ELUVIO INTERNSHIPS 2021
DS CHALLENGE - REPORT

PREPARED BY: THAMALI M ADHIKARI

1. Introduction

Coding Challenge Option 1: Data Science/ML

Given a large-scale dataset provide analytical insights or predictive modeling for business use cases.

2. Goals and Questions to be Answered from Data

The goal of this project is to perform both statistical analysis and building machine learning models.

2.1 Statistical Analysis

I have performed the statistical analysis in 4 major groups.

i. Upvote Related Analysis
ii. News Related Analysis
iii. Author Related Analysis
iv. Country Related Analysis

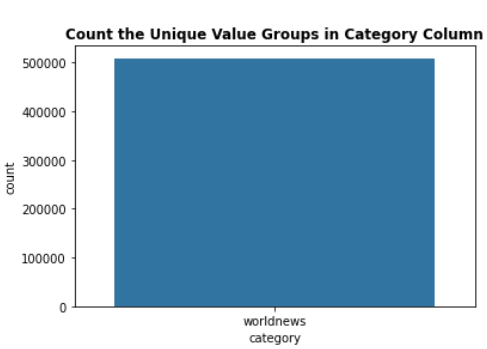
2.2 Machine Learning Models

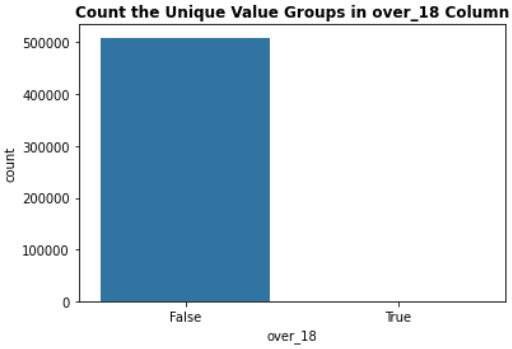
i. Build a machine learning model to cluster the news titles based on their content.
ii. Build a machine learning model to predict the upvote based on the content of the title.

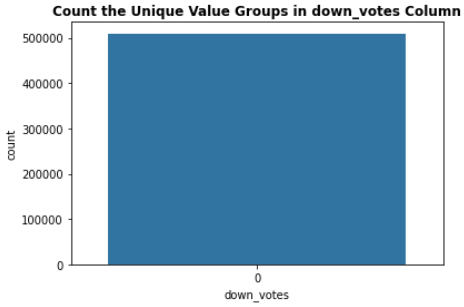
3. Methodology and Results

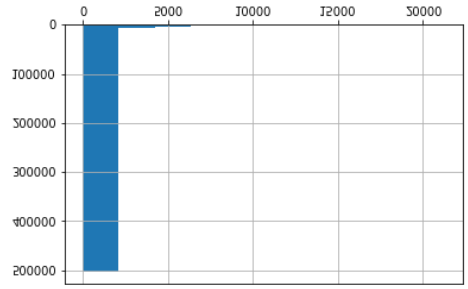
i. Basic Data Exploration and Cleaning Stage 1

Our data has 509236 rows with 8 columns with no missing values. Among the columns of the dataset, category, over_18 and down_votes columns have one unique value in each. Therefore, those columns do not provide valuable information. If we consider the up_votes column the data distribution is not symmetric. In the up_votes column there are 5782 unique vote values whereas the author column has 85838 unique authors. Also, the number of words in the title column range between 2-59 with 311 unique length values. Based on the data visualization graphs, category, over_18 and down_votes were dropped as they do not provide important information. So, we are left with only 5 columns and they are time_created, date_created, up_votes,title and author.









Upvotes Distribution

Fig 1: Feature Exploration

ii. Advance Data Exploration and Cleaning Stage 2

Now in this step we will dive into the data to explore the hidden information. Based on the time_created and date_created column we can extract the year, month, day and the post created time for each post. Also, by exploring the data in the title column we can extract the country names for each post as well. After the above exploration I have found that some posts didn't mention about any country and there 6 unique values in year column. As for the upvotes data visualization, there are some outliers. So, by considering the values greater than 20000 as outliers, the columns with upvotes> 20000 were dropped.

