Eluvio DS Challenge

Option 1 - Data Science/ML

As I analyzed and understand, the given dataset contains the information about news title and its related data, which includes time, date, up_votes, down_votes, author, over_18 and category.

Problem Statement

Here for this particular dataset, as I am free to choose problem statement according to my perspective, I would like to go for predictive modeling because if we talk about analytical insight, there are only few instances which have numerical data so it is difficult to get much more important information. Still, I am doing data visualization for some of the instances in this task.

As this task in based on NLP, ‘‘I am going to build a model which will predict rating for news title’’.

Implementation

- Import Required Libraries
- Load Dataset
- Drop unused columns
- Classifying up_votes in 5 different rating
- Extracting some of the data according to date in different variable (My PC is not able to process all the records but I am making a model which can process any number of records)
- Take same amount of data for each rating (For best data distribution and increase model’s efficiency)
- Create new column ‘Text length’, which contains length of title because it can be also important instance to predict rating for title.
- Data Visualization for rating and text length.
- Text preprocessing for ‘title’
- Creating NLP model
- Fitting data in model
- Evaluating the results
- Deployment of model

Note: we are not taking instances like ‘category’, ‘down_votes’ and ‘over_18’ because ‘category’ and ‘down_votes’ have constant value for all the records and ‘over_18’ has boolean value True and False. In that only 320 records out of 500,000+ records hold True value, which is not so good data distribution.

Future Work

We can put data in neural network to get better accuracy. We can also use text length as a predictor for more efficient model.