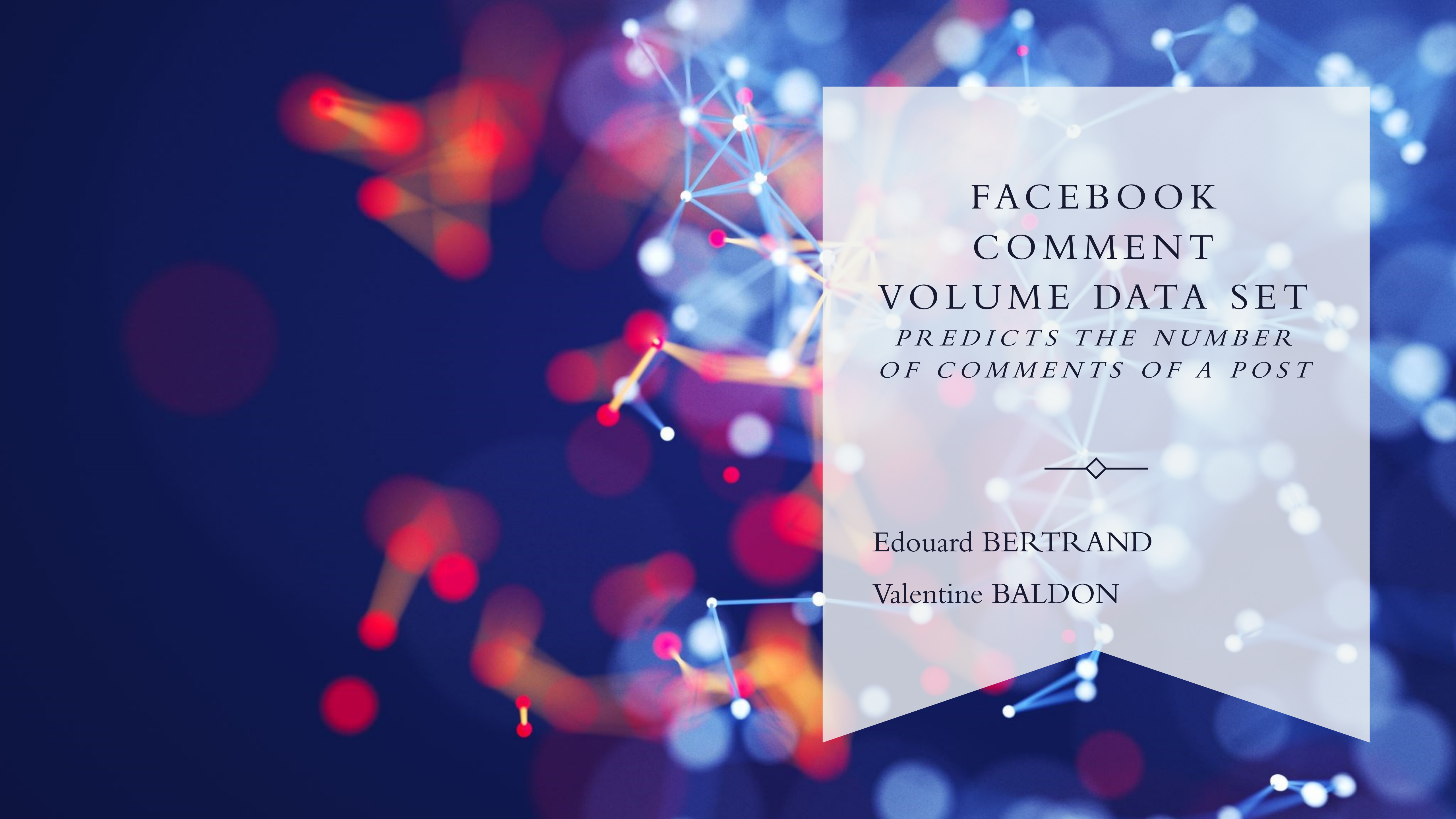


FACEBOOK COMMENT VOLUME DATA SET

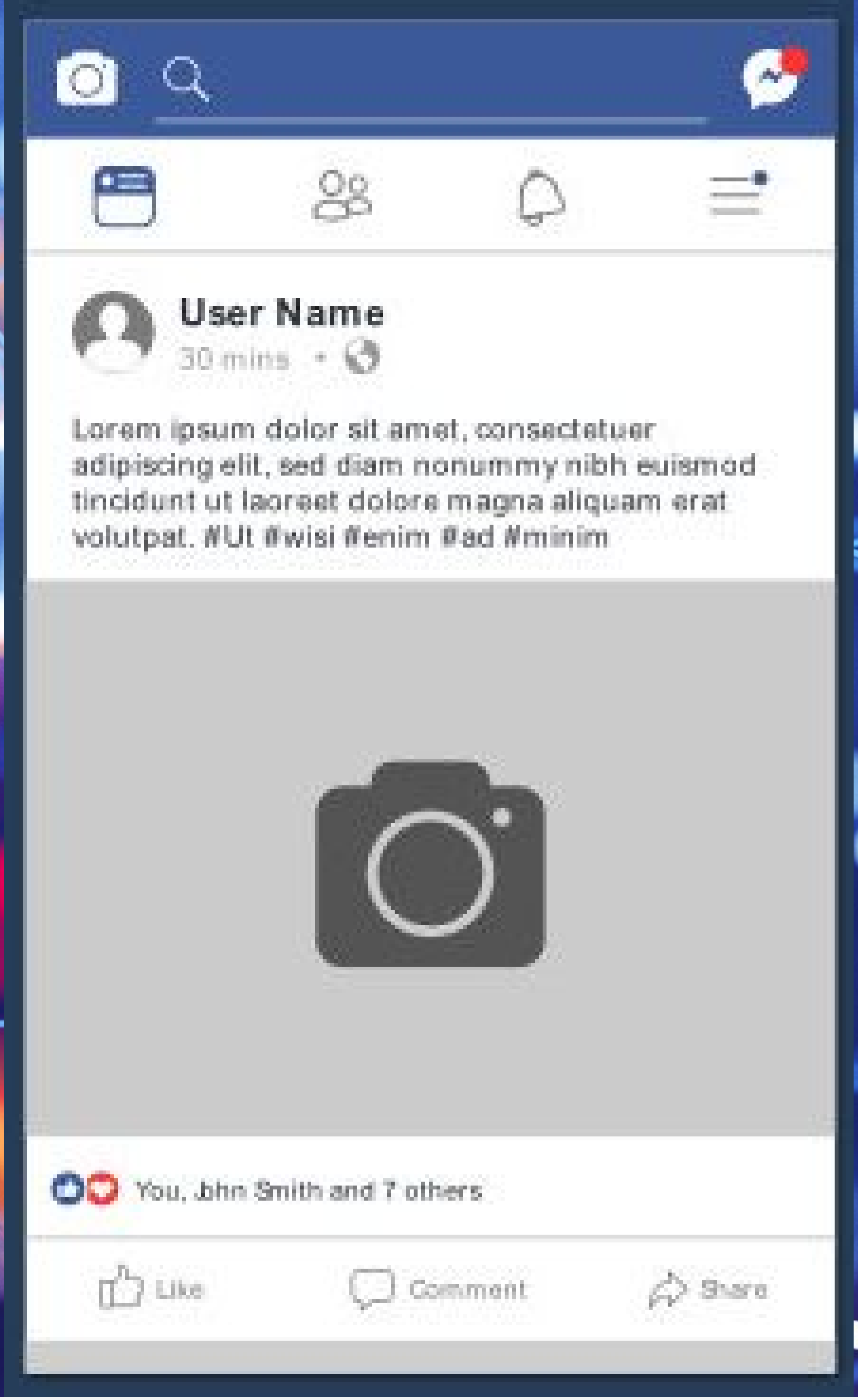
PREDICTS THE NUMBER OF COMMENTS OF A POST

Edouard BERTRAND

Valentine BALDON

The topic

The Facebook Comment Volume Data set is a really interesting dataset. It contains data from hundreds of thousands of Facebook posts. This data set allows us to predict with a machine learning algorithm how many comments a post will get in the next few hours.