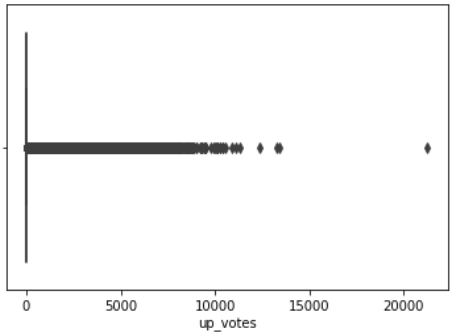


Fig 2: Scatter plot of Upvotes column

iii. Statistical Analysis

Let us consider a business scenario. Mr. X owns a publishing company, and he would like to know about some statistics. When he considers the statistics, he can expect information from 4 major criteria as mentioned above under section 2.1

By combining all the fields, we can prepare a set of statistical where Mr. X is interested in finding answers.

a. Average upvotes over the hour of the day
b. Number of news published over the hours of the day
c. Average upvotes over the week

d. Number of news published over the week
e. Average upvotes over the month
f. Number of news published over the month
g. Average upvotes over the year
h. Number of news published over the year
i. Rank the authors who published the highest number of posts
j. Rank the authors who got the highest number of upvotes
k. Rank the authors who published the highest amount of news in each year
l. Rank the authors who published the highest amount of news over the day
m. Rank the countries which were mentioned the most
n. country related news which received highest upvotes
o. Rank the country related news publish by year
p. Rank the country related news publish by month
q. Author articles based on country

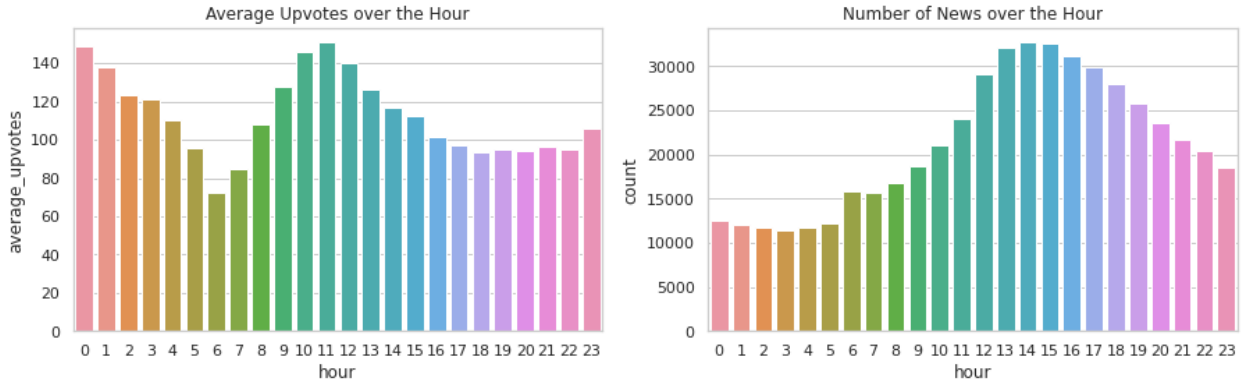


Fig 3: Average Upvotes and Amount of News over the hour (questions a and b)

If we look at the fig 3 above, we can see that around most of the news get published around 1pm but the most of the upvotes receive around 11am and 12am. Though there are many posts get published around 1pm, some posts might not reach to the users due to the number of posts. It seems around 11am and 12am there may be many users online. So, submitting news at a time when there are more users online but with less competition may have a higher chance of reaching the posts to more users and receiving more upvotes.

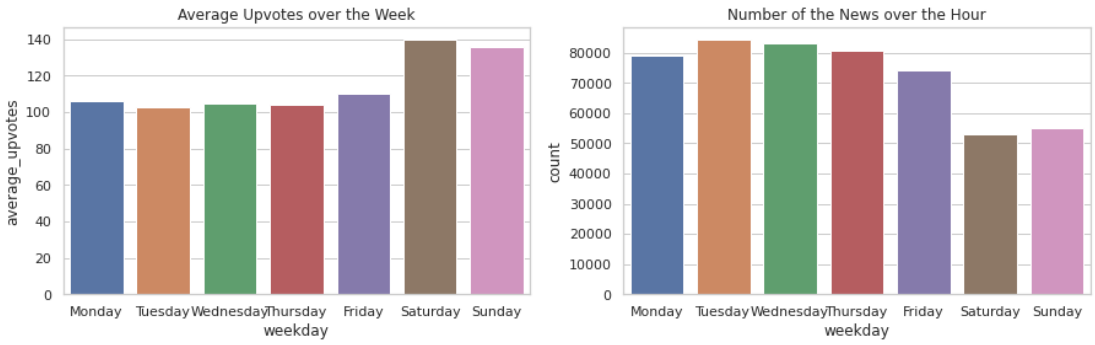


Fig 4: Average Upvotes and Amount of News over the day (questions c and d)

