Finding *Political* PAMPHLET POETRY.

by ECCO_Poetry, #DHH22

Background

For ordinary British readers in the eighteenth century, poetry would have often been encountered in the form of cheap, short pamphlets. Much of this was topical and political – the continuation of a tradition which had been popular in the previous century. With a large corpus of digitised documents, we set out to use computational methods to find and understand these political, poetical texts.

Our knowledge of the extent of text in this form across the corpus is limited. While until now we have had to rely on clues found in titles and metadata labels, developing a robust method for classification of poetry, particularly at the page level, is a valuable task we set out to tackle.

Research questions

We set out to investigate:

1. Which computational methods can best distinguish poetry from prose, specifically when faced with noisy OCR data?
2. What do page images tell about poetic features?
3. Which methods are applicable for extraction of textual or poetic features?
4. How prevalent were particular themes (in this case, political discourse) across the eighteenth century?

Data

We worked with three datasets:

- ECCO
  - Full texts for 200,000 documents of eighteenth century Britain. OCR'd texts and images available. Can be linked to the ESTC.
- ECPA & its Metadata
  - Smaller hand-curated and clean data set of eighteenth century poetry.
- ESTC
  - Metadata for early modern print products, half a million records in total. Linkable to ECCO and the ECPA.

OCR quality

The text of ECCO has been extracted using Optical Character Recognition (OCR) on digitised images. Because of this, the quality varies, particularly with more complicated layouts such as the two-column broadsheet below.

