Online Supplementary Information

Descriptive statistical model

Features

The original features engineered were:

- Response elaboration and complexity:
    1. Number of words: this measures the length of the response. Longer responses are in general more elaborate, and in harmony with our previous findings responses written to purportedly non-Roma clients tend to be longer.
    2. Variability of punctuation: the number of different non-alphanumeric characters in each message. The larger this number is, the more complex is the response.
    3. Moving average type-token ratio (MATTR): This feature measures the lexical diversity of the responses. We used a moving average to counteract the great variation in the length of the responses.
    4. Average length of sentences: This is a feature to measure the linguistic complexity of the responses. More sophisticated texts generally have longer sentences.
    5. Average length of words: This is also a complexity measure of the responses.
    6. Frequency of punctuation: The number of non-alphanumeric characters divided by the number of words.
- Information 'density' and structure of the responses.
    1. Ratio of stop words
    2. Ratio of address tokens
    3. Ratio of date tokens
    4. Ratio of email address tokens
    5. Ratio of time tokens
    6. Ratio of URL tokens
    7. Ratio of name tokens
    8. Ratio of number tokens
    9. Ratio of telephone number tokens
    10. Ratio of target town name tokens (name of the town or village to which the original email was addressed)
    11. Ratio of other town name tokens
- The linguistic structure of the responses:
    1. Ratio of punctuation
    2. Ratio of particles
    3. Ratio of verbs
    4. Ratio of proper nouns

5. Ratio of pronouns
6. Ratio of determiners
7. Ratio of adpositions
8. Ratio of conjunctions
9. Ratio of nouns
10. Ratio of auxiliaries
11. Ratio of adverbs
12. Ratio of adjectives

Feature selection

We used a two-phase feature selection method: the first one was faster, and the second one was more precise. For the first phase, we separated 200 random responses as a development set from the 1130-response train set. Then we built XGBoost models with two fixed hyperparameter sets. We used two hyperparameter sets to avoid optimising to only one point in the hyperparameter space. The two hyperparameter sets were (hyperparameters not mentioned were set to default value):

- reg_alpha=0.2, n_estimators=50
- subsample=0.9, colsample_bytree=0.9, max_depth=6, reg_alpha=0.3

In each step we built multiple models with different feature sets: using all remaining features, and leaving out 1, 2, 3, and 4 features from the remaining set. Next we removed the feature that was most often left out from the best performing models from the feature set. We repeated these steps until the best-performing model was not one of the smallest ones trained. We started the next phase of feature selection with the feature set of the best-performing model from the first phase. In this phase we used a two-layer cross-validation: We created a fixed four-fold validation method on the whole 1130-response train set. On each of these folds, we trained multiple models with differently sized feature sets (keeping all variables, leaving out one, and leaving out two) with a four-fold cross-validation to select the hyperparameters of the models. We then used the average performance of these models on for each left-out data to estimate how good each feature set is, and selected the best performing one for the next round. We repeated this until the best-performing feature set was the largest one.

Hyperparameters

To build the final model we used a 5-fold cross-validation to find the best hyperparameter set from a 1296-point grid, and refitted the model with these hyperparameters on the whole training set (for more details on learning models and hyperparameter tuning, see e.g. Eisenstein, 2019).

- colsample_bynode= 0.9,
- colsample_bytree= 1,
- learning_rate= 0.1,
- max_depth= 6,

- min_child_weight= 1,
- n_estimators= 50,
- reg_alpha= 0.3,
- subsample= 0.8

N-gram model

Hyperparameters

To find the best hyperparameters we used a 4-fold cross-validation search on the whole train set on a 93312-point hyperparameter grid. As a result, we found that the best hyperparameters are:

- Preprocessing:
    - Lowercased responses with stop words removed
- Vectorizer:
    - Number of most frequent features per corpus: 400
    - Ratio of most frequent features on the other subcorpus to discard non-distinctive features: 1
    - Token pattern: any alphanumeric token or punctuation
    - N-gram range: uni- and bigrams
    - Minimal document frequency: 0.005
    - Maximum document frequency: 0.99
- Booster:
    - Subsample: 0.9
    - Column sample by tree: 0.8
    - Column sample by node: 1
    - Maximum depth: 5
    - Regularization alpha: 0.1
    - Number of estimators: 50

Further results

Table 2 shows the results of the ablation study.

Dropped variable(s)	**Test accuracy**	**Test F1**	**Mean CV Acc**	**Mean CV F1**
ratio of pronouns	0.585	0.551	0.577	0.571

-	0.58	0.517	0.581	0.574
ratio of verbs	0.58	0.543	0.571	0.562
ratio of proper nouns	0.555	0.508	0.568	0.565
ratio of punctuation	0.55	0.521	0.581	0.577
ratio of determiners	0.545	0.48	0.571	0.565
variation in punctuation	0.54	0.5	0.577	0.572
linguistic structure	0.53	0.397	0.565	0.544
Number of words	0.525	0.438	0.575	0.57
ratio of adpositions	0.525	0.475	0.582	0.577
ratio of number tokens	0.515	0.464	0.583	0.581
complexity and information structure	0.48	0.381	0.575	0.565

Table 2: Performance of models with variables dropped

	overall	men	women	smaller settlements	larger settlements
descriptive stats	58%	62%	55%	60%	56%
n-gram	58.5%	68%	50%	66%	51%

Table 3: Accuracy of models by background variables

Figure 5 and 6: on the left, the markers tend to be at the bottom, and the crosses also tend to be at the bottom: discrimination is more easily detectable in emails sent to smaller towns and to men.