According to the figure 4, the highest number of votes receive during Saturday where the highest number of posts get published on Tuesdays. It seems may be users get online during weekends more than the weekdays. Though the number of posts get publish during the weekend is lower than the number of posts get publish during the weekdays, still the count of upvotes is higher than the weekdays. Therefore, it might be a good idea to publish a post during weekend at a less competitive hour to get more upvotes for the news posts.

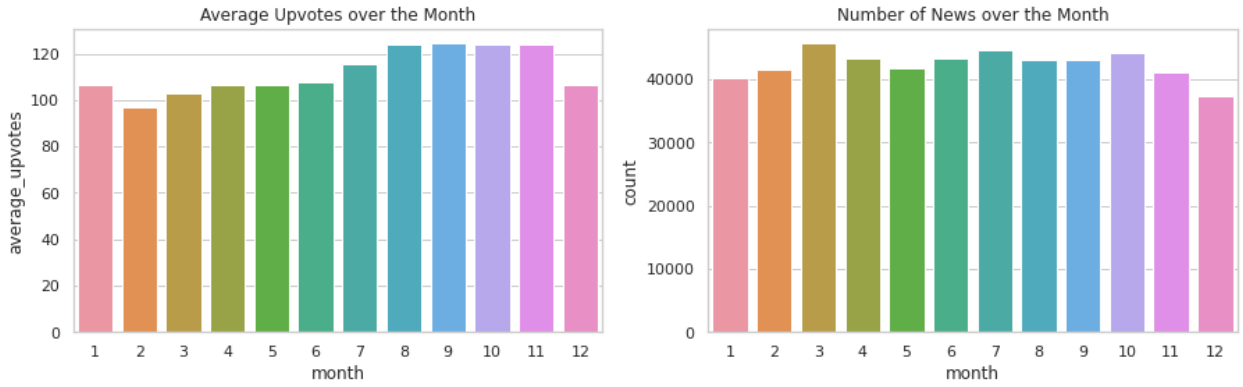


Fig 5: Average Upvotes and Amount of News over the Month (questions e and f)

According to the above figure the highest number of upvotes received during the September where the March has more news publications. By looking at the graph above we cannot make any clear assumption. Probably we should analyze the number of posts and count of upvotes in each month for few years to find a clear pattern from the data.

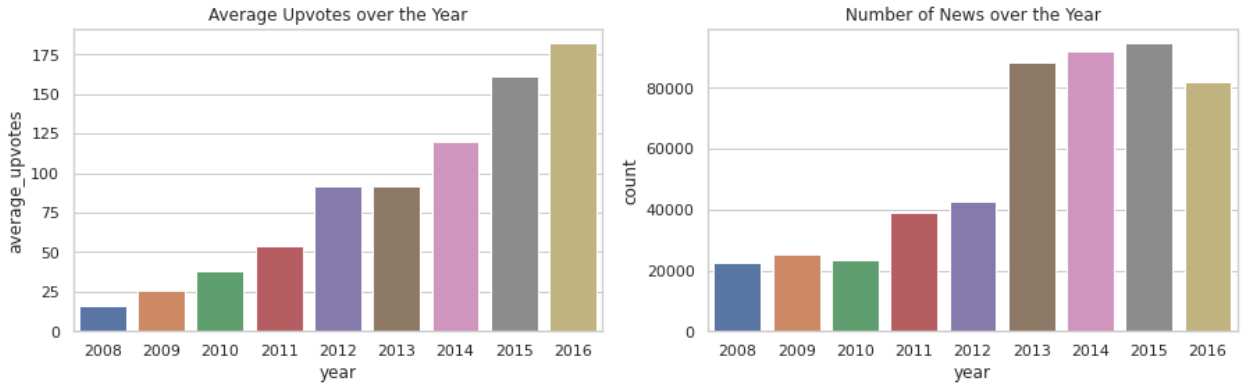


Fig 6: Average Upvotes and Amount of News over the Year (questions g and h)

If we look at the figure 6, we can see the number of upvotes increases over the years though the number of news posts get publish over the years has not the highest at that year. This may happen due to reasons such as: people have started to read news posts more over the years, the quality of the news posts content is high, and a special topic got attention during the year 2016.

