University of Helsinki
Department of Linguistics
Language Technology

Master's thesis

Document classification based on library catalogue metadata

Hege Roivainen
011511805

Supervisors:
Leo Lahti
Jörg Tiedemann
Mikko Tolonen

December 4, 2017

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto – Fakultet/Sektion – Faculty	Laitos – Institution – Department	
Humanistinen	Nykykielten laitos	
Tekijä – Författare – Author		
Hege Roivainen		
Työn nimi – Arbetets titel – Title		
Document classification based on library catalogue metadata		
Oppiaine – Läroämne – Subject		
Kieliteknologia		
Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä– Sidoantal – Number of pages
Pro gradu	11 / 2017	88

Tiivistelmä – Referat – Abstract

Kansalliskirjastojen metadataluettelot ovat hyviä informaatiolähteitä, sillä ne sisältävät tiedon lähes kaikesta tiettynä aikana ja tietyllä alueella julkaistusta aineistosta. Yleensä ne ovat kattavasti kuvailtuja, joten niitä voi käyttää kvantitatiivisen tutkimuksen lähteinä. Usein tutkimusta tehtäessä tutkimusaineisto kannattaa jakaa pienempiin osiin esimerkiksi genren perusteella. Monissa tapauksissa aineiston aukkoisuus kuitenkin vähentää aineiston käytettävyyttä. Tämä pro gradu -työ arvioi mahdollisuutta hyödyntää koneoppimista etsittäessä tutkimukselle relevantteja osajoukkoja kirjastoluetteloista. Esimerkkitapaukseksi valitsin English Short Title Cataloguen (ESTC) ja etsittäväksi osajoukoksi runokirjat. Runokirjojen genretiedon kuuluisi olla annotoitu, mutta todellisista kirjastoluetteloista tämä tieto usein puuttuu.

Käytin random forest -algoritmiä perinteisillä tekijän tunnistuksessa ja genreluokittelussa käytetyillä erityyppisillä piirrevektoreilla sekä metadatakenttien arvoilla parhaan tuloksen saamiseksi. Koska kirjastoluettelot eivät sisällä kirjojen koko tekstiä, piirteiden valinta keskittyi otsikoissa käytettyihin sanoihin ja lingvistisiin ominaisuuksiin. Otsikot ovat yleensä lyhyitä ja sisältävät hyvin vähän informaatiota, minkä vuoksi yhdistin piirrevektoreiden parhaiten toimivat piirteet yhteen ja tein lopullisen haun niillä. Tutkimuksen päätulos oli varmistus siitä, että otsikoiden käyttö piirteiden muodostamisessa on käyttökelpoinen strategia. Tutkimus avaa mahdollisuuksia määrittää osajoukkoja tulevaisuudessa koneoppimisen keinoin ja lisätä kirjastoluetteloiden hyödyntämistä kvantitatiivisessa tutkimuksessa.

Avainsanat – Nyckelord – Keywords
random forest, machine learning, genre classification, library catalogues

Säilytyspaikka – Förvaringställe – Where deposited
Keskustakampuksen kirjasto

Muita tietoja – Övriga uppgifter – Additional information

Contents

1 Introduction

Changes in printing reflect historical turning points: what has been printed, when, where and by whom are all derivatives of contemporary events and situations. Excessive need for war propaganda brings out more pamphlets from the printing presses, the university towns produce dissertations, which scientific development can be deduced from and strict oppression and censorship might allow only religious publications by government-approved publishers. History of printing is also important per se: for example, the invention of printing press enabled scientific revolution (Johns, 2001), likewise the use of Lucas Cranach's woodprints and framed title pages in the Martin Luther's books helped spreading the Reformation (Pettegree, 2016) and the shift from folio-sized books to octavos affected the everyday life of the literate bourgeois due to the ease of portability of the books, leading to introspective reading rather than reading aloud from the pulpits (Tolonen et al., FORTHCOMING). The history of printing has been extensively studied and numerous monographies do exist. However, most of the research has been qualitative studies based on close reading requiring a profound knowledge of the subject matter, yet still being unable to verify the extent of the new innovations. Close reading the library catalogues does not reveal, at least easily, the timeline of when the books, that Luther had published were printed, or how big a portion of books actually were octavo-sized and when the actual breakthrough happened.

One of the sources for these kinds of studies are national library metadata catalogues containing information about the physical book sizes, page counts, publishers, publication places and so forth. These catalogues have been researched for some time for quantitative analysis as well. The advantage of national library catalogues is, that they often are more or less complete, having records of practically everything published in a certain country or linguistic area in a certain time period. The computational approach to them has enabled researchers to connect historical turning points to the effect on printing and the impact of a new concept has been measured from the amount of republications and spreading of the book where the new idea was first introduced (Lahti et al., 2015; Tolonen et al., 2016, FORTHCOMING). Linking the library metadata to full texts of the books have made it possible to analyze the change in the usage of words in a massive corpus, but still limited to books only from the relevant field (Kanner et al., 2017). Computational methods work better, the more complete the corpora is. However, library catalogues often lack annotations for one reason or another: annotating resources might have been cut down at a certain point in time or the

annotation rules may have varied between different libraries in cases when the catalogues have been pooled together, or the rules could have just changed. The annotation could also be simply erroneous (Karian, 2011).

Genre information is especially important in defining subcorpora for research. The genre field, when annotated for each of the metadata records, could be used to restrict the corpus to contain every one of the books that are needed and nothing more. From this subset there is a possibility of drawing timelines or graphs based on bibliographic metadata, or in the case of full texts existing, the language or contents of a complete corpus could be analysed (Kanner et al., 2017).

Despite the significance of the genre information, that particular annotation bit is often lacking. A glance at the English Short-Title Catalogue (ESTC) shows, that the genre information exists for approximately one fourth of the records. This should be enough for teaching a model for machine learning and trying to deduce the genre information, rather than relying solely on the annotations of librarians. The metadata field containing genre information in ESTC can contain more than one value. In most cases this means having a category and its subcategories as different values, but not always. Because of the complex definition of genre in ESTC I decided to focus on one genre only: poetry. Besides being a relatively common genre, poetry is also of interest to literary researchers. Having a nearly complete subset of English poetry would allow for large-scale quantitative poetry analysis. Also, the ability to extend the methodology to include other genres, other fields and possibly other catalogues, would be a huge benefit for the book history project I am part of.

The main difficulties in applying machine learning techniques to the ESTC metadata lies in the absence of the complete unabridged texts. Almost all of the research on genre classification or author attribution is based on full texts. In library catalogues, the actual content is missing, and only the metadata exists. Metadata does contain the titles, which is textual and usable in a similar fashion to the full texts, but the amount of texts per record is merely a tiny fraction to that of a complete text: as machine learning is based on statistics, this might pose a problem. The titles themselves can be used as material for distant reading, as shown by Moretti (2009) in his article describing the style changes in the titles of English novels and Elliott (2014) analyzing the titles of romance novels. On the other hand, the metadata also contains additional information, such as topic and information about the physical book, typed by the librarians, which a full text edition does not.

The assumption behind this task in the first place is, that there does exist

some indicators on the title level or in the book format, that the content actually is poetry. I would assume, that an occasional buyer would have been able to interpret a book as a poetry book somehow, even without reading the contents: the author name could have been familiar, the small book size and page count, some keywords in the title or perhaps being grouped together with other poetry books.

There exists full text catalogues as well: Eighteenth Century Collections Online (ECCO) is even linkable to ESTC. ECCO, however does not contain the entirety of ESTC records and the selection criteria is unknown. Thus a quantitative analysis can not be reliably performed on ECCO. There is an on-going project on combining the two catalogues (Kanner et al., 2017).

I tackled the shortcoming of the missing full texts by creating several models each packed with similar features. The feature types ranged from part-of-speech tags and their trigrams to word frequencies from the book titles, which all are more or less typical in genre classification tasks. Additionally I tried some seldom used or arduous features, such as syntactical components, other catalogue fields and common proper nouns from the Antique literature. The Antique names I eventually had to discard completely as useless. From these feature sets I hand-picked the best performing features into one superset, which I then ran on the unannotated records.

My results were very promising: I found over 13,000 poetry books from the unannotated part of the corpus, with a precision of 95.0%. This provides a good basis for statistical analysis on a restricted subset of the corpus.

The extensibility of my methodology to include other genres or other fields of ESTC is quite good, provided, that there exists enough data for the desired genre or field. It will require manual supervision, but many process steps can be automated: quite a few can be omitted completely. The same applies to other metadata catalogues, but the amount of manual guidance is likely to be higher, because the catalogues differ in their field contents and levels and qualities of annotation.

About this thesis

This thesis belongs to the realms of language technology and digital humanism. The main methodological contribution will be the implementation of machine learning algorithm, namely random random forest, into library metadata catalogues. The methodology itself is familiar from earlier studies on author attribution and genre classification, but the material it is applied

on is significantly different from the previous research.

I had three separate main steps involving machine learning runs:

1. *qualification rounds*: the selection of the most predictive features from the various feature sets
2. the *semifinals*: validation of the selection of the feature sets and additional testing on different definitions of poetry
3. the *finals*: the search on the part without the genre annotated using the selected features.

The feature sets are essentially the same for the semifinal and final rounds.

My thesis consists of six chapters. The first one is this introduction, which establishes the context and the reason for this research.

In the second chapter I will describe earlier research. I focus mostly on the methodological aspects, describing relevant and interesting ways of extracting features for machine learning, but I will touch on the research made on library catalogues as well. I proceed to describe the data: the representation of MARC fields in ESTC is followed by the problematics of defining the poetry.

The third chapter is dedicated to the computational and more technical aspect of the methodology: it is highly descriptive with a low level of specificity to the classification task at hand. First I will describe the main learning method of my thesis: random forests. From that I will continue to the generic methodological principles I have applied to. After that I go through the measures comparing learning models, and finally I begin talking about features and feature selection methods.

In the fourth chapter I will talk about the test runs I made. The first subchapter is about the base features, which will be included in the final features too. The second subchapter contains a detailed description of the *qualification rounds*: the different types of feature sets I handle mostly in their own subsubchapters. After the qualification rounds come the *semifinals*. I explain, what kind of effect various parameters have on the results, if any. In addition I try different ways to define poetry, and see if there is any consequences.

The fifth chapter, results, is the *finals*. I will elaborate on what kind of results I got from the run on the records having unannotated genres and discuss the performance on different feature types.

In the sixth, and last, chapter I extend the discussion further. What kind of observations can be drawn and how well is this methodology adjustable to other areas in library catalogue research? What could have been done better and what are the next logical steps?

In this thesis I will *italicize* key terms, when they appear in a chapter for the first time. Field values will also be in italics.

The thesis will contain so many names of different kinds, that besides the aforementioned italicizations, I will use `fixed width` to denote R package names and SMALL CAPS for the predictor names in order for the reader to more quickly grasp the main category they belong to.

The program code used for and referred to in this thesis is available at https://github.com/hegroiva/Gradu. Although the code is open source, the library metadata catalogue is not. The code is of limited use without the metadata catalogue.

2 Background

Stylometry is the field of using various means to categorize texts by their stylistic properties. By stylometry in the scope of this thesis, I explicitly mean computational stylometry, which focuses on countable properties instead of linguistic anomalies resolved by close reading (McEnery and Oakes, 2000). For my purposes the interesting subfields of stylometry are authorship attribution and genre classification, by which I will base my classification methods on.

In works of author attribution the stylistic properties of a text are transformed into something that can be calculated. The main idea behind author attribution is, that a text contains intrinsic fingerprints left there by the author, either conciously or unconciously. Similar methods apply to genre classification as well, but the scope is different: instead of trying to find the author, the goal is to categorize the text into a pre-defined genre. The properties to track are hence a bit different: it is not the author's style that is under review, but rather the conventions within a genre. According to McEnery and Oakes (2000) the quantitative feature distinctions are perceived more clearly between different genres than between different authors. There are problems in author identification related to features, such as word or sentence lengths, being easily adjusted to conceal the identity of the author. Also, the style of an author may differ even within a single text and the school of writing may hide personal tendencies. (McEnery and Oakes, 2000)

The selection of properties varies case by case: there is no universally valid feature set, that would suit every occasion. The researches focus on features, that are appropriate for their own studies. In practice, this often requires hand-picking features, that could be considered especially relevant for the subject matter (McEnery and Oakes, 2000; Holmes, 1998).

2.1 Earlier research

There have been some earlier studies focusing on various aspects of library metadata catalogues: for example Lahti et al. (2015); Tolonen et al. (2016, FORTHCOMING) treat the library catalogues as a source for *distant reading* when researching book history and the turning points of intellectual history.

Previous stylometric studies on author attribution and genre classification have been based on varying types of features: n-grams of characters and amounts of punctuation marks (Kjell et al., 1994; Kessler et al., 1997; Stamatatos et al., 2000a; Peng et al., 2003), frequencies of part-of-speech tags (Argamon et al., 1998; Rayson et al., 2002), n-grams of part-of-speech tags (Argamon et al., 1998, 2003; Santini, 2004), bag-of-words or n-grams of words (Joachims, 1998; Aizawa, 2001; Tabata, 2014), prolific words within a genre (Argamon et al., 2007; Kim and Ross, 2008), common word frequencies from a huge corpus (Stamatatos et al., 2000a), intermediate steps of specific language processing programs (Stamatatos et al., 1999), stylometric features such as positioning (Kim and Ross, 2006), image classifiers (Kim and Ross, 2006, 2008), amount of nominalizations or topicalized sentences (Karlgren and Cutting, 1994), amount of affixes derived from Latin or terms of addressing (Mr./Miss/Mrs. et cetera) (Kessler et al., 1997), vocabulary richness (Stamatatos et al., 2000b), constituent frequencies and n-grams (Stamatatos et al., 2000a), and finally more refined methods involving semantics: modality, lexical units, comments (Argamon et al., 2007).

More commonly, the selected feature sets have been a combination of one or more feature groups.

It is notable, that almost all of these studies have been performed on a much smaller corpus than the ESTC catalogue.[1] That said, they often apply full texts, which means, that there often is much more material for each record, than I have. For example, Argamon et al. (2003) describes their corpus as large. It contains only 604 documents, but the amount of words is over 25

[1] An exception would be the gender and age attribution research by Argamon and Koppel (2010), which contained over 680,000 blogs entries. Methodologically it used a mixture of common and genre-specific words as features.

million. In comparison, my corpus contains over 450,000 documents, but the word count from the main titles and subtitles is approximately 19,2 million.

I will give a short description of the feature types and evaluate their suitability to the research question of finding poetry books.

Earliest features

McEnery and Oakes (2000) list the most commonly used non-lexical features. These include the average sentence length, average word length and the distribution of different word lengths, which have been discredited in later authorship attribution, but are well suited for genre classification (McEnery and Oakes, 2000). Average sentence length has already been utilized in the early days of stylometry by Yule (1939).

McEnery and Oakes also list the distribution of syllables and the length of gaps between the words of same syllabicity as having been considered promising features in the early years of stylometric research (McEnery and Oakes, 2000). Other features from early research include frequencies of carefully selected words (Mosteller, 1984), measures of vocabulary richness (such as *hapax legomena*, that is, words that occur only once in known literature) (Yule, 1944), and frequencies of common function words (Burrows, 1989).

The early studies of stylometry were varyingly successful: the Mosteller & Wallace and Burrows papers are considered groundbreaking, while most of the others gathered strong criticism. (Holmes, 1998)

Some of these features are easy to extract by modern computational methods, and despite the criticism, combined they actually do work surprisingly well. For my purposes they often are unfortunately unsuitable. For example *hapax legomena* would have been potentially able to seek out poetry books, if there only had been enough text material. The scarce titles make this particular feature type futile.

Character-level features

This group includes character *n-grams* and amounts of different punctuation marks.

Peng et al. (2003) have used rather successfully character *n-grams* in genre classification. They pointed out several drawbacks in regularly used methodologies, which predominantly are word-based: they are language dependent,

feature selection is dependent on the task one is about to perform, the cumulative effect of the uncommon features might be important, even if the feature itself is not significant, word-based methods require testing of arbitrary parameter thresholds and they ignore regularities in morphology, which could be handled by Natural Language Processing tools. Peng et al. created *n-grams* of all the characters in their material to overcome the language dependency problem and used every *n-gram* as their feature set. The resulting feature set was still relatively small. This approach works surprisingly well across language borders without additional modification to the process, but still Peng et al. reported a success of only 86 %. (Peng et al., 2003).

The same method was used in a more modest manner by Kjell et al. (1994). They used only the *bigrams* of characters to restrain the number of feature dimensions to $26^3 = 676$.

Another type of character-level features are the frequencies of punctuation marks. They are seldom, if ever, used on their own, but instead as a part in an ensemble of features. Such has been the case in (Stamatatos et al., 2000a) and (Kessler et al., 1997). It is clear, that author attribution benefits from these kinds of features. Provocative authors are more likely to use exclamation marks than some others; some authors use question marks distinctively and some prefer semi-colons while some like ellipses better. There is no reason, why this method could not be used with genre classification. In fact, both Stamatatos et al. and Kessler et al. were researching specifically genre detection, not authorship attribution. One would suspect, that some punctuation marks would be more common in particular genres: exclamation marks in pamphlets and commas in legal texts for example. However, the definition of *genre* in stylometry is not universal. Kessler et al. had a corpus of texts belonging to six predefined genres[2] and Stamatatos et al. had four predefined genres[3]. (Kessler et al., 1997; Stamatatos et al., 2000a)

POS frequencies and POS trigrams

A relatively common way of featurizing full texts for authorship attribution or genre classification is part-of-speech (POS) tagging and counting the frequencies of tags. This method has been applied for example by Biber (1992), Argamon et al. (1998, 2003), Rayson et al. (2002) and Santini (2004). Since POS tags are much more frequent than individual words, their use is more accessible than the word frequencies, when the corpus is small (Rayson et al.,

[2]Reportage, editorial, scientific writings, legal writings, non-fiction and fiction

[3]Editorials, letters to the editor, reportage and spot news

2002).

Biber (1992) was among the first to utilize the frequencies of POS tags in genre or text type classification. He performed pair-wise comparisons on five different corpora. His approach was to reduce the frequencies of POS tags into a single dimension by summing up similarly scaled frequencies and then compare the new value with the mean value of the genre. (Biber, 1992)

Similarly to Biber, Rayson et al. (2002) compared pairs of subcorpora against each other; their subcorpora was, however, apparently equal sized. They made pairwise comparison among others, between informative and imaginative writing achieving good results. (Rayson et al., 2002)

According to Rayson et al. (2002), coordinating conjunctions, except for the word *'but'* are relatively more common in informative writing, while *'but'* is more common in imaginative texts. Also, nouns in general and past participles of verbs were more common in informative writing, while modals and most forms of lexical verbs were preferred in imaginative writing. Furthermore, imaginative writing contained a relatively high number of adverbs and pronouns, and informative writing adjectives, prepositions and determiners. (Rayson et al., 2002)

In addition to using bag-of-words -features, Argamon et al. (1998) also used trigrams of the POS-tags. They claimed trigrams as being able to hold enough syntactic information while being computationally small enough to be practical. (Argamon et al., 1998)

In a subsequent research on gender and genre classification, Argamon et al. (2003) took 500 most common POS trigrams, 100 most common bigrams and all the 76 unigrams in addition to a selection of function words. From the large amount of resulting features less than 50 were then selected by an algorithm, and only these features they used for the classification. Argamon et al. claimed, that their research proved, that predicting writer's genre is possible by this method. (Argamon et al., 2003)

Santini (2004) tested whether punctuation should be included in the POS trigrams or not, but her tests were inconclusive. She also noticed, that the feature sets with less features tended to perform better than the full feature sets, but then again she was using Naïve Bayes classifier, which does not work well with redundant features (Santini, 2004; Witten et al., 2011).

Bag-of-words

Mosteller and Wallace's (first published in 1964) seminal research on Federalist Papers employed frequencies of a selection of content-free words as a means to differentiate the authors of certain passages. Their method was to seek word pairs of synonyms, the selection of which they deemed as distinctive for a given author. (Mosteller, 1984) This approach has often been successful on authorship attribution, but has also been considered as inadequate to solve attribution problems on a constant base (Goldberg, 1995; Argamon et al., 1998).

Joachims (1998) describes using Support Vector Machines (SVM) as a method in genre recognition. Joachims's approach involves getting the frequency of each appearing word in the corpus as a feature, excluding only the most infrequent words and the stop-words. This introduces huge feature sets, but is still possible because of SVM's high tolerance of nearly irrelevant features. Each feature does not need to be hand-picked separately, but can be automatically extracted from the corpus. (Joachims, 1998)

A more sophisticated method was proposed by Aizawa (2001), who used part-of-speech tagging to automatically extract terms to be used as features, including compound words. In Aizawa's research, the feature set grew up almost to 100,000 words. (Aizawa, 2001)

Stamatatos et al. (2000a) used a slightly variated method. Instead of using only the words from the corpus, they used a much larger corpora, from which they drew the feature terms. Their research showed, that on their corpus the results didn't improve after the amount of predictor words went beyond 30, due to data overfitting. (Stamatatos et al., 2000a). This is contradicted by Aizawa, who points out, that even words with low frequencies add up cumulatively (Aizawa, 2001).

