MASTER’S THESIS IN LIBRARY AND INFORMATION SCIENCE
FACULTY OF LIBRARIANSHIP, INFORMATION, EDUCATION AND IT

Classification of Fiction Genres

Text classification of fiction texts from Project Gutenberg

Rolf Bucher

UNIVERSITY OF BORÅS

English title:	Classification of Fiction Genres
Author(s):	Bucher, Rolf
Completed:	2018
Keywords:	text, classification, genre, machine, learning, supervised, Gutenberg, fiction

Heartfelt gratitude to Johan for tolerating my incessant idiocy.

Abstract

Stylometric analysis in text classification is most often used in authorship attribution studies. This thesis used a machine learning algorithm, the Naive Bayes Classifier, in a text classification task comparing stylometric and lexical features. The texts were extracted from the Project Gutenberg website and were comprised of three genres: detective fiction, fantasy, and science fiction. The aim was to see how well the classifier performed in a supervised learning task when it came to discerning genres from one another.

R was used to extract the texts from Project Gutenberg and Python script was used to run the experiment. Approximately 1978 texts were extracted and preprocessed before univariate filtering and tf-idf weighting was used as the lexical feature while average sentence length, average word length, number of characters, number of punctuation marks, number of uppercase words, number of title case words, and parts-of-speech tags for nouns, verbs, and adjectives were generated as the feature sets for the topic independent stylometric features.

Normalization was performed using the ℓ^2 norm for the tf-idf weighting, with the ℓ^2 norm and z-score standardization for the stylometric features.

Multinomial Naive Bayes was performed on the lexical feature set and Gaussian Naive Bayes on the stylometric set, both with 10-fold cross-validation.

Precision was used as the measure by which to assess the performance of the classifier. The classifier performed better in the lexical features experiment than the stylometric features experiment, suggesting that downsampling, more stylometric features, as well as more classes would have been beneficial.

Stylometrisk analys i textklassificering används oftast i författarskapsstudier. Denna avhandling använde en maskininlärningsalgoritm, Naive Bayes Classifier, i en textklasseringsuppgift som jämförde stilometriska och lexiska funktioner. Texterna extraherades från Project Gutenbergs hemsida och bestod av tre genrer: detektivfiktion, fantasi och science fiction. Syftet var att se hur bra klassificeringen utfördes i en övervakad inlärningsuppgift när det gällde att skilja genrer från varandra.

R användes för att extrahera texterna från Project Gutenberg och Python-skript användes för att köra experimentet. Cirka 1978-texter extraherades och förbehandlades innan univariat filtrering och tf-idf-viktning användes som lexikala särdrag, medan längd för ordlängd, genomsnittlig ordlängd, antal tecken, antal skiljetecken, antal stora bokstäver, antal titelord, och delar av talkoder för substantiv, verb och adjektiv genererades som funktionen för de ämnesoberoende stilometriska funktionerna.

Normalisering utfördes med ℓ^2-normen för tf-idf-viktningen, med ℓ^2 norm och z-poäng standardisering för de stilometriska funktionerna.

Multinomial Naive Bayes utfördes på den lexiska funktionen och Gaussian Naive Bayes på den stilometriska uppsättningen, båda med 10-faldig kryssvalidering.

Precision användes som en åtgärd för att bedöma klassificatorns prestanda. Klassificeringsenheten fungerade bättre i experimentet med lexikala funktioner än I

experimentet med det stilometriska siktet, vilket tyder på att nedsampling, mer stylometriska funktioner och fler klasser skulle ha varit fördelaktiga.

Table of Contents

1 List of definitions

Confusion matrix: a table used to describe the performance of a classification model

Downsampling: the process where a balanced dataset is created through the matching of a number of samples in the minority class with a random sample from the majority class

The F_1 value: the harmonic average of **precision** and **recall**. A value of 1 indicates perfect precision and recall.

Feature: a measurable property or characteristic of the phenomenon being observed that allows the machine to build an accurate predictive model

Feature scaling: a method to standardize the range of independent variables

***k*-fold cross-validation**: the partitioning of the sample into random *k*-sized subsamples where a single *k* sample is retained as the validation data to test the model and the rest of the *k*-1 subsamples are used as the training data

Naive Bayes classifier: a probabilistic classifier that assumes strong independence between the features

Precision: the number of true positives divided by the number of true positives plus the number of false positives

Recall: the number of true positives divided by the number of true positives plus the number of true negatives

Term frequency normalization: the process of making each term's weight proportional to the length of the document

