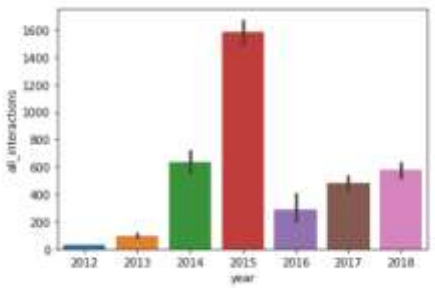
Goalcast Content Consumption Data Analysis

Goalcast provided a csv data file contains observations for the user engagements with the different types of posts (Video, Photo, Status and Link) stating the number of shares, comments and details of interactions (likes, loves, wows etc…).

The attached notebook contains the EDA as well as the prediction models (regression & NN), this summary contains a brief of the findings, description of the prediction models as well as my recommendations.

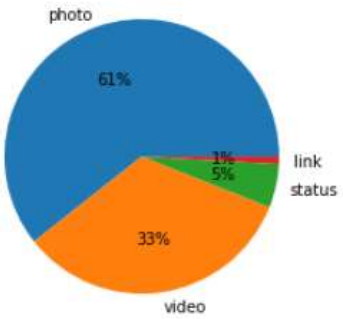
EDA (Exploratory data analysis)

- The data has 7050 observation with 51 duplicates (6999 remaining) with no missing data (n/a)
- 2012 had only photos, status and video started 2013, link started to appear on 2015
- The data shows a sudden drop in all user interactions in 2016, and that's when the reactions other than "like"s started to appear in the data, which indicates a site revamp or maybe the feature added was not very successful, the chart below shows the average interaction and how did it drop on 2016. Total number of interactions trended up resulting 4525% since 2012



Year	Total Interactins	Cumulative
2012	16552	16552
2013	49055	65607
2014	147095	212702
2015	533334	746036
2016	170390	916426
2017	1081592	1998018
2018	1437128	3435146

- Please note that 2012 data starts at June, and 2018 data is also not completed and ends with June, so specifically for 2018 the growth reflected in the data is partial.
- Majority of data are photos, the distribution is clarify in the pie chart below



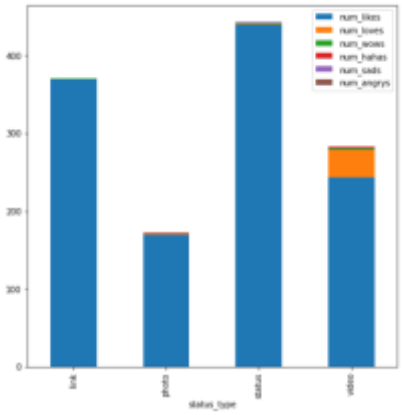
- Assuming that the num_reactions has the summation of different interactions (likes, loves, hahas ….) I found 9 records with a num_reactions do not match the summation. The difference is too small to worry about.
- There is a clear correlation between number of shares and number of comments, less with number of reactions

	num_reactions	num_comments	num_shares
num_reactions	1.000000	0.156190	0.259640
num_comments	0.156190	1.000000	0.640536
num_shares	0.259640	0.640536	1.000000

- High correlation between loves, waws and hahas and Medium correlation between loves and angrys
- Looking at the correlation between the different emotions and (num_comments, num_shares) we find that the highest correlation is between shares and (loves, then wows and hahas)

	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angrys	num_comments	num_shares
num_likes	1.000000	0.212715	0.200015	0.123129	0.057110	0.096276	0.105934	0.179664
num_loves	0.212715	1.000000	0.510148	0.507899	0.226574	0.394818	0.521320	0.820284
num_wows	0.200015	0.510148	1.000000	0.288468	0.091425	0.191672	0.163506	0.409803
num_hahas	0.123129	0.507899	0.288468	1.000000	0.154035	0.225083	0.325064	0.399864
num_sads	0.057110	0.226574	0.091425	0.154035	1.000000	0.143711	0.257545	0.218628
num_angrys	0.096276	0.394818	0.191672	0.225083	0.143711	1.000000	0.239557	0.332355
num_comments	0.105934	0.521320	0.163506	0.325064	0.257545	0.239557	1.000000	0.640536
num_shares	0.179664	0.820284	0.409803	0.399864	0.218628	0.332355	0.640536	1.000000

- Videos harvest the highest number of interactions, then photos, status and the smallest number of interactions goes to links
- Among the interactions, comments is the dominant type then comes the reactions
- Likes is the dominant type of user reaction in total, as it was the only available reaction before 2016, and mostly goes to status, then links, videos and finally photos
- Loves come after likes in dominance, but most of the hearts go to videos
- The counts of other reactions is not significant
- Although the emotions other than likes were added to Facebook in May 2017, we see these emotions appear before than in the provided data.