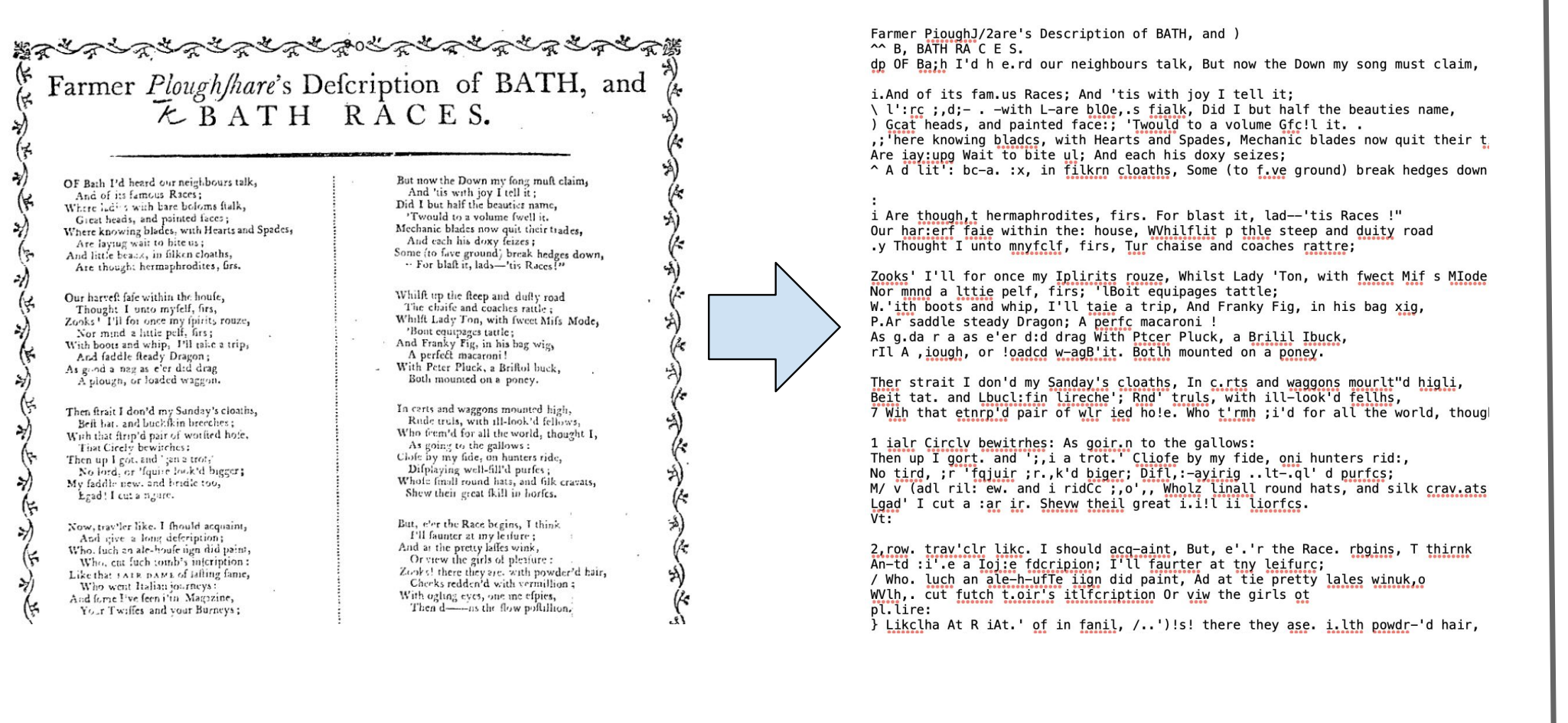


Figure 1. A sample image page and resulting OCR output.

How can we detect poetry?

The first step was to differentiate between poetry and prose. To do this we annotated a training set (see below) and identified three viable approaches. Poetry can be distinguished by **structural**, **lexical** and **visual** features, or even a **combination of several** of these. To test these assumptions, we designed a series of classifiers with the aim of comparing results and performance: a **text-based** approach and an **image-based** approach.

Results

- **On text**: We chose a random forest model using both Bag-of-words and linguistic features (roughly speaking these were rhyme, lines beginning with capital letters, characters per line, and number of empty lines). This method achieved ~90% precision, recall, and F1-measure on the test set.

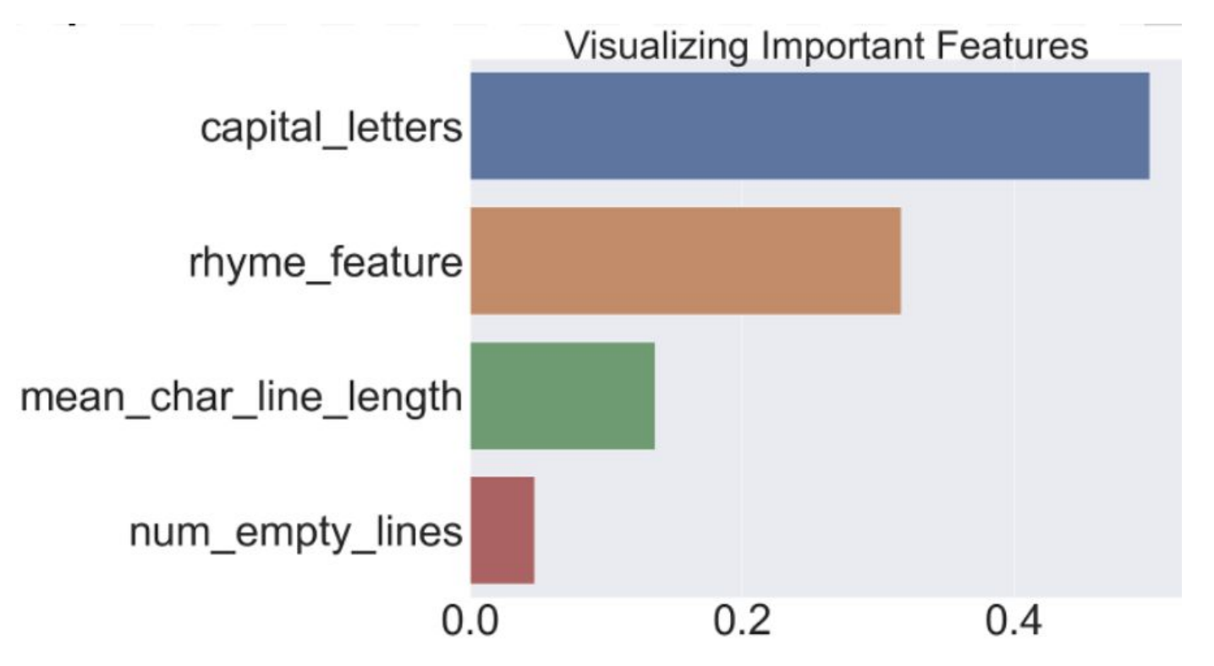


Figure 2. The Gini importance of features for the random forest model. It matters most whether lines begin with capital letters.

