Report for Week 10

I. Summary

1. I built textual information feature set and compared it with other Feature Sets
2. I performed Statistical Testing for our current results
3. I prepared and updated the repository containing all the work so far

II. Actions performed

1. Textual information

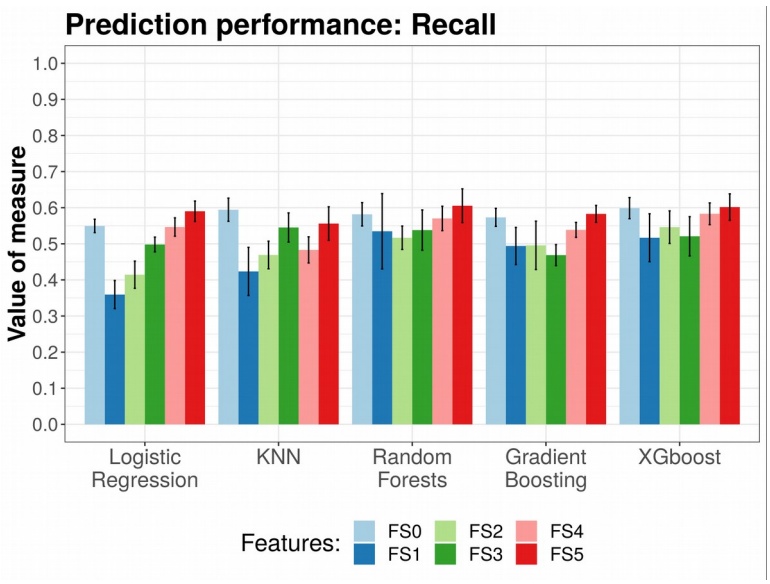
As discussed, I built a new feature set based on the textual information of the users' discourses. Thus, in order to obtain the new features, I detect the top 100 most frequent words (excepting stop words) and consider this to be my vocabulary.

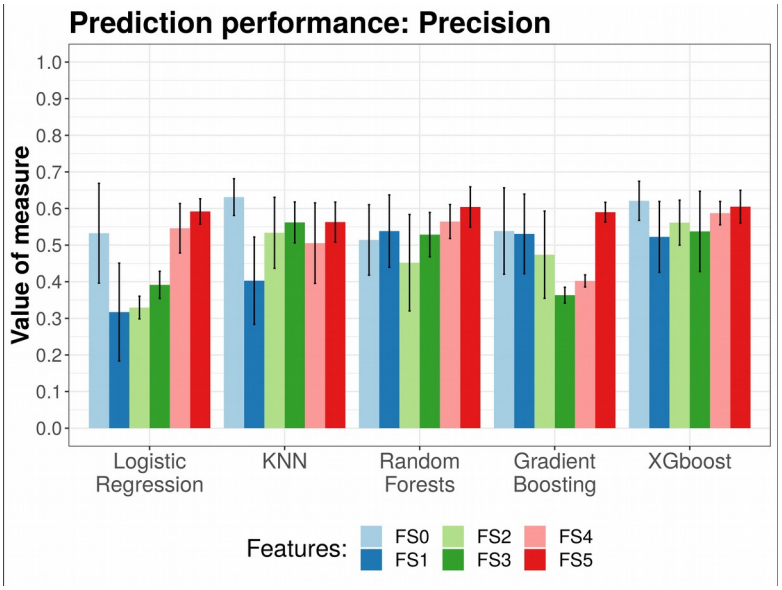
Then, for every time period, I aggregate the users' replies and build the discourse of each user (the aggregation of replies). Now I build a DTM, a Document Term Matrix, having as documents the users discourses and as terms the vocabulary formed earlier.