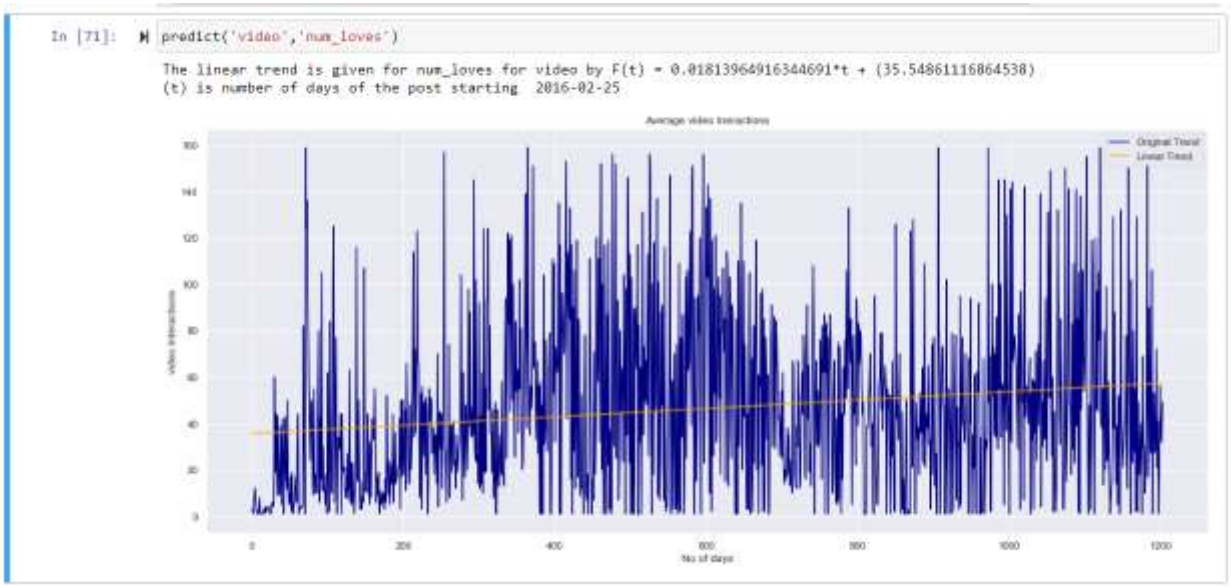
In the attached notebook you will find many heat maps and figures that support the theory above, and this is why I will only use the data from 2016 onward to train the 2 prediction models.

Prediction Models

As mentioned above, the data was sliced so only the observation in 2016 onward are used, I also removed the outliers using only the data between 0 and 95%, as outliers will act like a noise to the prediction models.

Regression model:

As user engagements differs among the different post categories, and although there are correlation between a few (likes, loves and angry's) as explained in the notebook, they are not directly linked, this is why I created a function that can be used to predict each category / emotion separately, the outcome of the model is a line chart for the chosen post category / emotion plus the regression line, and a formula that explains that



Neural Network

I was not as lucky in neural network, as it seems we need more data, but it will be able to predict (once trained) all different reactions in once shot.

The other reason was that the data that we have is in-balanced in terms of the different post categories, however there are solutions for that called SMOT (Synthetic Minority Oversampling Technique) which helps generating observation which look realistic and adding it to the training observation.

Recommendation

It is possible to proceed and deploy the regression model in the notebook, but I think there is much more that we can do

1- Having a bigger number of data is a must, the more data we provide to ML the better results we get.
2- Building data pipelines which will guarantee live data for the model,
    a. We can train and tune a time series prediction model, as it works best for predicting future data out of historical data (like weather / currency prices etc…).
    b. we can create a dashboard to show the real live view, I have created quickly a sample simple dashboard and attached it as GIF to see interactions
3- The data lacks the post meta-data, which can help to understand further the user engagement metrics, I noticed that posts have ID's and surely if have access to the title, description or article associated with each ID we can extract the keywords, categories the posts accordingly and maybe will help us to pick the title, and include the keywords that generate maximum engagement.
4- If we can get the user profile we can build a recommendation system that provide the user a content based on his previous interactions and/or his demographic and country, this will also guarantees a higher engagement by the user. Knowing that recommendation systems will mathematically extract a tenant characteristics for both posts and users that is beyond human ability to do manually.

October 2021

Hayan Al Mamoun