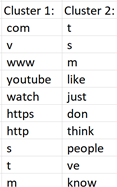
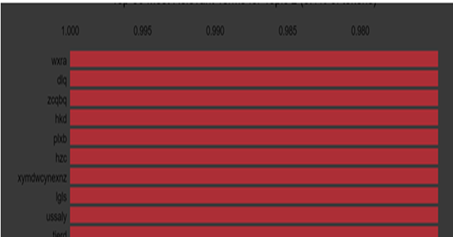
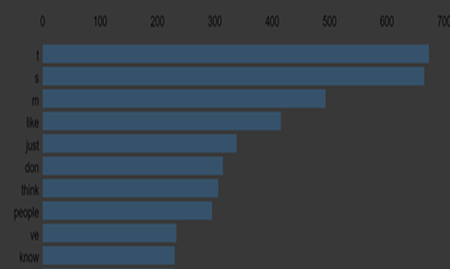
Two different clustering techniques were used during this analysis. The first was K-means clustering, and the second was Latent Dirichlet Analysis. Ideally, there would be 16 perfect clusters of topics, or one cluster per personality. However, any obvious pattern for cluster and one of the four personality traits could potentially be very powerful in our analysis. For example, if there are two cluster and they are noticeably different in extroversion and introversion. It is important to try and have distinguishable clusters and not any major overlaps for this or any cluster analysis.

The results of both clustering analysis were similar in the quantity of clusters, but differ in their respective top terms. Figure 9 shows the top words in the K-means clustering method. Figure 10 shows the top 10 words for the Latent Dierchlet Analysis, or LDA cluctering technique. Both results show case the same result of no real distinguishable clusters from our posts.



*Figure 9: K-Means top 10 terms*



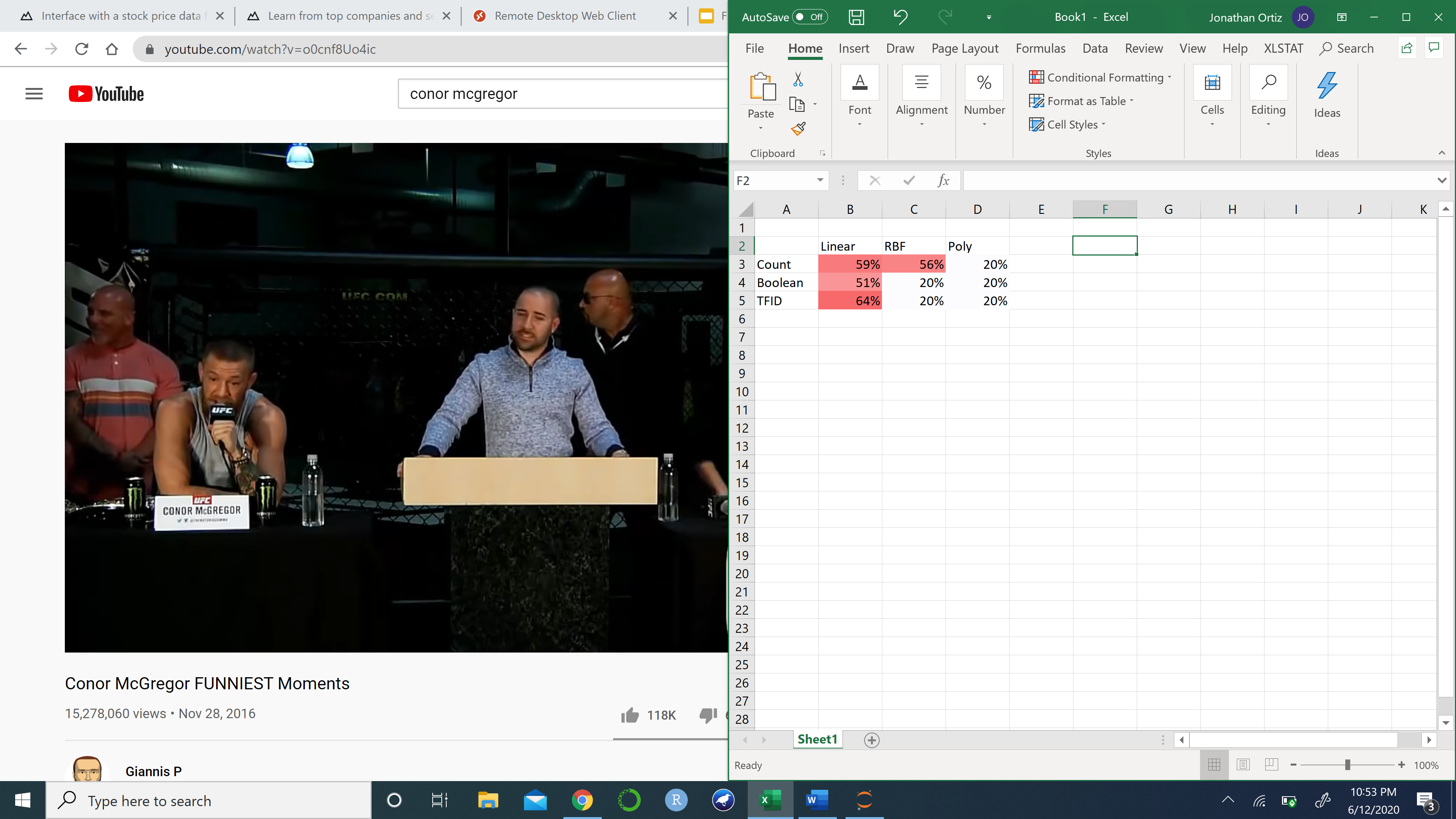
*Figure 10: LDA top 10 terms*

Models

Support Vector Machines were used to try and predict the personality. Three Kernels (Linear, Radial Basis Function, Polynomial ) , and three vectorizers (Count, TFID, Boolean) were used.

Results

The Results of the support Vector Machines are displayed in the table below. It is important to nots the models that have 20%, would predict all the personality as one personality type. The TFID vectorizer performed the best with 64 % accuracy.



*Table 10: Vectorizers and Kernels Results*