Tried logReg, RandomForest and SVM models - sometimes with kmeans clustering (normalized DocumentTermMatrix) and also with hierarchical clustering (only NewsDesk). RF performed best for me. Was #475 on public Leadership board and #523 on the private Leadership board. Fairly OK.

Engineering:

Filled missing NewsDesk, SectionName and SubsectionsName according to the pivot structure.

log + LogWordCounts + Weekday + WordCount Headline + WordCount Abstract + IsQuestion + RFD (binary for SubsectionName == Room for Debate) + Sentiment Abstract (Emotion + Polarity).

The sentiment analysis, in particular Polarity increased the AUC, not a high impact, but helpful.

The Boruta package was very useful for me to identify the most relevant classifiers for the Random Forest model. But depending on the number of variables it can run over an hour.

Would really like to know how to use ensemble model technically (I mean the code). Did not get it and had not too much time.

And: I was too stupid to select 2 submission models that would have been different to learn something. So, learning.

And stupid as well: I didn't figure out how the data set used to compute the final score differs from the one which was used during competition. Can anybody shed light on this?

Will clean up my code and then share.