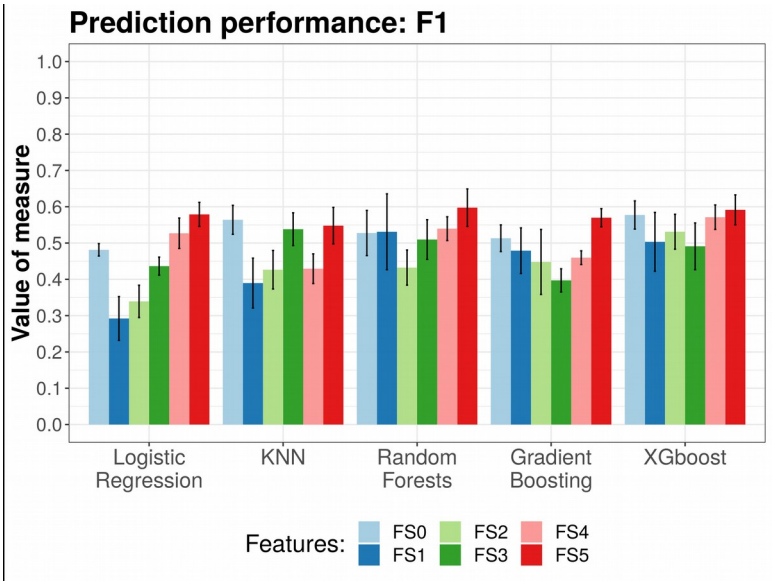
The weighting method applied is TF-IDF. Thus, now every user has 100 features corresponding to the tf-idf value of each vocabulary word present in his aggregated discourse. I also add the current political stance as a feature. The predicted / learned variable is the next political stance.

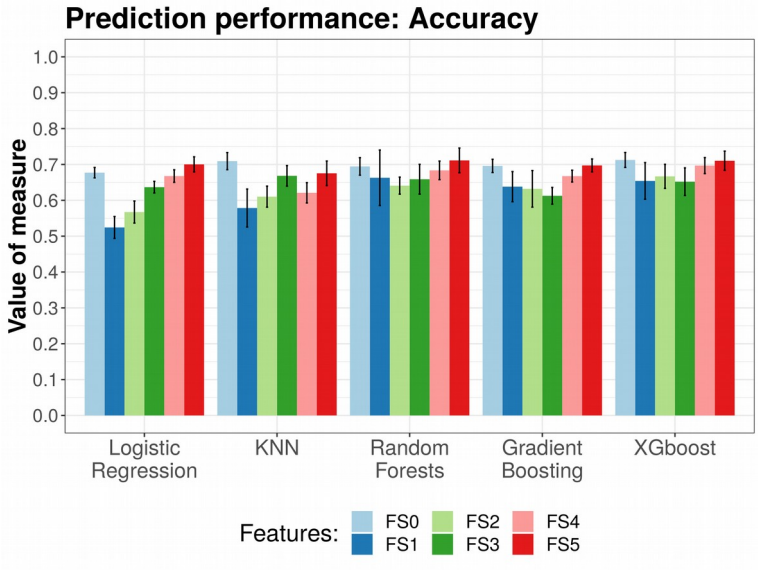
In the graphs below, the shortcuts mean:

Shortcut	**Description**
FS0	Textual Features
FS1	User activity (#posts, #initial posts, #replies per post)
FS2	User activity per group (same as FS1, just splitted per groups, ie #replies from A, B, N)
FS3	Structure of diffusions
FS4	Conversation Structure (FS1 + FS2 + FS3)
FS5	All = conversation structure + text (FS0 + FS4)


Prediction performance: Recall
Value of measure
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Logistic Regression
KNN
Random Forests
Gradient Boosting
XGboost
Features:
FS0
FS1
FS2
FS3
FS4
FS5