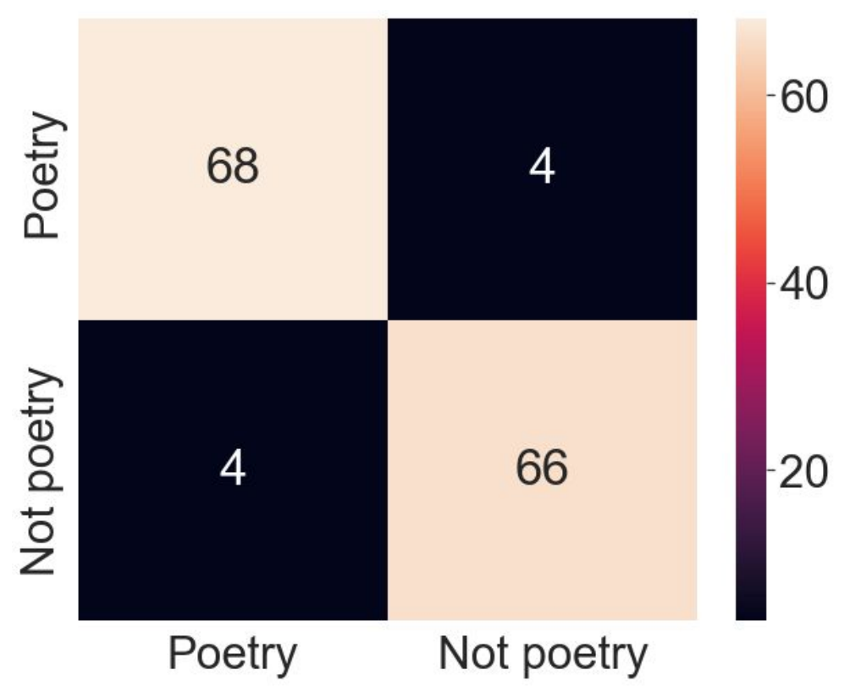


Figure 3. Confusion Matrix for the random forest model using linguistic features.

- **On images:** Additionally, we trained an image classifier built on the ResNet architecture. The output of the confusion matrix below shows that it is possible to detect poetry using a visual representation, although the model using this method achieves lower accuracy than text-based classifiers.

Class	Precision	Recall	F1 Score
Poetry	0.94	0.88	0.91
Not poetry	0.87	0.93	0.90

Table 1: ResNet classifier performance on val data.

Class	Precision	Recall	F1 Score
Poetry	0.86	0.86	0.86
Not poetry	0.86	0.86	0.86

Table 2: ResNet classifier performance on test data.

