Summary On News popularity Prediction

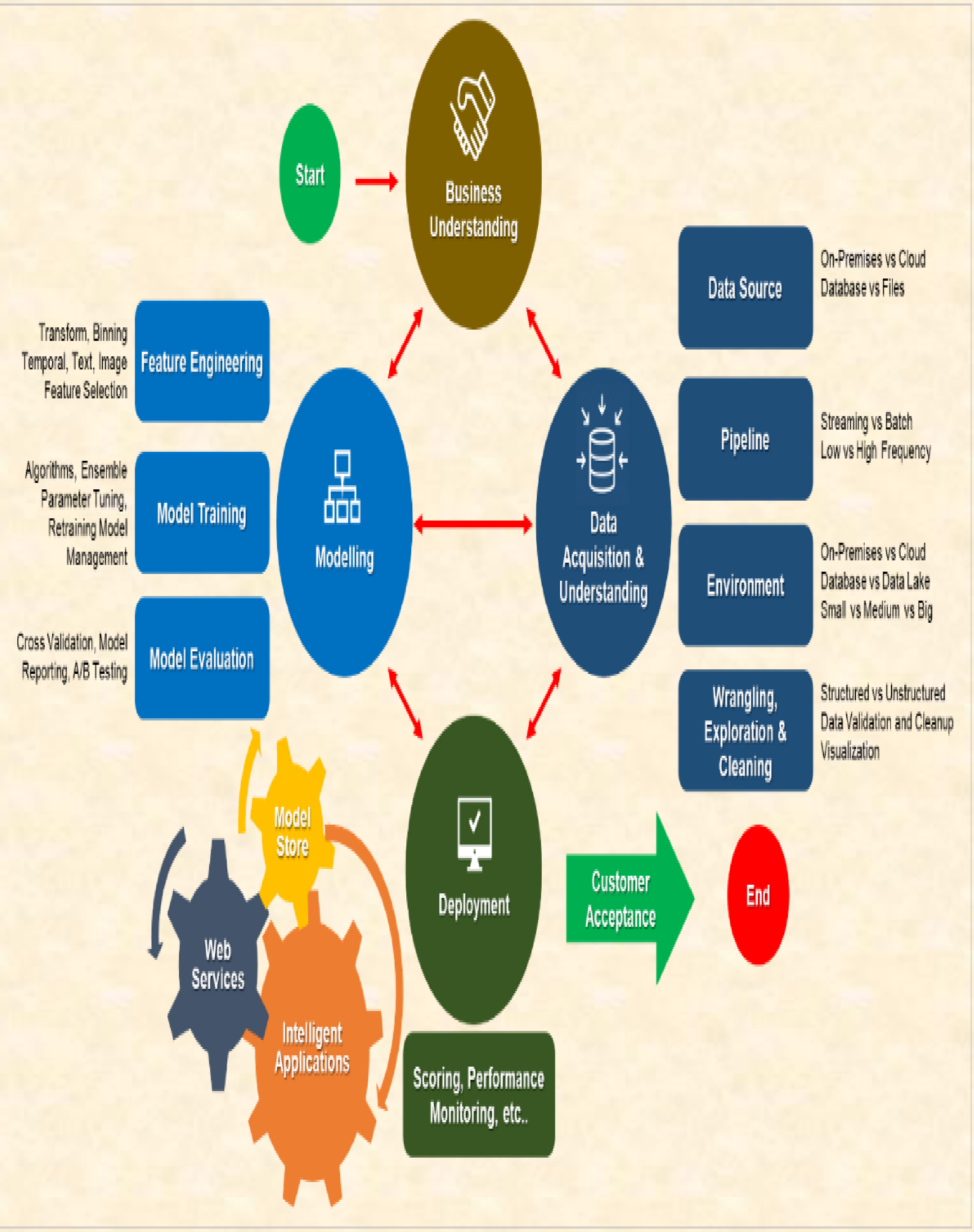
AlmaBetter Capstone project

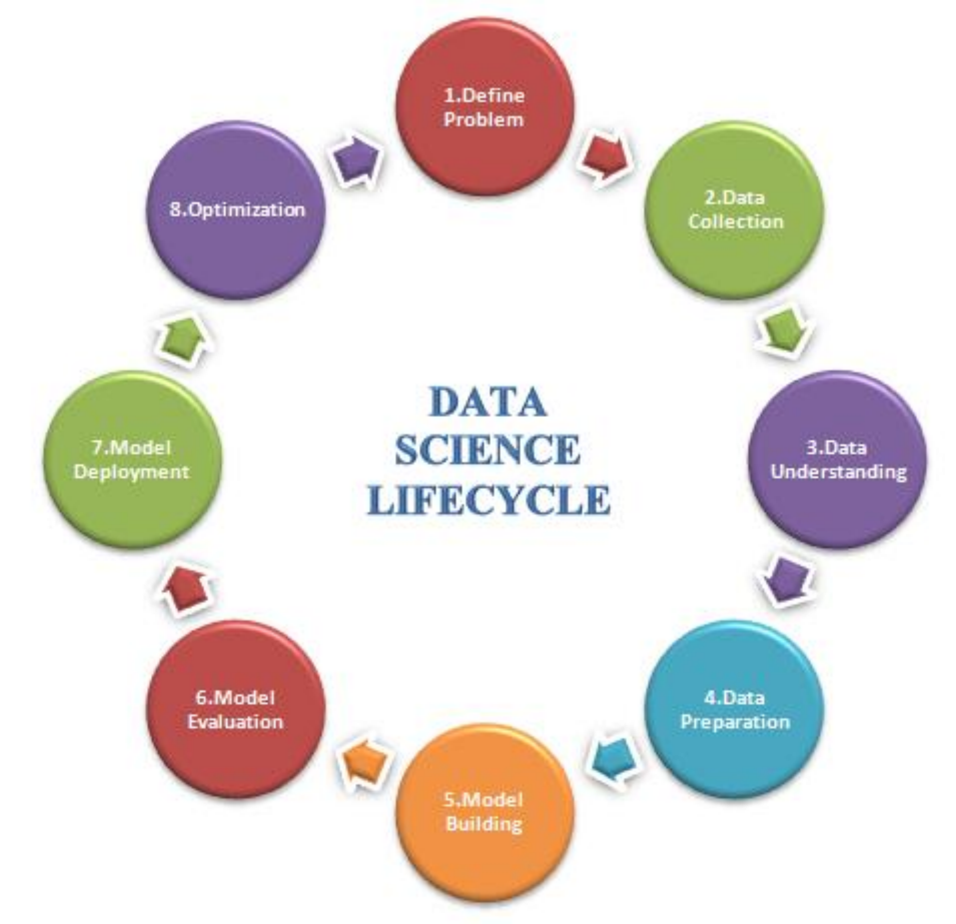
By,

Yamini Peddireddi

1: Introduction

1. Problem solving in Data Science





From the above diagrams it is clear about how to proceed to solve a problem in data science. Data science is a revolution to the fields and gives a lot of insights and transformative approach in problem solving for many problems. As the data is getting increased day by day in every field, organization and industry.

2. Methodology/Approach

2.1. Problem statement

- There is a large data set of news items and their respective social feedback on multiple platforms: Facebook, Google+ and LinkedIn. The collected data relates to a period of 8 months, between November 2015 and July 2016, accounting for about

100,000 news items on four different topics: Economy, Microsoft, Obama and Palestine.

- There are many variables which are impacting on the dependent variables. The problem here is find the number of likes or reactions generated on different topics discussed in different social media platforms using all other features available in given data with the news obtained from different sources.

2.2. Steps involved in problem solving

Here the problem technique is to design a system through which news popularity prediction is done for the given data or for next few weeks or days or seasons as per the requirement. So here requires the optimal solution for news popularity prediction on social media for given topics of news.

The steps taken in achieving or designing or solving this regression problem are:

- Understanding the data
- Missing value analysis
- Data Cleaning
- EDA
- Data Visualization
- Outlier detection from Statistical method or from boxplots
- Creating new features from News title and Headline column
- Feature engineering
- Dimensionality reduction
- Feature selection
- Train and test split
- Model building
- Hyperparameter tuning
- Model Validation
- Model Selection

2.3. Challenges

The challenges in solving the data science problem:

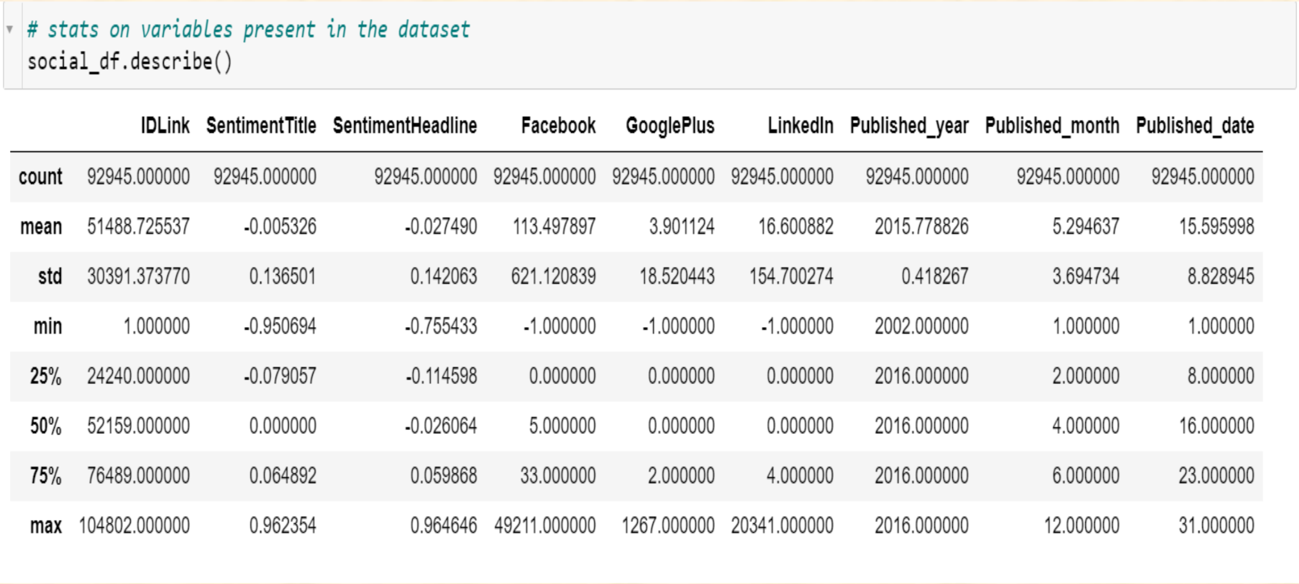
- Data collection
- Huge dataset
- Data limitation or more features like promotional activities… etc can also be used
- Manipulated data/ Missing values
- Security
- Computational cost and time
- Finding the target feature and understanding the problem
- Time taking
- Memory crash problems

3. Data Exploration

3.1. Data

The dataset contains 11 features and the dataset contains 93239 rows of data and this data consists of data from Nov'15-July'16 as per the given information through which test data or any other data to be trained and prepare a model which makes good predictions.

- There are a total of 93239 rows and 11 columns in the given data.
- There are 279 missing values in Source and 15 missing values in Headline.
- I have removed those rows as they are very less in percentage compared to the whole and avoids false information and also saves the memory which helps in avoiding memory problem.
- There are 5 object variables(categorical) and 6 numerical variables in the given data. This has been further processed and datetime variable is converted to datetime variable type, created new columns with publish date column.



```
# stats on variables present in the dataset
social_df.describe()
```

	IDLink	SentimentTitle	SentimentHeadline	Facebook	GooglePlus	LinkedIn	Published_year	Published_month	Published_date
count	92945.000000	92945.000000	92945.000000	92945.000000	92945.000000	92945.000000	92945.000000	92945.000000	92945.000000
mean	51488.725537	-0.005326	-0.027490	113.497897	3.901124	16.600882	2015.778826	5.294637	15.595998
std	30391.373770	0.136501	0.142063	621.120839	18.520443	154.700274	0.418267	3.694734	8.828945
min	1.000000	-0.950694	-0.755433	-1.000000	-1.000000	-1.000000	2002.000000	1.000000	1.000000
25%	24240.000000	-0.079057	-0.114598	0.000000	0.000000	0.000000	2016.000000	2.000000	8.000000
50%	52159.000000	0.000000	-0.026064	5.000000	0.000000	0.000000	2016.000000	4.000000	16.000000
75%	76489.000000	0.064892	0.059868	33.000000	2.000000	4.000000	2016.000000	6.000000	23.000000
max	104802.000000	0.962354	0.964646	49211.000000	1267.000000	20341.000000	2016.000000	12.000000	31.000000

3.2. Features in the dataset

Number of attributes:

- **IDLink (numeric)**: Unique identifier of news items
- **Title (string):** Title of the news item according to the official media sources
- **Headline (string)**: Headline of the news item according to the official media sources
- **Source (string):** Original news outlet that published the news item
- **Topic (string):** Query topic used to obtain the items in the official media sources
- **PublishDate (timestamp):** Date and time of the news items' publication
- **SentimentTitle (numeric):** Sentiment score of the text in the news items' title
- **SentimentHeadline (numeric):** Sentiment score of the text in the news items' headline
- **Facebook (numeric):** Final value of the news items' popularity according to the social media source Facebook
- **GooglePlus (numeric):** Final value of the news items' popularity according to the social media source Google+
- **LinkedIn (numeric):** Final value of the news items' popularity according to the social media source LinkedIn

VARIABLES OF SOCIAL FEEDBACK DATA

- IDLink (numeric): Unique identifier of news items
- TS1 (numeric): Level of popularity in time slice 1 (0-20 minutes upon publication)
- TS2 (numeric): Level of popularity in time slice 2 (20-40 minutes upon publication)
- TS... (numeric): Level of popularity in time slice ...
- TS144 (numeric): Final level of popularity after 2 days upon publication

3.3. Data understanding and cleaning

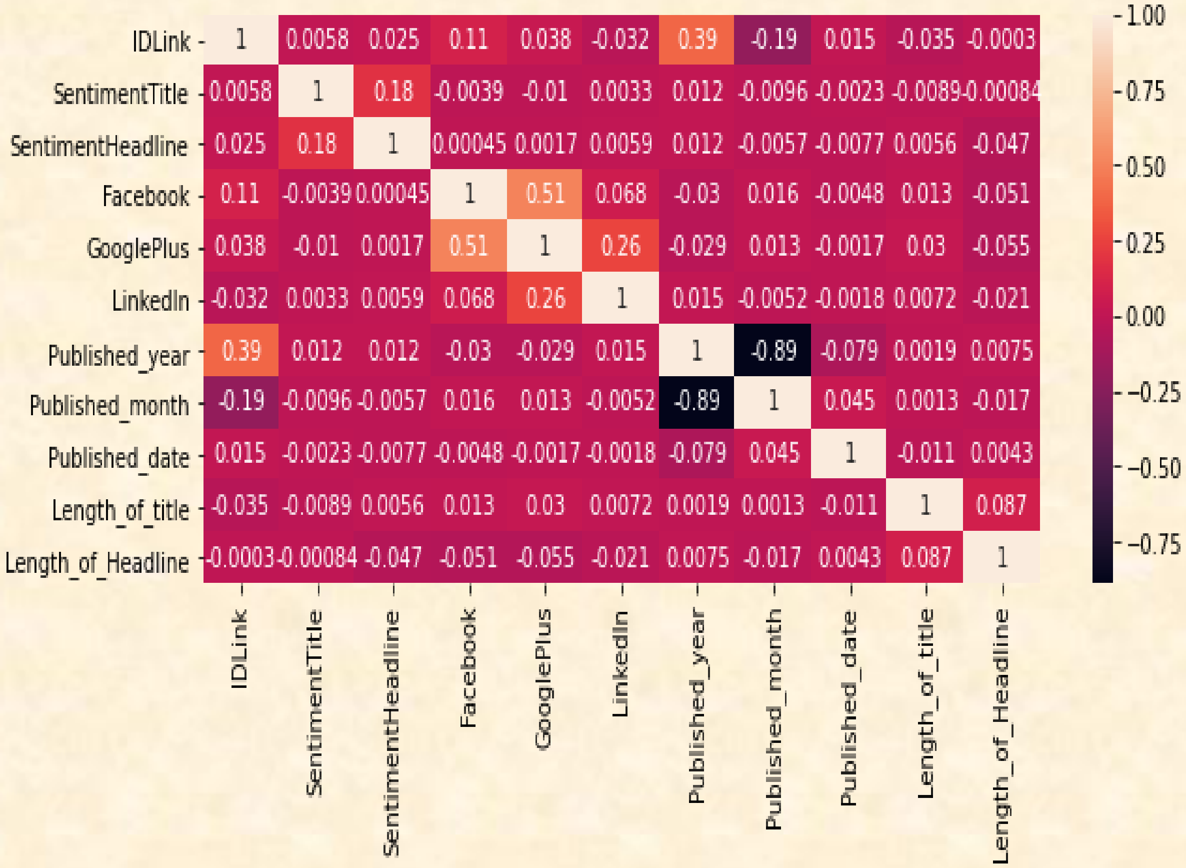
There are 279 missing values in Source and 15 missing values in Headline. I have removed those rows as they are very less in percentage compared to the whole data as imputation of those brings false/wrong information into the data.

3.4. Data and new columns

After exploring the data, we find that there are 10 features in the dataset which impacts the three dependent variable Fb, G+ and LinkedIn. The target can be any number and is of numerical type so it is a regression problem.

There is a Publish Date column which can be further divided to many columns as it has lot of details in one column. So, I split that data into the columns Year, month, date. Further I dropped the column Publish Date as I have taken the required data from that feature and that is no longer useful now. If you allow that data for further solving it show multicollinearity in the data which is not desired for solving the problem and increases the error rate. So that is removed.

Some correlation between the variables can be seen from the below diagram,



	IDLink	SentimentTitle	SentimentHeadline	Facebook	GooglePlus	LinkedIn	Published_year	Published_month	Published_date	Length_of_title	Length_of_Headline
IDLink	1	0.0058	0.025	0.11	0.038	-0.032	0.39	-0.19	0.015	-0.035	-0.0003
SentimentTitle	0.0058	1	0.18	-0.0039	-0.01	0.0033	0.012	-0.0096	-0.0023	-0.0089	-0.00084
SentimentHeadline	0.025	0.18	1	0.00045	0.0017	0.0059	0.012	-0.0057	-0.0077	0.0056	-0.047
Facebook	0.11	-0.0039	0.00045	1	0.51	0.068	-0.03	0.016	-0.0048	0.013	-0.051
GooglePlus	0.038	-0.01	0.0017	0.51	1	0.26	-0.029	0.013	-0.0017	0.03	-0.055
LinkedIn	-0.032	0.0033	0.0059	0.068	0.26	1	0.015	-0.0052	-0.0018	0.0072	-0.021
Published_year	0.39	0.012	0.012	-0.03	-0.029	0.015	1	-0.89	-0.079	0.0019	0.0075
Published_month	-0.19	-0.0096	-0.0057	0.016	0.013	-0.0052	-0.89	1	0.045	0.0013	-0.017
Published_date	0.015	-0.0023	-0.0077	-0.0048	-0.0017	-0.0018	-0.079	0.045	1	-0.011	0.0043
Length_of_title	-0.035	-0.0089	0.0056	0.013	0.03	0.0072	0.0019	0.0013	-0.011	1	0.087
Length_of_Headline	-0.0003	-0.00084	-0.047	-0.051	-0.055	-0.021	0.0075	-0.017	0.0043	0.087	1

- From the above correlation plot, it can be said that the target variables all the social media channels are being correlated to Length of Title, Published month, Sentiment Headline and there are some negatively correlated variables.
- There is some correlation with between Sentiment Title and Headline of 0.18.
- There are not any variables which are highly correlated with each other. So there is not much multicollinearity in the data.
- So, all the variables can be taken for further analysis.

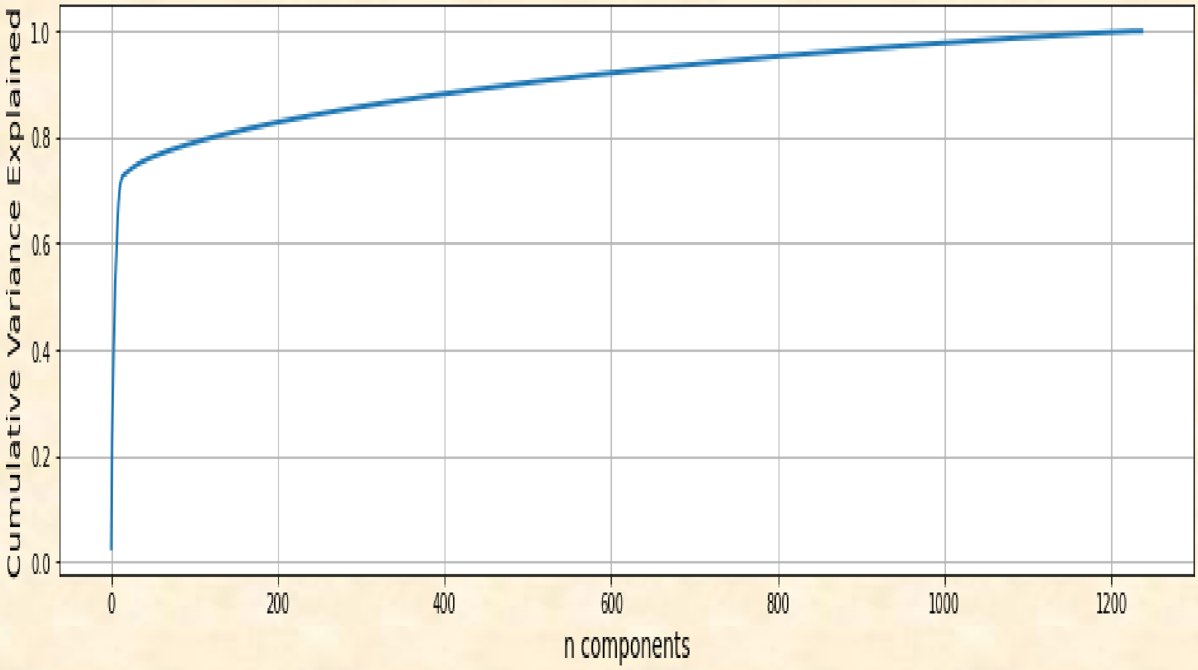
3.5. Feature engineering

- Here I have created dummies for categorical variables so that the model is able to distinguish them from other numerical variables present in the data
- I have used Hash encoding for Source column which gives some of important columns as specified in the parameters of encoding. This is useful because when other type of encoding is used it created n number of new features as there are n(>5000) different sources from where the new articles are obtained.
- I have standardized the numerical data using Standard scaler which brings all the data in it with mean 0 and standard 1. So that all the data exists in one scale otherwise there is a chance of misinterpretation while model building.

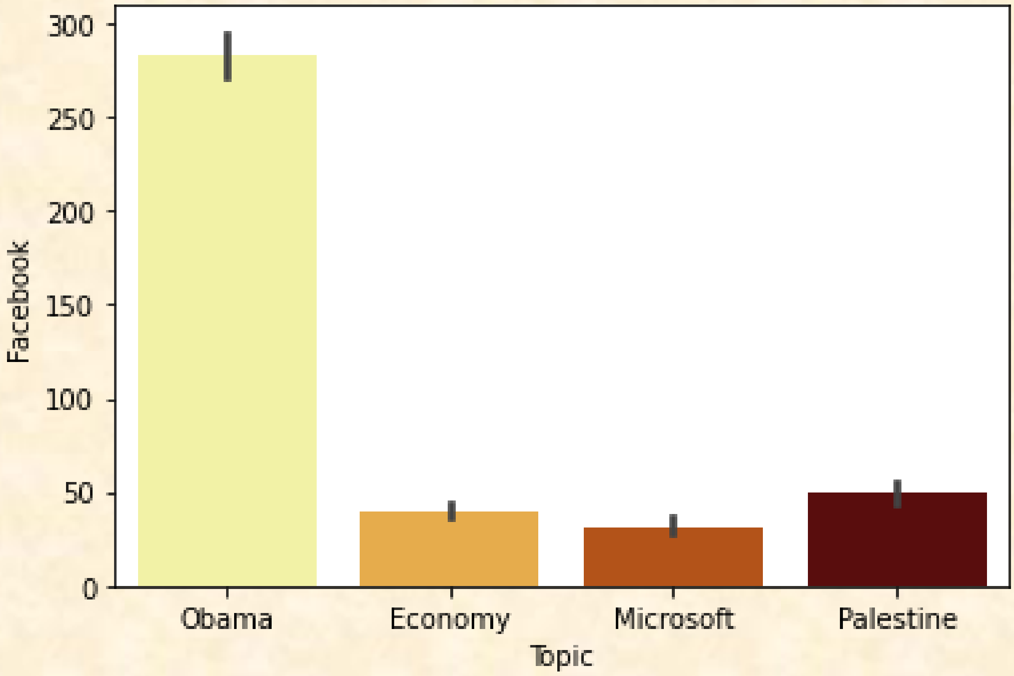
3.6. Feature selection

In Feature selection the important attributes from the dataset are extracted removing the redundant columns.

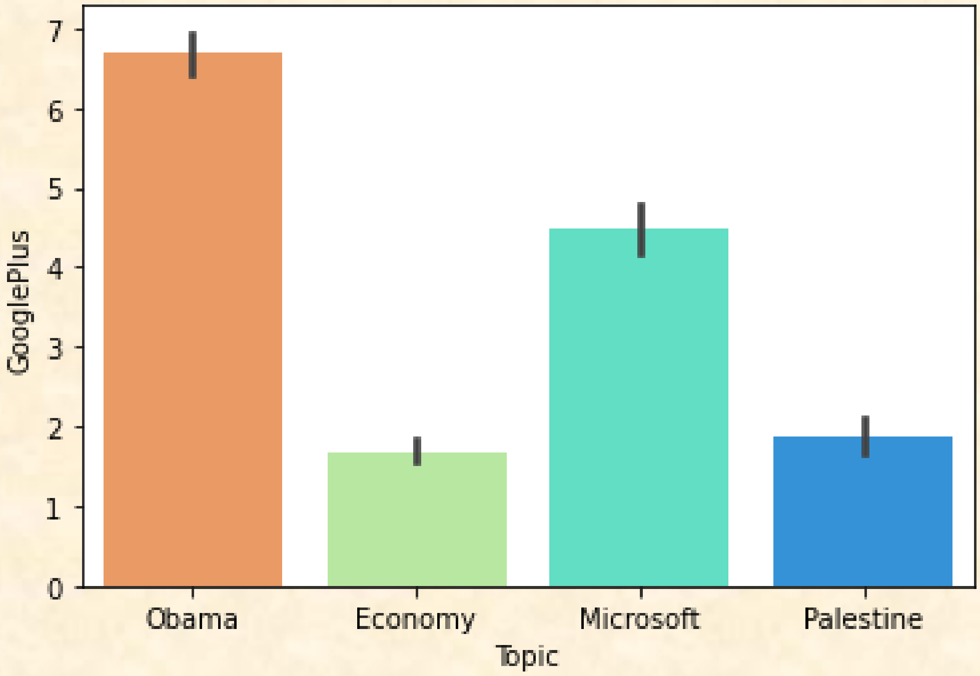
I used dimensionality reduction technique SVD to make the task easier for extracting the important features from the data without dropping the useful columns from the data.



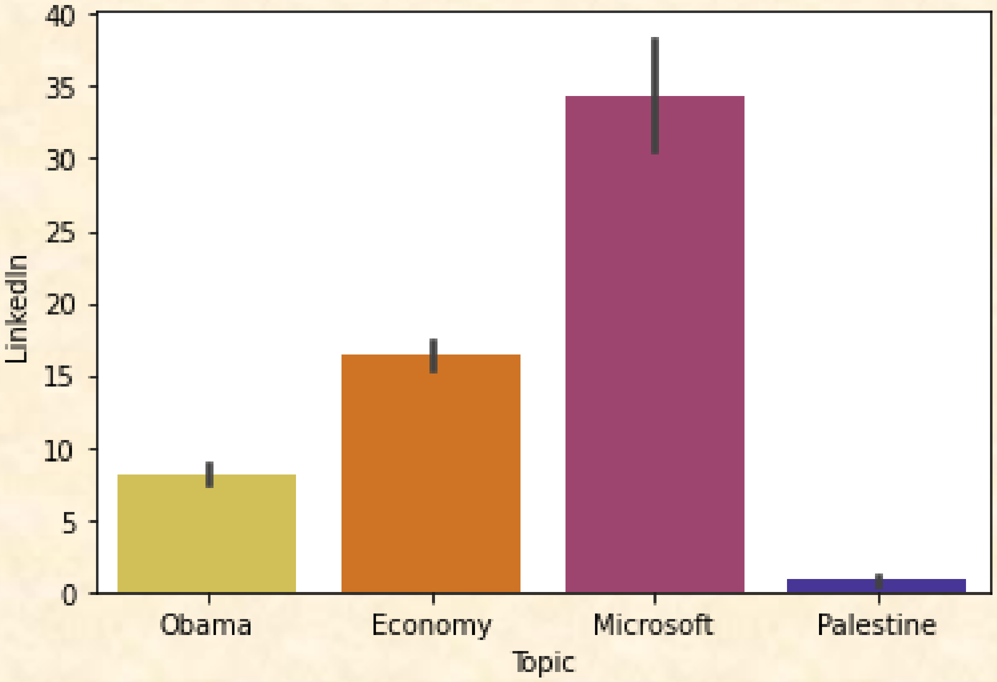
Chapter 4: Data Analysis and Data visualization



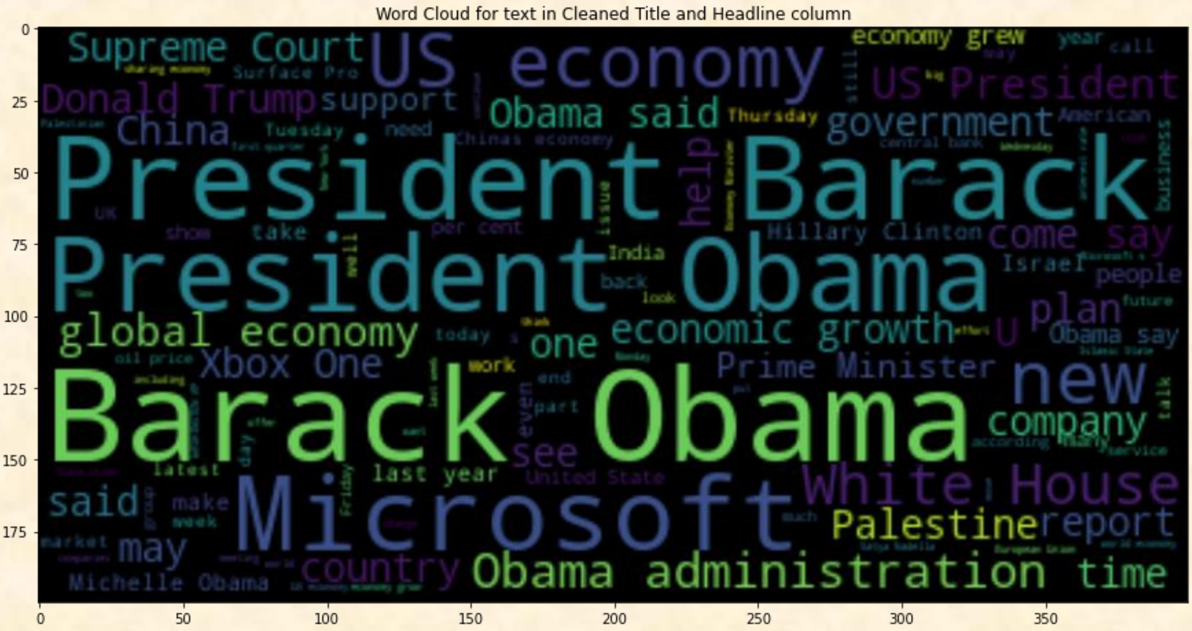
- From the above plot, the distribution of different topics on Facebook can be seen.
- It can be observed from the plot that there are more articles on Obama topic news articles than any of the other three topics.
- Obama topic is dominating on Facebook compared to other topics given in the data.



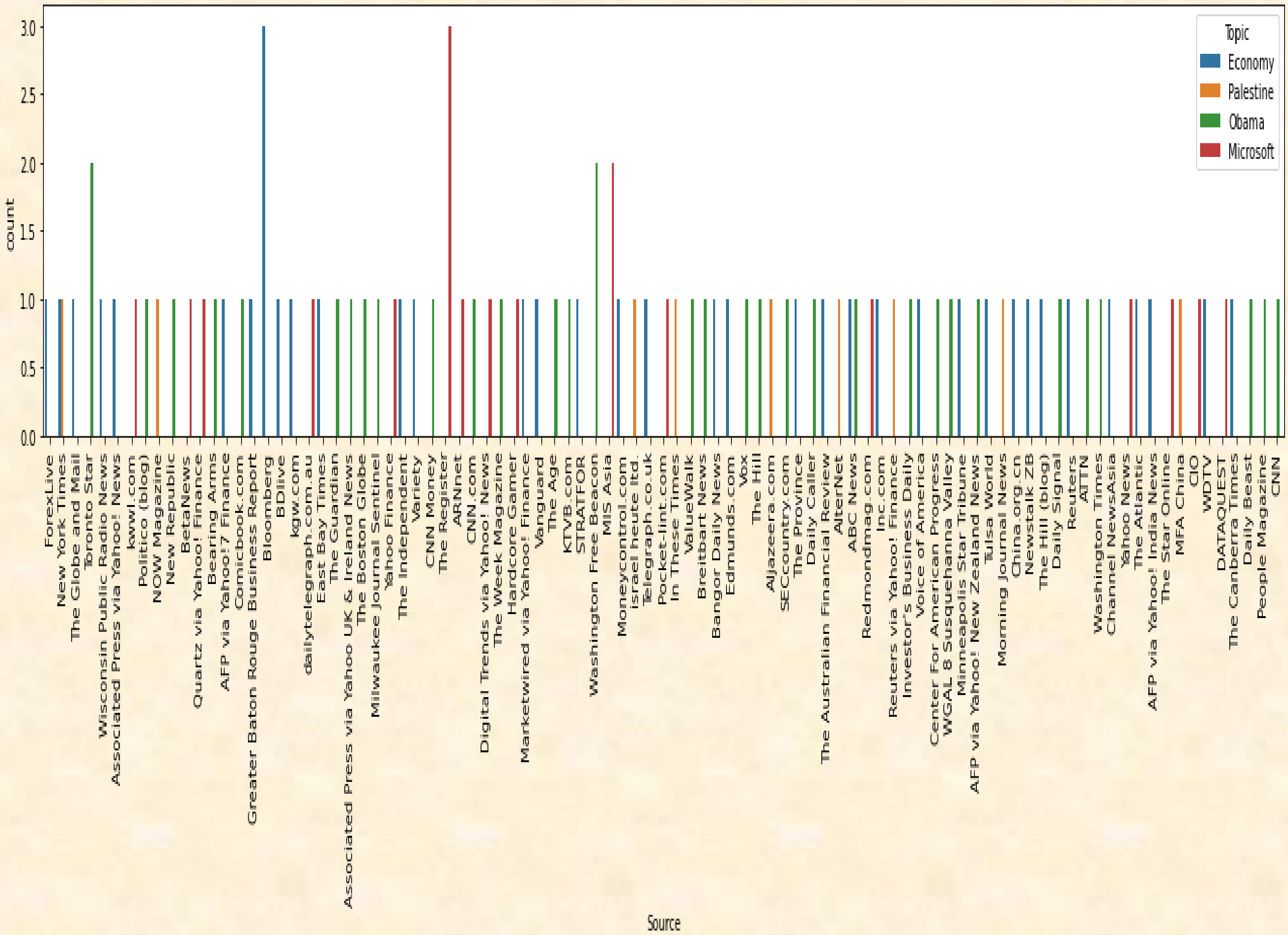
- The above plot talks about the distribution of four different topics of news articles on GooglePlus
- GooglePlus has more count of discussions happening on four different topics than Facebook where it has very few articles on other topics than Obama.
- Here Even on GooglePlus Obama news articles has more reach to audience than any other topic then it is not dominating where Microsoft news has also reached the GooglePlus users and which may have a good response on it with almost 3/4th of Obama's news articles reach.



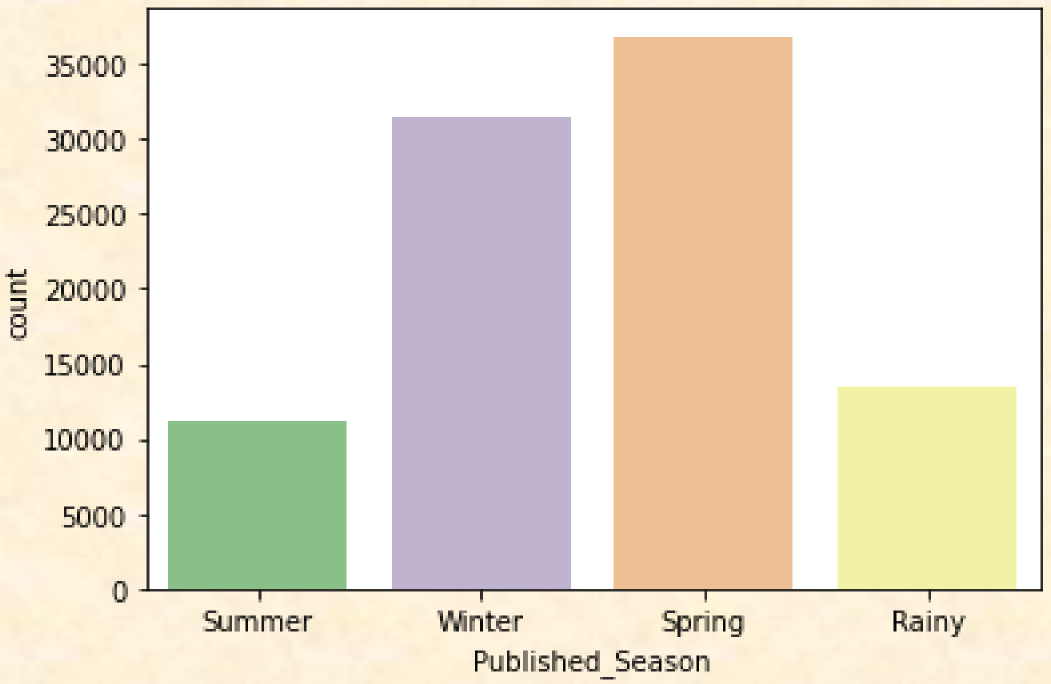
- Here is the plot of the topic distribution on LinkedIn where there are many news articles on Microsoft and also good number of articles on Economy than on Obama and very few on Palestine.
- As LinkedIn being the professional website, the discussions are more on Microsoft, and Economy than on Political topics or any other topics. There is a possibility.



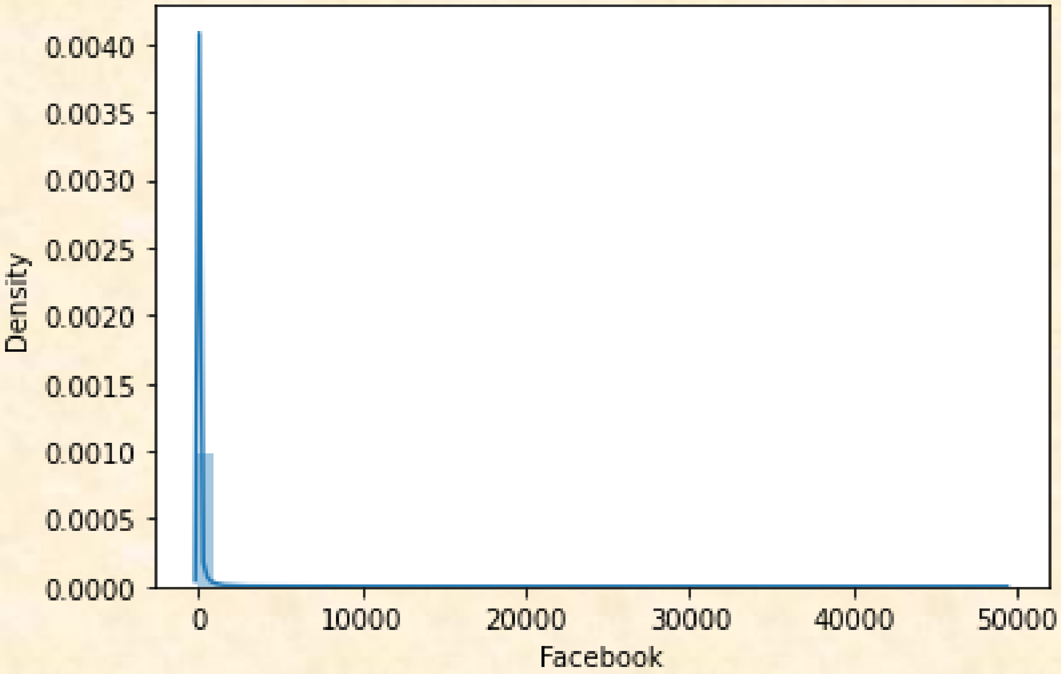
- From the above plot we can see the different names present in the Cleaned Title and Headline column.
- The highlighted names tell the frequency of the words present in the data.
- The words which occurred mostly in Cleaned Title and Headline are Obama, Microsoft, economy, Palestine, US economy, President Obama, new, Report, China, global economy, President Barack, Barack Obama, White House, US President, Obama administration, new, US economy, global economy, Thursday, may, first, Supreme Court and Support etc.
- There are good number of important words to be noted. Yes, they say a lot on the what the topic under discussion is. Oh Yes!!!



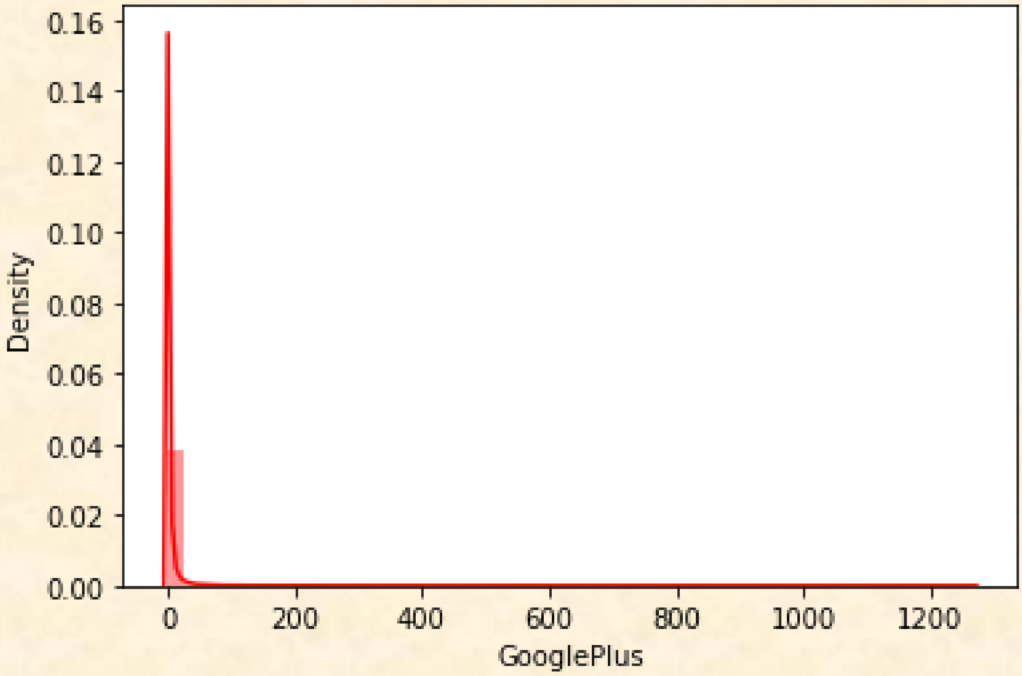
- The above plot is the sample distribution of Sources from where the news articles are extracted from and there are many unique sources taking all the sources is difficult to plot.
- So is the sample plot of Sources it be seen that there are some news sources where there are a greater number of times the news is obtained or collected from that place.
- Some of the repeated Sources or where there are high number of news articles published on social media are from Bloomberg, The Register, Washington Free Beacon, MIS Asia and Toronto Star.
- The news extracted from these 5 sources are on the topic Economy, Microsoft and Obama.



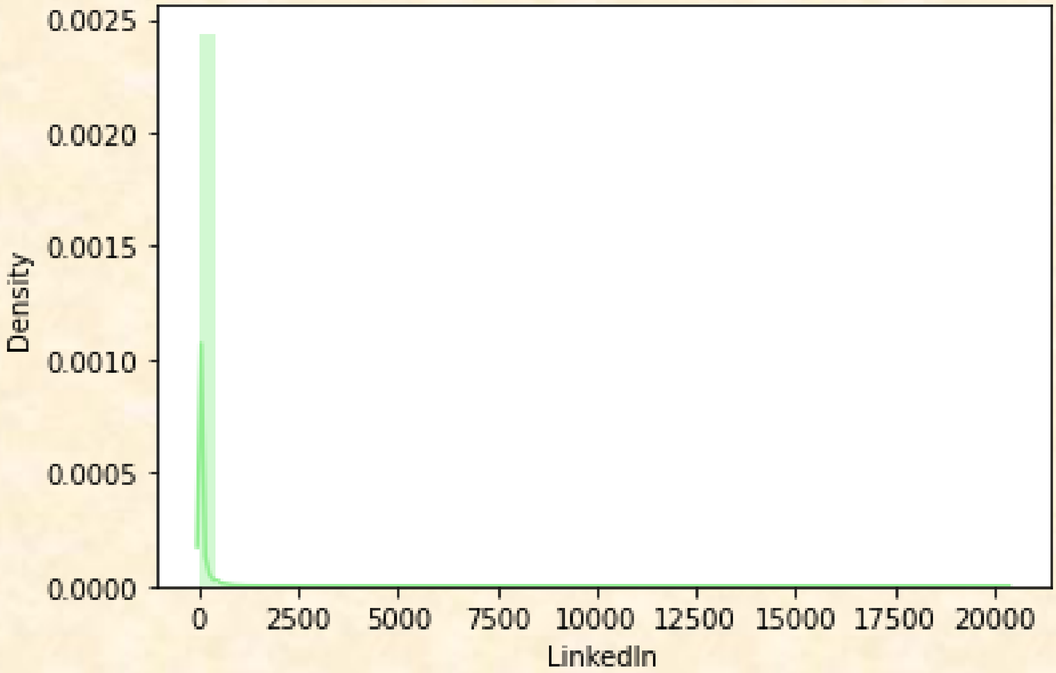
- This is one of the features created from the month column.
- From this it can be seen that there are news articles being published in Spring and Winter Season than in Rainy or Summer season.



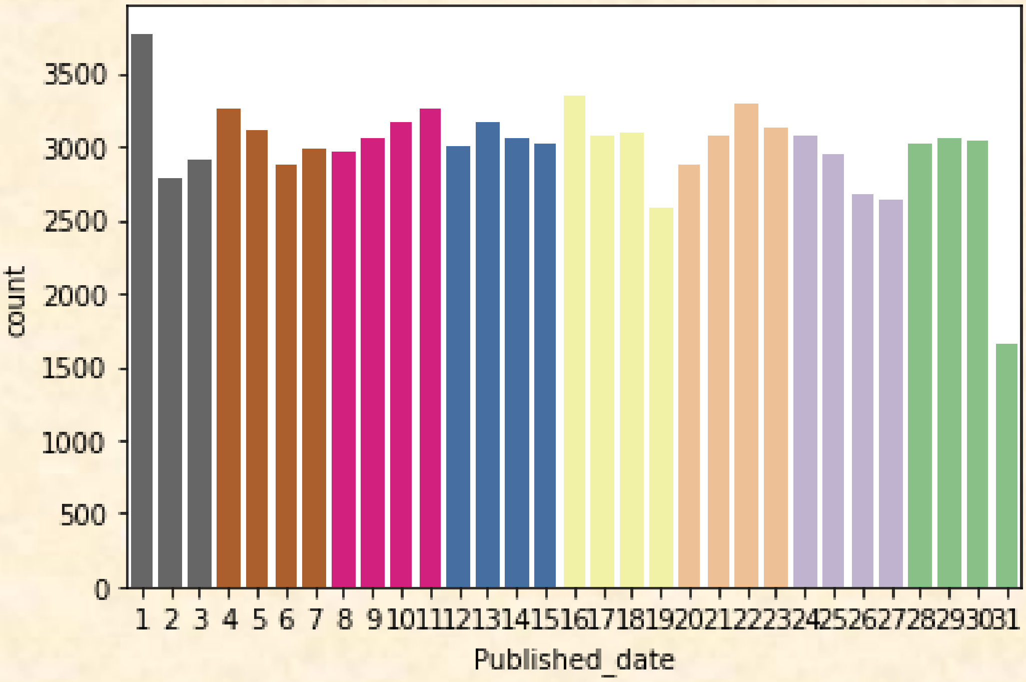
- From the above plot the target variable ‘Facebook’ reactions on the news articles can be seen.
- There are a maximum likes or reactions are app. 50000 and to a least of there are no likes with 0 on some news articles.



- This a plot of likes or reactions generated on GooglePlus on the different topics of news articles.
- The maximum or highest likes received on GooglePlus or app. 1300
- It can also be seen that this is right skewed distribution saying there are outliers with a greater number of likes obtained on few news articles than on other articles.



- Here is another plot giving the information on the reactions generated on LinkedIn.
- This is also a right tail distribution graph like Facebbok and GooglePlus.
- The maximum or highest number of likes generated on LinkedIn for the news articles are app. 21000.



- From the above plot it can be seen that there are more news articles being published on 1st of every month and least on 31st of any month.

5: Model Evaluation and selection

5.1. Model evaluation metrics

Evaluating your developed model helps you refine the model. You keep developing and evaluating your model until you reach an optimum model performance level. (Optimum model performance doesn’t mean 100 percent accuracy; 100 percent accuracy is a myth).

Regression Model Performance Parameters

Let’s talk about the regression model evaluation metrics. We usually check these parameters while developing linear regression models or some other regression models where the dependent variable is continuous (non-binary or categorical) in nature.

1. RMSE:

Root mean square error (RMSE) is the most used evaluation metric in regression problems. It follows an assumption that error is unbiased and follows a normal distribution. It avoids the use of absolute error and uses the square of the difference of actual and predicted, as an absolute value is highly undesirable in mathematical calculations. RMSE is highly affected by outlier values. Hence, make sure you've removed/treated the outliers from your data set before using this metric. If RMSE decreases, model performance will improve.

RMSE metric is given by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Where, N is Total Number of Observations.

2. R Square:

If we talk about MAPE and RMSE, we do not do any benchmark comparison. Hence, we use the R square statistic for that. R square limit is 0 to 1. Value more towards 1, tells us that the developed model is high on accuracy. R square metric is given by:

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

$$\frac{\text{MSE(model)}}{\text{MSE(baseline)}} \qquad \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (\overline{y_i} - \hat{y}_i)^2}$$

MSE (model): Mean Square Error of the predictions against actual.

MSE (baseline): Mean Square Error of mean prediction against actual.

3. Adjusted R Square:

Adjusted R-Square metric is a more advanced version of R-Square. On adding new variables to the model, the R-Square value either increases or remains the same. R-Square does not penalize for adding variables that add no value to the model. But on the other hand, adjusted R-Square increases only if a significant variable is added into the model. Adjusted R- Square metric is given by-

$$\bar{R}^2 = 1-\left(1-R^2\right)\left[\frac{n-1}{n-(k+1)}\right]$$

k: number of variables

n: number of observations

Adjusted R-Square takes the number of variables into account. When we add more variables in the model, the denominator n-(k +1) decreases, so the whole expression increases.

If Adjusted R-Square does not increase, that means the variable added isn’t valuable for the model. So overall we subtract a greater value from 1 and Adjusted R-Square will decrease.

Apart from these four above parameters, we have many other performance parameters, but these are the most commonly used.

- Yes, these are metrics used in this news popularity prediction evaluation.

5.2. Model selection

- From the results it can be seen that CatBoost is performing well on this than any other model with r2 score of 0.1143, 0.05419, 0.0643 for facebook, googleplus and linkedin predictions.
- Even adj r2 scores are high are for this model than any other. Rmse is relatively low for this model making this model best to be chosen for making popularity predictions.