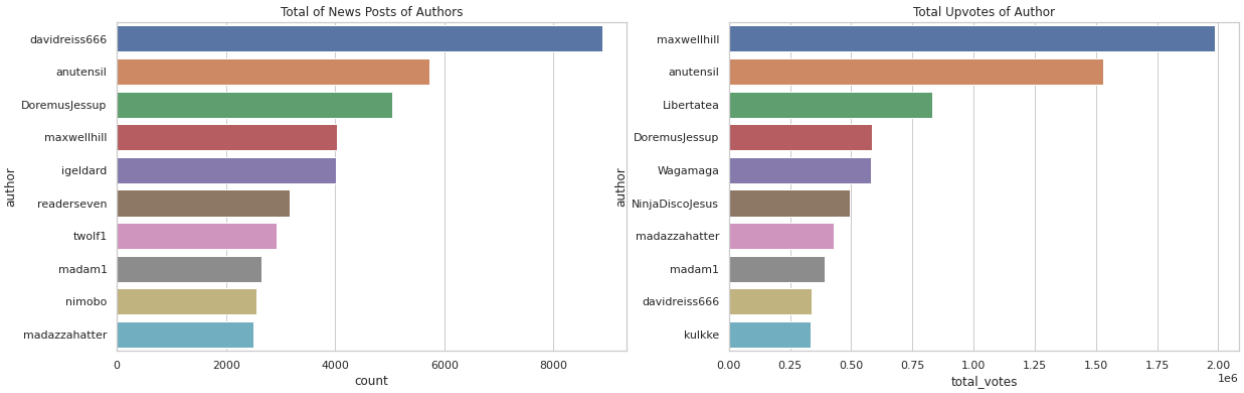


Fig 7: Rank First 10 Authors with Highest Amount of Upvotes and News Publish (questions i and j)

The figure 7 shows though an author publish higher number of news posts he might not get the highest number of votes for his/her work. This means based on the quality of the work or discussing about an attentive or popular matter during the time makes a post reach more to the users and might have higher chances of getting higher upvotes. So, Mr. X might consider hiring authors who are more responsible and maintain the quality of the work though they take bit more time to complete the task and train the employees to balance and work fast. On the other hand, there might was a famous situation during a special time, so the number of news posts got increased.