2 Introduction

The definition of genre has long defied consensus (Kwasnik and Crowston, 2005). Due to its polysemic nature, agreeing upon what a genre is, how it works with other so-called genres, and how best to study, construe, and identify genres is problematic, despite the term “genre” being recognized and used (Kwasnik and Crowston, 2005). While describing and defining genre is troublesome, it is clearly present in our everyday lives, and we see it used in the categorization of films or books in online bookstores such as Amazon (Amazon, 2018) and AbeBooks (AbeBooks, 2018) where genre is used to guide us in our selection process. A casual stroll through the fiction section of a library, for example, will reveal shelves of genre fiction categorized according to the readers’ expectations. The Oxford Dictionary of Literary Terms (2015) sheds some light on this statement with its definition of genre fiction as:

> “(t)he broad class of fiction that is easily identifiable as belonging within any of the recognized genres, especially of popular novel or romance, such as science fiction, detective story, thriller, western, historical romance, or love story. Genre fiction, then, is the kind of story that offers readers more or less what they would expect upon the basis of having read similar books before.”

While genre and its definition in relation to this paper will be looked at in more detail in later sections (particularly in the Genre subsection), this thesis is interested in whether or not the classification of genre fiction texts using a machine learning classification algorithm is possible. But first, let us take a step back and look at bibliographic classification.

Modern bibliographic classifications began to see the light of day in the late 1800s and early 1900s as a way to contend with the early stages of the print revolution (Satija and Martinez-Avila, 2015). Published in 1876, the Dewey Decimal Classification, or Dewey Decimal System, allowed new books to be added to a library based on subject. The Universal Decimal Classification emerged in 1905 and proved itself to be effective at the retrieval of documents from large collections, while the Library of Congress Classification sprung about at the same time and found itself eventually being utilized in research and academic libraries. A tool used for organization, classification and its development can be viewed as being synonymous with that of libraries.

Library classification systems bring similar items together through the grouping of objects/classes based on shared properties (Satija and Martinez-Avila, 2015). An integral feature is the organization and accessibility of documents in libraries, whereby related classes and subjects are brought together. This thesis is also interested in the relation of class and subject, and seeks to address whether documents classed in specific fiction genres can be distinguished as such through the style of writing. The method utilized in this experiment uses computational techniques to look at the writing style in texts and the relation between the style and genre.

Text classification assigns one or more classes to a document after analyzing the document's content. Through supervised learning, that is, with human intervention, a set of labeled training set of samples is generated and this set is sampled during training. A classification algorithm is then trained and applied to new texts to generate hypothetical class labels on its own. These classes are selected from a previously established collection of categories or classes. Classifiers are used for a range of text classification tasks, such as authorship attribution (Stamatatos, 2009; Halvani *et al.*, 2016), sentiment analysis (Ortigosa *et al.*, 2014; Hogenboom *et al*., 2015), topic classification (Varga *et al.*, 2014; Li *et al*., 2016), language detection (Zubiaga *et al.*, 2016), and genre detection (Mason *et al*., 2009; Lex *et al*., 2010).

A development of literary stylistics, stylometry is the statistical analysis of literary style (Holmes, 1998). Using computational distant reading methods, it concerns itself with the observation that, in writing, language is used in a consistent manner, for example through the use of function words (Stamatatos, 2009), the vocabulary that is chosen, sentence lengths, the use of punctuation, and so on. This consistency in language provides an opportunity to use stylometry as a means to see if literary style can allow a classification algorithm, through machine learning, to assign genre fiction texts to their respective genres/classes.

Stylometric analysis in text classification is an approach for producing features related to writing style. Large text collections are commonly analyzed in order to find similarities and differences that are undetectable by the human reader (Eder *et al*., 2016), and are often used in studies pertaining to authorship attribution. Features such as most frequent words, n-grams, part of speech counts, characters per word, and sentence length among others are selected and used to help train the classifier (Neal, Sundararajan, Fatima, Yan, Xiang, and Woodard, 2018). Stylometry is most frequently used for authorship attribution (Stamatatos, 2009; Koppel *et al*., 2009). It is also used for other writing style-related features such as stylochronometry (the date/time period the texts were written), gender identification of the author, and in the case of this thesis, the genre of the texts. The Oxford Dictionary (2018) defines stylometry as “(t)he statistical analysis of variations in literary style between one writer or genre and another.”

