Problem:

Providing analytical insights for a business use case.

Finding Most Popular Topics: In a headline a topic is not just any word. It’s the subject of a sentence. To find the subject of the sentence we look at the proper nouns in our title using nltk pos\_tagger to understand the context of the problem.

Finding Most Popular Topics based on sentiment: Looking at world news top topics by sentiment we can find out what world news obsesses over. We use the Vader sentiment analyzer of nltk and we process the word stems.

Analysed Article volume over its lifespan by analysing Frequency per month and Frequency per month with every overlapping year.

Analysed the influence of time of the day on upvotes. As you can see the best rated posts have been posted between 10:00 -15:00 hours.

Classification problem: After analysing the data we found out that there are 8 columns. The features that can affect the label could be"over\_18" as age will sometimes or to some extent determine their interest. And "author" could be one factor since some famous authors could be liked more. The column called "down\_votes" are found all "0" and the column named "category" are all shown as "worldnews", so those two features are useless, and just drop them. The effect of time\_created and date\_created is not that huge on the label. In my opinion, whether the news is attractive or not, and based on "up\_votes" values, the news can be easily split into two parts (attractive or non-attractive). If upvotes is high means news is attractive. Based on upvotes the label is whether a news is attractive or not on the feature title. We just use the [“title”] feature to solve the problem.

Used nltk package to pre-process the "title" and then use tfidfVectorizer to vector "title". Threshold set is 0.8 quantile. Split data into two parts, one part is train data, with 0.8 size of whole data, and the other is test data, with 0.2 size of whole data.

By applying multinomialNB model, I got the test accuracy to be 0.805. By applying logisticRegression, the test accuracy is 0.806.