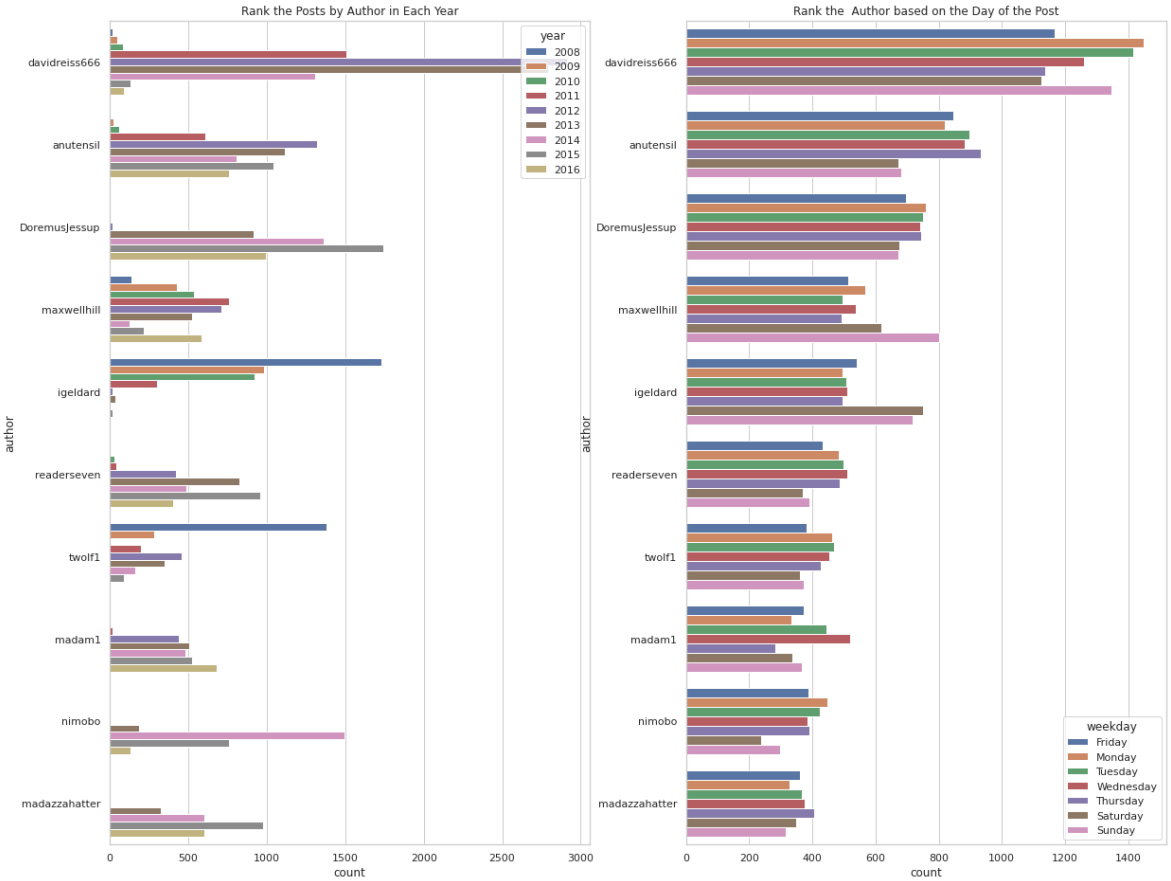


Fig 8: Rank First 10 Authors based on News Publish over Years and Day (questions k and l)

According to the figure 8, the author who created the highest number of posts overall has not actively submitted posts in later years. It seems during the year 2012 he was actively involved in writing posts but later it reduced fast. Also, the author who received highest number of upvotes seems to have a good set of skills though he has less publications but was able to collect the highest number of votes. Also, it seems there is no specific pattern among authors with the day, they publish their post. However, the author has highest votes, publish his higher number of posts during weekends. Therefore, his posts might have a higher chance of reaching to more users and less competitive due to the availability of number of posts and number of users during the weekend.

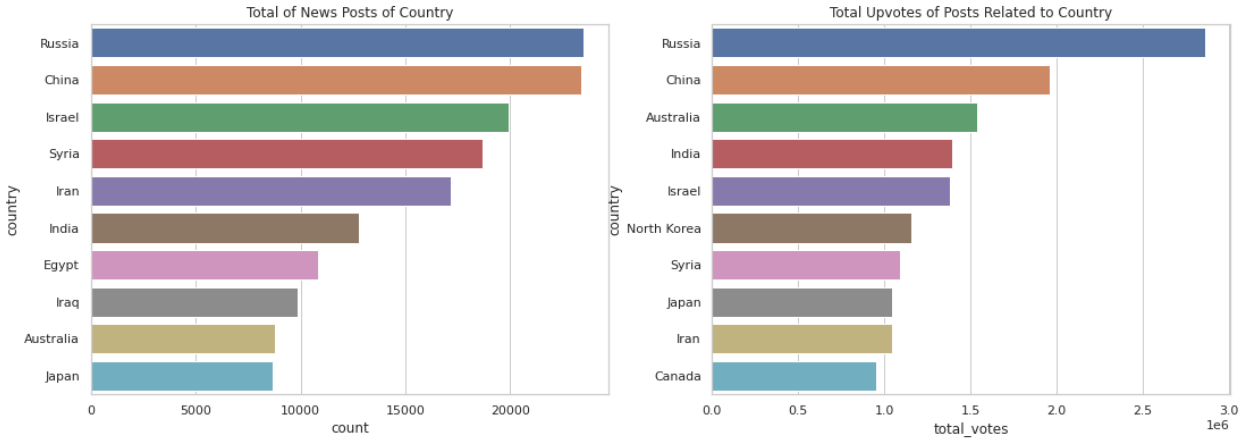


Fig 9: Rank First 10 Countries based on Number of News Publish and Number of Upvotes (questions m and n)

If we look at the figure 9, we can clearly see that the ranking order of the total number of upvotes is same as the total number of published news over the time.

