for each niche/flair. The following graph shows the distribution trend between the number of posts and the number of comments.



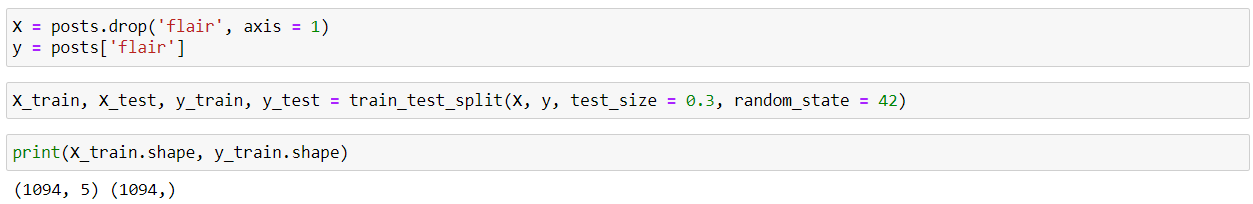
The following graph shows the top posts with respect to the number of comments on it.



To better understand how the features in the dataset correlate to each other and contribute to the dataset, a correlation matrix heatmap plotted.

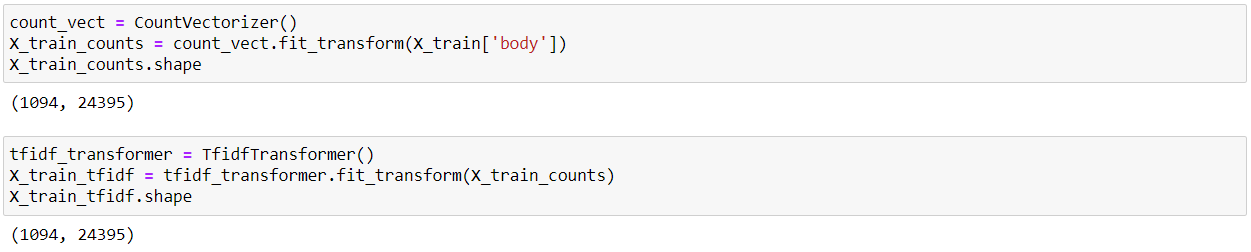


The next step is to process the data for the machine learning implementation. The dataset is split into training and testing dataset giving 30% to the test set and 70% to the training set.



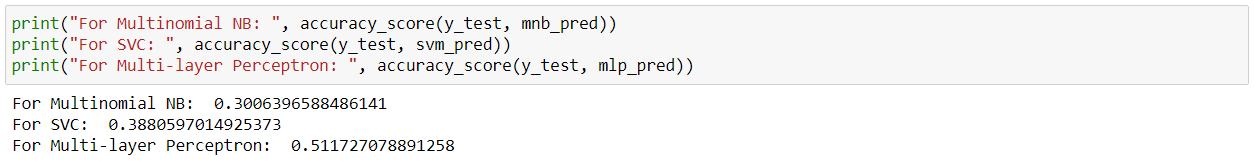
# Natural Language Processing

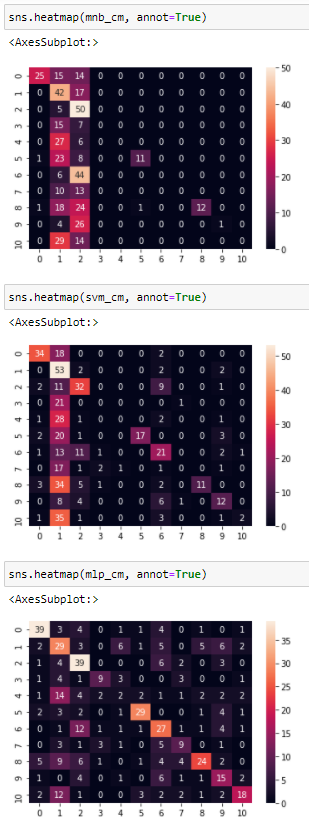
Using Natural Language Processing tools like the Count Vectorizer and the TF-IDF Transformer, features from the text data in the dataset are extracted and processed for further use by the algorithms applied.



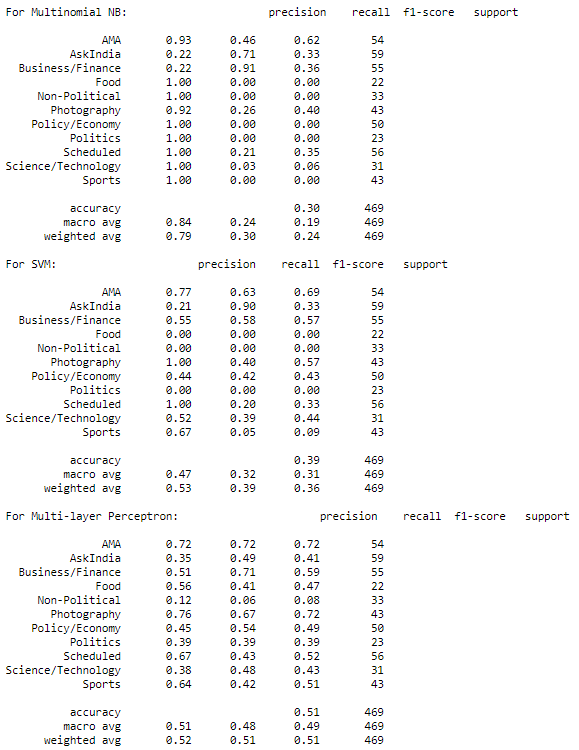
# Machine Learning

The objective of this code is to develop a machine learning model that can identify if a reddit post is of the science domain or not. To that extent, three algorithms will be applied and their performances compared to identify the best performing algorithm. The three algorithms chosen for this research are the Multinomial Naïve Bayes, Support Vector Machine and the Multilayer Perceptron. The results for these algorithms are as follows:





Confusion matrices for algorithm performances from top to bottom, Multinomial NB, SVM, MLP.



Results from 5-fold cross validation for each algorithm:

