Capstone Project - 2

Supervised learning algorithms for predicting article retweets and likes based on the title

Subitted by : Anshul Vyas

Mentor : Kevin Glynn

Index

Figures

Tables

Chapter - 1 : Definition

1.1 Project Overview :

Today we are surrounded by various Social meia platforms. These platforms are now integral part of human life. Research shows that on average different age groups are spending six to nine hours daily on social media platform. In these hours, people use these platform in various ways like some people enjoy watchng others updates, some updates their personal stuff or some chats with friends. One study shows that around 60 percent of the social media used for read news or gather information.

At the same time we can see the growth in number of article available on these platforms. Many writers have started writing about various topics as these platforms are providing enough audience for the articles. But it is thoughtful question that inspite of having good content and information , why some articles don’t perfrom well on social sites? Or is there any way to decide or predict that an article would be loved by people or not. What are the chances that a particular article would get good amount likes ond retweets on twitter or huge number of claps on medium. What are the attributes which can affect the popularity of an article. We are going to define a system in this project which can give the answers for all these questions. Some writers used a set of words in the title which force the audience to open and read it.

What are these set of wods which boost the curiosity of rear to that level where he cnt mis to open the article . IN machine learning there are various ways to explore this problem.

1.2 Problem Statement:

Due to enroumus amount of articles avaialbilty at social media , its been observed that so many good article lost the popularit battel agaist a average article. Why? While it is expected that a good article will inluence the thought process of the audience where it cant even reaching them or if reaches , people are not opening the link. Again question arises. why?

Reseach says that users take fraction of seconds to tdecide whether they are going to read the artcle or not. What is the deciding factor here. Getting the hit on a link of article is highly dependent on the title used by the autor. In this project several machine learning algorithms will be used to predict the chances of an article getting more like,retweets or claps on the basis of its title. This is a clssification problem and going to be solved by using various supervised learning algrithms.

1.3 Evaluation Matrics:

Every project must have an evaluation matrics to analyse the solution generated by the models. There are varius types of accuracy measusres available for data science projects. In this project a simple ration basd accuracy facotor will be used.

Accuracy = A / B

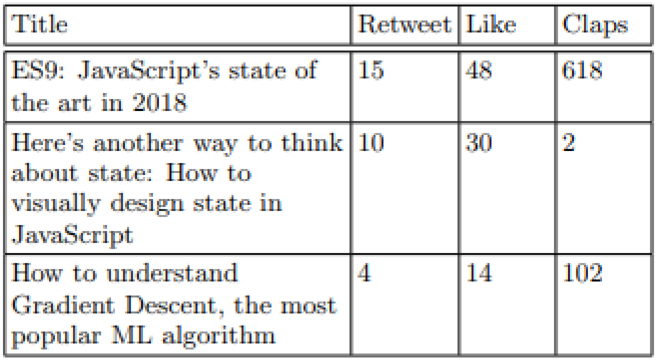
A = Number of correct predictions
B = Total number of predictions

The major restriction for using this formula for finding accuracy is it demands that there should be similar number of samples belong to every class. That is why rane of number of claps, likes and tweets are divide into few classes. Such division will remove thechances of imbalace class distribution problem .

Chapter - 2 : Analysis

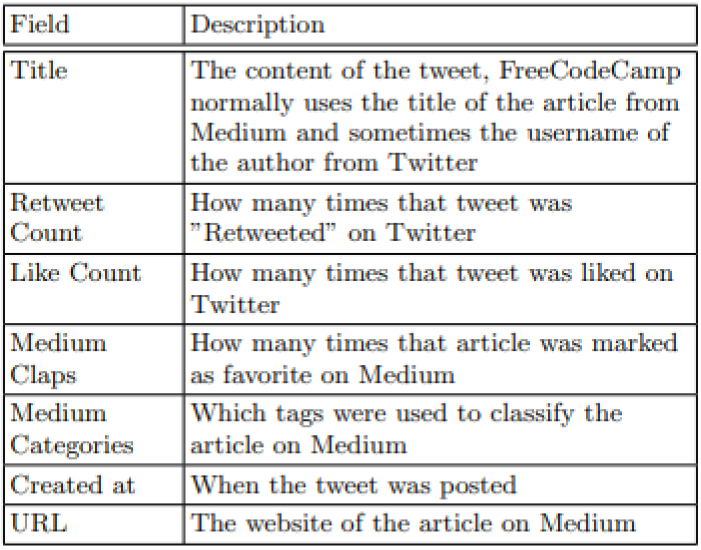
2.1 Data Exploration:

Data used to predict how titles will perform was gathered from the accounts of the non-profit organization FreeCodeCamp on Medium and Twitter. After getting the articles from FreeCodeCamp written on Medium and shared on Twitter, there is a dataset of 711 data points. Table I shows some examples of such correlation and table II explains the complete list of fields of the dataset.



Title	Retweet	Like	Claps
ES9: JavaScript's state of the art in 2018	15	48	618
Here's another way to think about state: How to visually design state in JavaScript	10	30	2
How to understand Gradient Descent, the most popular ML algorithm	4	14	102

Table 1 : Snapshot of Dataset



Field	Description
Title	The content of the tweet, FreeCodeCamp normally uses the title of the article from Medium and sometimes the username of the author from Twitter
Retweet Count	How many times that tweet was "Retweeted" on Twitter
Like Count	How many times that tweet was liked on Twitter
Medium Claps	How many times that article was marked as favorite on Medium
Medium Categories	Which tags were used to classify the article on Medium
Created at	When the tweet was posted
URL	The website of the article on Medium

Table 2 : Attributes of the data set

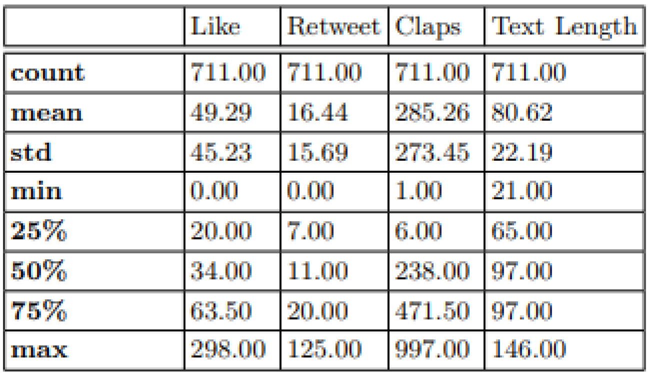
Chapter - 2 : Analysis

2.2 Exploratory Visualization:

In this section several techniques of Data visualization have been used. Histograms and box plot are used to demonstrate the datapoints. Scattter matrix os used to find the relatinship in all available features. This whole process is divided into following parts:

2.2.1 Overall Statistics
2.2.2 Histogram and Box plots
2.2.3 Scatter Matrix
2.2.4 Tite length that performed better
2.2.5 Categories that performed better
2.2.6 Words that performed better

2.2.1 Overall statistics:



	Like	Retweet	Claps	Text Length
count	711.00	711.00	711.00	711.00
mean	49.29	16.44	285.26	80.62
std	45.23	15.69	273.45	22.19
min	0.00	0.00	1.00	21.00
25%	20.00	7.00	6.00	65.00
50%	34.00	11.00	238.00	97.00
75%	63.50	20.00	471.50	97.00
max	298.00	125.00	997.00	146.00

Table 3: Statisctical Overview

From the above table it can be easily seen that there are total 711 data points. On average articles are liked 50 times , retweeted by 16 times and clapped 285 time. It can also observerd that article length ranges from 21 to 146 characters. Later in this project, we will try to find coraltion between these features to understand more about the problem.

2.2.2 Histogram and Box plots

Histograms and boxplot are very useful tools to find the distribution of data points. First we will plot histograms for the Retweets, Likes and Claps. In fig 1, we can see Retweet, likes and Claps , all these three features have positively skewed histogram. I means majority of the data points are performing at lower range and some of the data points are performing extra ordinary. These data points can disturb our analysis . So in later section we will remove these points which are also called outliers.

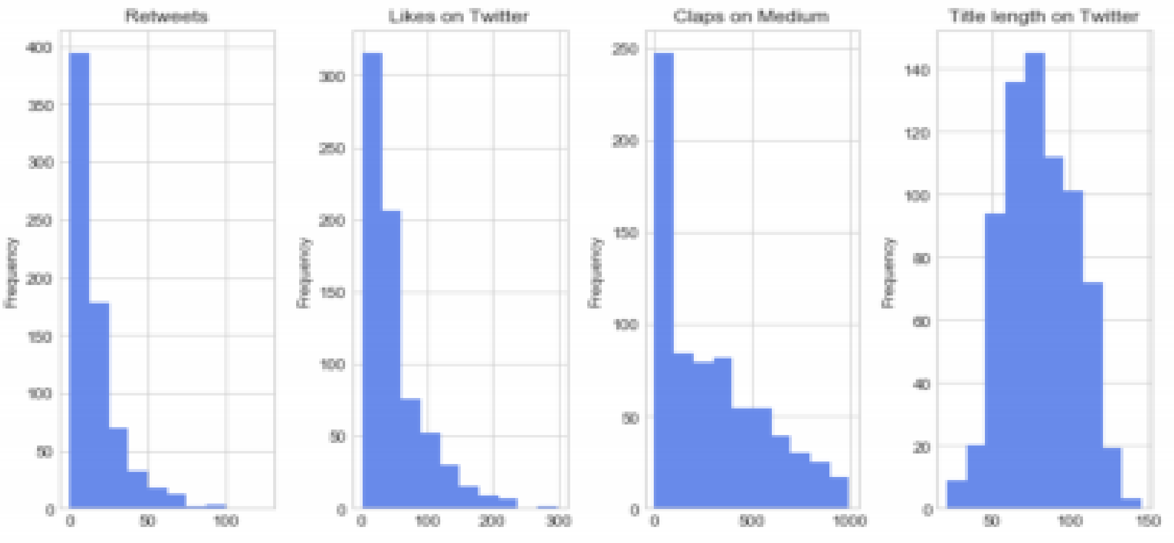


Fig 1 : Histogram with main features (Retweets, Likes and Claps)

Here are the box plots of the same features. Box plots help us to visualizre the statistics of the data points. In box plots, w can easily see minma and maxima along with all three quantile . Standard symbol for these quantiles are Q1, Q2 and Q3. There is one more keyword IQR which is called Inter Qurtile Range. We will use this term to remove the outlier in later section.

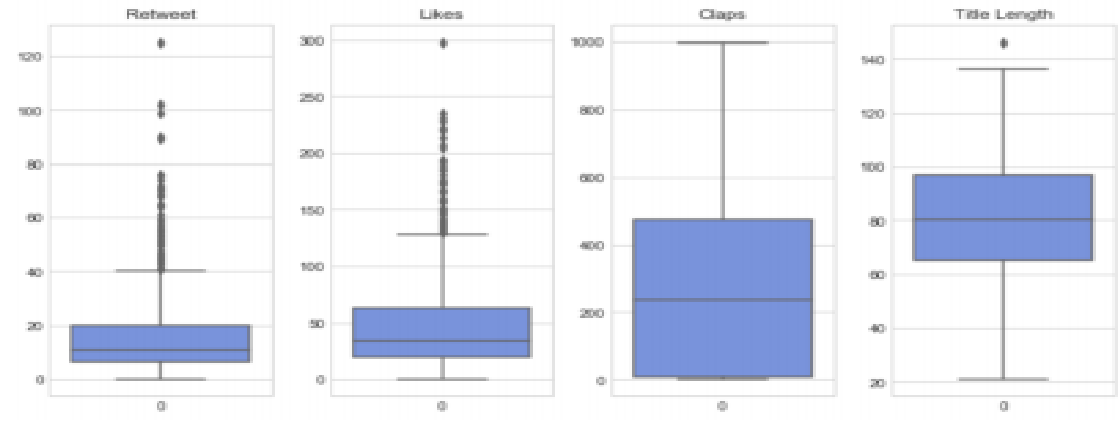


Fig 2 : Box plots with main features (Retweets, Likes and Claps)

We can understand from the above analysis that all three features i.e Likes, Retweets and Claps are positively skewed while data points belong to Title length feature are normally distributed .

2.2.3 Scatter Matrix:

This is the most sophisticated python function which produce an overall relationship matrix between all the features available in the dataframe. Dataset used in this project has several features like Retweets, Likes , claps and title length. Following is the matrx which produce the required visualiztion . We can conclude that that Retweets and likes are corelated . There is no direct corelation between other freatures so further analysis is required.

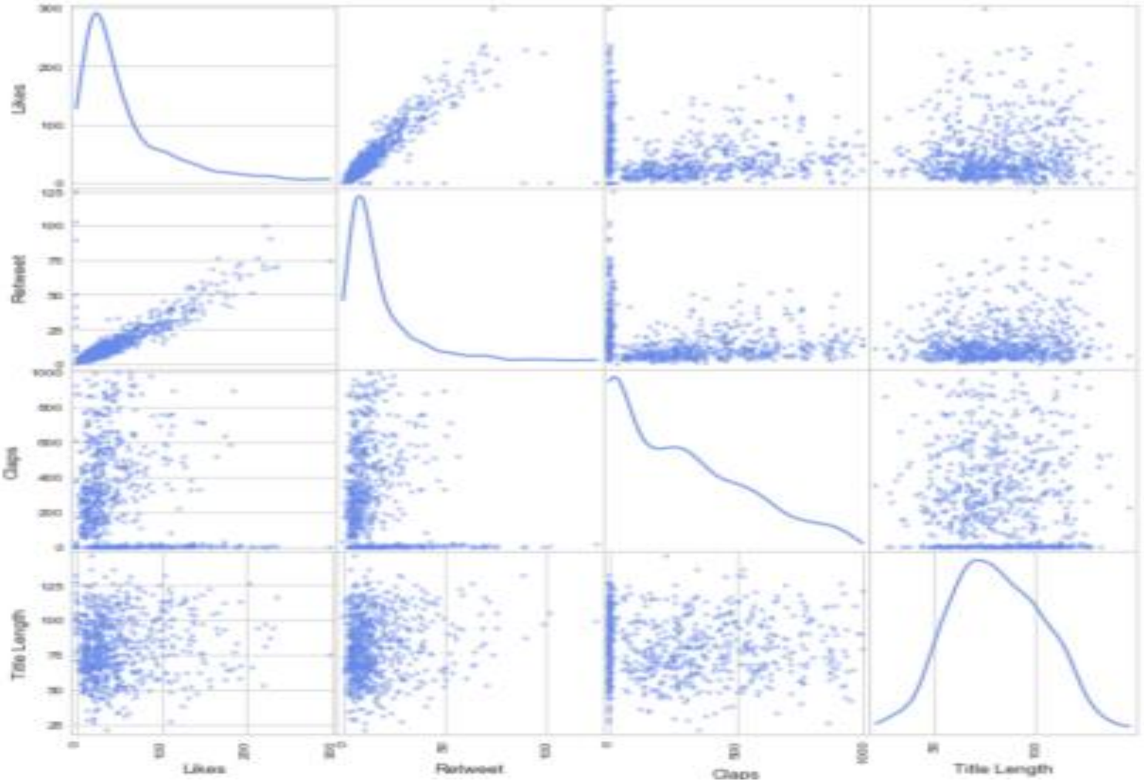


Fig 3 : Scatter Matrix between features (Retweets, Likes, Claps and Title Length)

2.2.4 Title Length that performed better:

As we did not get any concrete information form previous analysis, we are going to analyze the relation between title length and all other three features to visualize what exact relation these data points have. Follwing is the visualization of title length performace.

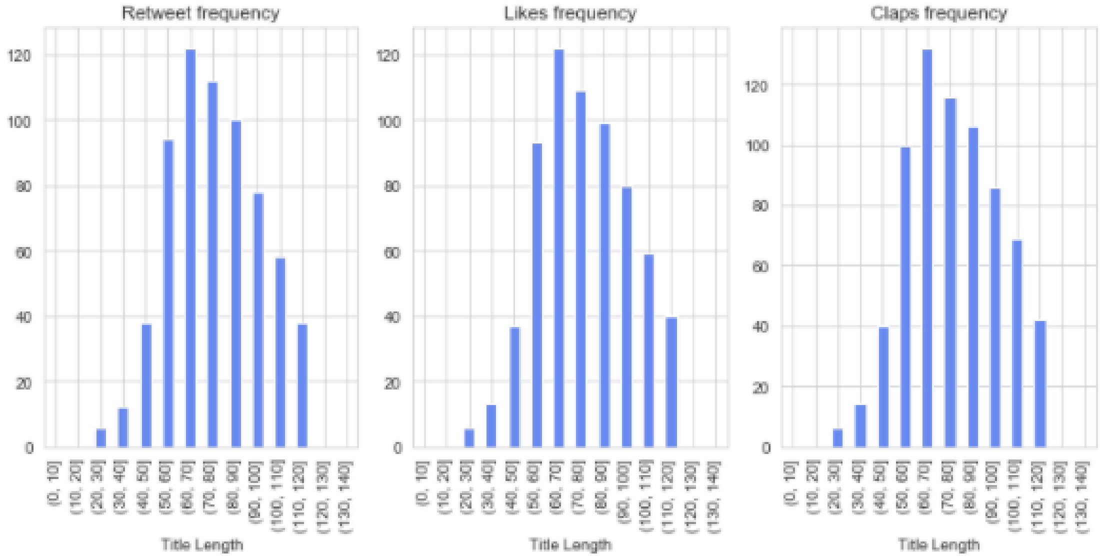


Fig 4 : Title length vs other features

To make the this anlysis bias-free, we are going to remove the outlier form the data set. There is a common practice to remove the biases using IQR method i.e Inter Quartile ratio. Followig is the way to cut down the outlier data points:

Outlier < Q1 – 1.5 * IQR
Outlier > Q3 – 1.5 * IQR

Q1 : First Quartile
Q2 : Second Quartile
Q3 : Third Quartile
IQR = Q3 - Q1

We can notice that title range from 50 to 110 performs better than others.

2.2.5 Catagories that performed better:

After removing the outlier, follwing graph was plotted to indentify the set of catagories which are peforming better than others.

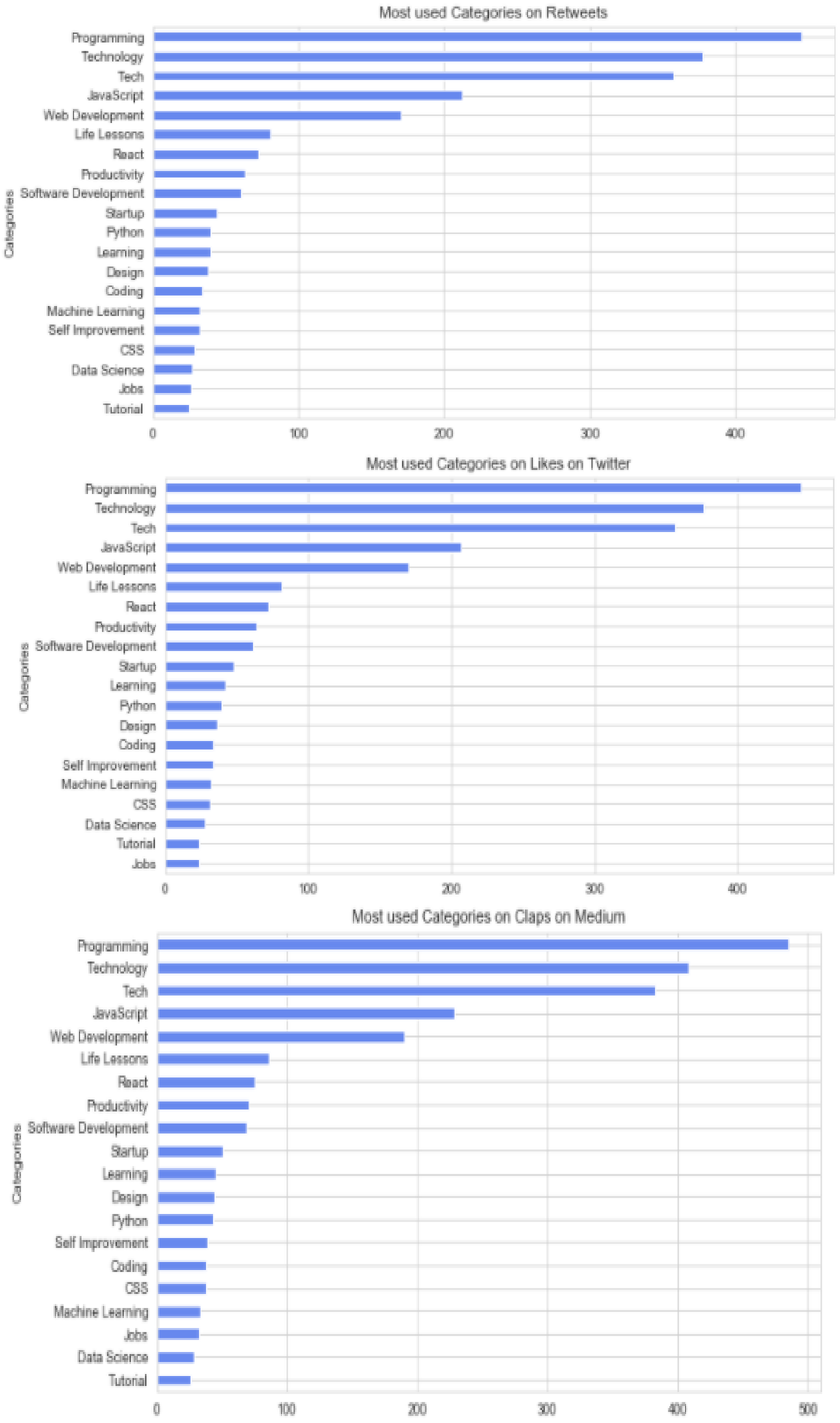


Fig 5: Category analysis on all three features

Chapter - 2 : Analysis

Previous analysis shows that there are good chances of getting more hits if an article has following catagories in the title:

- Programming
- Tech
- Technology
- JavScript
- Web development etc.

Chapter - 3 : Algorithms

Classification is a common task of machine learning (ML), which involves predicting a target variable taking into consideration the previous data . To reach such classification, it is necessary to create a model with the previous training data, and then use it to predict the value of the test data . This process is called supervised learning, since the data processing phase is guided toward the class variable while building the model . Predicting the number of retweets and likes of an article can be treated as a classification problem, because the output will be discrete values (range of numbers). As input, the title of the articles with each word as a token (t1, t2, t3, . . . tn), the title length and the number of words in the title. For this task, we evaluated the following algorithms:

- *Support vector machine (SVM):* SVM contructs a hyperplane (or a set) that can separate the points in the defined labels. The distance between the closest data points and the hyperplan is named margin. An ideal separation is defined by a hyperplan that has the largest distance to the closest points of any class, so the challenge is to find the coefficients that maximize this margin. This model was chosen, because it works well then big quantity of features and relatively small quantity of data and to deal well with linear and non-linear datasets.

- *Decision trees:* This model uses a decision tree to classifies the dataset into smaller subsets, and to define a conclusion about a target value. The tree consists of leaves, where the intermediate ones are the decision nodes and the ones from the extremes are the final outcomes. This model was chosen, because it can be easily interpreted, visualized and explained. Also due the fact that this model implicitly perform variable screening or feature selection.