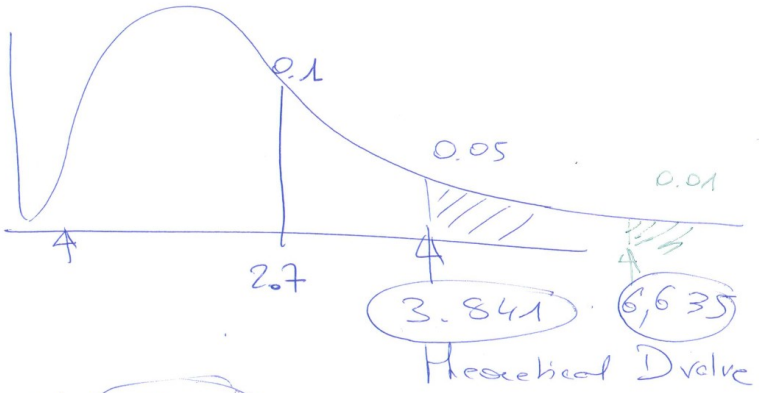

Prediction performance: Precision
Value of measure
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Logistic Regression
KNN
Random Forests
Gradient Boosting
XGboost
Features:
FS0
FS1
FS2
FS3
FS4
FS5

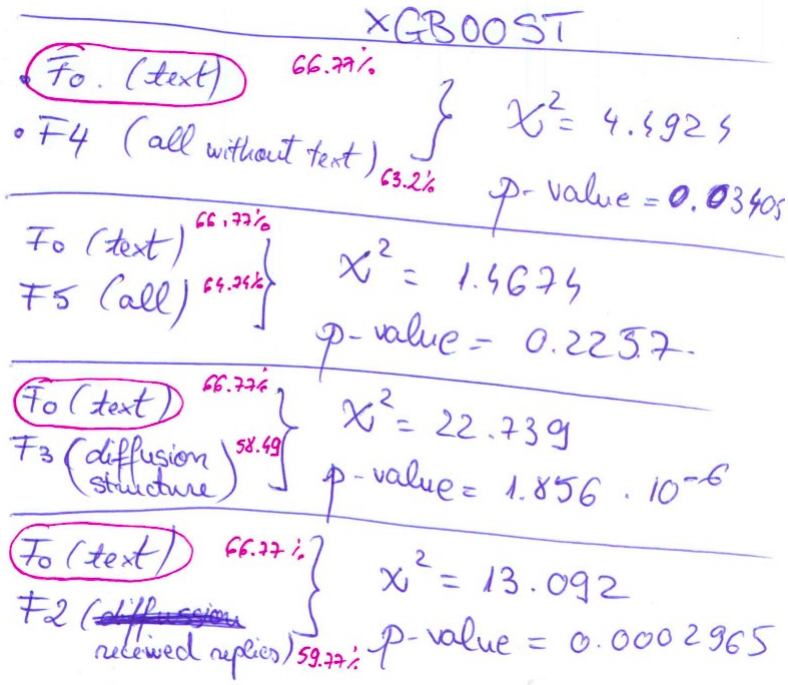

Prediction performance: F1
Value of measure
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Logistic Regression
KNN
Random Forests
Gradient Boosting
XGboost
Features:
FS0
FS1
FS2
FS3
FS4
FS5


Prediction performance: Accuracy
Value of measure
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Logistic Regression
KNN
Random Forests
Gradient Boosting
XGboost
Features:
FS0
FS1
FS2
FS3
FS4
FS5

2. Statistical Testing

I performed Chi Squared Test to determine if there is significant difference between the features sets when we add textual information. In the left figure, there is the probability density of the Chi Squared distribution, used to determine if the differences are significant or not.





The comparisons are made for XGBOOST between textual information features and the other feature sets. We can see that every time text info is used, there is a significant difference (thus a plus) and the accuracy is always better when we have text features. The only case when we don't have a significant improvement is at FS5 but that set already contains FS0.

3. Repository

I updated and completed the repository of this project:

https://github.com/andreimardale/InformationDiffusion.git

There I have the Final Documents, the Python scripts used for getting the reddit data and training the classifiers and ALL partial reports throughout this internship. In the README, there is a clean description of every code file and a link to the Data Sets.

Link to the Overleaf Report:

https://www.overleaf.com/4784255338pvmxffwqtpsz