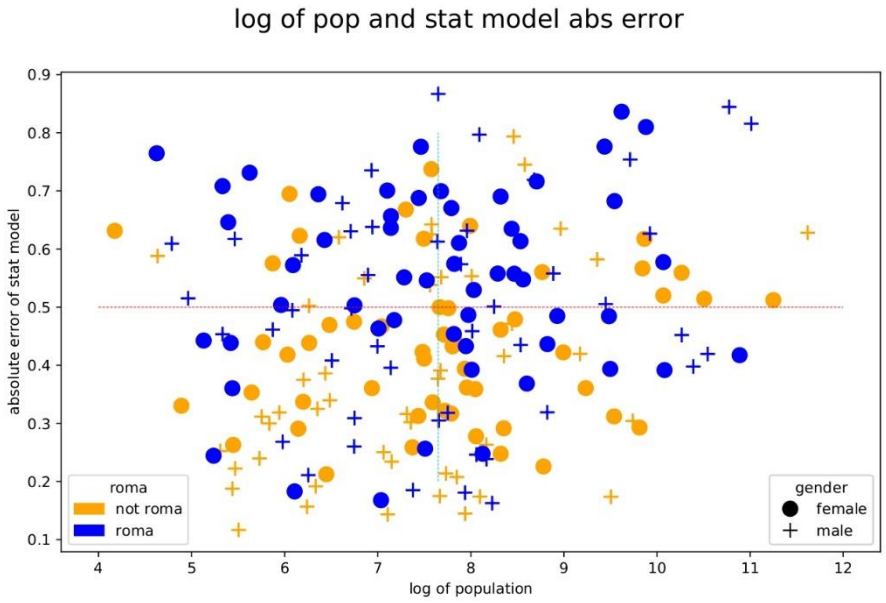

Figure 5: The absolute error of the descriptive statistics based model against the logarithm of the population of the town. The shape of the marker represents the gender and its colour represents the ethnicity of the addressee.

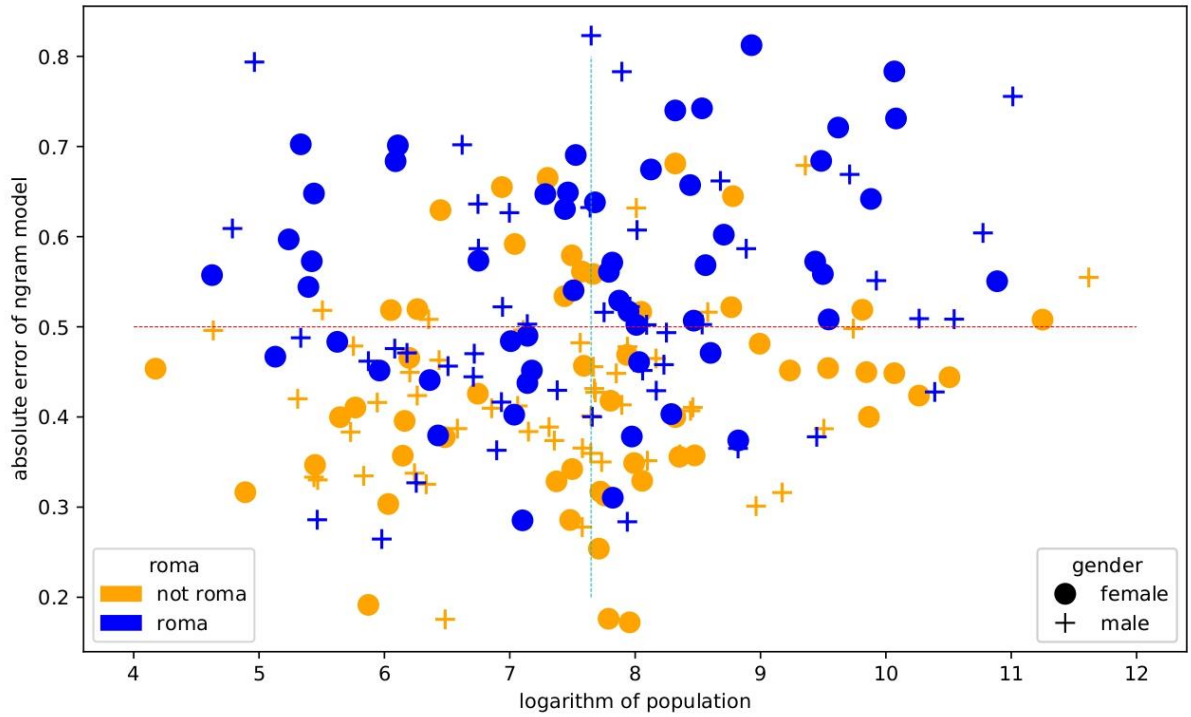

Figure 6: The absolute error of the n-gram model against the logarithm of the population of the town. The shape of the marker represents the gender and its colour represents the ethnicity of the addressee.