Kaggle Competition

Xavier Weber, Jeroen Vermazeren

April 2018

1 Preprocessing data

The data set contained no missing values. The summary of the dataset reveals there are some abnormally high values in the n_non_stop_unique_tokes, so those outliers were removed from the dataset. The summary shows no other obvious analomies. If one still decides to delete the outliers of every column using the standard boxplot method, the number of instances dwindles down to 140. So it was decided not to remove any more outliers.

2 Algorithms

A linear regression was fitted to the model to predict shares on all independent variables. It then shows that the null hypothesis of the following variables can be rejected:

- n_tokens_title
- num_hrefs
- data_channel_is_socmed
- num_self_hrefs
- num_imgs
- average_token_length
- data_channel_is_entertainment
- kw_min_avg
- kw_max_avg
- kw_av_avg
- global_subjectivity

A regression was then fitted to predict shares on the variables mentioned above. (figure 1)

A neural network (NN) was also applied, however due to time and computation constraints the NN could not be trained enough to provide any meaningful results.

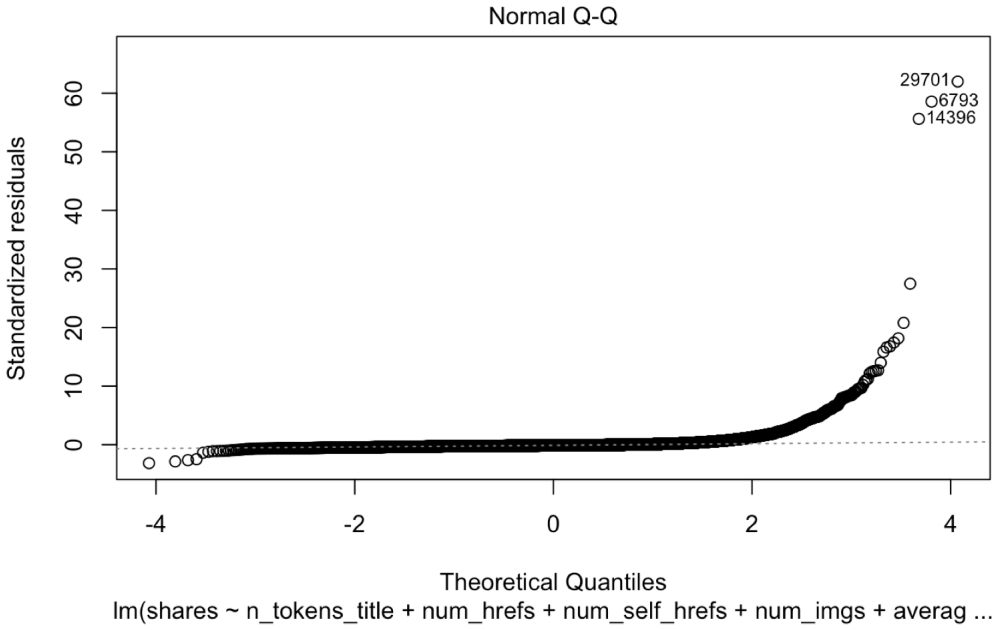

Figure 1: Regression