The use of merely the bag-of-words methods has been criticized for being applied blindly without any linguistic motivation behind it. Argamon et al. (2007) claimed, that

> The general methodology (...) is to find as large a set of topic-independent textual features as possible.

Their approach was to use (in addition to other types of features) bag-of-words, that were specific to the genre they were bound to detect (Argamon et al., 2007). They reprised their critic in a subsequent research (Argamon and Koppel, 2010).

The genre-specific bag-of-words approach combined with word n-grams has been used more recently by Tabata (2014) to analyze difference between the

language use of Charles Dickens and Wilkie Collins.

Other feature types

One of the more obscure feature types are features, which are dependent on specific language processing tools. Such features might be for example iteratively working chunking programs, and the extracted features the amounts of unprocessed words after each iteration. These feature types are completely dependent of the tools, and their working logic would is usually hard to back-track, effectively making them black boxes. Stamatatos et al. (1999) have used this kind of feature set along with more common phrase-level and token-level counts. Their motivation was to determine a measure for syntactic complexity, which they accomplished by using intermediate steps of their processing tool. (Stamatatos et al., 1999)

Kim and Ross (2006, 2008) had the idea of using image classifiers in genre identification. They used PDF files, from which they extracted pixel values. They then calculated the relative pixel count on sections, giving them language-independent values, in effect featurizing the layout of the pdf. (Kim and Ross, 2006) Also, they had the idea of using stylometric features, like positioning or font types as features (Kim and Ross, 2006).

Methods requiring more profound language processing, such as the amounts of nominalizations and topicalized sentences have also been used as predictors (Karlgren and Cutting, 1994). Stamatatos et al. (2001) used frequencies of the constituents, such as noun or verb phrases and their n-grams. Argamon et al. (2007) approached the classification task through semantic measurement. They resolved, among other features, the types and levels of modality, appraisal words ("quite" or "very") and the different types of commenting, such as admissive ("Frankly, ...") or assertive ("Certainly ...") and so on. Their mission seems to have been to bring back the linguistic motivation to stylometry. (Argamon et al., 2007)

Kessler et al. (1997) used the amounts of affixes derived from Latin and the terms of addressing persons, such as Mr. or Mrs.

Vocabulary richness has been rather commonly used indicator, which have been measured in several ways. These methods include *hapax legomena* (words occurring once in a corpus) and *dislegomena* (words occurring twice), as well as type/token -ratio (Yule, 1944; Stamatatos et al., 2000b). Type/token ratio is calculated by dividing the size of the vocabulary by the amount of tokens in the corpus (Stamatatos et al., 2000b). In addition to

these single measures, there have been efforts in using multivariate methods as well (Yule, 1944; Stamatatos et al., 2000b).

2.2 Data

My corpus is the English Short Title Catalogue (ESTC), which is a library catalogue maintained by British Library containing metadata records of books printed before 1801 either in Great Britain or in English language (ESTC). Each record is divided into metadata fields, which applies to MARC 21 standard[4]. The MARC standard defines, how the names containing the metadata information are named, and what kind of information can be annotated in them (MARC). The format, in which the metadata is actually input into the metadata fields, is in turn controlled by ISBD[5] (IFLA, 2007). In short, MARC standard describes, in which fields the information is input in the catalogue, and ISBD describes what the data format looks like.

Although MARC standard is meant as more or less universal, a closer acquaintance with several library metadata catalogues has shown, that each library data is unique in implementing it. This means, that the metadata fields, that are actually used, are catalogue specific. My thesis focuses specifically on ESTC and the following descriptions of MARC fields I can account for applying to ESTC only.

The ISBD definition is straight-forward, but the field itself still might contain huge amounts of information. In general, cleaning up catalogues to get uniform data is a huge effort, which might consume a lot of energy, depending on the field at hand. COMHIS/ESTC have done, and are still doing, a lot of field polishing in ESTC, and these results will be utilized in this thesis too. When the field has not been preprocessed yet, I have used the raw field values instead.

The copy of ESTC I have used in this research is not open to public, but my research group has access to it for research purposes. A public online database version exists, although with limited search capabilities[6]. I have used only the records printed in English in my research for two reasons: first, my methods are partially dependent on the language used and second, 93.8% of the records are in English anyway. ESTC contains 483.344 records in total, of which 454,488 are in English.[7]

[4]MARC stands for MAchine Readable Cataloging

[5]International Standard Bibliographic Description

[6]http://estc.bl.uk

[7]All the numbers of records and annotations of specific MARC fields and examples of

2.2.1 MARC fields in ESTC

Several MARC fields in ESTC might be considered important in finding the poetry books.

Genre field
The MARC field for *Genre/Form* (655a) is the most important single field for my thesis. In ESTC there are 115,023 (25,3%) records, in which the genre information is annotated The information in this field conforms to RBMS[8] standard for genres (RBMS). Genre information will be used to determine the response value: whether the book is considered as poetry or not. There are 512 different genres annotated in ESTC, and the genres vary from *Broadside poems* to *Prospectuses* and *Satires*. Some of the genres indicate the *form* (such as broadside or pamphlets) rather than the actual *genre*. Besides that, any book can have more than one value in this field.

Topic field
Field 650a *(Subject Added Entry - Topical Term)* usually contains the main subject of the book and is annotated for 266,188 (55.0%) of the ESTC records. Most of the poetry books are annotated as *Poetry* or as some subgroup of it, such as *Sonnets* or *Poetry, English*. However, not only books **of** poetry, but also books **about** poetry are annotated in the same manner. On the other hand, the field values often reveals, that the content most likely is other than poetry. Such values would include *Life skills*, *Christmas sermons* or *Jacobite Rebellion, 1715*, although someone could have decided to write an epic poem about Jacobite rising of 1715 or a poetic parody of a self-help guide. The topic field is a valuable addition to the other fields, but alone it is not enough; one must remember, that nearly half of the topic fields are unannotated.

Other topic related fields
In addition to the MARC field 650a, which is the principal field depicting subject or topic, there are also other more seldom used fields, which convey almost the same information. These fields declare a topic's subgroup, when subgrouping is based on the literary form of the work, such as in essays or sermons. I have included in the corpus fields *Subject uniform title: form subdivision (630v)*, *Subject corporate name: form subdivision (610v)* or

the field values in this thesis I have gathered from the metadata myself by program code.

[8]RBMS stands for Rare Books and Manuscripts Sections and it controls among others standard for genre classification.

Subject personal name: form subdivision (600v), in case they have been annotated, but the topic field 650a has not. There were 43,922 such cases in the corpus. Total amount of records containing some kind of topic or corresponding information is thus 310,110, which comprises 68,4%. It is good to note, that the contents of these fields actually are really close to the contents of *Genre/Form* field (655a), as they describe the record's form. They however seem to reflect the topic more than genre In any case, I could not have used them for defining the response value, since that decision should be kept as simple as possible and it definitely should not be dependent on multiple fields.

Title fields

Other potentially relevant fields include: *(Title statement)* (245a), and *(Title remainder)* (245b), which practically make up the contents of the books title page together with *(Edition statement)* (250a). These fields might contain keywords that are inherent to poetry books. As the *Title statement* (245a) only contains a small amount of words and the *Title remainder* often contains information about the genre, the premise is, that combining these fields would increase the possibility of success in classification. The *Edition statement* is not relevant for finding poetry.

Fields related to publishing

Certain book publishers might focus on certain kinds of publications, and therefore MARC fields related to publication 260a-c (*Publication place, Publisher* and *Publication time* respectively), might be considered relevant. There are some problems with these fields. The *Publication place* is *London* in over 300,000 records. This considerably reduces the usability of that field for this kind of task. The *Publisher* on the other hand contains huge amounts of information: there could be the name of the printer, the one who commissioned the book or the address of the book shop were it can be bought, or even all of these. This information, despite it conforming to ISBD standard, is not easily reduced into machine readable format, so that it would contain only (the unequivocal) name of the printer. In its raw form, the field contains over 200,000 publisher field values; the actual number of publisher is a fraction of this. Due to the various different pieces of information the cleaning up of the field is not an easy task.[9] The *Publication time* would be interesting, if it could be compared with the author's birth and death years;

[9]It is currently being done by the COMHIS group, but the project is still ongoing(COMHIS/ESTC).

in itself it is of less value.

Fields related to physical attributes of the book
Poetry books could be of more or less the same size and length, so MARC fields *Physical extent* (300a), which actually means page count, and *Dimensions* (300c) are definitely usable as properties. Even field *Other physical details* (300b) might contain a shared factor between poetry books. Relatively often this field implies whether the book is illustrated or not. In the ESTC there are 53,703 annotated cases of the field 300b. This in connection to the fact, that only a subgroup of the field values are linked to poetry led me abandon its use. The other two fields are nearly always annotated.

Author name
Author name, expressed in field 100a, is definitely one variable which must be considered for seeking poetry books. The logic here is that a person, who has written one book of poetry, has probably written other poetry books as well. That said, the same author might also have written something completely different, in which case a positive value might lead astray.

2.2.2 Definition of poetry

There are two primary traditions in the genre definition: a systematic tradition of categorizing literary works on some criteria starting from Aristotle and his *Poetics*, or empirically listing the existing genres without regard to the reasons they are differentiated from each other (Frow, 2006, 58-59).

It is worthwhile to note, that *genre* is often used interchangeably, when actually is *literary form* is meant. This is also true on RBMS and MARC descriptions as well. My take on defining poetry is based more on the systematic tradition prioritizing pragmatism above all, which allowed me to bypass the nitpicking of terminological inaccuracies with a clear conscience. *Poetry* does not exist as such in the RBMS list of values, so I had to create the category myself. I collected all the values from the Genre/Form field in the ESTC (MARC 655a), and mirrored them against the RBMS list of genres. The difficult part was deciding, which of the genres in the RBMS would represent poetry. Some of them are easy calls, since RBMS is an hierarchical system, and *Poems* has many subcategories. Some genres however are more problematic, because they contain poetical elements, but are not exactly poetry or they are hybrids of two or more conventionalized categories. Such problematic works include operas, dramas, music and emblems.

I did my main decision by the criterion of whether the book had been written in verse or not, first and foremost trusting the RBMS description of the genre. If the description mentioned the genre as having been written in verse or being poems or metrical, I judged the genre belonging to poetry. Conveniently, all the ESTC material is before the year 1801, so prose poems have not yet emerged.

Still not everything was clear: for example plays were written in verse before the emergence of Restoration comedy in the late 17th century (Albert, 1971, 179). Since there exists a conventionalized main genre division between *prose*, *drama* and *lyric* from the 17th century (Frow, 2006, 59), I decided that plays, even if written in verse, should belong to drama even more than lyric. This division is implicitly used by most of the literature historians, for example Albert (1971); Alexander (2000); Quennell and Johnson (1973); Fowler (1989). Note that (Frow, 2006, 59) names one of the three main literary forms as *epic* instead of *prose*, but still means roughly the same thing: the *epic* genre had evolved into meaning *narrative*, while still being written in verse. He also thinks, that the genre definitions are not constant, but change in time (Frow, 2006, 71). Together with discarding *Plays*, I discarded every subcategory it has in the RBMS classification, from *Heroic dramas* to *Masques*[10].

I hesitated on *Psalters*, which RBMS describes simply:

> Use for books containing only the Psalms.

The Psalms, as are the other lyrical portions of the Bible, definitely possible to classify as poetry. (See for example (Albert, 1971, 115).) Then again, because of the inevitable religious connotation of them, they could be dismissed as being primarily religious content. In the end I did include them as poetry, which had some repercussions presented in the Conclusions chapter.

Another difficult choice was to be made regarding music. My guideline was the amount of lyrical content in relation to the amount of other recognizable genres. For example *Opera*, *Operetta* and *Libretto* contain recitatives in prose. This was a situation that could not be resolved by RBMS descriptions. My decision was, that they essentially resemble plays more closely than poetry, and accordingly should be not defined as poetry. Conversely, I did include *Carol books*, *Song sheets* and *Songsters* as poetry, because their deviation from the archetypical poetry genre does not liken plays or prose,

[10] A masque is short dramatic performance with dialogue, lyrics, music and dancing. It is commonly regarded as a subcategory of drama (Albert, 1971).

but completely another art form instead. As a main literary form, they still are closest to *poetry*. I applied the same logic to *Emblem books* as well.

These were not the only complicated genre values. In RBMS there are many possible values denoting the format of the print rather than the literary form. These include *Broadsides* and *Chapbooks*. Ambiguous in other ways are for example *Paraphrases*, *Juvenilia*, *Acrostics* and *Memorial works*. From these it is impossible to tell for certain, whether they contain poetry or not. I ruled all of these out as non-poetry, on a perhaps silly assumption, that if they actually did contain poetry, another value containing the literary form would have been included in the MARC field. More of this on chapter 4.3.2. Records with those genre values, from which I was able to deduce the content as doubtlessly being other than poetry, I used later for feature extraction, as described in Chapter 4.1.

One additional aspect of the genre field is, that it is easier to apply for works of fiction than nonfiction. In RBMS there does exist categories like *Scientific works* or *Academic dissertations*, but these values are seldom used: For example, the *Scientific works* appears in the corpus nine times and *Academic dissertations* 489 times. These are really low figures compared to over 10,000 *Sermons* or over 30,000 *Poems* in the catalogue. The ESTC catalogue is definitely biased in this aspect.

The final note on the genre: regarding volumes with both poetry and some other works, I decided to include them as poetry. Complete list of the genre values, which I defined as poetry, is included as **Appendix A**

3 Methodology

3.1 Random forests

Random forest is a supervised learning algorithm developed by Breiman (2001), which can be used for both regression and classification analysis. Regression analysis is performed, when the information predicted *(response)* is *continuous*, such as integers whose distance can be used as a measure. Likewise, when the *response* is *categorical*, that is, it has a predefined set of possible values the random forest will perform *classification*. Breiman (2001)

Tan et al. (2006) describes random forest as an *ensemble method*, meaning it combines multiple learning algorithms to get more accurate predictions. In random forest the models are created from a number of randomly sampled

features, which are eventually aggregated for an optimal end result (Tan et al., 2006, 276-279).

For random forests the first requirement is a feature set. The features in a feature set is are vectors of *continuous* (numeric and scalable) or *categorical* (non-salable) values. These features are also called *predictors*, because they are used in predicting the *response*. Every feature describes some aspect of the record, such as the word count of the book title or a publication place.

Random forest creates *decision trees*, which are tree structures describing how features and their values are used to make predictions (Tan et al., 2006, 150).

The process described below is collected from Breiman (2001); Breiman and Cutler (2003); Liaw and Wiener (2002); Tan et al. (2006). In the beginning all the records of the training set are pooled in one node, so it contains *responses* with different values. Each feature is tried in turn, how good its values would be for splitting the nodes in such a way, that the resulting nodes would be as homogeneous in regard to the *response* as possible. For *continuous* features several cutpoints are tried, while for the *categorical* features the process is more complicated: the feature's values are split into two groups randomly to get an estimate of the splitting capability. This estimate is then multiplied by the $numberofvalues - 1$, and this number is compared to the estimate fo the continuous values. (Breiman, 2001; Breiman and Cutler, 2003) The most prominent feature is then used for the actual split by the optimal cutpoint specified by the algorithm (Tan et al., 2006, 153-172). The process is then repeated with the remaining features until all the features have been used for the splitting.

A predictor containing lots of different categorical values can also be reduced into a smaller amount of predefined categories. For example, publication place, which normally is a town or city, might be reduced into a smaller number of counties or countries and used instead of towns in the classification. The purpose for such a reduction would be *pruning* the *decision tree*, that is, removing branches that have little use in classifying the records (Tan et al., 2006, 184-185). *Pruning* is necessary, if there are many different values and only a few samples of each value (Witten et al., 2011, 195-197). A simpler decision tree results in the reduction of *overfitting* and improved prediction accuracy on the test data, that was not used for creating the learning model (Webb, 2010). *Overfitting* means using features, that are created from the noise or variance in the data, instead of the underlying distribution (Webb, 2010). For the random forest *pruning* is not used, and Breiman and Cutler (2003) state very emphatically, that random forest does not overfit,

although this has been contradicted (Hastie et al., 2009, 615). There may be a need for reducing the number of *categorical* values: the `randomForest` implementation in R is unable to handle more than 53 values.[11]

Liaw and Wiener (2002) and Breiman (2001) give a thorough account of the random forest: Each time a decision tree is built by *random forest*, only a random subset of predictors from the feature set will be used for building the *decision tree*. This method is called *bagging* (bootstrap aggregation). The number of predictors sampled for each *decision tree* is predefined by a parameter named *mtry*. Every time the features have been used for a tree, they are returned to the pool of features, so that they can be used again for the next tree. The number of decision trees to be grown is also parameterized, as *ntree*. Prediction of an individual record can then be gotten from the node with corresponding predictor values: the prediction will be a majority vote of the responses in the node. After the specified number of trees have been grown, the predictions are aggregated and the final prediction is again decided by a majority vote. (Liaw and Wiener, 2002; Breiman, 2001)

According to (Tabata, 2014, 30) random forest has outranked other machine learning methods, such as Support Vector Machine (SVM) and k-Nearest Neighbor in previous author attribution research. As my thesis was about the library catalogues more than the plain machine learning methodology, I decided not to conduct any tests between various algorithms. Instead I just selected one, which has been known to be efficient.

3.2 Generic principles

Training and testing sets

The basic machine learning manuals, such as Tan et al. (2006) and Witten et al. (2011) teach, that the *training* and *testing* sets should be kept separate. A common way is to split the records into two randomly selected subsets and use the the other one for training and the other one for testing. The training set is then used to create the learning model, which is then evaluated with the testing set. It is also possible to use a distinct *validation set* separated from the training set for assessing and fine-tuning the model before running it with the testing set. The testing set should not be in any way used in defining the features, because that might introduce an upward bias in the results. (Tan et al., 2006, 148-149; Witten et al., 2011, 148-150)

[11] https://stats.stackexchange.com/questions/157331/random-forest-predictors-have-more-than-53-categories

Cross-validation

Tan et al. (2006) describes *Cross-validation* as a model training approach, that ensures all the records being used an equal amount of times for training and only once for testing. A training set is split into k approximately partitions of equal size. Each of these partitions is used once as a *validation set* for the model, that is taught using the remaining $(k-1)$ partitions (*folds*) as a *training set*. The results of each run are finally averaged. This is called the k-*fold cross-validation* method. (Tan et al., 2006, 187)

Other kinds of cross-validation methods are described by Arlot and Celisse (2010). They are *exhaustive*, that is, they consume all the possible variations of splits into training sets of a desired size. In *leave-one-out* method each record is used once as a validation set, and in *leave*-p-*out* each subset combination of p records is used as a validation set exactly once. The methods are exhaustive also computationally, and for large corpora infeasible (Arlot and Celisse, 2010). The bias of the data decreases with higher k, as the training data is larger. At the same time, however, the variance increases. The proper selection of k is dependent on the quality and the size of the data. (Arlot and Celisse, 2010) A value between five and ten is seen as usually a good compromise (Hastie et al., 2009; Arlot and Celisse, 2010). Arlot and Celisse (2010) is more reserved on this value stating that other values may be even better, when the *signal-to-noise ratio* is low or high.

In my data, the ratio sure is low: singular features usually do not have much predictive power, especially in the qualification step. On the other hand, the training set is comparatively large. I decided to go with 5-fold cross-validation to save some processing time, because there were many different feature sets and test types, and I wanted to make the tests as uniform as possible.

My method

For each run of the qualification round I split my data into two: training and testing sets, as is the usual procedure. The training set I then further submitted to the 5-fold cross-validation process, which also is a standard procedure. What I failed to understand, was that my data was not correct in the sense, that my testing data should not have been the fifty percent partition of the genre-annotated records, but instead all the records without the genre annotation. Following this method, I only used 40% of the available material for training, 10% for validation and 50% for nothing at all. By the time I realized my mistake, I had already run most of the qualification rounds, and to keep the results comparable with each other and to not have to do

a significant number of runs again, I decided to simply make note of the mistake in my thesis.

As I had used the whole data with genre annotations as my source for the genre-specific features, I had also broken the other rule: keeping the test set away from the prediction creation. This means, that there probably is an upward bias in the qualification round. It must also be said, that the training set size was big enough anyway, that the predictivity of the features was still estimable relative to each other. Besides, doing frequency checks (for example to get the 100 most frequent words in the poetry books) in the qualification round could have resulted in different feature sets for each of the cross-validation runs making the evaluation of them unnecessarily complicated. If the sets would have been the same for each run, there was no harm done.

There were some empty values in the ESTC data. MARC field 300a, which describes the page count, was empty for 2,460 records in English language. As the page count is a continuous variable, I decided to supply the records with the median (30) value of the annotated page counts. The book lengths vary considerably from one-paged broadsides to thousands of pages for multi-volume books: the mean would have been over 95 pages. I figured, that the extremely long books add up to the book length considerably, and preferred median over the mean. MARC field 300c describing the physical dimensions was another field with missing values. In the ESTC there were 16,141 missing values for this field. The field being categorical in nature, I decided to create a separate category for those, that were not annotated. Assessing the most common value would have been guess work: the most common book size is *octavo* with 156,097 occurrences, but *folio* and *quarto* get close by 95,828 and 96,315 occurrences.[12] The most common value would have been the smallest size, despite that the other two combined are more common than the octavo alone. The empty author fields (MARC field 100a) I treated as authors, of whom there is no previous knowledge of having published poems.

3.3 Measures for comparing learning models

The prediction results fall into four categories:

1. True Positives (TP) means the number of correctly predicted response of the desired value

[12]The measures of book sizes describe the number of times a sheet is folded to create pages: folio once, quarto twice and octavo thrice. (Beal, 2008, sv. 'folio', 'quarto', 'octavo')

2. True Negatives (TN) means the number of correctly predicted response of the undesired value.

3. False Positives (FP) means the number of responses of the desired value, which have been incorrectly predicted as the undesired class.

4. False Negatives (FN) means the number of responses of the undesired value, which have been incorrectly predicted as the desired class.

In the case of finding poetry from the catalogues, the desired value is specifically *Poetry* and the undesired *Non-poetry*.

Several measures, which are commonly used for evaluating the learning models. *Recall*[13] is the amount of True Positives divided by the sum of True Positives and False Negatives:

$$recall = \left(\frac{TP}{TP + FN}\right)$$

In other words, *recall* tells, how well the records with desired classes have been retrieved from the corpus (Kuhn et al., 2017; Tan et al., 2006, 296-297).

The other main measure , which is frequently used, is the *precision*. It is calculated as the amount of True Positives divided by the sum of True Positives and False Positives:

$$precision = \left(\frac{TP}{TP + FP}\right)$$

It determines, how pure the category is with the response, that was sought after (Kuhn et al., 2017; Tan et al., 2006, 296-297).

Precision and *recall* tell only of the performance of one response value. To compare the performance of more than one response value with one figure, more metrics are needed. *Overall accuracy*[14] is a simple metric, where the sum of *recall* values of every class is divided by the amount of records (Tan et al., 2006, 298). An advanced version of this is the *balanced accuracy*[15], which is calculated as the average of the recall values of every class (Kuhn et al., 2017; Brodersen et al., 2010). *Balanced accuracy* weighs all the response values with equal value, whereas *overall accuracy* stresses the class

[13]Terminology is varied. Sometimes *accuracy* is used, sometimes *sensitivity* or *True Positive Rate* (TPR). For my thesis I will systematically use *recall*.

[14]To confuse the terminology even more, this might be called *weighted accuracy* or just *accuracy*.

[15]*Balanced accuracy* is also known as *average accuracy*.

with most records giving misleading results, when a biased classifier is used on an imbalanced data set (Brodersen et al., 2010). On an imbalanced set, where the less frequent classes are sought for, *balanced accuracy* is preferable. The formula for the balanced accuracy is:

$$balanced\_accuracy = \frac{1}{2} * \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right),$$

(Kuhn et al., 2017; Brodersen et al., 2010)

F1[16] measure is the harmonic mean of *precision* and *recall*, as descibed by Kuhn et al., 2017; Tan et al., 2006, 298. It is calculated with the formula

$$F1 = 2 * \left(\frac{precision * recall}{precision + recall} \right)$$ [17]

For this thesis the F1 measure is calculated by *confusionMatrix* function from `caret` package, which uses β value of 1 intrinsically which completely ignores the amount of True Negatives. Therefore F1 measure is usually not recommendable for sets with imbalanced classifiers. However, it is applicable here, because the interest is in the Poetry books and not the Non-poetry (Kuhn et al., 2017; Tan et al., 2006, 298)

H measure is a classification performance measure, which was designed by Hand (2009) as a replacement for *Area under a ROC curve* (AUC). AUC is a commonly used metric, which measures the performance of the model on a general level. While AUC treats a TRUE value predicted as FALSE as severe as a failed prediction vice versa, the H measure applies weights for both misprediction types. This addresses the problem AUC has with crossing curves, in which the AUC might be higher for a model, even though another model might perform better at certain points. H measure requires prior knowledge of the severity of a failed prediction. `randomForest` does not support the H measure, so I used the `hmeasure` implementation.[18] (Hand, 2009; Anagnostopoulos and Hand, 2012)

3.4 Feature sets

My approach involved keeping the amount of used predictors at minimum for the classification task in order to avoid the *overfitting* problem. For this

[16] F1 or Fβ or F measure

[17] This is actually slightly simplified; a more comprehensive formula would include a β variable for weighting *precision*.

[18] I borrowed a functional code wrapper for this from https://stackoverflow.com/questions/24509309/custom-metric-hmeasure-for-summaryfunction-caret-classification?rq=1

aim I created several different predictor sets containing either features that have either been commonly used in author identification or genre classification or features that I figured might be at least somewhat suitable for seeking the poetry books. Some of these feature sets I collected by external Natural Language Processing programs, some of them I created myself from the frequencies in the ESTC data.

Argamon et al. (2007) and Argamon and Koppel (2010) call for the use of genre related information, instead of relying on somewhat unsophisticated means of ignoring the class altogether and using the most common frequencies from the whole corpus. Different genres require different types of predictors (Kim and Ross, 2008). Keeping this in mind, the genre related information is precisely what I used with the lexical features for BASIC feature set, as well as for the other applicable feature sets: part-of-speech tags and their *trigrams*, bag-of-words, and topics. This idea was not uniquely new, when Argamon et al. made their remarks. Already Mosteller (1984) had used author specific features for attributing the Federalist papers. After their days the availability of brutal computing power has made the use of larger feature sets more tempting, and often without any regard to the classifiers at hand.

I tried to assemble the feature sets in such way, that similar types of features are grouped together for a qualification round, and those with the best predictive properties within each set are then brought to the actual classification rounds. I named the features uniquely for reference and prefixed each of them by a descriptive name of the bunch they were in. This way I came up with prefixes BASIC, POS, POS3GRAM, BOW, DEPREL, TOPIC, VARIA, PUNCTUATION and MARC.

3.4.1 Unused features

Some feature types I regarded as either inapplicable to my material or for some other reason rejectable.

Using tool dependent features like Stamatatos et al. (1999) easily leads to the features becoming hardly understandable and the research poorly reproducible.

Image classifiers and featurizing the layout of the books, as Kim and Ross (2006) and Kim and Ross (2008) did, might actually be a good way to pick poetry books from a large mass: the layout of poetry books often differs significantly from book layouts of the other genres. Once again, this would

require the full texts, or to be more exact, images of the books, as the library metadata catalogues contain text only and not any information about the images themselves. The same applies to the other idea Kim and Ross (2006) had, featurizing positioning or font types.

The nominalizations and topicalized sentences also sound like an interesting effort (Karlgren and Cutting, 1994). Their use would be however limited to the book titles only. The titles often in practice being incomplete sentences, the identification of such things would be too unreliable.

Kessler et al. (1997) used the terms of addressing, which I seriously doubt would have any significance in searching the poetry books. Their other major innovation was the amounts of Latinate affixes (Kessler et al., 1997). I would expect that to work reasonably for this type of material. At least scientific material probably could be excluded from the poetry quite neatly with this feature alone; that said, I do not think that the scientific material would be the most difficult material to distinguish anyway. Unfortunately, currently no tool seems to exist for extracting the Latinate affixes reliably and creating one would have been beyond the scope of this thesis.

The ease of adapting vocabulary richness surely tempted to trying them also. Research has however shown, that the text length is critical for them to work properly; the results are unstable for texts, that are shorter than 1,000 words (Tweedie and Baayen, 1998; Stamatatos et al., 2000b). This certainly rules out their use for this material.

It would also have been interesting to make use of constituents in a similar vein to the research by Stamatatos et al. (2001). The short texts of the titles effectively prohibit this, as I suspect, that the high level of homonymy in the English language exposes ungrammatical titles to false estimates, in regard to what is a word chunk.

The semantic approach by Argamon et al. (2007) is also interesting, and would be even more so, if full texts had been available. Their modality detection approach likens sentiment analysis, but for the poetry book titles there is not enough text to express modality, appraisal or commenting. They make a note themselves, that the irrelevant features reduce performance, due to *overfitting*.

3.4.2 Feature selection

One important aspect regarding features is removing redundant and irrelevant features from the feature set. Redundant features are features, that

describe more or less the same thing. An example could be page count of a book in addition to average number of characters on a page, as compared to total number of characters in a book. Irrelevant features are simply meaningless for the task and can be removed without affecting the results (Liu, 2010). The more features there are in the data set, the bigger samples of training data are needed to ensure, that there exists enough samples for each combination of feature values. This is the *curse of dimensionality*, which means the data analysis becoming increasingly difficult, as the data contains more and more features. (Keogh and Mueen, 2010)

Liu (2010) describes three approaches to do the feature selection: *embedded*, *filter* and *wrapper* approaches. Embedded method is handled by the data mining algorithm itself, while selecting the best feature to use for each of the splits. An example of an algorithm using embedded methods would be *random forest*. Filter method applies some external method to find out the correlation between the features disregarding the actual data mining algorithm and the result set. This information can then be used to remove redundant features from the set. Wrapper approach is to take a subset of features and do the regular classification task with the algorithm. (Liu, 2010)

My approach to feature selection was to use embedded approaches in `randomForest` and `party` packages for a number of subsets of possible features. The results of the metrics I then evaluated manually.

3.4.3 Measures in `randomForest` package

The `randomForest` package in R records four different measures for variable importance when applied for classification problems with two result classes (Liaw and Wiener, 2002). The first two measures are mean decrease accuracy for both the class values respectively. The third measure is the mean decrease in accuracy over both the classes combined, while the fourth one is the mean decrease in Gini index. (Liaw and Wiener, 2002)

Mean decrease in accuracy (MDA) [19]

The process of calculating the MDA is described to some extent by Liaw and Wiener (2002) and Hastie et al. (2009). An out-of-bag (OOB) error rate is calculated for each subset not included in the bootstrap sample, which was used in making the decision tree by getting a prediction for every record. The error rate is then calculated and that error rate is compared to a new

[19] This refers to the implementation in R's `randomForest` package

error rate, which is gotten by permuting each variables. The average of the differences over all the trees is then summed and the result is finally normalized by the standard deviation to get the MDA. (Liaw and Wiener, 2002; Hastie et al., 2009, 612) In other words, the accuracy does not mean the accuracy of the whole forest, but the mean accuracy of the single trees, which is normalized by standard deviation.[20] This differs from the standard definition of MDA. MDA has been shown to overestimate the importance of correlated variables, although some studies have been unable to confirm this notion. (Louppe et al., 2013; Strobl et al., 2008)

Mean decrease in Gini index (MDG)
According to Breiman et al. (1984) *Gini impurity* or *Gini index* describes the homogeneity or heterogeneity of each splitting of the node. Gini impurity of a split node is the ratio of class values within it compared to the assumption. When the amount of different class values are equal, the Gini impurity is thus 0.50, and when there is only one value, the Gini impurity is 0. Breiman et al. (1984) The mean decrease in Gini index for each variable is calculated by comparing the Gini impurity of the parent node to that of the child nodes whenever that variable is used for doing the split. My empiric testing showed, that this Gini impurity seems to be weighted by the amount of records in the node, but Breiman et al. (1984) is not explicit on this. (Breiman et al., 1984, 146-148)

Measure from `caret` package
`Caret` package Kuhn et al. (2017) claims to use the importance metric from `randomForest` package, but this clearly is untrue.[21] The most relevant features differ in some cases from the ones that have been produced from the `randomForest` code. I did not wish to delve too deeply into the mathematical consequences of the differing implementations, so I simply did an additional variable importance check with the `caret` package.

Conditional importance
Gini importance as a variable importance measure has been strongly crit-

[20]The process is described in the `randomForest` manual, but the elaboration by username Jianyu in user community Stack Exchange was necessary to understand the procedure. https://stats.stackexchange.com/questions/197827/how-to-interpret-mean-decrease-in-accuracy-and-mean-decrease-gini-in-random-fore

[21]This was demonstrated by username Shape on user community Stack Overflow. https://stackoverflow.com/questions/37888619/difference-between-varimp-caret-and-importance-randomforest-for-random-fores

icized by Strobl et al. (2007), because it favors those features with many values over those with only a few. The explanation for this is, that the variables with more different values provide more possible cutpoints by chance for the decision tree splits. Further, unordered categorical predictors provide an exponentially growing number of cutpoints. (Strobl et al., 2007)

In their article and also in Strobl et al. (2008) they suggest another measure to be used instead of Gini importance: the *conditional variable importance*.

The `randomForest` package is strictly based on Gini index and that can not be overridden: the conditional variable importance simply is not implemented. Therefore I ran extra checks for each feature subset with the `cForest` function, which is included in the `party` package (Hothorn et al., 2006a). Their approach was to separate the variable selection from the splitting procedure. They standardized all the covariables with permutation tests, so that they were able to compare them independently between each other without bias (Hothorn et al., 2006b).

4 Test runs

4.1 Basic set

This set contains primarily features, that are renowned from the earliest research on author identification or genre classification or otherwise easily calculable, the so-called low hanging fruit. These features often produce good or at least decent results on full texts and a small group of different class values. In this corpus the features are based on full titles or only the main title instead of full texts, which results in stylistic variation being bound to be noticeable more infrequently than if the whole text had been available. I intended the basic set as a bunch of features providing the basis of the combined features, and as a way to set a benchmark to evaluate the other feature sets against. I dubbed the feature set as BASIC, because of the diversity of the feature set: there is no good denominator, except for the ease of use and the fact, that these types of features are commonly used in most modern classification tasks.

The features I have selected for this set were:[22] the average number of characters in words (BASIC_WORD_LENGTH), amounts of charac-

[22]all the amounts in these features have been collected from the full title (concatenated main and subtitle), unless otherwise stated

ters (BASIC_CHARS), words (BASIC_WORDS), verbs (BASIC_VERBS), interjections (BASIC_INTERJECTIONS), adjectives (BASIC_ADJECTIVES), adverbs (BASIC_ADVERBS), foreign words (BASIC_FOREIGN_WORDS), proper nouns (BASIC_PROPER_NOUNS), pronouns (BASIC_PRONOUNS), gerunds (BASIC_GERUNDS), verbs in past tense (BASIC_VERBS_PAST), characters in the main title (BASIC_CHARS_MAIN_TITLE), words in the main title (BASIC_WORDS_MAIN_TITLE), characters in the subtitle (BASIC_CHARS_SUBTITLE), words in the subtitle (BASIC_WORDS_SUBTITLE) and sentences (BASIC_SENTENCES).

The number of sentences I counted by summing up all the periods, exclamation marks and question marks.

All the amounts of parts of speech I extracted with `tm`, `NLP` and `openNLP` packages (Feinerer and Hornik, 2017; Hornik, 2017, 2016, respectively). First, the corpora were created by `tm`, and after that split to tokens with `openNLP`. `NLP` was then used with default settings to do the actual annotation.

The method of summing frequencies of part-of-speech tags into singular features was introduced by Biber (1992). I followed this approach, but not to the full extent: due to the shortness of the titles I chose to use raw values of the tags instead of the relative frequencies, and hence I did not compare them to the genre mean value either.

I collected into one feature BASIC_VERBS the words annotated as *VB*, *VBP*, *VBZ*, *VBD*, *VBN* or *VBG* (Base form, first or second person singular present, third person singular present, past tense, past participle and gerund, respectively). textscBasic_adjectives were tags *JJ*, *JJR* and *JJS* (positive, comparative and superlative) and BASIC_ADVERBS tags RB, RBR and RBS (positive, comparative and superlative).

Of BASIC_PROPER_NOUNS there existed tags *NNP* (singular) and *NNPS* (plural), while for BASIC_PRONOUNS there was *PRP* (personal) and *PRP$* (possessive).

Verb tags *VBD* (past tense) and *VBN* (past participle) I collected into BASIC_VERBS_PAST.

BASIC_INTERJECTIONS, BASIC_FOREIGN_WORDS and BASIC_GERUNDS I got directly by counting the occurrences of one single tag for each (*UH*, *FW* and *VBG*, respectively).

Argamon et al. (2007); Argamon and Koppel (2010) call for the use of genre related information, instead of relying on somewhat unsophisticated means of ignoring the class altogether and using the most common frequencies

from the whole corpus. The genre-related information is precisely what I used for to get the lexical features for BASIC feature set. Additionally I got some predictors based on lexicon: amounts of the most frequent Poetry and Non-poetry words (BASIC_POETRY50 and BASIC_NONPOETRY50, respectively), amounts of the most common Poetry words in relation to Non-poetry words and vice versa (BASIC_POETRY50_COMPARED and BASIC_NONPOETRY50_COMPARED, respectively) and the amounts of the next one hundred most common words within the class in relation to the other class (BASIC_POETRY100_COMPARED and BASIC_NONPOETRY100_COMPARED).

I also created one feature from the whole corpus, instead of limiting the corpus by Poetry genre: amount of the most frequent words (BASIC_COMMON_WORDS).

The procedure for getting BASIC_POETRY50 was really simple: I took fifty most common words in full titles of the books classified as Poetry. From that set I excluded those words, which were in the list of fifty most frequent Non-poetry words. Further, I also excluded the English stopwords.[23] The only processing which was done on the word forms, was conversion to lower case. This was the default setting in `tm` package's `termFreq` function. Finally I counted the occurrences of any of the remaining words in the titles and thus bundled the sum as one single feature. The same method with reversed class values was used for extracting BASIC_NONPOETRY50; for BASIC_COMMON_WORDS I omitted only the stopwords and used the frequencies of all the words regardless of the class. I thought that the possibility of ruling out Non-poetry words was worth looking into.

To draw the other lexicon-based features I used a more complex method. For each word I counted their occurrences in Poetry and Non-poetry classes separately and divided the occurrences of Poetry words with the Non-poetry words to get an index for comparison. I disregarded all the terms with frequencies of hundred or less, and took fifty Poetry words with the highest comparison index. I summed up the occurrences of these words into only one predictor BASIC_POETRY50_COMPARED. The sums of words of the next one hundred indices after the previous fifty I gathered as BASIC_POETRY100_COMPARED. Similarly I gathered the BASIC_NONPOETRY50_COMPARED and BASIC_NONPOETRY100_COMPARED merely by reversing the classes.

The two methods to create the lexicon-based predictors described above

[23] List of stopwords was the default list in R's `tm` package.

might introduce some overlapping words between different features. Due to the relatively high number of accumulative word counts, I considered this as such a small problem, that I would not need to separate the predictors into different feature sets.

One note about the BASIC_COMMON_WORDS: usually the main motivation for using them in the author attribution and genre classification is, that as being the most common words in the vocabulary, they are less likely to be under conscious control by the author (Argamon and Koppel, 2010). In this particular case this is not necessarily so: the book titles almost certainly are under conscious control, if not by the authors themselves, by the book publishers trying to come up with book titles, that would attract people to buy the books.

Results

Inspection of the variable importance metrics show, that there was one feature above all the others: BASIC_POETRY50. This was hardly a surprise, as this feature was composed using the most common words from all the Poetry books, which were not present in Non-poetry.

Close behind came - according to *caret* and *conditional importance* measures - BASIC_CHARS, the number of characters in the whole title. The average length of Poetry book titles is approximately 145 characters, while the same figure for Non-poetry titles is 250 characters. It seems, that there clearly is a difference between Poetry and Non-poetry title lengths. All the word or character length related features did well in the measures.

Of the Poetry or Non-poetry words predictors, only the aforementioned BASIC_POETRY50, BASIC_NONPOETRY50 and BASIC_POETRY100_COMPARED were listed in the better performing half of the variables by the *caret* and *conditional importance*. Contrary to my previous assessment, BASIC_POETRY50 and BASIC_POETRY50_COMPARED (and similarly the corresponding Non-poetry features) might be too overlapping after all. This could be concluded from the fact, that the fifty most ordinary Poetry words were outperformed by the next one hundred. Also, BASIC_SENTENCES was ranked high.

I realized only after extensive qualification runs, that the method, which the feature was created with, counts ellipses (marked as three consecutive periods) as three sentences. An another possibly distracting element is the use of initials. I did not wish to redo everything, so I made a decision to divide the BASIC_SENTENCES into two separate features for the

actual classification: the actual sentences excluding ellipses and initials (BASIC_ACTUAL_SENTENCES) and the ellipses (BASIC_ELLIPSES. I figured the use of initials is in no way connected to Poetry, so I did not make a separate feature for the initials. A high or low amount of initials would be an indirect consequence of a title containing more names in general.

The features performing worst were part-of-speech related. BASIC_INTERJECTIONS and BASIC_FOREIGN_WORDS have practically nothing to do with classifying the books. Decent rates were achieved by BASIC_ADJECTIVES, BASIC_VERBS and especially BASIC_PROPER_NOUNS, which even reached top five rank with *conditional_importance*. The other predictors of this type were low ranking, although BASIC_PRONOUNS had mixed results. There will be a separate qualification round for part-of-speech (POS) tags, so that there will be a more thorough testing for them. The reason for having these features here in the BASIC set in the first place is, that I needed a compilation of different approaches to get a baseline for the performance. Besides, this way I was able to test the effect of groups of several tags (such as BASIC_VERBS instead of bundling them together with individual tags thus creating unnecessary overlaps in the predictor set. The possible overlapping between BASIC and POS features I will resolve later.

4.2 Qualification rounds

4.2.1 Selecting features from bunches

Perhaps the most arduous task in creating machine learning models is selecting appropriate features. As I already pointed out in chapter 2.1, there exists numerous studies incorporating different types of feature sets. I considered this specific task rather troublesome due to the unusually short snippets of available text, so I decided to try a whole lot of different feature sets and to collect the best of the breed for this specific classification problem.

4.2.1.1 Character-level features

Adopting the use of character-level *bigrams* from Kjell et al. (1994) for searching poetry from a library metadata catalogue is hindered by the fact, that the actual text in the title of the metadata catalogue is really short, and does not contain enough text for this kind of comparison. In case full texts of books had been used, the method would have been more plausible: rhymes and alliterations would probably be more pronounced in poetry than in other types of material. I have such a strong disbelief towards poetical book titles

manifested in unusual bigrams, that I did not wish to indulge in trying this method at all.

As stated above in chapter 2.1, amounts of punctuation marks has traditionally been used for genre detection, for example by Stamatatos et al. (2000a) and Kessler et al. (1997).

In the case of this poetry book research, there are only two genres: Poetry and Non-poetry. To be more precise: the genres are actually *poetry titles* and *non-poetry titles*, as the full texts are not available. Nevertheless, there is a possibility of a common style of poetry titles expected by the public and this style could very well be manifested, at least partially, in the use of punctuation in the titles. Besides, as this predictor type is relatively common on this field, I could not completely ignore it. The usability of this method would better fit full texts, rather than the short titles of the catalogues, as there simply is not that much punctuation in the short titles. Although the main title of the books is really short and often consist of noun phrases only thus lacking any punctuation altogether, the combination of main and subtitles usually do contain some. Also, non-poetry titles being a mixture of a whole bunch of unspecified differing genres, I consider it unlikely, that there would be common denominator for that class.

I have included the punctuation counts as one feature set, but even so, my hypothesis is, that they do not contribute much to the results. I dubbed this feature set as PUNCTUATION.

I gathered from the full titles the counts of all the punctuation marks: question marks, exclamation marks, colons, semicolons, commas, hyphens, single quotes, parentheses and periods; double quotes did not exist in the ESTC data at all. I decided to ignore the ellipses, for at least in ESTC ellipsis denotes omission: some text considered irrelevant or moved to a different MARC field has regularly been replaced with an ellipsis. The count of ellipses would still be a possible candidate as predictor, but it would not measure the count of a specific punctuation mark; instead it would measure the count of unsuitable passages in the titles. These omitted sections could be a lot of different things, such as lists of honorifics of the author, the date of thesis defense or publication time and place.

After collecting the punctuation counts I created a unified set from them without bothering to convert the integer counts into numbers relative to word count, because I did not intend to ever use the predictors independently, but instead always in addition to the basic feature set containing the word count as well.

Results
The results can be seen in **Figure 12**. Every variable importance measure has increased from the predictor set, which excluded the punctuation features. The rise is not great, only approximately one percentage point for each parameter category, but the gain is clear.

Closer look reveals, that features PUNCTUATION_QUESTION_MARKS, PUNCTUATION_EXCLAMATION_MARKS and PUNCTUATION_PARENTHESES are simply inferior to the basic and bag-of-words predictors. In fact, the first two of the aforementioned in addition to BASIC_INTERJECTIONS composed a trio of worst performing features by every variable importance measure I used with the exception of Mean decrease accuracy for Poetry class. The figures in that are insignificant however.

PUNCTUATION_PERIODS was the only feature of the PUNCTUATION features to rank really well within the set. `Caret` variable importance and `conditional inference` ranked all the rest of the PUNCTUATION predictors somewhere in the lower half. The other two predictors ever ranking by any of the other variable importance measures in the upper half were PUNCTUATION_SINGLEQUOTES and PUNCTUATION_HYPHENS.

I suspect these two are exceptions deriving more from the instability of these variable importance measures themselves than from the applicability of these features. I did include them in the final rounds anyway, except for the best performing feature, PUNCTUATION_PERIODS. That one I had to leave out, because it is essentially the same feature as BASIC_SENTENCES, for in the ESTC corpus there are only 3,361 records having at least one exclamation or question mark. Another way to look at this, is that the sentences almost always terminate with periods. For the actual classification step I have already dedicated BASIC_ACTUAL_SENTENCES and BASIC_ELLIPSES from the redefined BASIC features..

The other PUNCTUATION features were always evaluated somewhere in the lower half of the set and consequently will be eliminated from the final predictor set.

4.2.1.2 POS frequencies

The common usage of POS frequencies in the fields of author attribution and genre classification makes it almost an inevitable feature type to try. Some POS frequencies I already have used for the BASIC features. Whereas for BASIC features I followed the method of Biber (1992) with summing up the raw except for not using relative frequencies, I applied here the frequencies

relative to the word count of the title. Instead of repeating the unification of several features, I decided to create a completely new predictor set for singular POS frequencies only, similarly in the style Rayson et al. (2002) had in their research. Their good results in predicting imaginative and informative writing provide optimism on the usability of POS tags of predictors. It must be noted, however, that the distinction between informative and imaginative writing is not the same as the distinction between poetry and non-poetry; there are imaginative genres within non-poetry as well. Yet another point to remember is, that Rayson et al. had full texts as their corpus, while the corpus of this thesis is composed only of titles, in which one often encounters incomplete sentences.

There is a vague hope, that some imaginativeness has influenced the poetry titles too, which in turn would be reflected in the POS tag frequencies. Of primary interest are the POS tags, which are more frequent in imaginative than in informative writing, as the frequent tags within informative writing (eg. conjunctions and determiners) are bound to be missing or at least more scarce even in the individual non-poetry titles. This is due to the tendency of having titles, that would be ungrammatical as sentences. I made a separate preliminary qualification round with all the POS tags to test the potential of them.

Part-of-speech (POS) tagging was done using the combination of three R packages: `tm`, `NLP` and `openNLP` (Feinerer and Hornik, 2017; Hornik, 2017, 2016, respectively). First, the corpora were created by `tm`, and after that split to tokens with `openNLP`. `NLP` was then used with default settings to do the actual annotation. I had two separate versions of POS tag sets: one extracted from the whole title (that is, the main title concatenated with the subtitle), and another one with the main title only. There were 42 different POS tags in the corpus and all of them were gathered into the feature set.

I gathered all the occurrences of POS tags within the titles divided by the word count of the title into the POS feature set, whose features I prefixed with POS. The counts of occurrences were divided by the word counts of the titles.

The qualification runs for sieving the most prominent POS tags were done using `randomForest` package (Liaw and Wiener, 2002), with *ntree* value 250. There was a total of four test runs: with POS values from the main title only (MARC field 250a) with *mtry* values 5 and 10, and POS values of both the title fields (245a and 245b) having the same *mtry* values.

Results

The overall results were moderate at best, as can be seen in **Figure 1**. A clear sign can be seen: POS tags, which were extracted from the whole title, were more informative than those which were extracted from the main title only. This applies for both *precision* and *recall*, as well as the *H measure*.

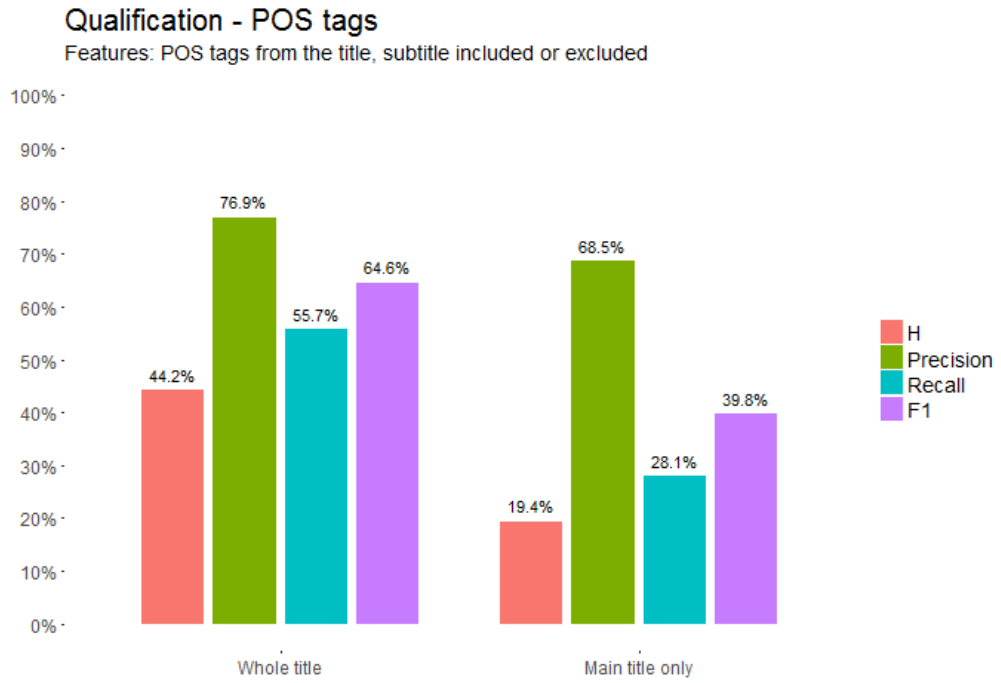

Figure 1: Results of POS tagging features

As a consequence, I completely ignored the feature set, which had been extracted using only the main title.

It is evident from **Figure 2**, that different *mtry* values have no significant effect on the relevant values. Therefore I ran all the variable importance checks by using only a single *mtry* value without even trying the other one. In this particular case I chose *mtry* value 10.

The better of the two feature sets - the one from concatenated main and subtitles - I examined closely to obtain the most significant features within it. All of the several variable importance measures lifted up almost precisely the same subset of most relevant features.

I qualified the features with *Mean decrease accuracy* over all classes (*MDA*) or *Mean decrease gini* (*MDG*) value over the threshold of 100. The other measures, which I used, were *conditional importance* with a threshold of over 1.5%, or the `caret` *variable importance* of threshold value of more than 1.0%.

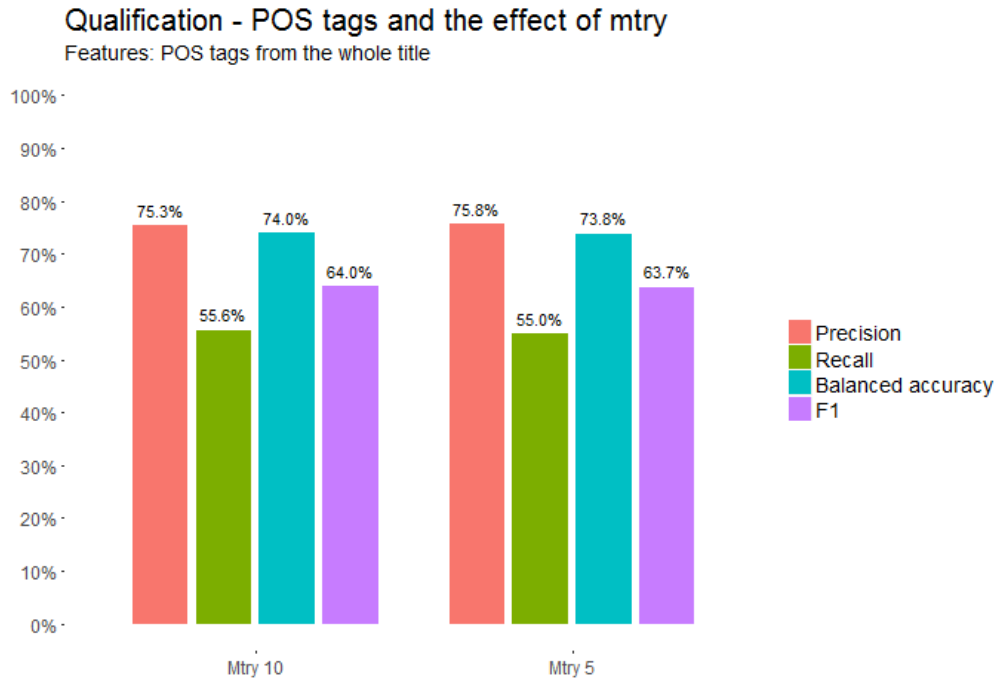

Figure 2: Results of POS tagging features: effect of mtry

Even though the thresholds themselves were arbitrarily chosen as they were neat round numbers or figures, there is not much variation in the top feature lists. The thresholds were able to separate the same subset of features almost always.

Running variable importance checks on the feature set revealed easily six features protruding above the others: POS_IN, POS_DT, POS_NN, POS_NNP, POS_COMMA and POS_JJ.[24] All these features had values over the specified thresholds in *conditional importance*, `caret` threshold and *MDG*. The threshold of *MDA* was reached only by POS_IN, while the *Mean decrease accuracy* for individual class values (that is, for Poetry and Non-Poetry) were so low, that I ignored those results completely.

Second wave of best performing POS features were POS_PERIOD, POS_CC, POS_CD and POS_NNS. These feature values were over the *MDA* and `caret` thresholds, but not the *MDG*. As described in chapter 3.4.3, the two measures are supposed to be alike. There was also one additional feature above the *MDA* threshold, namely POS_VBN, and another one above the *conditional importance* threshold, POS_TO.

[24]See **Appendix B** for the explanation of the part-of-speech tag abbreviations.

In total, there were six undisputed features to qualify to the next rounds, and six additional values, which were picked by one or two variable importance measures only.

4.2.1.3 POS trigrams

Part-of-speech *trigrams* have often been used as features in author identification and genre classification (see chapter 2.1. However, the titles in the ESTC are really short, and as such they do not contain many useful POS *trigrams*. POS *trigrams* being rare do not possess high prediction capability. I seriously doubt the claim by Argamon et al. (2003) to POS *trigrams* being able to predict the genre. Despite my concern, I made an additional qualification round on POS *trigrams*, so that I could test the possibility of adding at least some of them into the feature set for the actual classification.

I simply selected to prefer the punctuated POS *trigrams* over those without the punctuation, as my premise is, that the results would be reasonably equal, although not optimized. Repeating the tests between the versions including and excluding the punctuation by Santini (2004) would have not brought significant enhancement for the result. I barely note, that there is a distinct possibility, that there might have been a slighly better result attainable, even though that is improbable.

I used the POS tags (described in previous paragraph) as a base to form the POS *trigrams* from. First I selected the most common *trigrams* from the records classified as Poetry and then created two subsets of features from: one for POS *trigrams*, that are extracted from the main titles only, and another one for those extracted from the concatenated main and subtitles.

For the feature set extracted from the whole title I filtered out the POS *trigrams*, which occurred less than 2000 times. This resulted in a set of 89 predictors. For the main title features I used a smaller limit of 500 occurrences, as the total amount of words and hence also of POS *trigrams* is considerably smaller. The feature set contained 108 variables.

To avoid computing the POS *trigrams* on-the-fly at the training stage I did this gathering of features using the whole training set. This might introduce bias favoring the use of POS *trigrams* as features.

Results

Figure 3 shows clearly, that POS *trigrams* derived only from the main title had poor performance, especially when they are compared with the POS *trigrams* from the whole title. The same phenomenon applied to POS tags

as well. Also similarly to POS tags, the change in *mtry* value has little or no value at all for the performance (see **Figure 4**). Again, I gathered the variable importance metrics using *mtry* value of 10, for no particular reason ignoring *mtry* 5. The thresholds I used were the same as earlier with the POS tags.

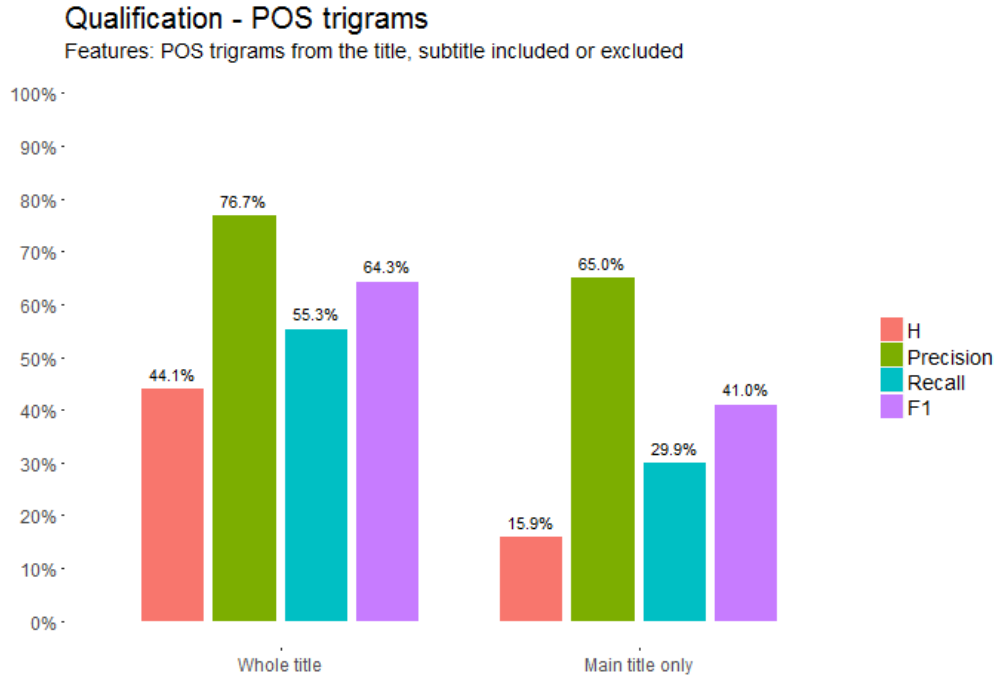

Figure 3: Results of POS tagging features

The two POS *trigram* features that emerged from *MDG*, *conditional importance*, and `caret` *variable importance* were POS3GRAM_IN_DT_NN and POS3GRAM_PERIOD_TO_DT.[25]

Neither *MDA*, or either of the class-value specific measures reached the threshold level. *MDG* and CARET *importance* shared five features: POS3GRAM_IN_NNP_COMMA, POS3GRAM_IN_DT_NNP, POS3GRAM_NN_IN_NN, POS3GRAM_IN_NNP_NNP and POS3GRAM_DT_NN_IN.

Furthermore, there were seven more features, which exceeded the threshold of one of the variable importance measures, but not of the others.

A comparison between POS tag features and POS *trigram* features shows, that the measures are disturbingly similar (**Figure 5**). This might indi-

[25]See **Appendix B** for the explanation of the part-of-speech tag abbreviations.

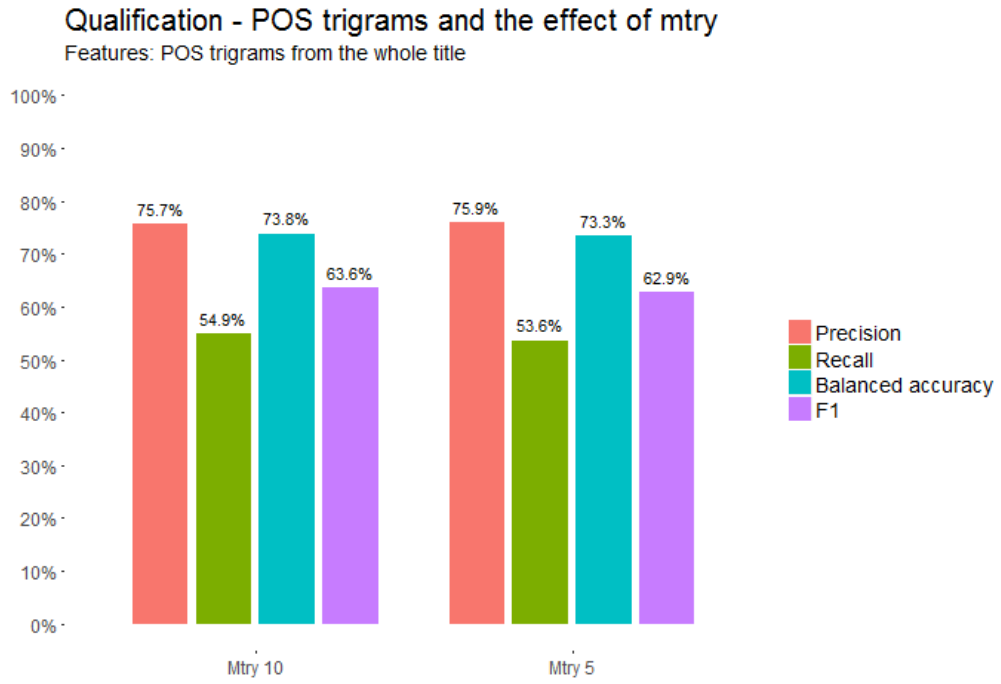

Figure 4: Results of POS tagging features: effect of mtry

cate that there is some limit, beyond which the sets comprised only of POS tag derived features can not tread without the support from other types of features.

As it is, there is an error margin in using the part-of-speech tags: the assignment of POS tags to words has been reported to go over 95%. (Argamon and Koppel, 2010) To achieve the figures that high, the sentences would most certainly need to be complete.

It is notable, however, that those two feature sets still performed better than a random sample would have: the *prevalence*, which measures exactly that, of 29.2% was cleared by all of the measures.[26]

From the POS *trigrams* qualification round I collected the seven most outstanding features for the final rounds.

4.2.1.4 Bag-of-words

Probably the most common features sets in author identification and genre classification tasks have been bags-of-words (see chapter 2.1).

[26]The *prevalence* is not shown in the picture. It is calculated by dividing the amount of positive classes (in this case, Poetry) by the amount of the total records.