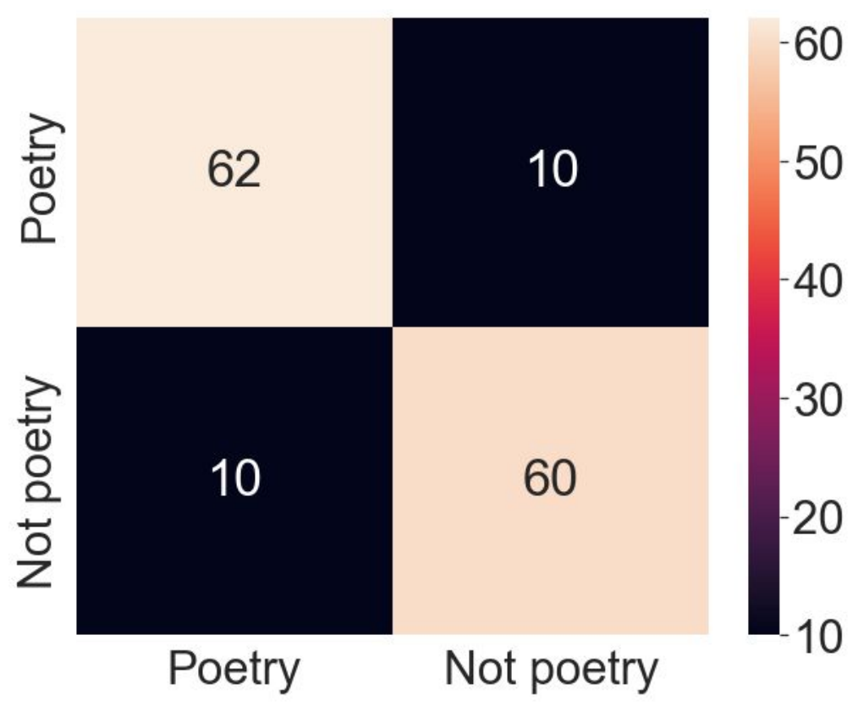
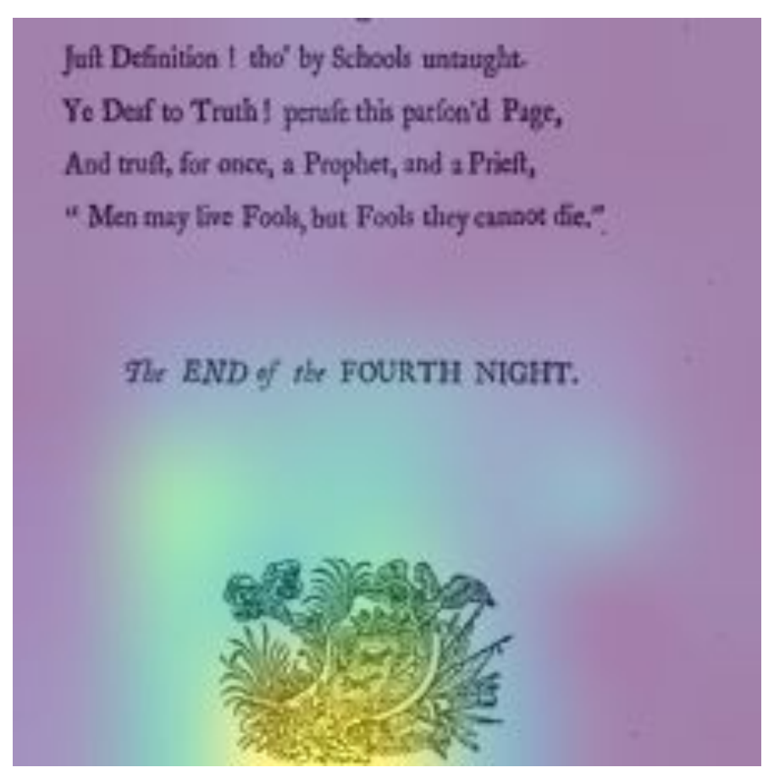


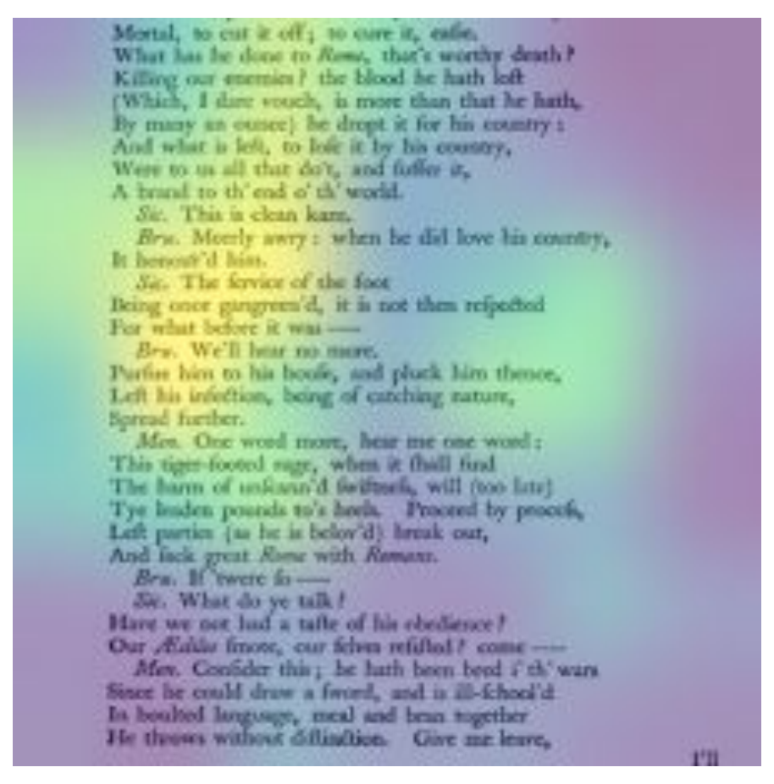
Figure 4. Confusion Matrix for the ResNet classifier model.

