Part 1

Report a table containing the accuracy of your classifier, precision, recall, F1, specificity, and AUC.

Note on metrics reported

- Precision and Recall are reported for the class with label True i.e. for successful requests.
- Specificity is the recall for unsuccessful requests.
- Classifier configuration used for all model results reported in this section

```
SVC(kernel='linear')
```

Table 1: Unigram and Bigram model

Accuracy	Precision	Recall	F1	Specificity	AUC
0.72	0.39	0.19	0.26	0.90	0.54

Table 2: Activity and Reputation model

Accuracy	Precision	Recall	F1	Specificity	AUC
1.00	1.00	0.99	1.00	1.00	1.00

Table 3: Narratives model

Accuracy	Precision	Recall	F1	Specificity	AUC
0.74	0.00	0.00	0.00	1.00	0.5

Table 4: Moral Foundations model

Accuracy	Precision	Recall	F1	Specificity	AUC
0.74	0.00	0.00	0.00	1.00	0.5

Part 2

Present a discussion of the performance of the above four models:

a) **Which of the four classifiers performed the best; which one performed the worst?** The classifier with the activity and reputation features performed the best. The worst classifier cannot be deduced from the above experiment as both model 3 and model 4 were not able to predict any successful request.

However, I further tried to build the model using the following configuration. class_weight parameter accounts for imbalance in the dataset with only $\sim 25\%$ of the posts being successful.

```
SVC(kernel='linear', class_weight='balanced')
```

Table 5: Narratives model with balanced class_weight

Accuracy	Precision	Recall	F1	Specificity	AUC
0.67	0.24	0.14	0.18	0.85	0.5

Table 6: Moral Foundations model with balanced class_weight

Accuracy	Precision	Recall	F1	Specificity	AUC
0.74	0.00	0.00	0.00	1.00	0.5

The above classifier gave some better results for model 3 but not for model 4. In the light of above results, model 4 can be said to have the worst performance.

b) **Describe your anticipated reasoning driving these differences in performance of the classifiers.**

After reading the paper, I expected model 1, 2 and 3 to give good results as they capture textual, social, temporal and narrative feature discussed in the paper. However, model 3 does not give good results. Some narratives do have a bigger impact on success however they cannot be used as a standalone model, but can be helpful in complimenting model 1 and 2. I wasn't sure how will model 4 perform but I expected some predictive power. This is not the case as can be observed from the results.

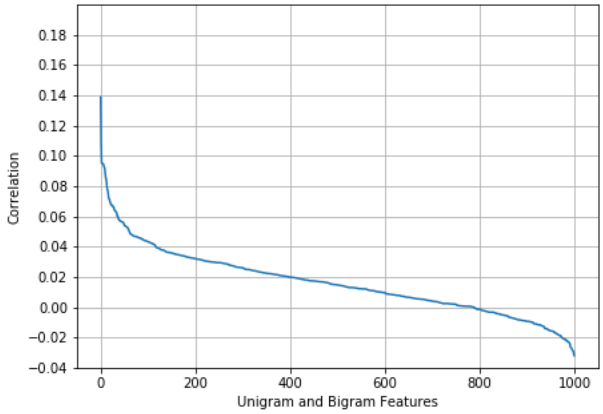
Model 2 outperformed my expectations. As I will show later in the assignment this is due the badge feature on Reddit.