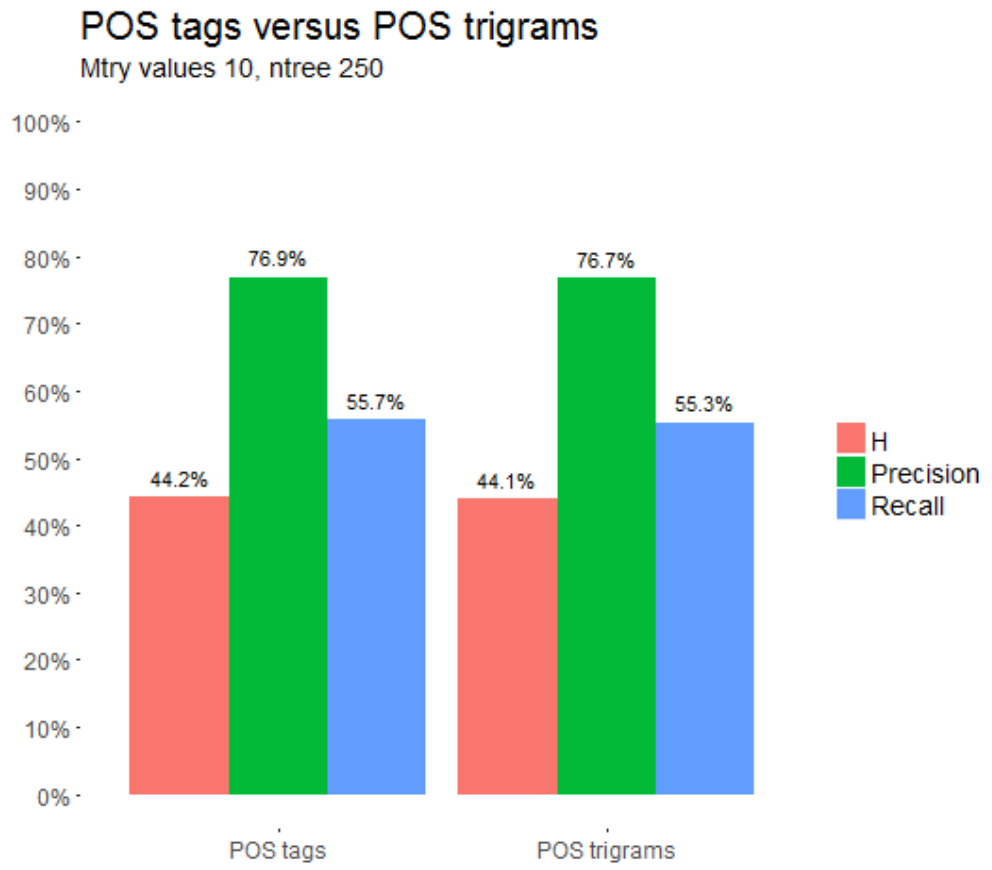

Figure 5: Comparison of POS tags and POS trigrams

However, the bag-of-words approach implemented in the same manner as Joachims (1998) and Aizawa (2001) I deemed unsuitable. First of all, I felt discomforted by the huge amount of features and consequently of processing power required for them. Secondly, I suspected the scarcity of the words in singular titles would make the feature values created from over 235,000 unique words in titles mostly zero, which would lead to the infamous *curse of dimensionality*. The same thing applies even more so to the method of drawing bag-of-words from a substantially larger corpora, as done by Stamatatos et al. (2000a).

As the full texts of the books are not available, the words can be extracted only from the main title or the full title (containing the main title and the subtitle). With severely restricted length of texts it surely is better to focus on words, that are known to exist in the corpus, than on words that are known to exist in the universe. For this reason I decided there was a need

to preselect the best words to be used as features, so that the amount of features would be at minimum. Even if the predictability would benefit from additional features, as stated by Aizawa (2001), there was a threshold, after which the improvement would be minimal. (Stamatatos et al., 2000a)

While a large feature set would make the usage of bag-of-words impractical, a small one would make the likelihood of a bag-of words method standing on its own without the support from the other types of features at all very small. At no point was my intention to find an optimal feature set comprising only of the frequencies of the most predictive words. Instead I aimed to find the best words and to use them in addition to the other types of features.

I tested three different approaches to determine the most significant words for predicting poetry titles. The prefix I set for the features in the testing was BOW. I was able to use the same prefix for all of them, as the feature sets were never to be used in conjunction of one another.

1. Method 1

   My first attempt to capture the most common words in specifically poetry book titles was to extract the n words from the titles labeled as Poetry and not included in the most m common words of Non-poetry book titles.

   Because the titles are short, possibly only one word long, even simple lemmatization might easily introduce errors: there simply is not enough context. For this reason I did not wish to reduce the words into lemmas, but simply extracted the word forms from the titles. The only preprocessing step I did was conversion to lower case.

   I used both the main title and the subtitle to draw the words from, just to get more base material to extract the words for the feature set. In order to test multiple thresholds I created multiple feature sets using different sizes of feature sets.

   In the **Figure 6** can be seen, that the more predictors there are, the worse the results become, the peak being at $n = 50$. Normally, when using random forest, the higher number of predictors should indicate also better results. The reason for this happening is simple: there is a flaw in this method. The size of the exclusion list should not be tied to the size of bag-of-words itself. As the number of predictors grows, also the number of words in the exclusion list increases. The best Poetry words were rejected, because they were common enough in Non-poetry as well. It is interesting to note, that the figures tend to stabilize after

bag-of-words size 100. This is most likely an indication, that the words introduced after that are really rare. The more common words have been excluded, for there is a slight drop in the figures.

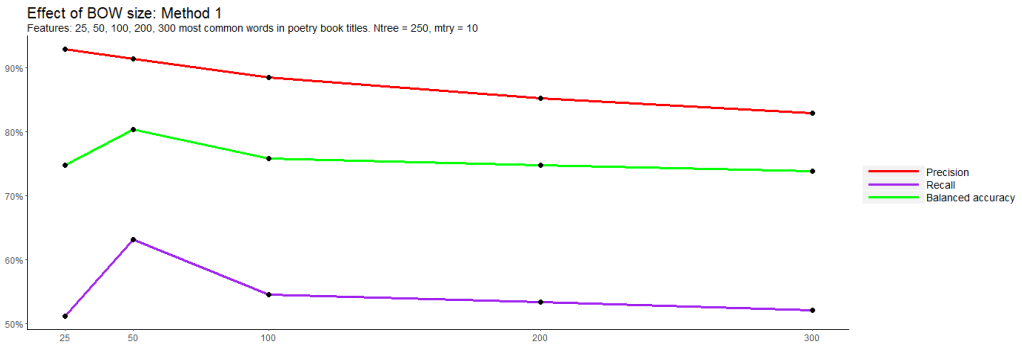

Figure 6: Effect of bag-of-words size

2. Method 2

   Shortly after doing the qualification rounds using the first method described above, I began thinking, that the way I had been eliminating the words was definitely suboptimal. Many of the most common Poetry words were rejected, because they were included in the most common Non-poetry words as well. The trouble with that method was, that it resulted in getting relatively rare words for the data set. Therefore I tried another approach, which took the word frequencies within dissimilar classes into account.

   First I extracted all the most common words from the titles of books that were classified as Poetry in the corpus. As I had no idea of how many predictors would be required for proper analyses, I created several different feature sets and tested them all.

   After deciding the amount of variables for a set I converted the titles into lower case, but did not do any other preprocessing. I then extracted all the letter sequences, which were separated by one or more non-alphabetic characters and sorted them in frequency order from most to least common into a table, excluding those, with less than three occurrences. I got the frequency of each word by dividing the number of occurrences by the number of all the words in the titles.

   I repeated the same process to create an exclusion table from all the titles to weed out those words, which are the most common in books classified as Non-poetry. I deselected any words from the first table having a frequency less than three times the frequency of the same

word in the exclude table. From the resulting words I took the n most frequent words according to the original decision.

The words were extracted from the main and subtitles, just the same as with the first method.

Figure 7 shows more typical behavior of increased feature set size than described above for the first method. The *recall* and *balanced accuracy* clearly increases as new predictors are introduced, although there is a steady drop in *precision*.

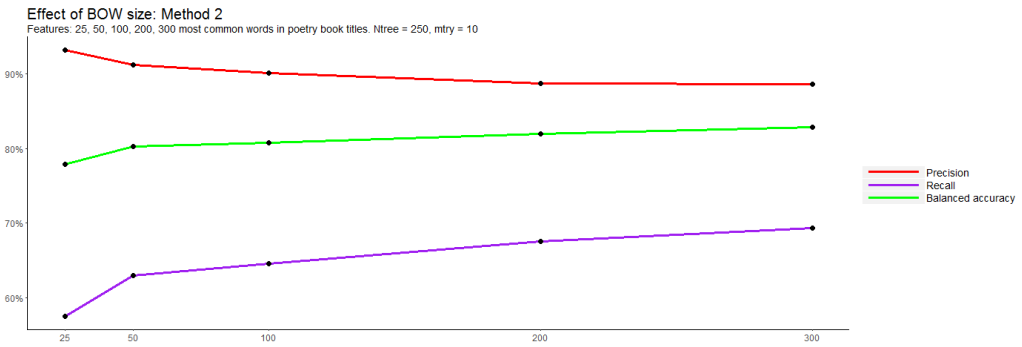

Figure 7: Effect of bag-of-words size

3. Method 3

   This method was the same as method 2, with the sole exception of the features being extracted from merely the main title instead of the full title. **Figure 8** indicates a huge increase in all the main measures (*precision*, *recall* and *balanced accuracy*) as the number of features are increased. A comparison to the other features suggest however, that this is due to *overfitting*, as the same phenomenon occurs only for the method with the least amount of words to draw the bag-of-words from.

A comparison between the three methods shows unambiguously their mutual ranking. **Figure 9** indicates, that the second method outranks in every respect the other two, while the third method is always the weakest.

Because the results were so clear between the three methods, I decided to qualify features only from the feature sets created by the second method.

I used the same variable importance measures and thresholds as previously for the POS tags and POS trigrams: MDA for the individual classes or combined above 100, MDG 100, `caret` importance 1.5% and conditional importance 1.0%. I did not measure variable importance for bag-of-words

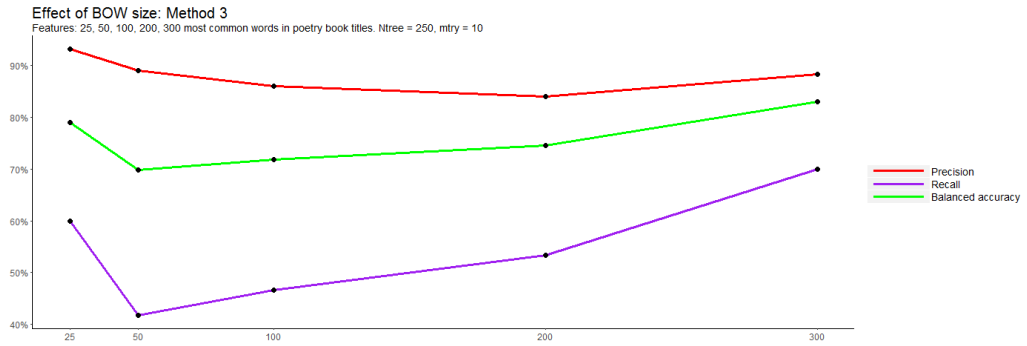

Figure 8: Effect of bag-of-words size

size 300, as I suspected it might be useless, since the most relevant words are already contained in the other feature sets. One must note, that the second method omits words only by their relative infrequency, that is, if the words are not frequent enough in Poetry titles as compared to those in Non-poetry titles. The ratio of individual Poetry and Non-poetry words remains the same in the corpus, no matter what the size of bag-of-words is. Therefore increasing bag size only adds new words to the feature sets, but does not remove the previous ones. That said, the effectiveness of an individual word might be different within different feature sets.

Results

There were three predictors above the rest: BOW_TUNE, BOW_POEM and BOW_SONG. They were always in the group of three most significant predictors and they exceeded the thresholds in all the measures. The next three exceeded at least four of the thresholds, regardless of bag-of-words size: BOW_POEMS, BOW_SONGS and BOW_SUNG. As noted before, I did not lemmatize the words, so these word forms are separate from BOW_POEM, BOW_SONG and BOW_SING. Other commonly occurring features above the thresholds were BOW_BALLAD, BOW_ELEGY, BOW_ODE, BOW_POETICAL, BOW_GARLAND, BOW_LAMENTATION, BOW_VERSES, BOW_HYMN and BOW_HYMNS.

Barely a couple of times over a threshold were BOW_MAID, BOW_PSALMS, BOW_EPISTLE and BOW_LOVE.

From this qualification round I took a set of 18 predictors: all those, which had been over the threshold at least a couple of times, with the exception of BOW_MAID. I considered it too infrequent in this material to be able to make a difference for the predictions: the word *maid* appears in the titles in

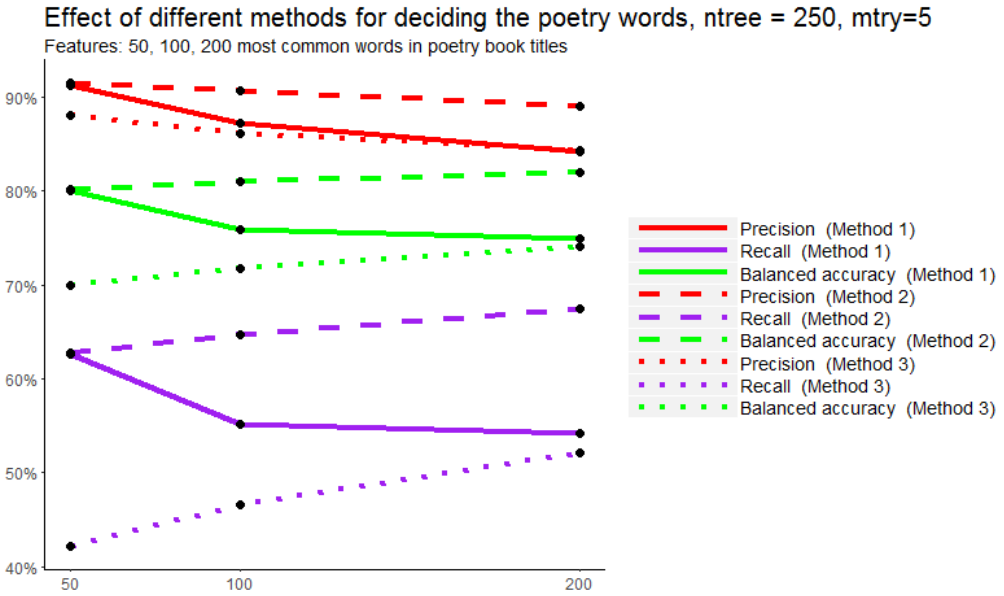

Figure 9: Effect of different methods in determining poetry words

its singular form merely 1,426 times.

However, I did use BOW_MAID in the qualification round as the 19th bag-of-words predictor together with the BASIC features, to test the effect of features, that were insufficient to form a feature set on their own. (See Chapter 4.2.1.7).

4.2.1.5 Dependency relations

To get a bit more syntactical point of view of the poetry titles for my research I decided to do a little test with features connected to dependency relation.

I experimented with R's `coreNLP` package (Arnold and Tilton, 2016), which is a wrapper package built around Stanford CoreNLP Java Library (Manning et al., 2014) and tried to pick some low-hanging fruit from dependency parsing (Arnold and Tilton, 2016; Manning et al., 2014; De Marneffe et al., 2006). Basically, I took everything, which I could reasonably well turn into either a numerical value or a factor containing only a decent amount of classes. There were not many features I came up with this way. I simply used `coreNLP`'s `annotateString` function on the concatenated main title and subtitle with default English language settings and modified as little as possible for them to be usable as predictors. Manning et al. (2014) Besides that, I tokenized the annotated string with `getToken` function.The `getToken` also lemmatizes

and performs part-of-speech tagging. (Toutanova et al., 2003)

The predictors I acquired through this procedure were the number of root elements (DEPREL__NO__OF__ROOT), the number of dependents of the first *root* element in the title (DEPREL__NO__OF__DEPENDENTS), the number of sentences in the title (DEPREL__NO__OF__SENTENCES), part-of-speech value of the first *root* element (DEPREL__ROOT__POS), the distance of the first *root* element from the beginning (DEPREL__ROOT__OFFSET__CHARACTERS) and the number of inflected words in the sentence (DEPREL__NO__OF__INFLECTED).

`CoreNLP` also offers named entity recognition (NER), but I found this was too slow and provided too scarce and arbitrary results at best on this particular data set and thus I deemed those features unsuitable for my purposes, although location and person names would have been an interesting predictor set. (Finkel et al., 2005).

Sentiment analysis in `coreNLP` seems to be on experimental level (Socher et al., 2013): at least the results are even more random than with NER, which was the reason I had to abandon that as well.

First results

Results of the qualification round with dependency relations were disappointing. Comparing the measures in **Figure 10**, one can see, that the measures were the weakest of all for this feature set. It must be noted, however, that the predictor bunch was also easily the smallest one, with only seven variables in all.

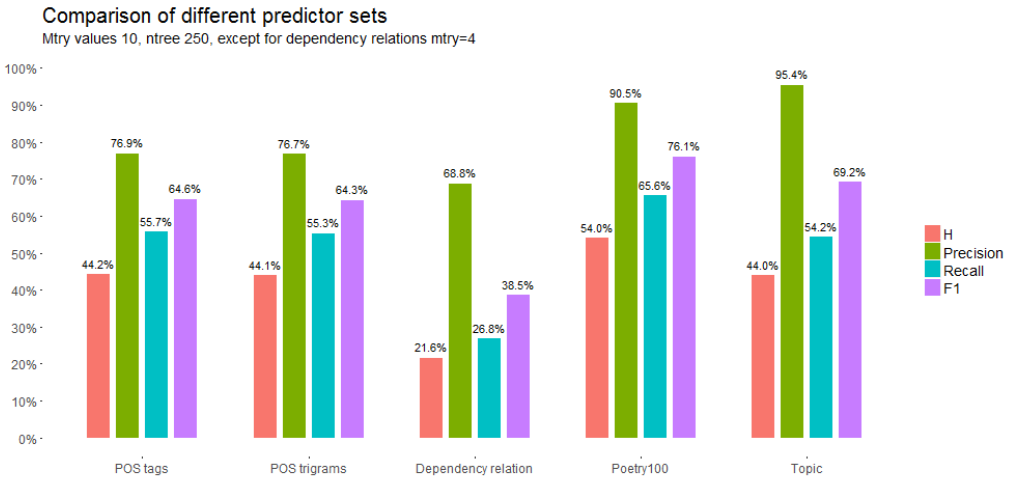

Figure 10: Comparison of qualification rounds

The results for dependency relations are disturbingly low, as the *H measure* is only barely over 23.3% and *recall* (25.4%) is even lower than the baseline (29.9%) of a completely random sample. This is not the whole truth, as the *F1* measure does exceed the baseline, which means that the *precision* compensates for the negative *recall*.

Variable importance measures with dependency relation features does reveal some features being better than the others, but I preferred not obeying any explicit thresholds, because the predictor set was so small. Conditional importance suggested DEPREL__NO__OF__DEPENDENTS, DEPREL__NO__OF__ROOT, DEPREL__NO__OF__INFLECTED and DEPREL__ROOT__OFFSET__CHARACTERS more important than the other predictors, and the other measures besides Poetry MDA did not contradict this. The MDA for Poetry class however considers only the DEPREL__NO__OF__ROOT, DEPREL__NO__OF__INFLECTED and DEPREL__NO__OF__SENTENCES as significant at least to some extent. The figures were however really low, which raises some suspicions to the power of these features actually finding poetry books at all.

It must be noted, that DEPREL__NO__OF__ROOT was listed as the primary feature by all the measures.

Considering that the result was remarkably low for this qualification round I decided to conduct a secondary qualification round for dependency relations with the addition of basic and nineteen selected bag-of-words features to see if the dependency relation predictors actually contribute to the results or not.

For this test I created two new subsets from the dependency relations: the first containing the four predictors suggested by conditional importance measure (NLP4) and the second containing only DEPREL__NO__OF__ROOT (NLP1). I also took the original NLP set containing all the tryout predictors (NLP7) into the comparison.

Second results

Figure 11 shows, that the set with most predictors, NLP7, performed worse than the reference point without NLP features at all. Therefore I skipped the variable importance measure for NLP7 set completely, but instead performed it separately for NLP1 and NLP4, in combination with the basic and bag-of-words features. NLP1 and NLP4 have more or less equal *Balanced accuracy* and *F1* values than the stripped feature set. Variable importance measures place DEPREL__NO__OF__ROOT relatively high: `caret` measure as ninth and

conditional as sixth by both NLP1 and NLP4 analyses. The other measures regard it as mediocre, but definitely not a low performer.

The rest of the DEPREL measures were ranked in the upper half as well, DEPREL__NO__OF__DEPENDENTS being easily the next highest after DEPREL__NO__OF__ROOT.

The low increase in the outcome with the addition of three decent to good predictors suggests that there might be some issues with the original feature set. The lowest performing POETRY features have little to no significance. Anyway, as for the DEPREL predictors, I decided to include the four in NLP4 for the actual classification task.