- One key contribution comes from the analysis of misclassified examples from the test set. This allows us to understand the visual patterns on which the convolutional neural network makes decisions. Warm colours indicate a higher correlation.



Predicted class: not poetry (99%)
Ground truth: poetry

(Possible explanation: due to the higher representation of images that contain visual content, the model predicts new images wit a similar distribution as the not poetry class. In the future it would be better to annotate by highlighting poetry pieces rather than on a page level.)



Predicted class: poetry (85%)
Ground truth: not poetry

(Possible explanation: the model follows a pattern of text that is seemingly similar to the shape often found in the representation of the poetry class. A similar structure can be observed in, for example, printed dramatic works.

Content Analysis

- Further text analysis was carried out on the 3,200 poems of the *Eighteenth Century Poetry Archive*. Fig. 5 informs us on the "canonical" metrical forms implemented most frequently by the poets. Though we hoped to apply these methods to our 'found' poetry, this will have to be saved for a future project.

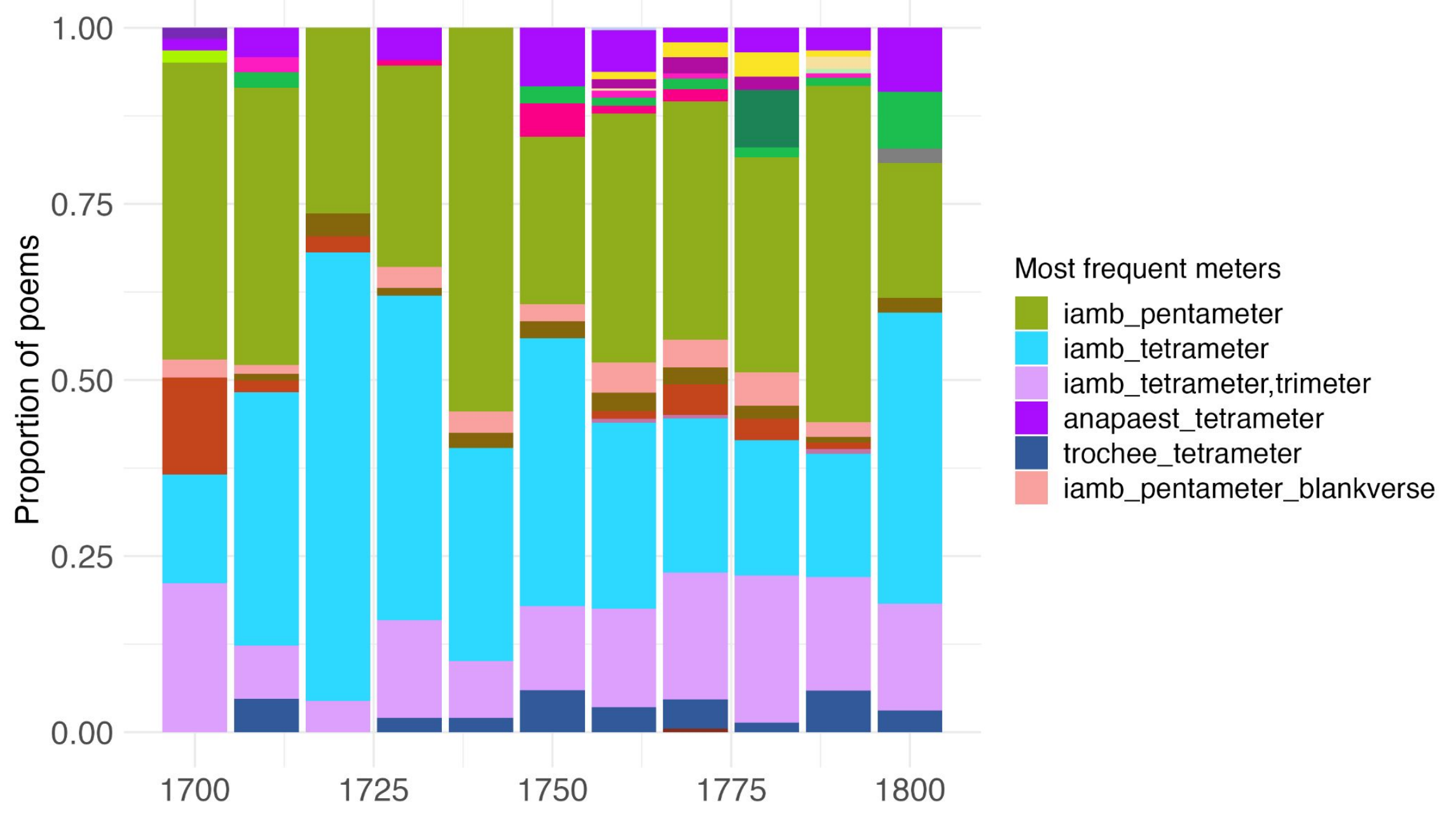


Figure 5: content analysis of the ECPA poetry, showing the proportion of most frequent meters used by decade.

Examples

Iambic: unstressed — stressed pattern
Iambic pentameter (5 stresses)

Then h**o**me, with sh**a**king l**i**mbs and qu**i**ckened br**ea**th,
His f**oo**tsteps **u**rges fr**o**m the pl**a**ce of d**ea**th.

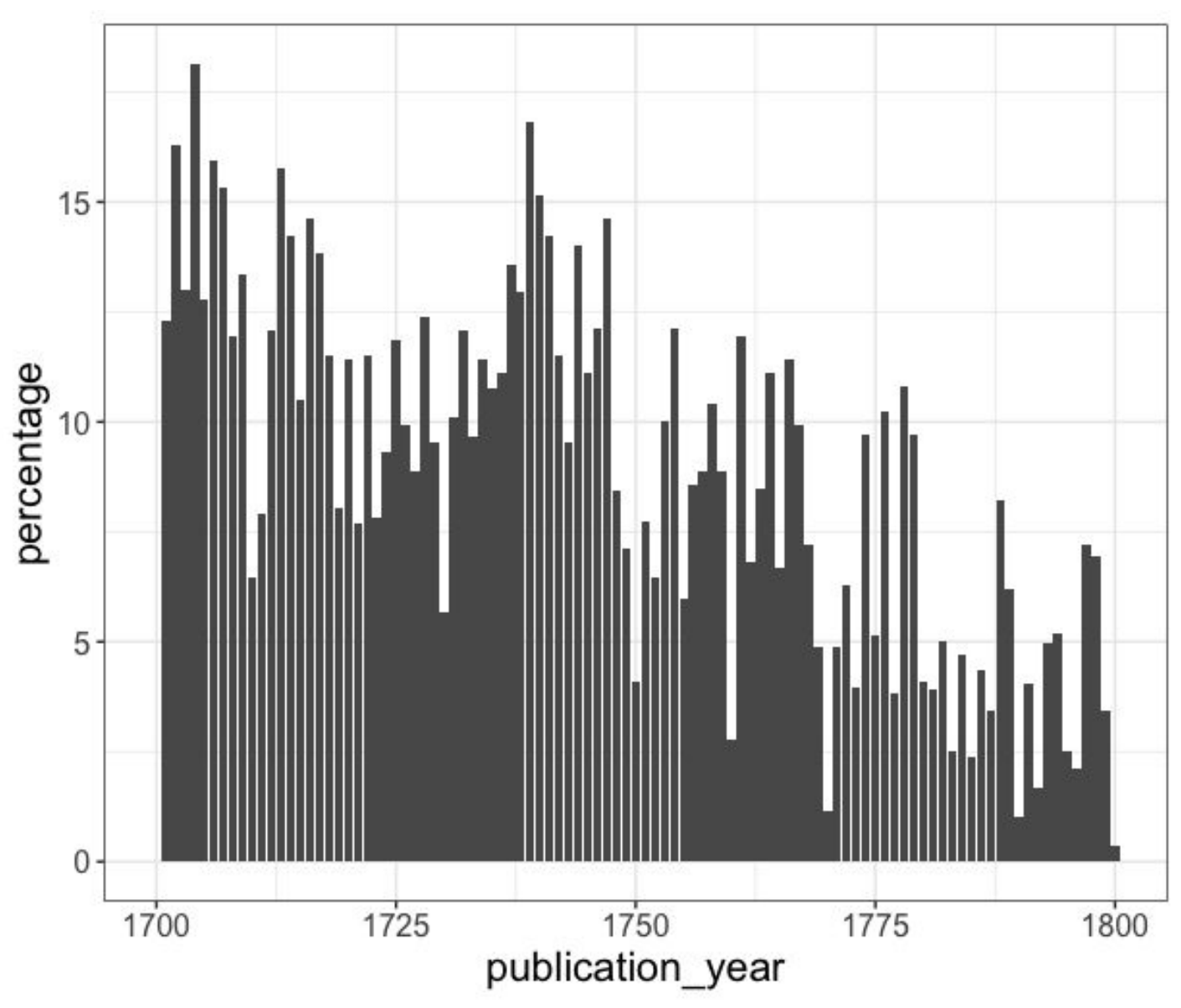
Iambic tetrameter and trimeter (4 and 3 stresses)