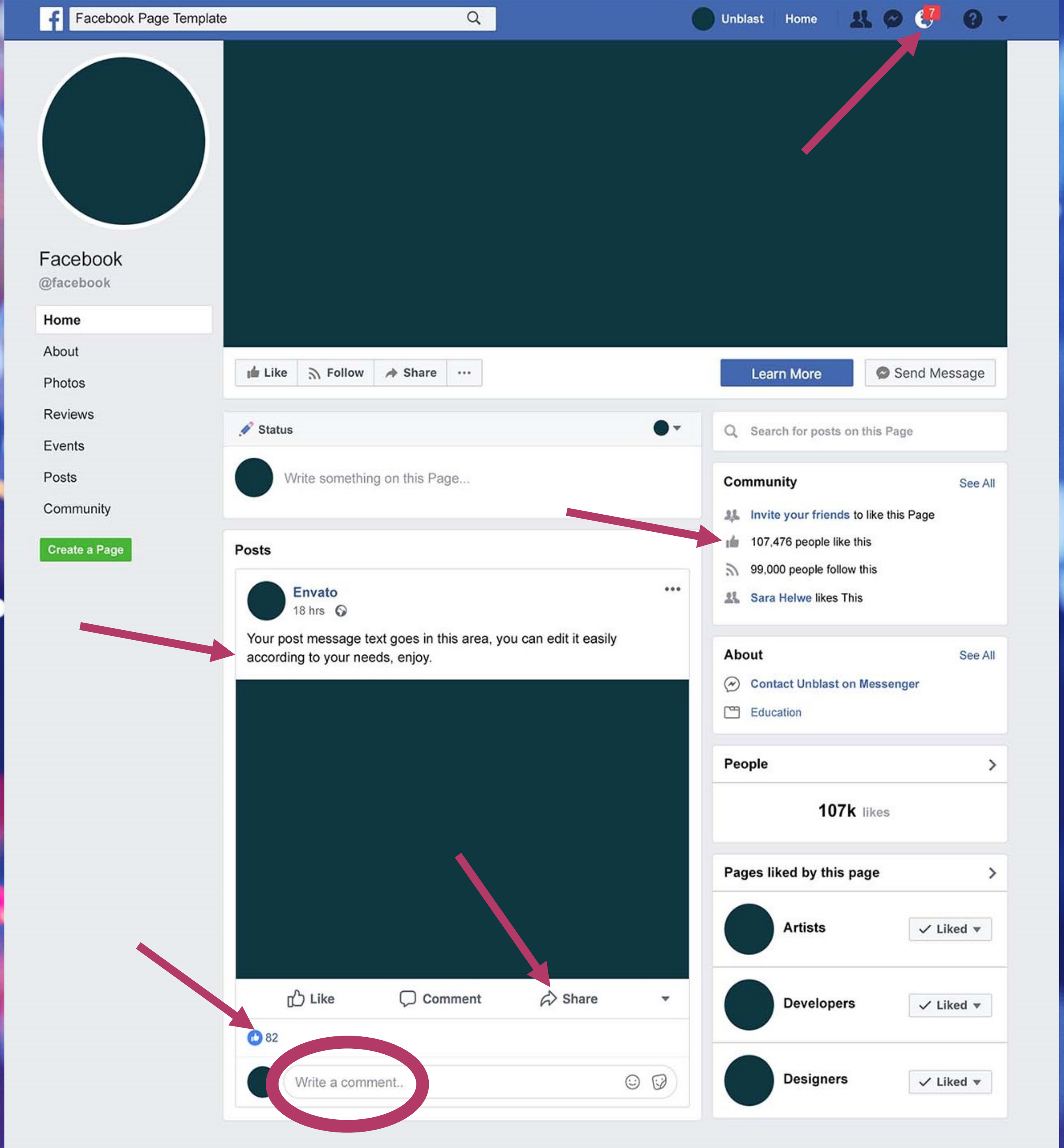


Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. #Ut #wisi #enim #ad #minim

The variables

Our data set has 54 variables. They give information about the posts and the page that posts them. This information is, for example, the number of likes of the page, the daily interactions, the category of the page, the number of comments at different times...