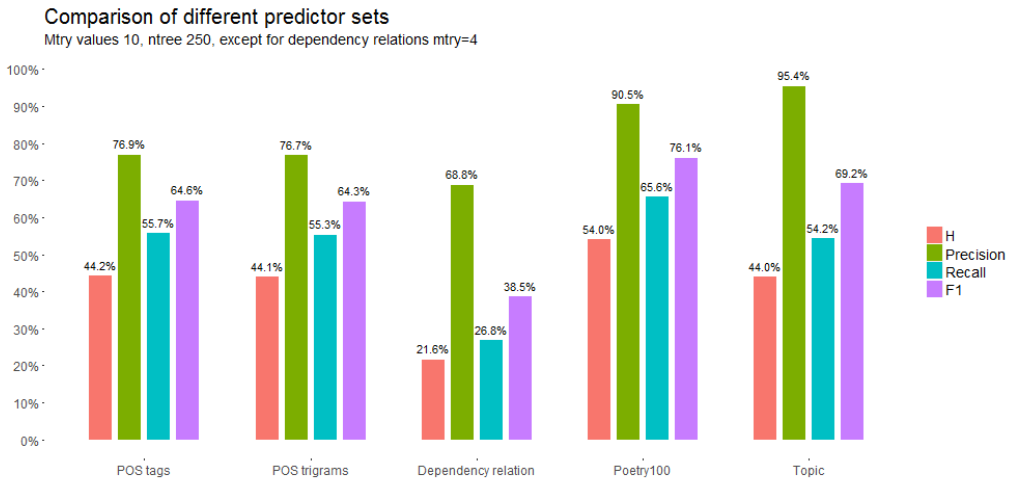

Figure 11: Comparison of NLP feature set variants

4.2.1.6 Topic

A topic annotated by a librarian is much more reliable indicator of the record's topic than a keyword in the titles. MARC field denoting the topic *Subjecttopic* (650a) had been annotated for 266,188 records, comprising 55.1% of total. I added *form subdivision* annotations from three *Subject added entry* (6XX) subfields suffixed with letter 'v' of *Uniform title* (630v), *Corporate name* (610v) and *Personal name* (600v), iteratively to the topic in this order, if the topic was missing. Despite its name, *form subdivision* is often used for annotating topic, when its parent field denotes subject instead of topic. This way the annotated record count was risen to 310,110 (64.2%).

I applied the Method 2 of the Bag-of-words feature selection process for topics, except for treating the topical phrase as the measurable unit instead

of singular words. Therefore each feature could only have a value of zero or one, as each individual predictor could be represented only once by a single record. Each record could have multiple topic values, though. Also, I limited the feature set to include only one hundred most common Poetry topics and prefixed the features with TOPIC.

Results

The results compare well to the other feature sets (see **Figure 10**). The *precision* value is high, which is easily explained by the fact, that the existing topic values have been manually annotated for each individual book. The relatively low *recall*, comparable to *recall* values of POS and POS trigrams sets, is likely to be due to most of the Poetry topics having only a small amount of books annotated onto them. In other words, when a book is assigned one of the features from the set, the predictability is usually really good.

I used the same measures and thresholds for variable importance as for the other predictor sets in the qualification round. Three of the predictors were the most relevant by every measure: English poetry (TOPIC_ENGLISH_POETRY), Ballads, English (TOPIC_BALLADS_ENGLISH) and Songs, English (TOPIC_SONGS_ENGLISH), a fourth variable Poetry (TOPIC_POETRY only barely missing the *conditional importance* limit. These four predictors together cover 17,999 records (5.8% of the annotated and 3.7% of the whole catalogue). This means, that these predictors are useful in classification for only a fraction of the records. Adding subsequently the next seven predictors, increases the coverage to 22,231 records (7.2% of the annotated and 4.6% of the total). It is easily deducible, that the four best predictors also have the most numerous occurrences. This is also logically sound, as the high level of accuracy means that the variables having most values would also be the most predictive ones.

The cumulative coverage of the seven next predictors is thus 4,232 records. I chose to ignore these seven completely for the final round, as even combined they possess the potential of identifying less than a percent of the whole catalogue. Similarly, I dismissed all the subsequent, even more rare, topic features. For the final rounds I took only the four most prominent ones.

4.2.1.7 Other features

There were also some unique features and feature sets, which were too few to conduct a meaningful qualification without the addition of other features.

In order to test them I used the basic features and the nineteen bag-of-words predictors as a base level, onto which I appended the qualifying features in separate steps. Also, I did variable importance check to see, how the additional features ranked within the feature sets. I did not blend these unique features with each other to create a truly diverse group of predictors, so that in this qualification round I would be able to judge their performance to by their effect with the basic and the bag-of-words collection on a feature set level.

Author

MARC field denoting the personal name of the author (100a) is a very potential source for poetry book detection: an author, who has published one lyric book, is likely to have published more. At first I was reluctant to using this kind of feature as it might seem like cheating, but I thought this over and decided for the inclusion anyway. In many cases the same author has written different types of books: it is not uncommon for a poet to write novels or plays as well, nor for a playwright to write poetry.

This is the only feature, which I decided to check on-the-fly, because a poetry author list is possibly too powerful to be drawn from the poetry book authors. The list would have potential of generating too biased feature sets, if it contained every author of the known poetry books. For the purpose of understanding the predictive power of the author, I created a separate poetry author list for each subset in the cross-validation procedure.

One must keep in mind, that for the actual classification task on the unannotated material, this feature is not as predictive as with the relatively small subset of the catalogue, which contains only the annotated records: at the qualification step, the share of records used in modeling is 80% of total amount for each cross-validation round, but in the actual classification the share will be slightly above 25%. For the qualification step this amounts to approximately 18,000[27] unique authors and the final rounds 21,859 unique authors. The total amount of unique authors in the testing set however increases from approximately little below 4,000 authors to 45,449. In other words, during this qualification round most of the poetry authors are already known before the prediction, but this is not the case in the final rounds when prediction will be made for the part with the unannotated genre.

I added a prefix VARIA for this predictor, so it could be easily seen as a stand-alone feature and not belonging to any set with similar characteristics.

[27]This estimate is based on a random 80% sample of the records having genre annotated.

VARIA_AUTHOR was the best performing variable with a clear margin in the feature set as indicated by all the variable importance measures besides *conditional importance* and *Poetry MDA*; those two ranked the feature as sixth and third, respectively. As stated above, I suspect the performance of this measure to drop somewhat on the final round, but still it should be among the most predictive ones.

Antique names

In my first attempts to understand the catalogue contents I noticed a potential difficulty in catching certain types of book titles: short titles containing proper nouns. I hypothesized, that this might possibly be overcome by having a list of proper nouns, that are often used in poetry. Common sense tells, that the most frequent names might be taken from Antiquity or Arthurian legends, or then again, maybe from the British folklore.

After a quick glance at the book titles with Arthurian names I opted not to use them at all. It appeared, that *Merlin* was easily the most common of the names I tried in the titles, but it did not correlate with poetry strongly. *Rider's British Merlin* was an almanac with astronomical and astrological information and *Merlin* was used as a personification to almost anything relating to future and prophecies.

On the other hand, I could not detect the same phenomenon in the proper nouns from the Antiquity. I searched for as comprehensive list of names used in epics of Antiquity as possible, and found an online version of an English translation of Ovid's Metamorphoses including a mythological index. Metamorphoses is a collection myths containing the majority of the most common names from Greek and Roman pantheon. However, some addition was required, as many of the heroes are missing from the index. Also, the glossary often provides the names only with the Latin orthography, as the original text from Latin had already transformed the Greek names to Latin equivalents. Therefore I joined an mythology index with an online Iliad to the Metamorphoses index. (Ovid, 2000; Homer, 2009)

I extracted all the proper nouns from the index and manually removed known homonyms, which might exist in English book titles even without any connection to Greek or Roman literature, such as planet names and anglicized impersonations like *Furies*, *Graces* or *Dawn*. This way I ended up with 2,341 unique names from Antiquity, although most of them are definitely way too obscure to have entire book collections based upon.

Finally I summed up all the occurrences of any of the names for each record

into just one variable. Still this variable has a non-zero value in only 2,378 cases, which means its potential for identifying anything is very low, especially when the same themes have been used extensively in plays and prose as well.

Nevertheless, I made a quick test combining the Antiquity predictor with the bag-of-words and basic features to see, if there actually was any effect on the results. I dubbed this feature as VARIA_ANTIQUE in a similar fashion to VARIA_AUTHOR, as this too is a stand-alone feature and decidedly different from the author features. I had hopes, that this particular predictor could resolve some cases otherwise doomed as impossible to detect. Figure **12** shows in the values only such a small increase, that it is not clear, whether the gain comes simply from adding a single variable more.

The variable importance measures rank the VARIA_ANTIQUE variable somewhere in the mid-range: not the worst and definitely not the first. Due to the extremely low occurrence rate of the Antiquity non-zero values, I decided to let go of the feature anyway. It will not be included in the final rounds.

Other MARC fields

As interesting it would be to see, if there was any correlation between author's age and himher writing and publishing poetry, the low annotation level of relevant MARC field describing the time the author lived or flourished (100d) will not allow that. The same field would be required to determine the posthumousness of a work.

Two other fields, however, are frequently annotated. The MARC field depicting page count: *Physical extent* (300a) is almost always filled and so is the one which tells the size of the book *Physical dimension* (300c). From these two fields I extracted features MARC_PAGECOUNT and MARC_SIZE, respectively.

The book size and page count are more or less standard for poetry: one does not expect regular books of lyric to contain hundreds of pages or for them to be folio size.

Figure 12 shows the capability of the MARC fields. In comparison with no additions at all, all the scores have risen. The two measures, which depict the overall result, *balanced accuracy* and *F1*, have both gained over two percentage points, *F1* over three and a half, just by the inclusion of two additional features.

Comparing the variable importance measures of this run places MARC_PAGECOUNT in the top-performing group: `caret` and *MDG* as high

as the second best predictor. MARC_SIZE gets situated within the ten most predictive features (with the sole exception of *Poetry MDA*). As this set has already been putrefied from the poor features, these rankings can be considered really high.

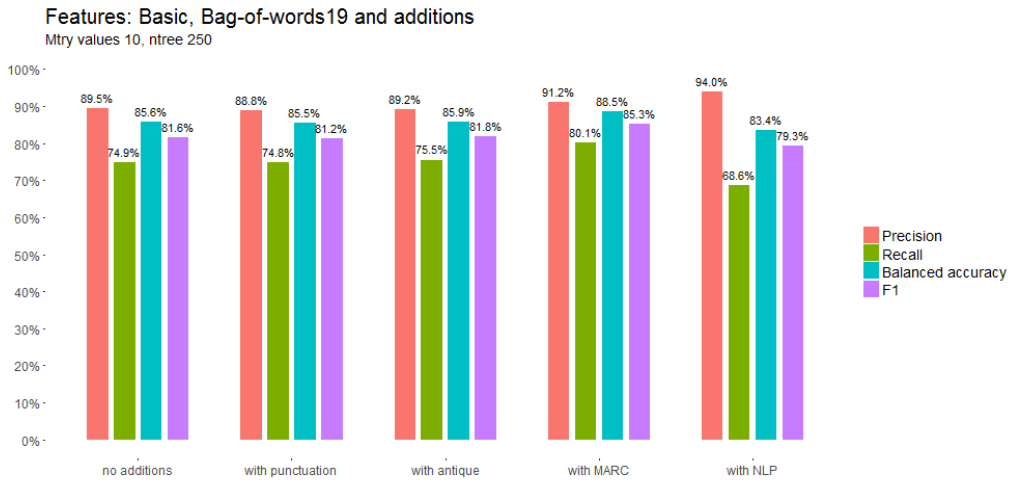

Figure 12: Comparison of other feature sets

4.3 Semifinals

The phase, when the ultimate feature set was already decided, but still some additional testing was to be done, I called Semifinals. First I made some testing with the random forest parameters and then with trying several redefinitions of the responses.

4.3.1 Finding the optimal settings

4.3.1.1 Effect of mtry on the final feature set

After finishing the final feature set I wanted to discover the optimal random forest settings for the feature set in order to see the optimal results. For the qualification round I had tried *mtry* values five (5) and ten (10) for some of the test sets. The default value in `randomForest` package is five, but the sample tests showed slighly better performance with a value of ten. I used the value of ten as an assumption, but also added an extra classification round using different *mtry* values with the final feature set in order to discover, if there actually was a more optimal value available.

As can be seen in **Figure 13**, there actually was. The most important parameters (*balanced accuracy* and *F1*) have been highlighted. The highest values happen to occur with *mtry* value eighteen (18). The difference between values, however, is marginal: all the *balanced accuracy* and *F1* results from *mtry* values of fourteen (14) and upwards fit within one tenth of a percentage point range (92.33%-92.42% for *balanced accuracy* and 90.43%-90.53% for *F1*). None of the *mtry* values beyond five gives results farther than one percentage point from the maximum.

In other words, the *mtry* parameter is not really that important for the results; nevertheless I decided to stick with the highest numbers.

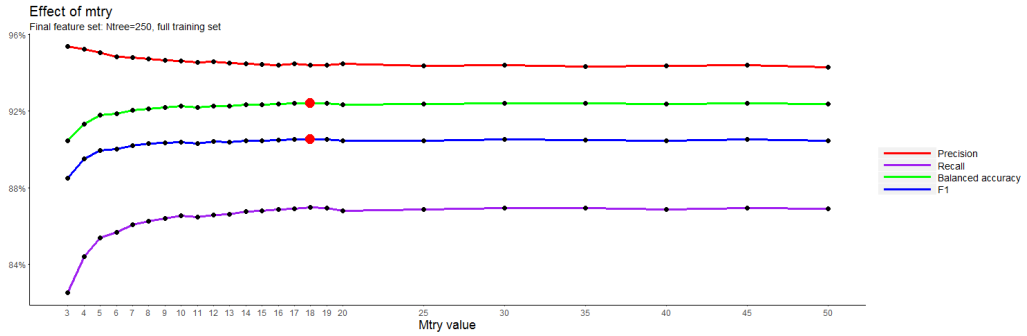

Figure 13: Results of final feature set: effect of mtry

4.3.1.2 Effect of ntree on the final feature set

The other `randomForest` parameter to test was *ntree*. The default value is 500, but I had used value 250 during the qualification rounds to cut out the processing time, as there were quite a lot of different feature sets to test. *Ntree* means the amount of decision trees the function is growing, from which an average performance is calculated. Basically it means that a high *ntree* value is supposed to only stabilize the results, so that there should not be peaks in the results. There may be achieved a bit higher values with lower *ntree* values, but these are anomalies from the normal performance. Another round with the same feature set and the same *ntree* value might produce a peak in the opposite direction. As the value is grown higher, there will be peaks less and less often. Therefore a higher *ntree* value should be kept at as high a level as is reasonably possible.

In **Figure 14** the highest *balanced accuracy* and *F1* values occurred when using *ntree* values 150 and 500, respectively. Similarly to *mtry* values, all the

ntree values above 150 produce results, which all fit nicely within a tenth of a percentage point.

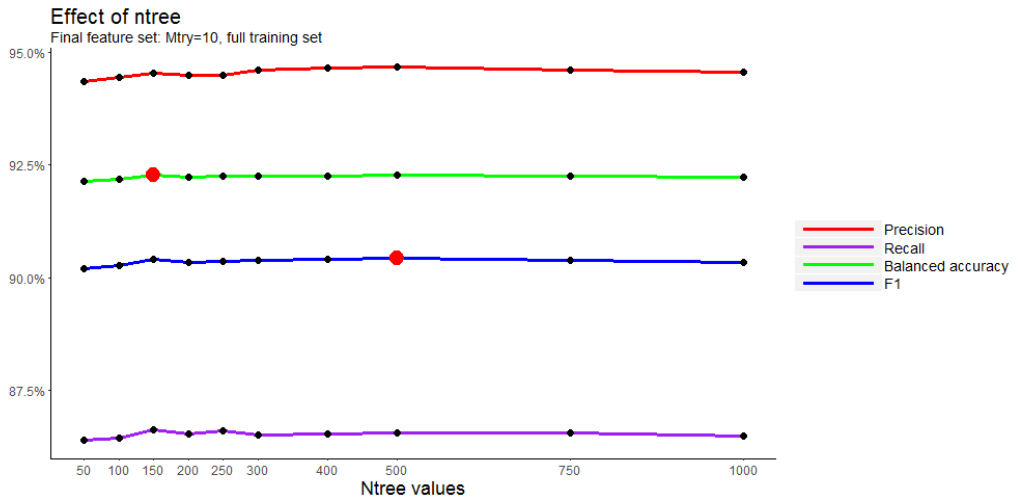

Figure 14: Results of final feature set: effect of ntree

4.3.1.3 Effect of training set size on the final feature set

The whole qualification rounds I split the corpus comprised of English-language records containing an annotated genre field into two equal sized halves, of which only the other part was used in training the data. Another part I reserved for testing the model, although I did not actually do it.

Further, I used cross-validation to sample the training set into five parts. From these five parts I combined four for teaching the model and tested the model against the fifth one. I repeated this process for each of the five parts and aggregated the results into the final results.

I did not realize until it was too late, that I would not have needed to do the original split into two, as the cross-validation procedure intrinsically includes the division into training and validation sets. In other words, the figures for the qualification round would have been higher, if I had not done the extra split. By the time I understood my mistake, I had already done so many classification rounds, that I had no interest in redoing everything. The effect on the final feature set can be seen in **Figure 15**. Surprisingly or not, the difference between the first sample being complete or half is really small, barely less than one percentage point for both *F1* and *balanced accuracy*.

On the other hand, thanks to my mistake, I had more processing time avail-

able, as all the model learning steps were significantly faster than they would have been without my mishap. Besides, the whole purpose of the qualification round was to find the best possible features for the actual processing, and the split size being the same for all the qualification steps there was no real harm done: the mutual order of the predictor sets and the rank of predictors within them would more or less have remained the same.

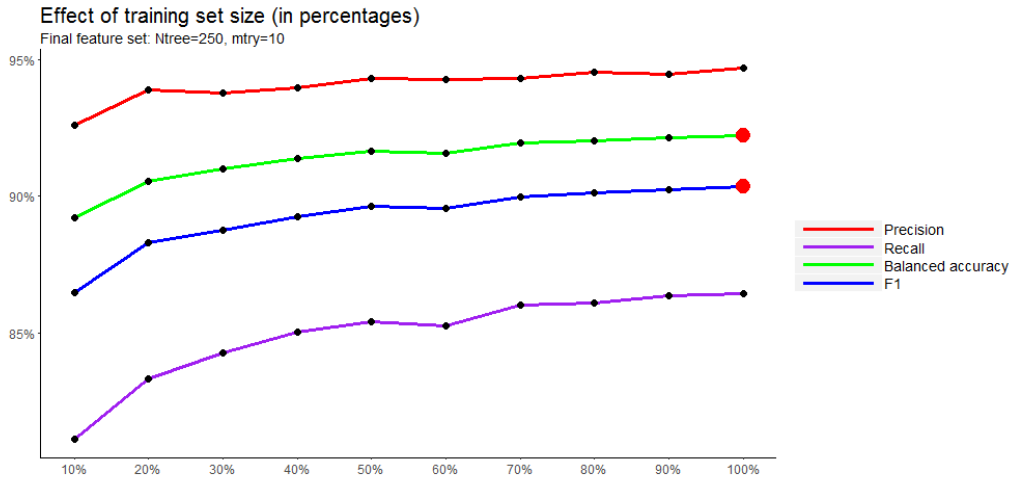

Figure 15: Results of final feature set: effect of training set size

4.3.2 Expanding to the unknown genres

4.3.2.1 The effect of changing the genre definition

The definition of poetry was also one possible source of suboptimal results. For the duration of the qualification steps I maintained the same definition as described in chapter 2.2.2. I wanted to make some additional tests to see, how a further division into records, that could be classified as poetry without a doubt and those with a slight hesitation, would affect the outcome. I dubbed the two groups into *hardcore* and *fringe*, respectively.

I made the split in the same manner as previously (description in chapter 2.2.2). This time I divided genres denoting *Poetry* into two new files. The *hardcore* class I chose from genres explicitly stating "Poetry", "Poems", "Verse" or a well known subcategory of poetry, such as "Ballads" or "Odes"; *fringe* was defined as a *Poetry* class without the explicit statement.[28]

[28]Full list of the 27 *hardcore* genres is viewable in `https://github.com/hegroiva/Gradu/blob/master/poetry_hardcore_genres.txt`

There was a potentially problematic decision of which class - *hardcore* or *fringe* - would be more dominant, as often the records would contain a genre or more from both categories. This I resolved by trying both approaches. In the *HC prevails* approach the class is considered *hardcore*, whenever the genre field has at least one value belonging to the *hardcore* group. The *fringe prevails* approach is defined in a similar way: any *fringe* value triggers the record as *fringe*.

An additional twist in regard to the new division is the definition of poetry authorship, which is done on-the-fly for each classification round. Since the authorship was clearly the most significant predictor, redefining the authorship might introduce great changes in the classification. Therefore I decided to run extra rounds - just in case - for both the scenarios: in the first one the authorship of *fringe* value is not considered as authorship of Poetry, while in the second one it is. This means, that the feature set is different only for one of the predictors, albeit the most informative one.