When r**e**cent **i**n the w**o**mb I l**ay**,
Ere **ye**t my l**i**fe beg**a**n,
Thy c**a**re pres**e**rv'd the sl**ee**ping cl**ay**,
And f**o**rm'd it **i**nto m**a**n.

Trochee: stressed — unstressed pattern

Sh**e**pherd! s**ee**k not w**ea**lth nor p**o**wer,
L**e**t the v**e**rdant w**oo**dbine b**o**wer

What about political poetry?



britannia
⋮
pow
⋮
war
⋮
state
⋮
britain
⋮
triumphs
⋮
pride

This graph represents the proportion (out of the poetry pamphlets detected in the previous step) of documents with a high probability of two topics relating to war or political discourse (see highly probable words in these topics in the table). The method used was a Contextualized Topic Model. The graph suggest that its proportion is significant but decreasing over time.

Discussion

- A larger set of training images would be desirable: because of the variety of texts and layouts, our annotated data only contained a proportion of the possible examples of the classes.
- We wish to investigate combining the image and text models, and evaluate the results.
- Ultimately, we hope these methods will allow us to extract most of the text in the poetic form found in the ECCO corpus, which will then be available for analysis.

Annotation & Labelling

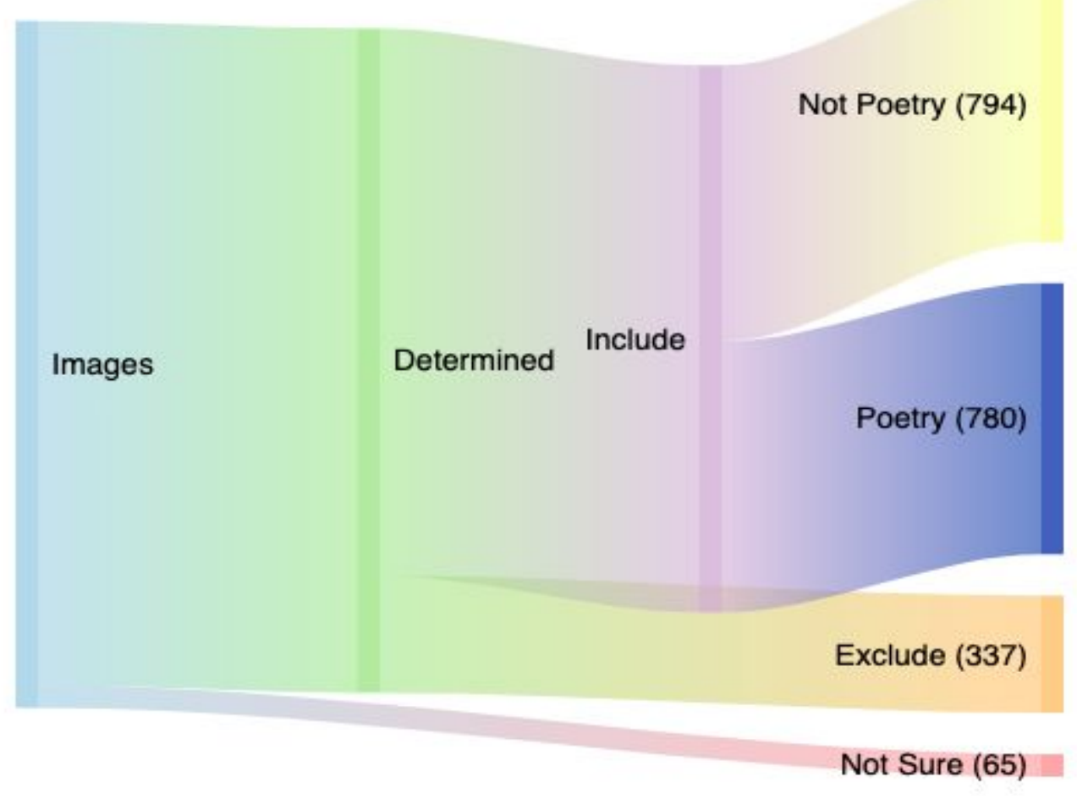


Figure 5. Annotation workflow of the first group.