- Even after hyperparameter tuning for these models Catboost is best to choose for making news popularity predictions on the four different topics on three social media platforms Facebook, GooglePlus, LinkedIn.

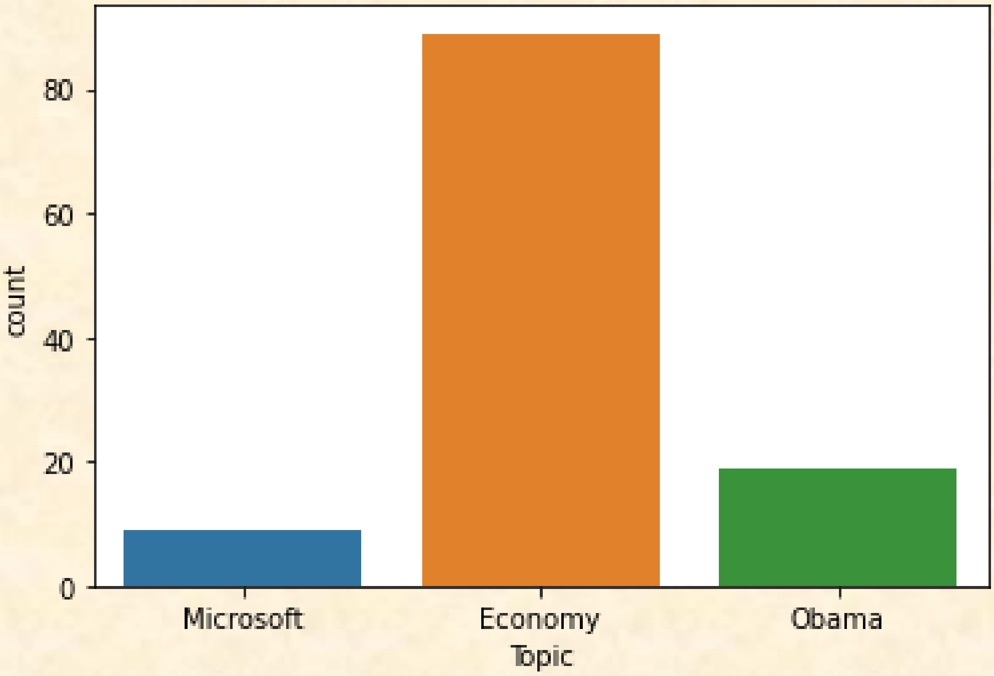
6: Case studies

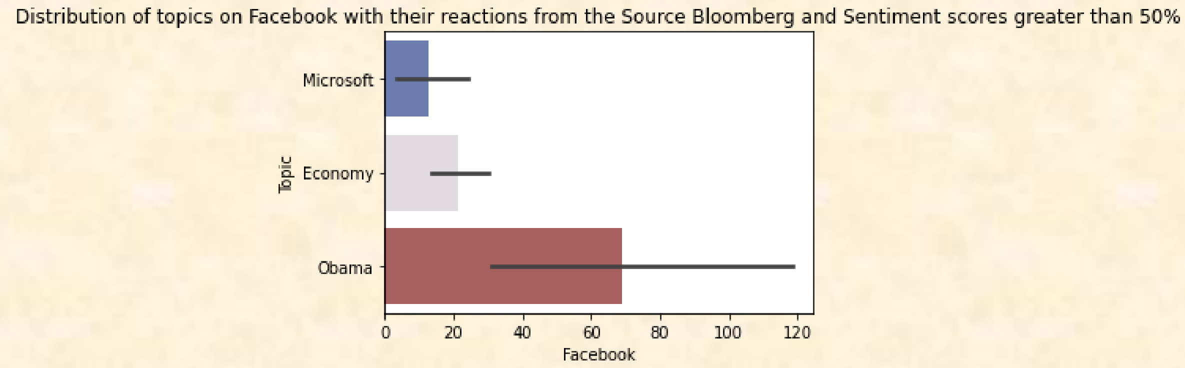
1. Find the reactions on facebook, googleplus and linkedin when Sentiment Title and Headline is >0.70 and from the year 2015

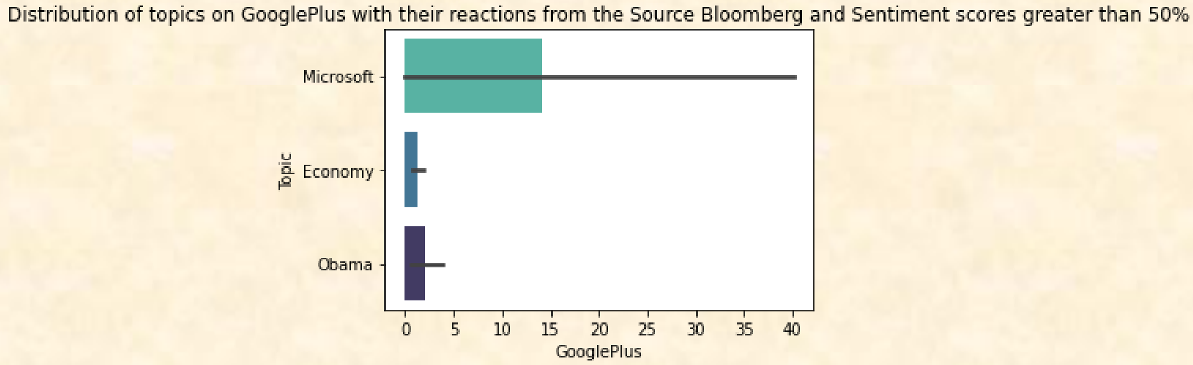
Observation:

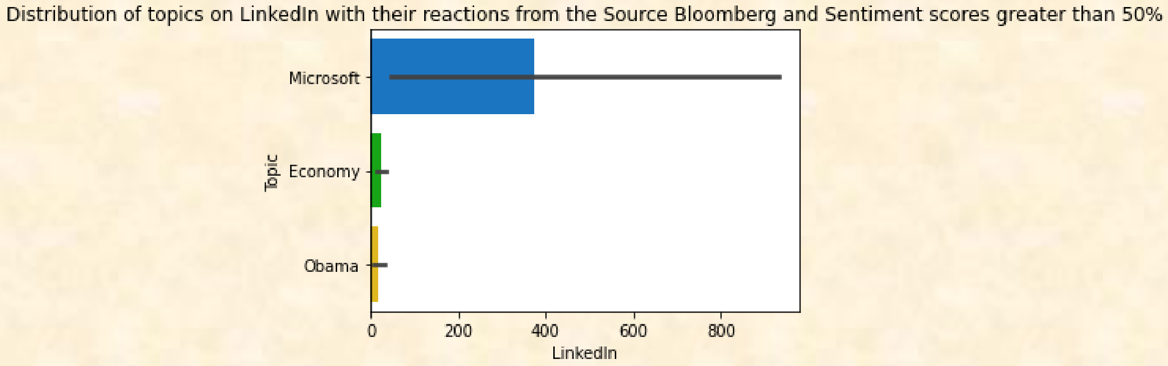
1) Here the sentiment is scores for Title and Headline greater than 70% and from the year 2015 is on 'Economy' Topic sourced from 'The Nation'.
2) There is only 1 news or 1 row of data which is from the last month of the year 2015 and also the last days ie..., from second half of the month Dec.
3) The number of likes or reactions generated on this news on Facebook are 8 and on GooglePlus and LinkedIn it is 0

2.Bloomberg reachtions on Fb,g+,ln with sentiment >0.50 on all the topics




Distribution of topics on Facebook with their reactions from the Source Bloomberg and Sentiment scores greater than 50%