Interpreting the results from the multiclass prediction was more tedious, than I expected. The CARET package's *confusionMatrix* function calculates measures for each class separately (Kuhn et al., 2017). In this particular case the *hardcore* and *fringe* classes are comprised of the standard definition *Poetry*, which means, that the figures can not be compared without modification. In **Figure 16** the figures for *hardcore* and *fringe* have been calculated using the formulas from the `caret` package's help file.

For example: to get the *precision*, the formula is

$$A/(A+B),$$

where A equals the amount of correctly predicted Poetry values, and B equals the amount of Non-poetry predicted as Poetry. When using two correct classes instead of only one, the A and B must calculated differently. The A is actually the amount of either *hardcore* or *fringe* predicted either as *hardcore* or *fringe*. B is the amount of *Non-poetry* records, which have been classified as *hardcore* or *fringe*.

Figure 16 shows a slim preference in favor of the standard division into *Poetry* and *Non-poetry*, but the difference is so minimal, that it might occur by random chance. It is safe to state, that there is no significant effect. One must also keep in mind, that the features were optimized using this particular poetry definition. Hence it would be truly surprising, if sudden change in class values would surpass the standard definition. Limiting the poetry authorship to include only the *hardcore* poets did not have notable effect, if any, on the results.

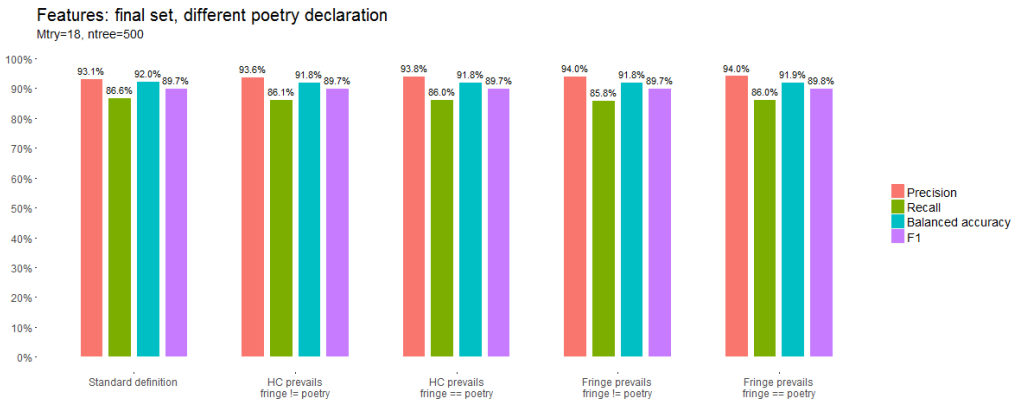

Figure 16: Results of final feature set: different ways of defining poetry

A closer examination of the *fringe* class reveals, that the amount of records belonging to it was barely 3,881 (*hardcore* prevailing) or 5,738 records (*fringe* prevailing). The corresponding amounts for *hardcore* were 31,414 and 29,557. It should come as no surprise, that the difference between the two methods deciding whether the record is *hardcore* or *fringe* does not really matter, and neither does the change of poetry authorship.

4.3.2.2 Introducing the *Unknown*

Earlier on I mentioned, that the division to genre classes can be problematized. There definitely exists the *Poetry*, and there also definitely exists the *Non-poetry*, although the definitions can sometimes be considered a matter of taste. In addition, besides the unannotated records, there does exist a large group of records, which I refer to as *Unknown*. They are the group of records, which have the genre field annotated in ESTC, but the field value is ambiguous in regard to the genre being *Poetry* or *Non-poetry*. Such field values denote for example the relation to the timing of the work *Juvenilia* or the physical form of the print *Broadsides* instead of the literary genre. In MARC the genre and form are declared in the same field.

For the purpose of extracting features I have treated these fields impartially in this thesis: they have been considered neither *Poetry* nor *Non-poetry*. For classification I have treated them as Non-poetry, which of course is a fallacy. Some of these records surely belong to poetry books.

I made a quick test for comparison, for which I considered the *Unknown* records as unannotated and removed them altogether and another one, for

which I declared the *Unknown* as its own group.

Figure 17 shows, how it turned out: The two leftmost bars belong to the same test run without the *Unknown* and the three on the right to the second run including the *Unknown*.. It can be seen, that *Poetry* is as easy to detect in both scenarios. *Non-poetry*, on the other hand, blends eagerly with the *Unknown* and is found with less difficulty, when the *Unknown* class is not present. For the purpose of finding the *Poetry* books, it has no relevance, whether the *Unknown* class exists or not.

Another interesting part in the figure is, how low the rate for classifying the *Unknown* itself is. A closer look in the *confusion matrix* shows, that of the misclassified 2,560 *Unknown* records 499 have been classified as *Poetry* and 2,261 as *Non-poetry*. See **Figure 18**. The proportion of *Poetry* in the misclassified *Unknown* as compared to the *Non-poetry* (0.22%) is less than half the proportion of the two classes in the training material (0.48%) This suggests, that the *Unknown* contains at least some *Poetry* books. I am inclined to think, that for large part those records misclassified as *Poetry* are actually classified correctly. The *Unknown* is after all an obscure class containing both *Poetry* and *Non-poetry*, and none of the predictors in the set are based on its values.

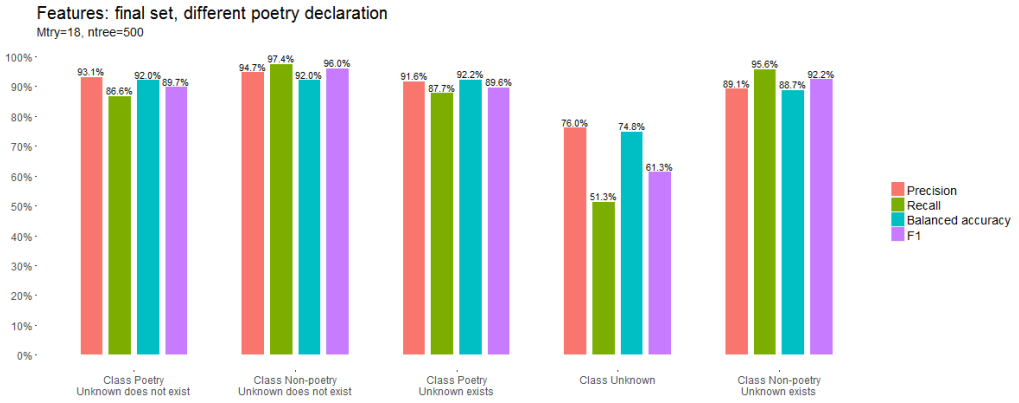

Figure 17: Predictability of classes, with or without the Unknown class

	Real value		
Predicted.value	Poetry	Nonpoetry	Unknown
Poetry	14810	857	499
Non-poetry	1832	33425	2261
Unknown	245	673	2909

Figure 18: Confusion matrix: Poetry, Non-poetry, Unknown

5 Results

5.1 Overall performance

Eventually I taught the superset model to conduct a search for the unannotated poetry books in the catalogue. The run predicted as Poetry 18,870 records in English language. I sampled a random subset of these records, and inspected them manually to see if the prediction had been correct. The manual inspection was rather tedious, as the library metadata often did not give any clues, whether the record actually should be classified as Poetry or not. I checked the first one hundred records either from the ESTC or simply by doing an Internet search with the book title and I was able to determine each of these records.

The inspection revealed, that 27 of the 100 records actually were not poetry. Provided that the subset sample is proportionate to the rest of the records, this would mean $\left(\frac{18,870}{100}\right) * 27 = 5,095$ False Positives and $\left(\frac{18,870}{100}\right) * 73 = 13,775$ True Positives.

The same run predicted 319,586 English-language records as Non-poetry. I created a random subset from these and inspected it closely, just as I did for the records predicted as Poetry. Similarly under the assumption, that the subset is in proportion to the rest, there would be $\frac{319,586}{100} * 2 = 6,392$ False Negatives and $\frac{319,586}{100} * 98 = 313,194$ True Negatives in the whole set.

Together this would translate to *precision* of $\left(\frac{73}{100}\right) = 73.0\%$ and *recall* of

$\left(\frac{13,775}{13,775+6,392}\right) = 68,3\%$.

As I read the records more closely to find out, what kind of records were misclassified, I noticed, that most of them were two-page pamphlets. I ruled out all the records containing two pages, to see the effect, if any, that would have. This way the number of the records was dropped to 13,985 categorized as Poetry and 282,895 as Non-poetry. I continued the manual inspection from the same subset samples as earlier, simply ignoring the two-paged records, until I reached one hundred records from both categories.[29] With this method I found merely five non-poetry books in the Poetry corpus, the proportion of Poetry books in Non-poetry books remained unchanged. The new figures without the two-paged records are: $\left(\frac{13,985}{100}\right) * 5 = 699$ False Positives, $\left(\frac{13,985}{100}\right) * 95 = 13,285$ True Positives, $\frac{282,895}{100} * 2 = 5,657$ False Negatives and $\frac{282,895}{100} * 98 = 277,237$ True Negatives.

Precision and *recall* are thus $\left(\frac{95}{100}\right) = 95.0\%$ and $\left(\frac{13,285}{13,285+5,657}\right) = 70,1\%$, respectively.

5.2 Variable importance

The table of the variable importance (**Appendix C**) is rather interesting, as it clearly shows dissimilarities between the variable importance metrics. Generally, the MDA ranks seems to be quite close to class specific MDA (Poetry and Non-poetry) ranks, which of course is not a surprise, for MDA is a combination of the two. While some of the predictors can almost unanimously be considered among the most important, there is not one feature, that would have been ranked in the top ten by all of the measures. In fact, examining only at the lowest ranks of each of the features, it can be seen, that TOPIC_ENGLISH_POETRY was ranked nineteenth, which is the best position in this comparison. In other words, its rank was somewhere between the first and nineteenth.

MARC_PAGECOUNT performed well by all of the measures, except for the *conditional importance*. This is particularly important, because the problems in detecting the poetry books predominantly seems to originate from the high

[29]My original subset sample size had been one thousand, but I close-read only the first one hundred records.

MDG ranking of this feature, as the `randomForest` package uses a modified splitting method based on Gini measures. (Liaw and Wiener, 2002)[30]

Another good performer was TOPIC__ENGLISH__POETRY, which was a clear winner in the TOPIC family, although TOPIC__ENGLISH__SONGS was ranked first by the *MDA Poetry*. Both of them were far superior in performance to their umbrella topic *topic_poetry*, which ranked in the lower half by all of the measures.

VARIA__AUTHOR, BASIC__POETRY50 and BOW__TUNE were successful with all the measures except for the three *MDA* measures. The same goes for MARC__SIZE, except for it doing well with *MDA Poetry* as well.

Part-of-speech -based predictors were mostly not well ranked. There were however features POS__IN, BASIC__VERBS BASIC__PROPER__NOUNS and POS__JJ. Prepositions and subordinating conjunctions (POS__IN) ranked relative well, four of the measures ranking it between seventh and fourteenth. This seems to be due to the fact of the them seldom appearing in poetry books: the worst rank POS__IN received, was from the MDA Poetry measure. The other measure not ranking it in the top 15, was MDA, which is calculated using MDA Poetry as a component. *Basic_verbs* ranked well by *conditional importance* only, whereas BASIC__PROPER__NOUNS was ranked in the top 20 by all of them, all the *MDA* measures even ranked it in the top ten. Clearly, the amounts of proper nouns and verbs were of importance, but in what manner, it is difficult to say. One possibility is, that this can at least partially deduced from the fact, that the feature values are raw values, instead of being standardized by the amount of words. The verbs and proper nouns might be represented with equal proportions in the titles, which would give smaller amounts for those with less words, in this case Poetry. A sidenote supporting this theory is, that BASIC__WORDS is ranked low by the *MDA* features, but not the rest. If the word count has already been indirectly involved in feature ranking, the BASIC__WORDS would not have the same predictive power anymore. This could be a a chance behaviour, because of an ill sample or the measures behaving differently, or then again, it is possible, that proper nouns and verbs simply appear more often or rarely in Poetry than Non-poetry. The same speculation applies to POS__JJ, which ranked well by the *conditional importance*, but not by the others. POS3GRAMS were low-ranking, with the exception of POS3GRAM__DT__NN__IN, which was ranked in the top 15 by all three of the MDA variable importance measures.

[30]This method is more thoroughly described by user Soren Havelund Welling in https://stats.stackexchange.com/questions/144818/does-breimans-random-forest-use-information-gain-or-gini-index.

The dependency relation measures performed surprisingly well. DEPREL__NO__OF__ROOT and DEPREL__ROOT__OFFSET__CHARACTERS even got top five ranks and DEPREL__NO__OF__ROOT three top 15 rankings.

A surprise was the relatively poor performance of bag-of-words based predictors. Besides the aforementioned BOW__TUNE, BOW__POEM was the only really high ranked predictor. This might actually show a bias in the annotations. It is possible, that titles containing magic words like *poems*, *songs* or *sung* have been easy marks for the annotators and having more readily had the genre annotated as *poetry* thus making their unannotatedness a rarity; it could also be, that for the same reason, being an easy target, they have had their topic annotated as *English poetry*, and the predictive power has then slipped to the other features. On the other hand, the high position of BASIC__POETRY50 (the compilation of poetry-related words) may have deterred better ranks for the other BOW features.

6 Conclusions

As is evident in the **Appendix C**, the weakest predictors in the final sets were features, which had been selected among the last ones from the bag-of-words qualification round. Two of these features are especially noteworthy: BOW__EPISTLE and BOW__PSALMS.

Epistle is a word, that has been used in English literature for epistolary poems, that is, poems written in the form of a letter. This poetry subgenre was flourishing in England during the 18th century, which is covered by the ESTC. On the other hand, *epistles* refer to letters directed at someone or someones in general. Besides that, there is also an even more precise meaning: the letters of New Testament by the Apostles. The titles in the catalogue disambiguate poorly this polysemy, and in many cases it is simply impossible to determine if a book contains poetry or not without seeing the text itself or a metadata description from some other source. Also *garland* is polysemic, meaning either a wreath of flowers or a collection of poems. In this material, this polysemy was not a problem, since *garland* was used almost exclusively in the meaning of a poetry collection.

Conversely, the word *psalms* proved problematic. The trouble with that is not related to polysemy, but instead to the definition of whether psalms should be counted as poetry or not. The psalms have been written in verse, and often the English translations have been in verse too. If a Bible contains psalms translated in verse, does that make the Bible a book of poetry? Or

better yet, if a book contains psalms translated in prose, would it still count as poetry? One way to look at this, is the purpose of the text: a complete Bible is intended for religious purposes, not poetical enjoyment, whereas psalms as a stand-alone publication would probably be meant to be sung in a hymn. Disambiguating the two potentially distinct purposes, which itself incorporates an intentional fallacy[31], is beyond the scope of this thesis. An easier solution would have been ruling psalms (RBMS value *Psalters*) as Non-poetry in the first place.

In chapter 5.1 I mentioned the poor performance regarding two-paged printings. More accurate would be to say, that many of these actually are broadsides, which are printed on one side only. As one sheet always contains two sides, even if only one side was actually having anything printed on it, the library catalogues might list them as two-paged. Alternatively, the two-pagedness in my data can be derived from the preprocessing described in Chapter 2.2. Broadsides were commonly printed when the print orders were low, so that the printing machines would not have to stay idle. The contents were often ballads or other poetry with a nice layout, so that it could be hung on the wall. Also broadsheets have two pages. They are printed on both sides, and their contents regularly were announcements or pamphlets. Random forest handles the page count as linear, so the exact page count is irrelevant, what matters is whether the cutpoint determined by the algorithm is below or above the page count of the record. Now that these two major types (broadsides and broadsheets) have the same amount of pages, they can not be distinguished reliably from each other. As it happens, there seems to be a bias in the annotations: the genre has been annotated from a much larger proportion of Poetry sheets than the other books. In the data there are 18,503 two-paged Poetry records, and 16,756 Poetry records with more pages. The corresponding figures for Non-poetry are 19,543 two-paged and 65,100 with more pages. In the remaining portion, where the genre information is missing, there are 41,153 two-paged records and 319,519 with over two pages. Of the two hundred record samples I found a total of 35 two-page records, 26 predicted as Poetry and nine as Non-poetry; of the 26 Poetry guesses there were only four correct identification. This indicates, that the two-paged poetry broadsides are much rarer in the unannotated part of the corpus than in the annotated part. Also, the relatively small numbers of poetry books in the predictions suggest, that Poetry in general has been annotated more thoroughly than the rest of the material.

[31]Intentional fallacy is a term by New Criticists from the field of literature studies meaning, that the authorial intent is irrelevant for the interpretation of the text, and hence it should be disregarded.(Wimsatt and Beardsley, 1946)

It is good to keep in mind, that the samples from the final round I examined, were really small. The estimate of 6,392 Poetry books predicted as Non-poetry was calculated from two False Negatives in a sample of one hundred. A small fluctuation in the sample would have altered the estimate considerably. One False Negative less, and the estimate would have been 3,196 Poetry books and one more 9,588, setting *recall* somewhere between 59.0% and 81,2%.

The results however brought forth over 13,000 poetry books (longer than two pages) with a good *precision* of 95.0%. The broadside hunt was not as successful, but it could be made better by focusing on the two-paged publications and building the feature sets specifically for them. Also, a return to the original, unmodified ESTC data might give access to the onesidedness of the poetry books.

A subset can be deduced by machine learning, but it does require close attention.

Extensibility to other genres

My research method is relatively easy to extend to other genres. The basic feature set, punctuation and the MARC features are usable as they are, but the relevant topics, bag-of-words and POS-tags are dependent on the genre definition. These genre specific feature sets for the qualification rounds can be picked by program code automatically, provided that the genre itself is cautiously defined manually. The most difficult part would be the automation of variable selection: different variable importance measures lift up different predictors. The ranks for each of the measures can be extracted, and with some custom algorithm the best ranking ones be selected for the final rounds. Some caution is to be used, however. My research showed a bias in the annotation level of the poetry in the corpus and it is possible, that there exists other biases regarding other genres as well. Remember the low annotation level of the *Scientific writings*: additional sources denoting scientific writings would have to be sought from other fields and possibly even from external sources.

As extensive testing as in this thesis is not necessary in the future. It is clear from the additional tests on the material, that the difference between different algorithm variables were marginal Also, the effect of adding more text to the learning process uniformly adds the predictivity: by this I refer to both the extraction of features from main titles including or excluding the subtitles, and the effect of the learning subset size.

It would also be possible to conduct a semi-automated feature selection skipping the qualification rounds completely, and gathering for example, the ten most frequent topics, bag-of-words and POS trigrams. This coarser approach would definitely save time, but it would also be riskier, possibly leading to ignoring prominent features completely. The method itself should be usable for any genre, if there only exists enough training material, the more the better.

Extensibility to other fields

Obviously, the topic of the record could be decided by a similar way, by making relevant topics the responses of the fields, and replacing the topic feature set with a genre feature set. In the ESTC, the language, page count, book size and publication place and time are almost always annotated anyway. Other relevant fields, such as Edition statement (MARC 260a), can not be deduced from the other fields. The use of this method is somewhat limited, as it demands a comparatively large training set. It would be interesting to modify the method for finding authors, where author information is missing, but that would require a more heuristic approach.

Extensibility to other catalogues

The approach is fairly easily transferrable for usage with other catalogues. The important thing is to take care that enough training material exists. As I have stated earlier, the amount of annotated fields, field contents and even which fields are used, varies from catalogue to catalogue. In some catalogues the genre field (MARC 655a) hardly exists: for example, the Finnish National Bibliography, Fennica, contains only 1,667 of 71,919 records published before 1917 having the genre field annotated. However, there are MARC field 080a and 080x for Universal Decimal Classification Number (UDC) and its auxiliary subdivision, respectively, from which the literary genre can be deduced to some extent. The field 080a is common in Fennica: 21,145 annotations exist, while 080x is annotated 8,904 times.[32] The library catalogues usually contain some kind of shelving information, because their function is to assist in finding the books. Provided that the books are arranged in some

[32]This information I got from examining an XML dump of the Fennica metadata provided by the National Library of Finland (Fennica).

meaningful order, that considers the literary genres or topics, this information should exist in the metadata. Using it does require close understanding of the contents of the data, as the data might be fixed codes.

Next steps

Some things could have been better in this thesis. First clear enhancement would be separating the one-paged broadsides from the two-sided broadsheets. There exists mentions in the topic fields of *Broadsides*, and some of the page count fields actually have a value of one page before the preprocessing phase, which transforms one-paged prints into having two pages. A separate process would be possible. This process might be adjusted a bit more by combining the MARC field 110a, *Main Entry-Corporate name* with the author name field (100a). I suspect, that the usage of the corporate name would be beneficial in differentiating the broadsheets apart from the broadsides.

Another one would be repeating the same process, but treating *Psalters* as non-poetry this time.

Publication place (260a) and publisher (260b) could be tested in a similar fashion as the bag-of-words of topics and the titles. The hindrance right now is the still ongoing preprocessing.

The most fruitful development might be combining the ESTC with ECCO, to access the full texts of the records. Although ECCO is not as vast as ESTC, many more books could definitely be found with a bigger text corpus. This might also require a recheck on the used feature sets.