- *Gaussian naive Bayes (GaussianNB):* This model is a classification technique based on the Bayes' Theorem. It assumes the independence among the involved features. Nevertheless, this approach performers well even on data that are dependent between them. This algorithm was created by Bayes to prove the existence of God. It relies on the probability of an event, based on prior knowledge of conditions that might be related to the event.

- *K-nearest neighbors (KNN):* This algorithm takes in consideration the k closest points (neighbors) around the target and use them learn how to classify the desired point. This model was chosen, because its simple to implement, no assumption about the data is necessary and the non-parametric nature of KNN gives an advantage in certain settings where the data may be highly unusual.

- *Logistic regression:* This model is named after the core statistical function that it is based on, the logistic function. The Logistic regression estimates the parameters of this function (coefficients), and as result it predicts the probability of presence of the characteristic of interest. This model was chosen, because provides probabilities for outcomes and a convenient probability scores for observations.

- *Naive Bayes classifier for multinomial models (MultinomialNB):* This model is similar to the Gaussian naive Bayer, but the difference is that it was a multinomial distribution of the dataset, instead of a gaussian one. This model was chosen, because it works well for data which can easily be turned into counts, such as word counts in text. However, in practice, fractional counts such as TF-IDF may also work.

Chapter - 4 : Methodology

4.1 Data Preprocessing

4.1.1 Data Cleaning:

The first part of Data Processing is to clean the data set. We need to take some set of actions on the dataset so that we can analyse efficiently and make it compatible to provide as an input for various algorithms. User names from the tweets was removed . '@' symbol was kept intentionaly to make statistical analysis easy. There were non ascii caracters in the tile those were also removed by using regex. Outliers wee aslo removed by using IQR method. After following all these steps , totla 711 daa points wee ready to analyse further.