Despite this definition, stylometry has been used primarily for authorship profiling/verification and not much attention has been to paid toward genre, despite suggestions that factors such as style, sentiment, topic, and genre might be highly correlated (Neal *et al*., 2018). Studies that have performed stylometric classification on texts have often used texts from the Internet in order to discern between web genres (Mason *et al*., 2009; Lex *et al*., 2010). For their use of the term “web genre,” Mason *et al*. (2009) used a 7-genre data set comprised of blog, eshop, FAQ, online newspaper front page, listing, personal homepage, and search page, while Lex *et al*. (2010) divided the genres into news related blogs and “rest.” While the authors of the two papers did not provide an actual definition of what constitutes a web genre, Rosso (2008) defines it being as user-based; that is, as web users exist they are the user groups of various genres who, through their knowledge of what a genre instance looks like, can in turn recognize it as belonging to a genre. A user, then, through their knowledge of what a specific web page should look like and contain, will thus perceive the genres of the web pages they encounter. An Internet user, according to Rosso (2008), when searching for something (such as a ‘how to ski’ search query) would recognize a web page as belonging to that genre (a ski-related page) or not (an online tax form). With this definition in mind web users identified 18 web genres, which included articles, blogs, poetry and personal websites. It is the lack of fiction genre classification using stylometry in studies, however, that makes this area one that is ripe for exploration, particularly in a digital library context. In addition to the stylometric features, the experiment in this thesis includes classification using lexical features as well (see Methodology), whereby a comparison in performance between the two can be made. This expands upon the work by Lex *et al.* (2010), where web genre classification using both stylometric and lexical features was undertaken with the performance of the aforementioned features compared.

The classification of fiction in libraries is more problematic than that of non-fiction classification (Ward and Saarti, 2018). Where the labeling of non-fiction by human minds is rather straightforward due to non-fiction being subject-based, fiction operates on literal, symbolic, and thematic levels (Hayes, 1992), with the classifications being more open so that readers can make their own interpretations as well (Saarti, 2000). In public libraries, for example, the Dewey Classification System is one of the most common schemes, yet has also been found to be limited in usage (Rafferty and Hidderly, 2005). While the definition of genre is a multifaceted one, it remains the predominant format in which works of fiction are classified (Ward and Saarti, 2018), and leads the user to expect commonalities between the texts of items in the same genre (Wagers, 1981).

With the advent of digital libraries, the classification of fiction can – and already has – been expanded. Users can find recommended books based on similar authors, for example. These recommendations are often in the form of what other users have read or bought, and there is a lack of documents that are categorized according to writing style and lexical features. With supervised machine learning the possibility of computers picking up similarities between text documents that the human eye cannot is increasing. Through this supervised method it might be possible to address this situation and add more options for discerning readers looking for texts of a similar nature to read. This thesis seeks to explore how well the machine can distinguish between genres based using stylometric and lexical methods. It will look at writing style in a quantifiable style sense and compare it with a lexical classification. The problem statement is as follows:

Numerous studies have used stylometry for authorship attribution machine learning learning tasks, but few have concerned themselves with using stylometry for genre fiction classification, while fewer still have compared such a task with one using lexical features for the classification.

The research question related to this problem statement is in the next section. Following that there is a literature review that summarizes some of the studies done using text classification for genres, corpora from the Gutenberg website, and stylometry in their research. The section after the literature review will look at the concepts behind genre and stylometry and how they are to be used in this experiment, after which the methodology section addresses the corpus used and the preprocessing steps taken on it. The algorithms used in the experiment are then addressed, followed by the feature selection section to look at what features were selected for the stylometric and lexical classifications. The evaluation section guides the reader through how the results of the experiments were evaluated, followed by the results. The discussion section summarizes the experiments from a critical perspective before the conclusion finalizes and ends the thesis.

3 Purpose and research question

This thesis aims to investigate whether genre can be distinguished by elements related to style as well whether the same can be done with lexical elements, and how they compare in performance. Further exploration in this area could expand on this and use the same methods to distinguish various sub-genres from one another, such as hard science fiction, soft science fiction, New Wave science fiction, and then with these boundaries established, look for distinct stylistic features within the sub-genres. In this sense genre would be subcategorized by means of style, an enticing prospect.

The comparison of lexical and stylometric features by Lex *et al*. (2010) for web genre detection proved vital in the formulation of the design of this experiment. The authors made the distinction between lexical features and stylometric features, a distinction often cast aside in favor of regarding lexical features as falling under stylometric features (Abbasi and Chen, 2008). The distinction is a necessary one as this paper considers lexical features as being either genre specific terms or just terms that one way or another find themselves used more in a specific genre, whereas stylometric features are the ones that come through the authors' style of writing. It provides an interesting question: are there quantifiable style-related features, regardless of author, for specific genres, and if so, how do stylometric features and lexical features perform in a text classification task? Will the features chosen for both classification tasks result in an outcome that minimizes false positives? These questions then lead to the following research questions:

Q1: How well do lexical features perform in a text classification task for genre fiction with regards to precision, recall, and the F_1 value?