Acknowledgements

I have been able to do this research working as a research assistant at University of Helsinki in a book history project COMHIS, which has received funding from the Academy of Finland. I wish to express appreciation to my supervisors and my research group, the members of which are partially overlapping. Also warm thanks to everyone at the loosely organized research community COMHIS Collective. The deepest gratitude goes to my family.

References

Akiko Aizawa. Linguistic techniques to improve the performance of automatic text categorization. In *Proceedings of NLPRS-01, 6th Natural Language Processing Pacific Rim Symposium*, pages 307–314, Tokyo, JP, 2001. URL `http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=EC36820A9B26C6921F2930198F24D01A?doi=10.1.1.2.6476&rep=rep1&type=pdf`. Accessed 2017-11-16.

Edward Albert. *A History of English Literature*. George G. Harrap & Co, London, fourth edition, revised and enlarged edition, Feb 1971. ISBN 0245504702.

Michael Alexander. *A History of English Literature*. Foundation Series. Palgrave Macmillan, Basingstoke, 2000. ISBN 0336722667.

Christoforos Anagnostopoulos and David J. Hand. *hmeasure: The H-measure and other scalar classification performance metrics*, 2012. URL `https://CRAN.R-project.org/package=hmeasure`. R package version 1.0.

Shlomo Argamon and Moshe Koppel. *The Rest of the Story: Finding Meaning in Stylistic Variation*, pages 79–112. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-12337-5. doi: https://doi.org/10.1007/978-3-642-12337-5_5.

Shlomo Argamon, Moshe Koppel, and Galit Avneri. Routing documents according to style. In *In Proceedings of First International Workshop on Innovative Information Systems*, 1998. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.52.688`.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *TEXT*, 23:321–346, 2003. doi: https://doi.org/10.1515/text.2003.014.

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. Stylistic Text Classification Using Functional Lexical Features: Research Articles. *J. Am. Soc. Inf. Sci. Technol.*, 58(6): 802–822, 2007. ISSN 1532-2882. doi: https://doi.org/10.1002/asi.v58:6.

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4:40–79, 2010. doi: https://doi.org/10.1214/09-SS054.

Taylor Arnold and Lauren Tilton. *coreNLP: Wrappers Around Stanford CoreNLP Tools*, 2016. URL `https://CRAN.R-project.org/package=coreNLP`. R package version 0.4-2.

Peter Beal. *A Dictionary of English Manuscript Terminology 1450–2000*. Oxford University Press, 2008. ISBN 9780191727955. URL `http://www.oxfordreference.com/view/10.1093/acref/9780199576128.001.0001/acref-9780199576128`. Published online 2011. Accessed 2017-11-17.

Douglas Biber. The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities*, 26(5):331–345, 1992. ISSN 1572-8412. doi: https://doi.org/10.1007/BF00136979.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 08856125. doi: https://doi.org/10.1023/A:1010933404324.

Leo Breiman and Adele Cutler. Manual–Setting Up, Using, And Understanding Random Forests V4.0, 2003. URL `https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf`. Accessed 2017-11-16.

Leo Breiman, Jerome H. Friedman, Richard. A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. The Wadsworth Statistics/Probability Series. Wadsworth International Group, Belmont, CA, 1984. ISBN 0-534-98054-6.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, ICPR '10, pages 3121–3124, Washington, DC, 2010. IEEE Computer Society. ISBN 978-0-7695-4109-9. doi: https://doi.org/10.1109/ICPR.2010.764.

J. F. Burrows. 'an ocean where each kind. . .': Statistical analysis and some major determinants of literary style. *Computers and the Humanities*, 23 (4):309–321, Aug 1989. ISSN 1572-8412. doi: https://doi.org/10.1007/BF02176636.

COMHIS/ESTC. ESTC project on GitHub. URL `https://github.com/comhis/estc`.

Cross Validated. URL https://stats.stackexchange.com.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, volume 6, pages 449–454, Genoa, 2006. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/440_pdf.pdf. Accessed 2017-11-17.

ECCO. Eighteenth Century Collections Online. URL http://www.gale.com/primary-sources/eighteenth-century-collections-online. Accessed 2017-11-14.

Jack Elliott. *Patterns and Trends in Harlequin Category Romances*, pages 54–67. Palgrave Macmillan UK, London, 2014. ISBN 978-1-137-33701-6. doi: https://doi.org/10.1057/9781137337016_4.

ESTC. English Short Title Catalogue. URL http://estc.bl.uk/. Accessed 2017-11-16.

Ingo Feinerer and Kurt Hornik. *tm: Text Mining Package*, 2017. URL https://CRAN.R-project.org/package=tm. R package version 0.7-1.

Fennica. Fennica - The Finnish National Library. URL https://www.kansalliskirjasto.fi/en/node/161. Accessed 2017-11-17.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, 2005. Association for Computational Linguistics. doi: https://doi.org/10.3115/1219840.1219885.

Alastair Fowler. *A History of English Literature*. Basil Blackwell, revised paperback edition edition, 1989. ISBN 0-631-171479.

John Frow. *Genre*. The New Critical Idiom. Routledge London ; New York, 2006. ISBN 0-415-28063-X.

J. L. Goldberg. Cdm: an approach to learning in text categorization. In *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*, pages 258–265, Nov 1995. doi: https://doi.org/10.1109/TAI.1995.479592.

David J. Hand. Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning*, 77(1):103–123, Oct 2009. ISSN 1573-0565. doi: https://doi.org/10.1007/s10994-009-5119-5.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, NY, 2 edition, 2009. ISBN 978-0-387-84858-7. URL `https://web.stanford.edu/~hastie/Papers/ESLII.pdf`. Accessed 2017-11-16.

David I. Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117, 1998. doi: https://doi.org/10.1093/llc/13.3.111.

Homer. *The Iliad. A Translation into English Prose with Index by A. S. Kline*. Poetry In Translation, 2009. ISBN 978-1532975240. URL `http://www.poetryintranslation.com/klineasiliad.htm`. Accessed 2017-11-16.

Kurt Hornik. *openNLP: Apache OpenNLP Tools Interface*, 2016. URL `https://CRAN.R-project.org/package=openNLP`. R package version 0.2-6.

Kurt Hornik. *NLP: Natural Language Processing Infrastructure*, 2017. URL `https://CRAN.R-project.org/package=NLP`. R package version 0.1-10.

Torsten Hothorn, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro, and Mark Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006a.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, 2006b. doi: https://doi.org/10.1198/106186006X133933.

IFLA. International Standard Bibliographic Description (ISBD). 2007. doi: https://doi.org/10.1515/9783110263800.

Thorsten Joachims. *Text categorization with Support Vector Machines: Learning with many relevant features*, pages 137–142. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998. ISBN 978-3-540-69781-7. doi: https://doi.org/10.1007/BFb0026683.

Adrian Johns. The birth of scientific reading. *Nature*, 409(6818):287, Jan 2001. URL https://search.proquest.com/docview/204498622?accountid=11365. Copyright - Copyright Macmillan Journals Ltd. Jan 18, 2001; Last updated - 2017-10-31; CODEN - NATUAS. Accessed 2017-11-15.

Antti Kanner, Jani Marjanen, Ville Vaara, Hege Roivainen, Viivi Lähteenoja, Laura Tarkka-Robinson, Eetu Mäkelä, Leo Lahti, and Mikko Tolonen. OCTAVO – Analysing Early Modern Public Communication [poster]. Presented in Digital Humanities at Oxford Summer School. 2017. URL https://comhis.github.io/posters/octavo/. Accessed 2017-11-16.

Stephen Karian. The limitations and possibilities of the estc. *The age of Johnson*, 21:283–297, 2011. URL https://literature-proquest-com.libproxy.helsinki.fi/pageImage.do?ftnum=3718860841&fmt=page&area=criticism&journalid=08845816&articleid=R05228409&pubdate=2011&queryid=3021719542796. Accessed 2017-11-16.

Jussi Karlgren and Douglass Cutting. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 2*, COLING '94, pages 1071–1075, Stroudsburg, PA, 1994. Association for Computational Linguistics. doi: https://doi.org/10.3115/991250.991324.

Eamonn J. Keogh and Abdullah Mueen. Curse of dimensionality. In *Encyclopedia of Machine Learning and Data Mining*, 2010.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. Automatic Detection of Text Genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 32–38, Stroudsburg, PA, 1997. Association for Computational Linguistics. doi: https://doi.org/10.3115/976909.979622.

Yunhyong Kim and Seamus Ross. Genre Classification in Automated Ingest and Appraisal Metadata. In *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'06, pages 63–74, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 978-3-540-44636-1. doi: https://doi.org/10.1007/11863878_6.

Yunhyong Kim and Seamus Ross. Examining variations of prominent features in genre classification. *Proceedings of the Annual Hawaii International*

Conference on System Sciences, (Dcc):1–10, 2008. ISSN 15301605. doi: https://doi.org/10.1109/HICSS.2008.157.

Bradley Kjell, W. Addison Woods, and Ophir Frieder. Discrimination of authorship using visualization. *Information Processing & Management*, 30 (1):141–150, 1994. ISSN 0306-4573. doi: https://doi.org/10.1016/0306-4573(94)90029-9.

Max Kuhn, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. *caret: Classification and Regression Training*, 2017. URL `https://CRAN.R-project.org/package=caret`. R package version 6.0-76.

Leo Lahti, Niko Ilomäki, and Mikko Tolonen. A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470-1800. *LIBER Quarterly*, 25(2):87, Dec 2015. ISSN 2213-056X. doi: https://doi.org/10.18352/lq.10112.

Lexalytics Salience 6 Dev Wiki. URL `http://dev.lexalytics.com/wiki/pmwiki.php`. Accessed 2017-11-03.

Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002. ISSN 16093631. doi: https://doi.org/10.1177/154405910408300516. R package version 4.6-12.

Huan Liu. *Feature Selection*, pages 402–406. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: https://doi.org/10.1007/978-0-387-30164-8_306.

Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 431–439. Curran Associates, Inc., 2013. URL `http://papers.nips.cc/paper/4928-understanding-variable-importances-in-forests-of-randomized-trees.pdf`. Accessed 2017-11-16.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP

natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL https://pdfs.semanticscholar.org/2f51/02ec3f70d0dea98c957cc2cab4d15d83a2da.pdf. Accessed 2017-11-17.

MARC. MARC Standards Homepage. URL https://www.loc.gov/marc/. Accessed 2017-11-16.

Tony McEnery and Michael Oakes. Authorship Identification and Computational Stylometry. In Robert Dale, Harold Somers, and Hermann Moisl, editors, *Handbook of Natural Language Processing*, Chapman & Hall/CRC Machine Learning & Pattern Recognition. CRC Press, Jul 2000. ISBN 978-0-8247-9000-4. doi: https://doi.org/10.1201/9780824746346.ch23.

Franco Moretti. Style, inc. reflections on seven thousand titles (british novels, 1740–1850). *Critical Inquiry*, 36(1):134–158, 2009. doi: https://doi.org/10.1086/606125.

Frederick Mosteller. *Applied Bayesian and classical inference : the case of the Federalist papers*. Springer series in statistics. Springer, New York, NY, second edition, 1984. The first edition of this book, Inference and disputed authorship : the Federalist previously published by Addison-Wesley 1964.

Ovid. *The Metamorphoses. A Translation into English Prose by A. S. Kline*. Poetry In Translation, 2000. ISBN 978-1502776457. URL http://www.poetryintranslation.com/klineasovid.htm. Accessed 2017-11-16.

Fuchun Peng, Dale Schuurmans, and Shaojun Wang. Language and Task Independent Text Categorization with Simple Language Models. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 110–117, Stroudsburg, PA, 2003. Association for Computational Linguistics. doi: https://doi.org/10.3115/1073445.1073470.

Andrew Pettegree. The great reformers. *Apollo*, 184(646):78–83, Oct 2016. URL https://search.proquest.com/docview/1822029239?accountid=11365. Accessed 2017-11-15.

Peter Quennell and Hamish Johnson. *A History of English Literature*. Weidenfeld & Nicolson, London, 1973. ISBN 0297765779.

Paul Rayson, Andrew Wilson, and Geoffrey Leech. *Grammatical Word Class Variation within the British National Corpus Sampler*, pages 295–306.

New Frontiers Of Corpus Research. Rodopi, Amsterdam, 2002. ISBN 9042012374. URL https://search-proquest-com.libproxy.helsinki.fi/docview/85581914?accountid=11365. Accessed 2017-11-16.

RBMS. Genre Terms : A Thesaurus for Use in Rare Book and Special Collections Cataloguing, 1991. URL https://rbms.info/vocabularies/genre/alphabetical_list.htm. Accessed 2017-10-24.

Claude Sammut and Geoffrey I. Webb, editors. *Encyclopedia of Machine Learning*. Springer, 2010. ISBN 978-0-387-30768-8. doi: https://doi.org/10.1007/978-0-387-30164-8.

Marina Santini. A Shallow Approach To Syntactic Feature Extraction For Genre Classification. *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, (1996):229–230, 2004. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.387.2093&rep=rep1&type=pdf. Accessed 2017-11-17.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Stroudsburg, PA, Oct 2013. Association for Computational Linguistics. URL https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf. Accessed 2017-11-17.

Stack Exchange. URL https://stackoverflow.com/. Accessed 2017-10-25.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Automatic Authorship Attribution. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pages 158–164, Stroudsburg, PA, 1999. Association for Computational Linguistics. doi: https://doi.org/10.3115/977035.977057.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Text Genre Detection Using Common Word Frequencies. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, pages 808–814, Stroudsburg, PA, 2000a. Association for Computational Linguistics. doi: https://doi.org/10.3115/992730.992763.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35 (2):193–214, May 2001. ISSN 1572-8412. doi: https://doi.org/10.1023/A:1002681919510.

Efstathios Stamatatos, George Kokkinakis, and Nikos Fakotakis. Automatic text categorization in terms of genre and author. *Comput. Linguist.*, 26 (4):471–495, Dec 2000b. ISSN 0891-2017. doi: https://doi.org/10.1162/089120100750105920.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1), Jan 2007. ISSN 1471-2105. doi: https://doi.org/10.1186/1471-2105-8-25.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(307), 2008. doi: https://doi.org/10.1186/1471-2105-9-307.

Tomoji Tabata. *Stylometry of Dickens's Language*, pages 28–53. Palgrave Macmillan, Basingstoke, 2014. ISBN 978-1-137-33699-6.

P.N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson International Edition. Pearson Addison Wesley, 2006. ISBN 9780321420527.

Mikko Tolonen, Niko Ilomäki, Hege Roivainen, and Leo Lahti. Printing in a Periphery: a Quantitative Study of Finnish Knowledge Production, 1640–1828. In *Digital Humanities 2016: Conference Abstracts*, pages 383–385. Jagiellonian University & Pedagogical University, Krakow, 2016.

Mikko Tolonen, Leo Lahti, Hege Roivainen, and Jani Marjanen. A Quantitative Approach to Book Printing in Sweden and Finland, 1640–1828, FORTHCOMING.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, 2003. Association for Computational Linguistics. doi: https://doi.org/10.3115/1073445.1073478.

Fiona J. Tweedie and R. Harald Baayen. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352, Sep 1998. ISSN 1572-8412. doi: https://doi.org/10.1023/A:1001749303137.

Geoffrey I. Webb. Overfitting. In *Encyclopedia of Machine Learning*, page 744. 2010. doi: https://doi.org/10.1007/978-0-387-30164-8_623.

W. K. Wimsatt and M. C. Beardsley. The intentional fallacy. *The Sewanee Review*, 54(3):468–488, 1946. ISSN 00373052, 1934421X. URL `http://www.jstor.org/stable/27537676`. Accessed 2017-11-17.

Ian H Witten, Eibe Frank, and Mark A Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 3rd edition, 2011. ISBN 9780123748560.

G. Udny Yule. *On Sentence-length as a Statistical Characteristic of Style in Prose: With an Application to Two Cases of Disputed Authorship*, volume 30. Biometrika, Jan 1939. doi: https://doi.org/10.2307/2332655.

G. Udny Yule. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, 1944. URL `https://books.google.fi/books?id=-R09AAAAIAAJ&printsec=frontcover&hl=fi&source=gbs_atb#v=onepage&q&f=false`. Accessed 2017-11-17.

Appendices

A List of genre values denoting Poetry content

[33]

Alphabet rhymes	Juvenile poetry
American poetry	Lyric poems
Ballads	National songs
Ballads, English	Neo-Latin poems
Begging poems	Nonsense verse
Broadside ballads	Occasional poems
Broadside poems	Odes
Broadsides poems	Paraphrases, Metrical
Carol books	Pastoral poems
Carriers' addresses	Penny poems
Eclogues	Poems
Elegiac poetry, American	poems
Elegiac poetry, English	Poetical miscellanies
Elegies	Poetry
Emblem books	Poetry of places
English poetry	Psalters
Epics	Single sheet verse
Epigrams	Song sheets
Epistolary poetry	Songs
Epistolary poetry, English	Songs and music
Fabliaux	Songsters
Hymnals	Verse
Hymns	Vocal scores without accompaniment

[33] Values are in the same format, as they appeared in the ESTC data.

B Part-of-speech tag set

[34]

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle

[34]Penn Treebank tag set documentation is incomplete. This list was compiled from Lexalytics Salience 6 Dev Wiki -page http://dev.lexalytics.com/wiki/pmwiki.php?n=Main.POSTags

VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb
comma	comma
period	period
semicolon	semicolon
single quote	single quote
-LRB-	left opening parenthesis
-RRB-	right closing parenthesis

C Variable importance ranks: final round

feature name	Caret	Conditional	MDA	MDG	MDA Poetry	MDA Non-poetry
marc_pagecount	1	35	3	1	4	1
topic_english_poetry	2	7	15	2	19	3
varia_author	3	1	25	6	30	28
basic_chars	4	32	17	3	21	6
bow_tune	5	4	21	8	23	31
marc_size	6	11	26	4	25	4
basic_words	7	2	39	14	28	55
basic_poetry50	8	10	35	9	34	48
bow_poem	9	57	2	45	3	2
basic_nonpoetry50	10	13	24	20	35	20
deprel_no_of_root	11	14	38	19	44	15
basic_ellipses	12	18	29	10	16	30
basic_chars_main_title	13	9	49	21	47	54
pos_IN	14	12	22	7	10	33
basic_proper_nouns	15	20	4	15	6	8
pos3gram_IN_DT_NN	16	24	37	18	29	37
topic_songs_english	17	16	7	13	1	16
basic_chars_subtitle	18	19	13	11	12	25
deprel_no_of_dependents	19	17	9	5	5	11
basic_word_length	20	15	10	12	9	13
basic_actual_sentences	21	29	6	28	8	7
bow_song	22	25	32	16	38	17
basic_words_main_title	23	30	51	33	46	41
basic_words_subtitle	24	23	20	22	27	9
basic_verbs	25	8	64	24	62	65
deprel_root_offset_characters	26	21	5	17	7	10
pos_DT	27	45	43	34	45	40
topic_ballads_english	28	3	67	26	68	66
basic_common_words	29	49	14	42	15	19
pos_NNS	30	44	46	25	51	32
pos_NN	31	22	41	23	43	27

bow_sung	32	26	1	30	2	5
bow_songs	33	28	54	32	60	62
pos_NNP	34	50	53	49	55	49
bow_psalms	35	43	60	46	64	35
pos_JJ	36	5	66	36	66	67
bow_poems	37	40	36	38	37	52
bow_elegy	38	52	55	39	56	44
pos_period	39	6	69	37	53	71
deprel_no_of_inflected	40	61	56	47	49	50
topic_poetry	41	42	44	51	42	47
pos_CD	42	38	62	41	65	61
basic_adjectives	43	60	63	54	61	60
basic_poetry100_compared	44	39	57	48	57	53
pos_comma	45	34	12	29	13	23
bow_ode	46	51	42	44	48	14
pos_TO	47	59	58	53	59	58
pos3gram_IN_DT_NNP	48	54	61	50	50	63
bow_hymns	49	33	18	31	18	24
basic_poetry50_compared	50	41	27	27	26	26
pos_CC	51	36	47	40	41	34
basic_verbs_past	52	53	48	55	52	45
pos_VBN	53	31	19	43	20	29
pos3gram_IN_NNP_NNP	54	66	52	64	54	59
pos3gram_IN_NNP_comma	55	62	34	52	36	36
basic_nonpoetry50_compared	56	47	31	35	24	21
basic_nonpoetry100_compared	57	48	71	61	71	64
basic_pronouns	58	55	8	58	11	22
pos3gram_DT_NN_IN	59	58	11	56	14	12
basic_gerunds	60	67	33	63	40	46
bow_verses	61	46	45	57	33	56
pos3gram_NN_IN_NN	62	27	40	62	39	57
punctuation_singlequotes	63	63	30	59	32	38
punctuation_hyphens	64	37	23	60	22	39
pos3gram_period_TO_DT	65	56	70	66	67	70
bow_poetical	66	71	50	67	58	43
bow_epistle	67	64	16	65	17	18
basic_adverbs	68	69	59	69	70	51
bow_garland	69	70	65	71	63	68

bow_ballad	70	68	28	68	31	42
bow_hymn	71	73	72	70	72	72
bow_lamentation	72	72	73	73	73	73
bow_love	73	65	68	72	69	69