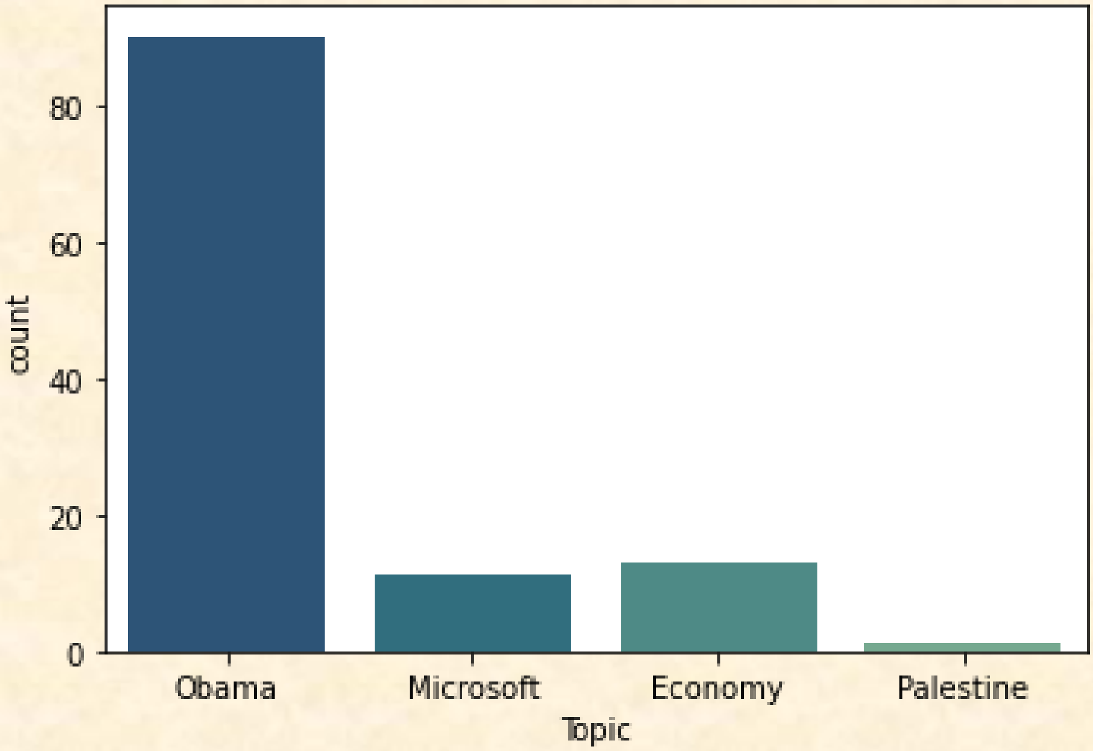

Distribution of topics on GooglePlus with their reactions from the Source Bloomberg and Sentiment scores greater than 50%

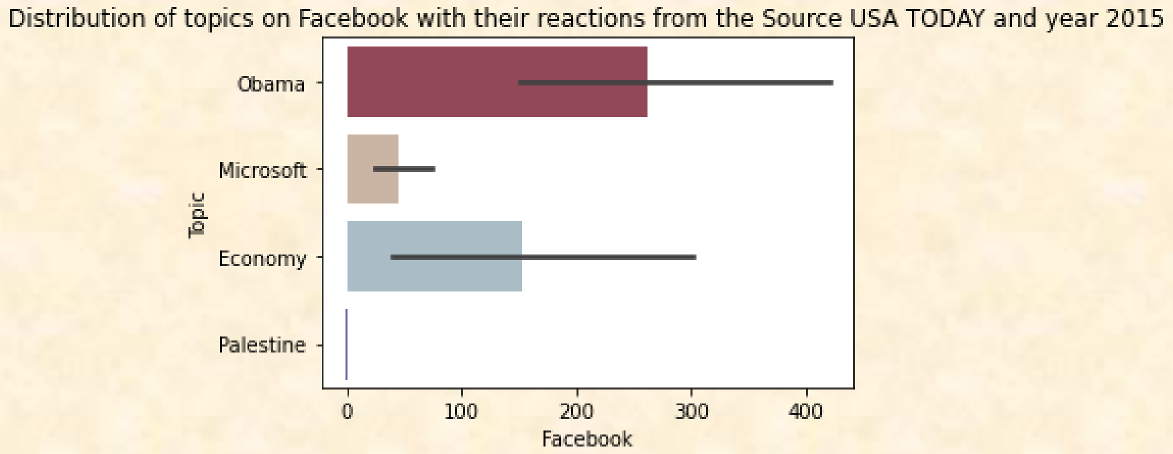

Distribution of topics on LinkedIn with their reactions from the Source Bloomberg and Sentiment scores greater than 50%

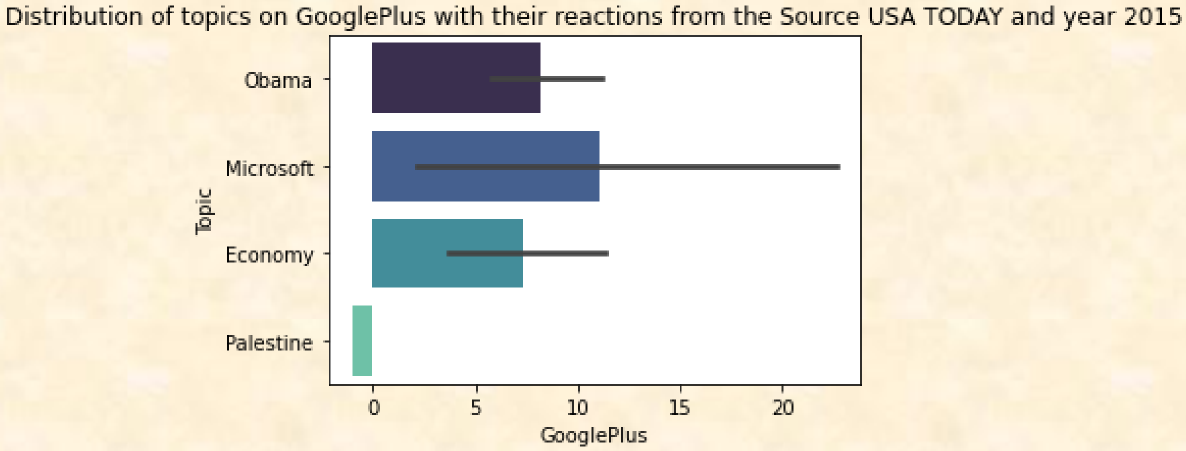
Observation:

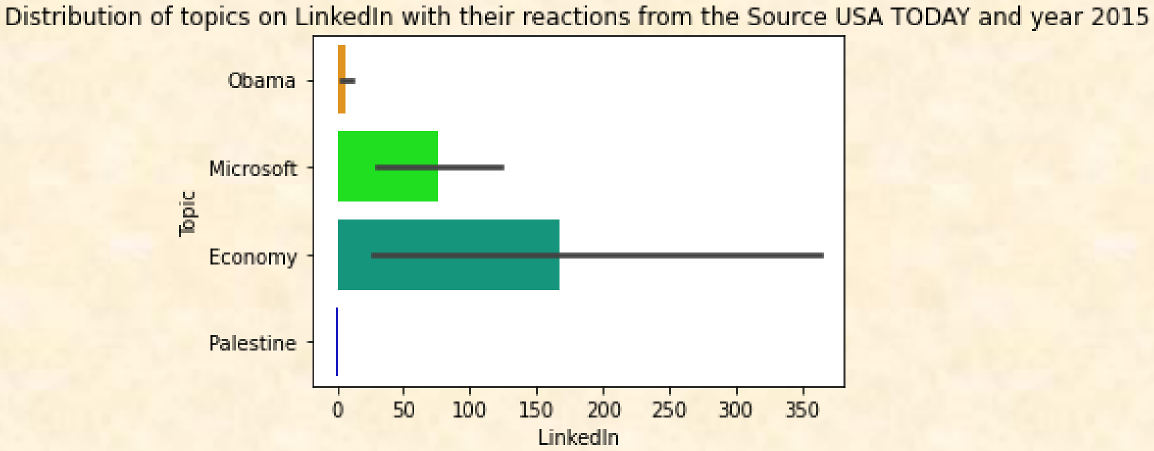
1) It is known from the analysis that from the sample data Bloomberg source news are more than any other.
2) So, from Bloomberg news where the Sentiment of Title and Headline are greater than 50% the more news is on Economy topic(89) and other news are less from Blooberg where the sentiment is greater than 50%.
3) The reactions generated on Bloomberg where sentiment is greater than 50% are 437 likes on Facebook, 118 likes on GooglePlus and 2572 likes on LinkedIn.
4) It can be observed that this Source with Sentiment score greater than 50% has more reactions on LinkedIn than on facebook and googleplus with more than the double reactions of Facebook and GooglePlus. 😯

3. News reactions from Source USA TODAY from the year 2015









Observation:

1) Even though the news published from this source are less on the social media the likes or reactions generated are relatively more.
2) There are 115 news articles published on social media from the Source 'USA TODAY' consisting of various topics with 90 from Obama, 13 from Economy, 11 from Microsoft and 1 from Palestine.
3) There are more likes on facebbok for this news published from the Source 'USA TODAY' than on Googleplus and LinkedIn.

4) The likes generated on Facebook for the news from this Source are 5367, 64 likes on Googleplus and 1242 on LinkedIn.

7: Conclusion

- Predicting the exact number of likes or interactions based on different topics of news articles on the social media platforms can be hard then approximate or near estimation can be done.
- May be this the reason behind the low r2 scores on the models compared to other regression problems where we can get upto 70-95% scores with tuning and choosing different models.
- There is an imbalance in the news articles topics given in the training data. Having more data on the different topics can increase the performance of the models. Like news on Palestine is very less compared to other topics
- Based on the news articles people have seen it can be said that there are good number of people who have interacted or discussed or have responded to the news on given social media.
- People are using Facebook more than GooglePlus and LinkedIn also the reactions generated on Facebook are relatively more than other two social media platforms available for analysis.
- More business news like Microsoft, Economy topic news are more discussed on LinkedIn than on Facebook.