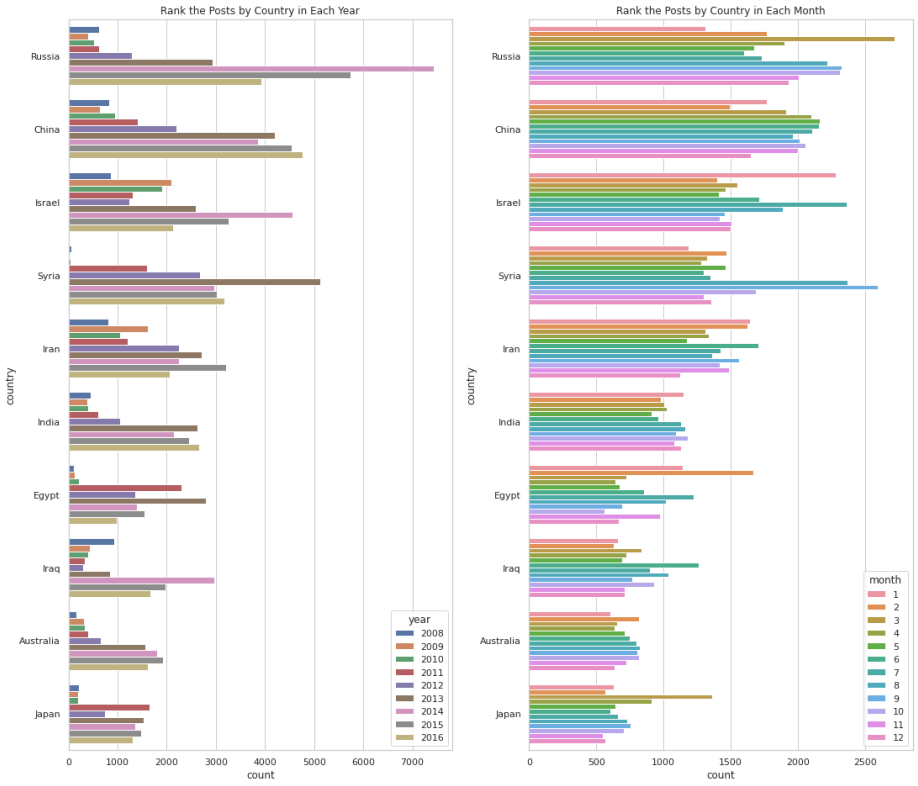


Fig 10: Rank First 10 Countries based on Number of News Publish by Year and Month (questions o and p)

According to the figure 10, some countries were highly mentioned in posts in some months of a particular year. This might occur because of an incident occurred in those country which took the attention of the world.

	country	author	Time
23625	China	bob21doh	716
119020	Russia	vigorous	685
24067	China	davidreiss666	647
77982	Japan	madazzahatter	562
137863	Syria	uptodatepronto	447

Fig 11: Rank Authors by the Country of their News Articles (question q)

Figure 11 shows us the China is famous among the authors than other countries. It ranked two times among top 3 ranks above. It seems the distribution of the posts mentioned about Russia is better than China because all the posts didn't come from one or two authors. So, it is better to create news posts about the countries which were not mentioned a lot and also, ask more authors to work on country related posts by dividing the post creating task.