The problem is that we did not get a clear description of variables 5 to 29. The link to the document is dead and on the internet we did not find anything on this part of the data.
We also lacked a valuable piece of information throughout our study: the date and time of publication. Without this we could not realize some of our ideas.



Your post message text goes in this area, you can edit it easily according to your needs, enjoy.

Added variable

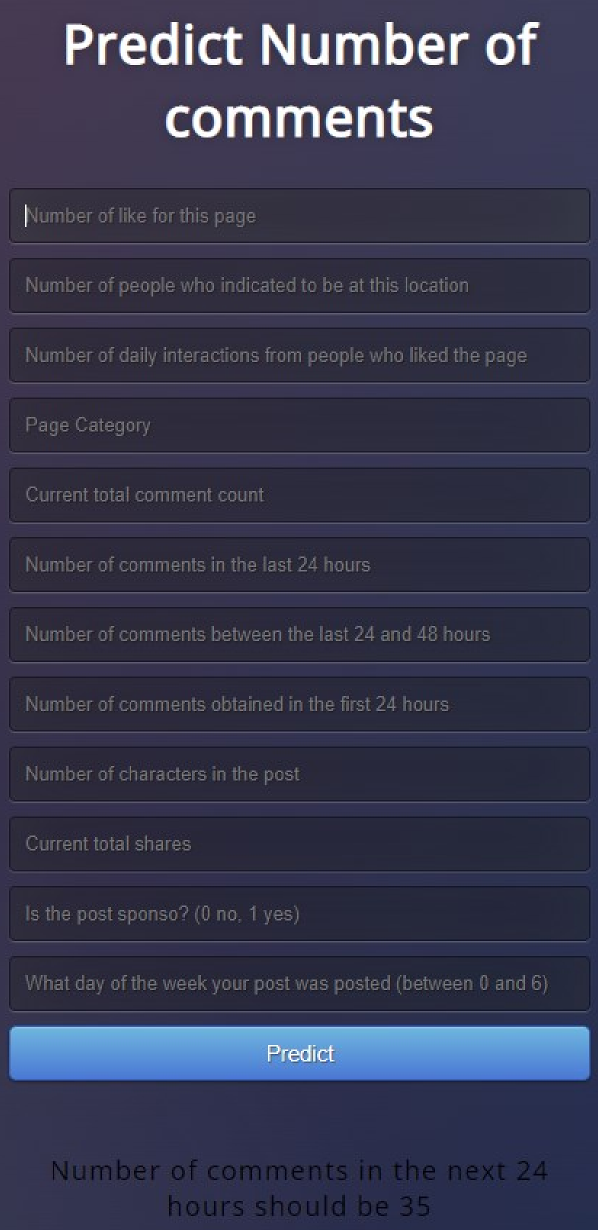
We have chosen to add a variable which gathers the variables 40 to 46. These variables allow us to know which day the post was put online, 1 if it is the day of the variable 0 otherwise. The problem is that they are very hard to use. We had to combine groupby, count_values and do each day independently. So we decided to create a new variable ranging from 0 to 6 with each number corresponding to a day of the week. This variable is called DayWeek.

Cleaning dataset

To be able to correctly use the variable NbCom48hBeforeBD which corresponds to the number of comments received between 24 and 48h before the time of prediction (BaseTime) we removed all the posts having been posted less than 24h before BaseTime. For this we looked at all the posts that had received 0 comments in this interval and where the number of comments 24 hours after the post and 24 hours before BaseTime were equal so as not to remove posts that had only received 0 comments.

We also chose to only keep post on which the prediction was done on 24h for consistency reason. We did not want to have unnecessary variation related to the time span given to the prediction. We do not want to compare a post that got 5 comments in 1 hour with a post that had 5 comments in 7 hours, for example.

Our Flask application



Predict Number of comments

Number of like for this page

Number of people who indicated to be at this location

Number of daily interactions from people who liked the page

Page Category

Current total comment count

Number of comments in the last 24 hours

Number of comments between the last 24 and 48 hours

Number of comments obtained in the first 24 hours

Number of characters in the post

Current total shares

Is the post sponso? (0 no, 1 yes)

What day of the week your post was posted (between 0 and 6)

Number of comments in the next 24 hours should be 35