To detect poetry among the digitised pages, four classes of labelling was applied to 2,000 page images for the creation of a clean dataset. The categories were: *Poetry*, *Not-poetry*, *Not sure*, and *Exclude*.

The annotation task was carried out by two groups simultaneously. We found that by focusing on a narrow but clear definition of what consisted of a 'page of poetry', we ended up with high agreement between annotators.

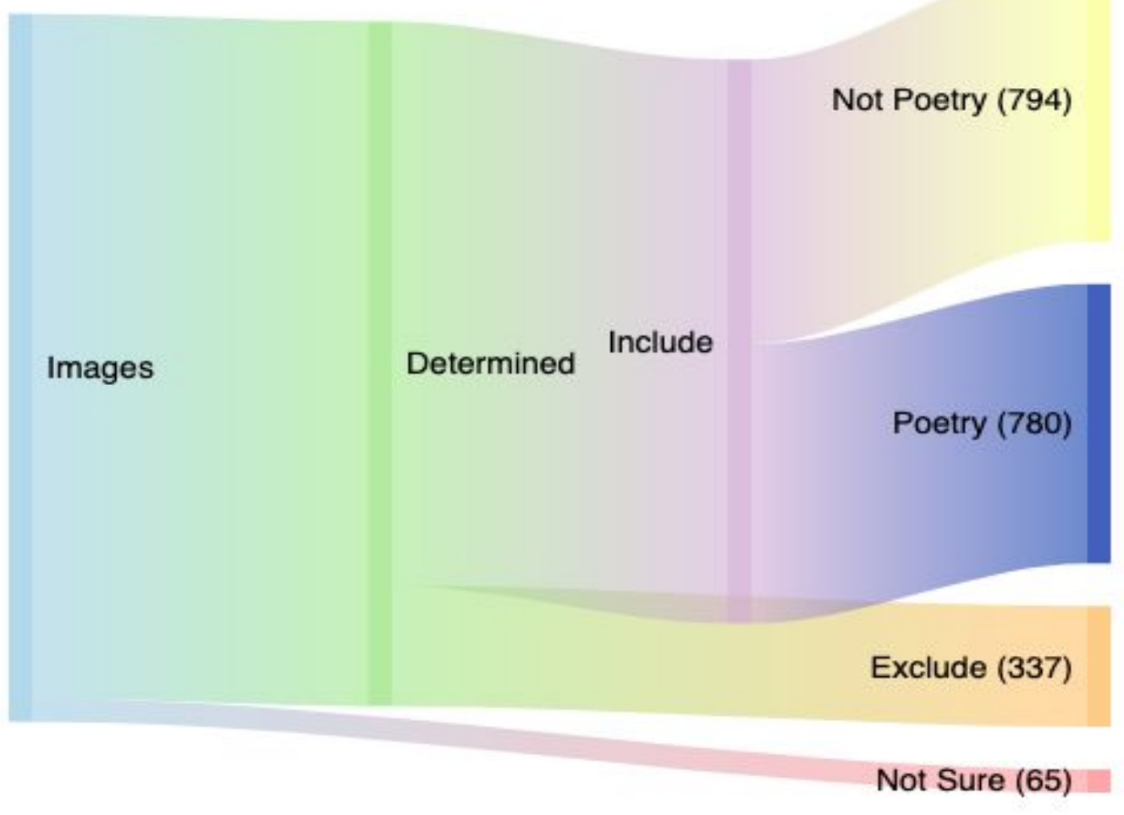


Figure 6. Annotation workflow of the second group.

Group Members: Antonina Martynenko, Daniel Kvak, David Rosson, Khanim Garayeva, Lassi Saario, Varvara Arzt, Yu An.
Group Leaders: Iiro Tiihonen, Lidia Pivovarova, Yann Ryan.