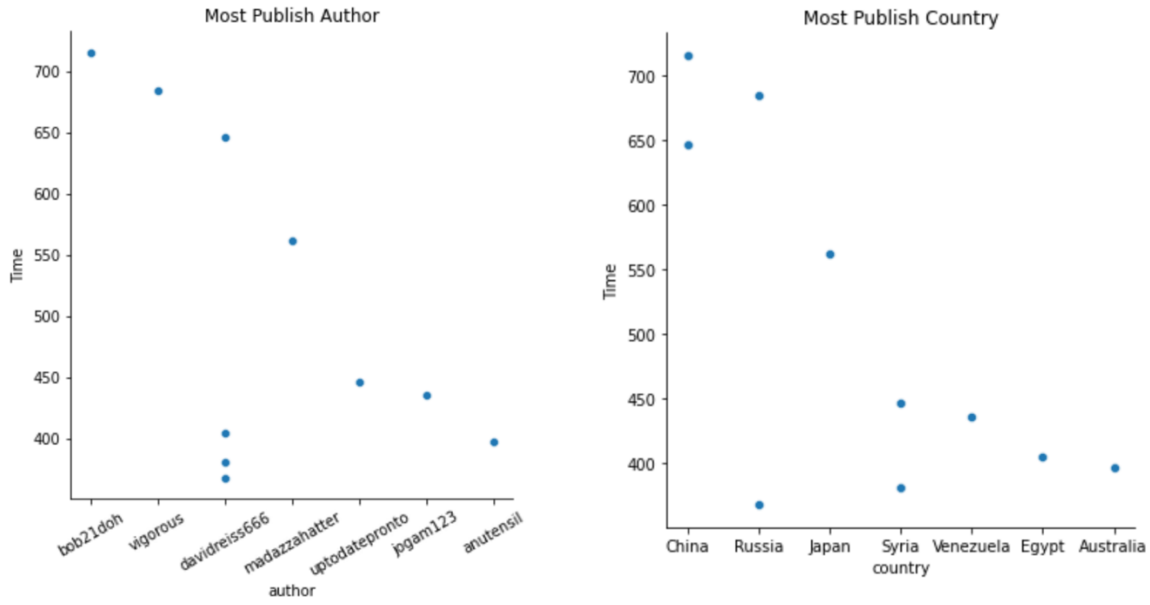


Fig 12: Comparison between most publish author and country

If we consider the author rank from the figure 11 and perform an analysis on like on figure 12, we can see the first author has only created posts related to China. So, all of his 716 posts are about China. So, this author might be an area reporter.

iv. Machine Learning Model Building

a. Title Clustering Based on the Content

In this section first of all we need to clean our text data because the more you clean the more accurate model you can get from it. Text cleaning occur in few steps such as:

- Remove punctuation
- Correct format
- Correct short forms of words
- If there are any countries mentioned by their abbreviations correct them to the original country name
- If there are symbols like $ and other currency we need to remove them
- Remove digits
- If there are information about citizens of a country, we can extract the name of the country from it. (Ex: convert Italians to Italy)
- Remove stopping words
- Remove pronouns
- Remove emoji symbols
- Convert all the letters to lowercase
- Lemmetize the words
- Remove links(https)
- Sigularize the plural words
- Convert verbs in any form to their present tense format

In this project I have executed all the above-mentioned cleaning steps. After the cleaning there might be some titles without no content. This might occur because of some noisy data points. Some titles might have only symbols or numbers. As an example, in our dataset we had, titles like 69, 127 and sequence of symbols and I dropped those rows after cleaning the data.

After the data cleaning next stage is to build a frequency term dictionary to find the words which has higher frequency.

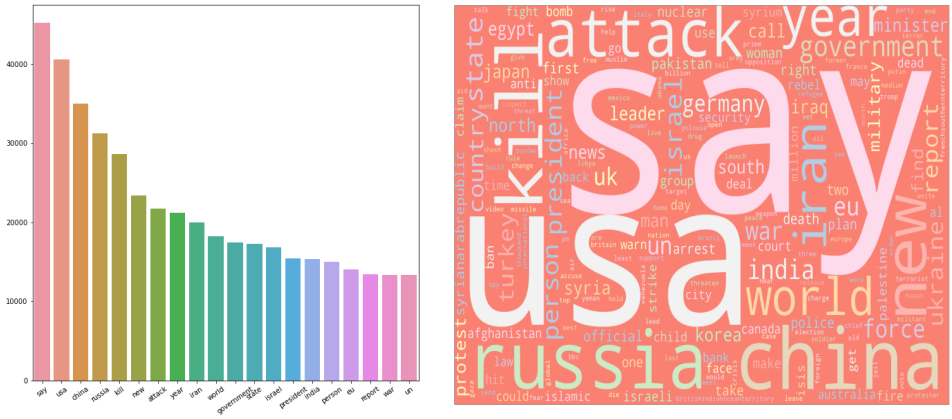


Fig 13: Word Frequency

After this not it is time to clean the word frequency dictionary to make a model which is more accurate. Therefore, I removed the words which occur less than 5 times in the entire documents and remove words which occur more than 50% of all documents (filter words very rare and very common). Then convert the frequency dictionary to bag of words. Then create a tfidf model object for further training. Now we need to cluster the documents. Before that we need to decide how many cluster groups give us the best results based on our data. We can look at the coherence scores according to the different number of topics and pick the number of topics which gives us the highest coherence value. Ideal number of topics will maximize coherence and minimize the topic overlap.