4.1.2 Assgning classes to the dataset:

As discussed earlier in the Overview section, to utilize the full potential of supervised learning algorithms, we need to divide the data into classes as supervised learining algorithms work more efficiently whn all clasess have same number of data data points. Following graphs shows the proper class division between the data features. To decide the clasees, following ranges were introduced:

1. Retweets: 0-10, 10-30, 30+
2. Likes: 0-25, 25-60, 60+
3. Claps: 0-50, 50-400, 400+

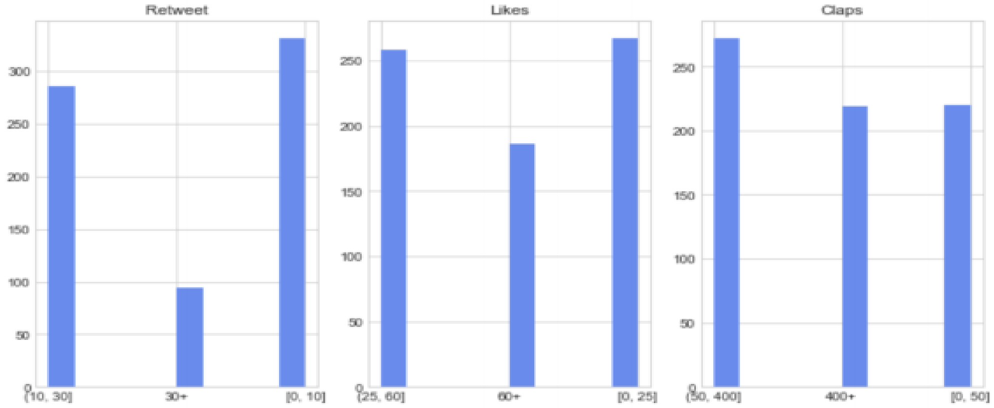


Fig 6 : Range Distributio for equal classes

4.1.3 Bag of words :

To be possible to analyze the title in each data point, we need to map each word into a number. This is necessary because machine learning models normally don’t process raw text, but numerical values. To reach this, we used a bag of words model . In this model, it is taken into consideration the presence and often the frequency of words, but the order or position is ignored. For the calculation of the bag of words, we will use a measure called Term Frequency, Inverse Document Frequency (TF-IDF) . The goal is to limit the impact of tokens (words) that occur very frequently. At this step, we processed the collection of documents and built a vocabulary with the known words. We reached a vocabulary of 1356 words for retweets, 1399 words for likes and 1430 words for claps.

4.2 Implementaion:

4.2.1 : Training and Testing Data:

Before starting the training and the evaluation of the models, we split the dataset into test and training sets. Retweets and likes have a total of 658 data points each, with 526 (80% approximately) as training and 132 as testing points. Claps has 711 data points, with 568 (80% approximately) as training and 143 as testing points.

4.2.2 : Model Performance Metric:

We separated the dataset into learning and validation set. A validation set is important to reduce the risk of over-fitting of the chosen model. To avoid discarding relevant data points, we used a cross-validation strategy. Cross-validation splits the training dataset in k folds, being k - 1 folders used to train the model and the last one to test it. This strategy will repeat multiple times and the overall performance is the average of the computed values. To estimate the model's accuracy, we used a 5-fold cross validation that split the dataset into 5 parts, 4 of training and 1 of testing. 4.

Outliers: We made some tests to discover if we should keep the outliers for the training or remove them. During the tests, we discover if we keep the outliers, the accuracy was always worse.

Bag of words: To create the bag of Words, we had the option of choosing the CountVectorizer or TfidfVectorizer. During the simulation we got better results with the last one, TfidfVectorizer.

Model's parameters: For each model tested, we calculated the accuracy for the default model

Chapter - 5 : Result

After applying all previously mentioned supervised algorithms, follwong results were achieved.

ID	Model	Accuracy(Like)	Accuracy(Retweet)	Accuracy(Claps)
1	Logistic-Regression	45.45	59.09	**46.85**
2	GaussianNB	46.21	49.24	41.25
3	DecisionTreeClassifier	49.96	50.00	31.46
4	SVC	46.96	56.06	41.95
5	KNeighborsClassifier	36.36	48.48	37.76
6	MultinomialNB	40.90	**60.60**	31.46
7	GradientBoostingClassifier	**50**	59.09	43.35

Table 4 : Results

We can notice the best accuracy for each feature are :

- Like is 50% using GradientBoostingClassifier
- Retweets is 60.60 % using MultinomialNB
- Claps is 46.58% using Logistic-Regression

Chapter - 6 : Future Scope

For future work we can think about some additional improvements like adding more features to the original dataset making possible to relate more information to the success of the article. For example, we can correlate the words of the title, with trendy words of the month.Bring more data points to train our model, would also increase the accuracy of the solution.Try to use the position of the word on the title to classify its importance. All models are used with the default parameters in this analysis. By using parameter tuning , results can be further improvised.