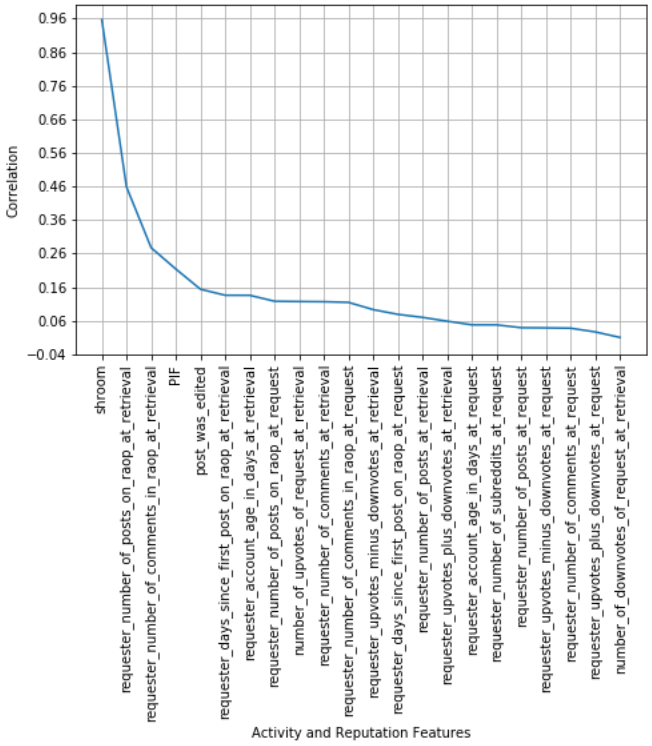
Model 1 does not perform as well as expected from the baselines mentioned in paper owing to pre-processing details not mentioned in the paper and use of a different model with strict parameters.

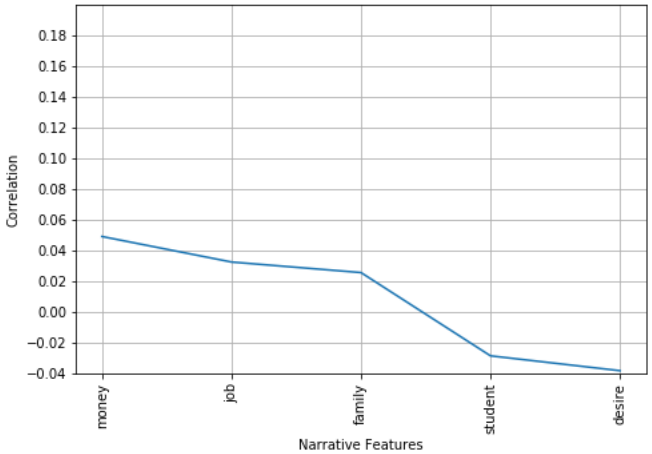
c) **For models 3 and 4 in particular, describe their performance compared to models 1 and 2. Why do you think they perform better or worse than models 1 and 2? Between models 3 and 4, which one is better? What could be the reason behind this observation?**

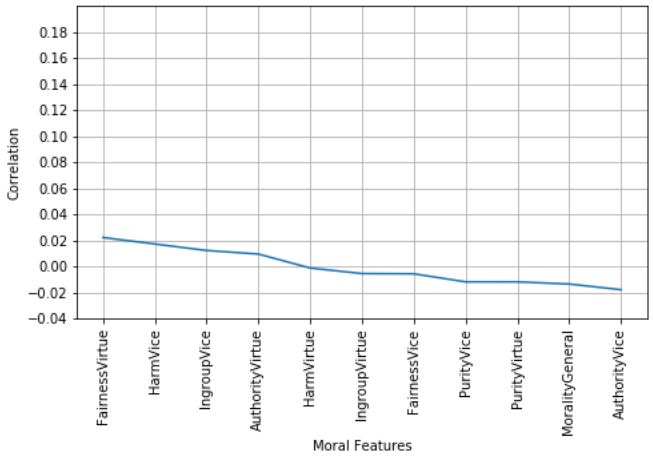
Models 3 and 4 have very low predictive power compared to models 1 and 2. For the default Linear SVM classifier models 3 and 4 are not able to predict successful requests at all giving a recall and precision of 0.

Quantitatively, the features of models 1 and 2 have a higher correlation with the outcome which can be observed from the following plots. Infact, the correlation exhibited by features is in order of model performance, i.e. model2 > model1 > model3 > model4. Also, models 3 and 4 have comparitively lesser number of features compared to 1 and 2 which can also account for the performance.