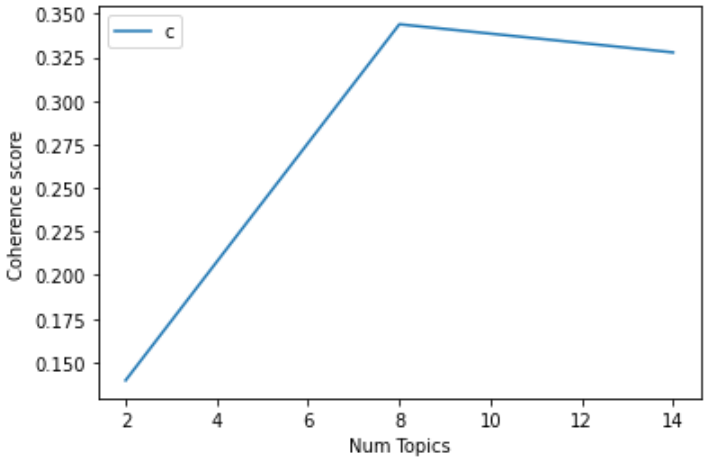


Fig 14: Coherence Values vs Number of Topics

Based on the figure 14, we can see the number of best cluster groups is 8 and we will use this to build our model. Here, I have used the LDA model available in Gensim. Running LDA using bag of words and tfidf vector gave two different results. The Bag of Words and Tfidf vector gave coherence score with C-v model of 0.352 and 0.294 respectively.

b. Predict the Upvotes based on the Content

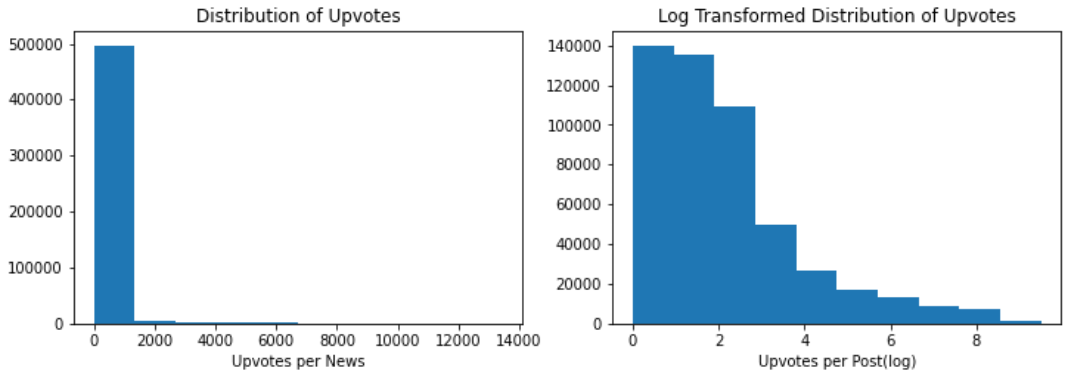


Fig 15: Upvotes Distribution

According to the figure 15, the upvote distribution is not balanced. Therefore, I transformed the upvotes column into its logarithmic format.

The mean absolute error from linear, ridge and lasso models are 1.53, 1.46 and 1.37. So, among all the models ridge regression gave the lowest MAE score. The values are in logarithmic format. So, we can see the MAE scores ranges between 23.44 – 33.88.

4. Discussion and Future Works

This project was an interesting challenge. However, we need to make these models stronger by doing more experiments. As an example, we know that the best number of topics is 8. But we still don't know the topics we get are the best in our model. Therefore, we need to run the process by changing the parameters like iterations and pass values. Then we should visualize the results of the given clusters using library called PyLDAvis. So, we can see if there are any clusters which are too closer or overlap with each other. As another approach we can apply KMeans clustering to perform our clustering task. So, as for the future work we should build models to predict the headline category using both deep learning and KMeans. Once we have all these models, we can compare the performance of each and pick the best model which gives us the highest accuracy. When we measure and compare the accuracy between models, we need to think in few aspects like coherence score, which data points group together in all the methods, how well the clusters separate from each other and if the selected word for the topics helps to identify cluster groups.
And for the upvotes prediction task we need to train and test machine learning models on different algorithms like random forest and we should use cross validation to get a better performance. Once we have strong models, we can perform more depth analysis such as: the type of the news comes from each country, what are most popular news cluster group among the authors and authors duty on publishing news on each category. Another important thing is we can apply more text filtering techniques in this project to make more stronger results. In our data some posts mention the name of the state with their abbreviation, and some has the complete name. In my analysis I did not correct the state wise, but I corrected the country wise. So, we can apply the same transformation to states as well. Also, we can try different trials with the words we keep on the dictionary to create the vector.

Given the time duration it was challenging but there is more to explore in the future. I have used python Colab and python libraries in these tasks.