Correlation
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
0
200
400
600
800
1000
Unigram and Bigram Features


Correlation
0.96
0.86
0.76
0.66
0.56
0.46
0.36
0.26
0.16
0.06
−0.04
shroom
requester_number_of_posts_on_raop_at_retrieval
requester_number_of_comments_in_raop_at_retrieval
PIF
post_was_edited
requester_days_since_first_post_on_raop_at_retrieval
requester_account_age_in_days_at_retrieval
requester_number_of_posts_on_raop_at_request
number_of_upvotes_of_request_at_retrieval
requester_number_of_comments_at_retrieval
requester_number_of_comments_in_raop_at_request
requester_upvotes_minus_downvotes_at_retrieval
requester_days_since_first_post_on_raop_at_request
requester_number_of_posts_at_retrieval
requester_upvotes_plus_downvotes_at_retrieval
requester_account_age_in_days_at_request
requester_number_of_subreddits_at_request
requester_number_of_posts_at_request
requester_upvotes_minus_downvotes_at_request
requester_number_of_comments_at_request
requester_upvotes_plus_downvotes_at_request
number_of_downvotes_of_request_at_retrieval
Activity and Reputation Features


Correlation
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
money
job
family
student
desire
Narrative Features



Qualitatively, I think activity and reputation features are very evident, more trustworthy compared to other features which can account for the results of model 2. Model 1 captures language features in the form of bigrams and unigrams. It can capture social and textual features like reciprocity, politeness and evidentiality which are found to have an impact on the success of requests. While model 3 tries to capture the narratives using keywords, they serve more like categorization of a request and not as feature standalone with some narratives like 'money' yielding more successful requests. Regarding model 4, in general when people are presented with many indicators like activity, reputation, narrative and language, the moral dimension of the requester would not be explicit, noticeable and difficult to judge from short requests. Therefore, I think model 4 performs the poorest.

d) **Present your reasoning if your models indicate that language is able to predict success of altruistic requests – other than model 2, all of the other models rely on language.**

Model 1 and model 3 (with balanced class_weight) indicate that language can be used as a tool to predict success of altruistic requests. However, language based models are comparitively weaker than activity and reputation. This can be attributed to Reddit's features like flair, karma, upvotes and downvotes which are decided by the community and not by the individual requesting. To further test this idea, I dropped the feature 'requester_user_flair' from model 2. This actually led to degradation of model 2 indicating that badge feature on Reddit was driving its success. It shows the sheer impact of badges on the success of altruistic requests and language as a predictive tool in the absence of such apparent features.

Table 7: Activity and reputation model without feature requester_user_flair

Accuracy	Precision	Recall	F1	Specificity	AUC
0.26	0.25	0.99	0.41	0.00	0.5

Part 3

Presentation a comparative discussion of the performance of all of your classification models and the performance metrics (AUC) reported in Table 4 of [1]:

a) **In what ways are your models similar or different from those in Table 4 of [1]?**

Model	AUC
Most frequent unigrams and bigrams	0.54
Activity and Reputation	1.0
Narratives	0.5
Moral Foundations	0.5

Models 1, 2 and 3 are quite similar in terms of the features that they capture compared to models mentioned in Table 4.

Model 1 uses top 500 unigrams and top 500 bigrams as features and models in the paper also build unigram, bigram and trigram baseline. However, they measure their performance separately and do not combine them. Model 1 also implicitly captures textual features like politeness, evidentiality, reciprocity and sentiment owing to unigrams and bigrams.

Model 2 captures the temporal, status and similarity features discussed in the paper in the form of upvotes, number of days, subreddits posted.

Model 3 captures the narratives discussed in the paper, limiting to 5 narratives only.

Model 4 is different from all the models used in paper. It tries to capture the moral dimension of the requester using language and using it to predict success. It tries to capture the innate characteristics of an individual unlike the social and textual which can be acquired and manipulated.

b) **Where and why do they perform better or worse compared to [1]?**

Model 2 performs the best among all models built in this assignment and the paper. This as observed above can be attributed to the flair feature on Reddit which measures the 'status' of a requester.

Model 1 though composed of unigrams and bigrams does not yield as good results as the ngram baselines in paper. This can be attributed to the fact that we are using a different classifier with default configuration. Indeed the performance betters on using Logistic Regression, however I could not achieve the values mentioned in paper owing to preprocessing of features.

Model 3 gives some results on accounting for the class imbalance. The paper does not use the narratives in its final model. It only shows the impact of different narratives on the success of a request. We can observe from the correlation curve that the correlation mirrors the impact of different narratives on success mentioned in the paper.

Model 4, does not give any results on the dataset. The features used in this model also have a very low correlation with the success. This indicates that these features are not a good measure of prediction in this context.