Q2: How well do stylometric features perform in a text classification task for genrefiction with regards to precision, recall, and the F_1 value?

Q3: How do stylometric features perform in comparison to lexical features in a text classification task for genre fiction with regards to precision, recall, and the F_1 value?

While authorial style is specific to the individual author, often the genre influences the style of writing. Fantasy, for example, could push authors toward a more florid prose while science fiction could contain more technical vocabulary. For example, made-up words and names would be expected to be more commonplace in the fantastical genres, while the detective genre would contain more crime-related terms. These use of genre-specific terms are generally made on the conscious level. On the subconscious level is the use of certain features in their writing that the authors are not aware of. This is where stylometry comes into play, in quantifying these features and seeing if they can be used to distinguish between genres. Some examples of such features are punctuation mark counts and average word and sentence lengths. These will be explored in more detail in upcoming sections.

This experiment aims to see whether or not supervised machine learning will be able to differentiate between the classes (genres) of detective and mystery stories, fantasy, and science fiction from a large corpus of texts extracted from the Project Gutenberg website. If a classifier can classify genre fiction based on the writing style, then not only can a machine learn to classify books according to their genre but could also lead to the possibility of text classification being used to sort books by their writing style, into sub-genres, and so forth. The first step is to see how well the classifier performs using readily-quantifiable writing styles to differentiate between genres, and the research questions in this thesis intend to explore the machine’s capability when doing so.

To see how well the classification algorithm performs with the tasks at hand failing test cases are to be treated as negative and passing cases as positive samples. Thus with the results there are four possibilities:

- true positives whereby positive samples are correctly labeled as passing
- false positives in which negative samples are incorrectly labeled as passing
- true negatives where negative samples are correctly labeled as failing, and
- false negatives where positive samples are incorrectly labeled as failing

In order to gauge the effectiveness of the classification algorithm precision is to be used as the main determiner of success, with recall and the F_1 value also taken into consideration. Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives. Related to precision is recall, which is simply the number of true positives divided by the number of true positives plus the number of true negatives. As this study would like to minimize the number of false positives a high precision score is desirable and would be a good measure of the classification algorithm's "success."

The following section looks at previous studies conducted in the same or similar areas. When choosing these studies, focus was given to their methods, use of stylometry, and genre classification tasks. It should be noted that studies of this ilk can vary quite remarkably in how they conducted their experiments, through choice of classifier/s and feature selection to the types of corpora used. The aim, however, is to provide a relatively broad picture of studies in the area and how they relate to the one in this thesis.

4 Literature review

While stylometric tasks are related to authorship attribution, numerous studies have focused on genre classification using a variety of methods (such as lexical features) and some have been presented in the literature review here. The criteria for choosing the studies were based on features used, the classification algorithm, and if they used genre detection/discrimination. As text genre classification is not as well a researched field as, for example, authorship attribution, the following research does not include all of the aforementioned criteria. This study was concerned with readily quantifiable stylistic features (such as parts-of-speech counts, for example), some of which are present in the studies below, studies which were deemed to be related or relevant to the research done in this thesis.

4.1 Classification of web genre documents

Numerous studies have focused on what they term ‘web genres.’ In the introduction of this thesis a definition by Rosso (2008) was given where the genres are already known in the users’ minds which allows them to recognize the web genre when they encounter it. Rosso (2008) also views genre itself as being a document type that is based on form, content, and purpose (elaborated upon in the Genre sub-section). A study by Dewe, Bretan, and Karlgren (1998) also took into account how web users perceived the material they interacted with online, and considered the style of web-based documents, as well as the source, to determine their genre. Lim, Lee, and Kim (2005) see web genre documents as being comprised of subject (for example, golf games) and style (homepages, news articles, image collections) without considering form or purpose. Besides Rosso (2008), the studies using web genre documents tend to assume that genre (as used in a web-based sense) is implied. What the web-based documents share with genre fiction texts as used in this thesis when it comes to genre is that they can be said to have content, form, and purpose.

Applying the same methodological approach to genre fiction as it has been applied to, for example, web genres would be problematic because despite the ‘genre’ tag in both, they are clearly of a different form and content. The methodology used for criteria such as feature selection were approached as guidelines (in the sense that related/the same feature selection could be used, as well as the distinction made for lexical features in both studies) for the similarity (classification of genres, for example) in the studies. Studies using web genres in classification tasks are discussed below.

Five document genres were defined by Michos, Stamatatos, Fakotakis and Kokkinakis (1996), namely literary, public affairs, journalistic, scientific, and everyday communication. The documents were also represented by four main features (elegance, formality, syntactic complexity, and verbal complexity), after which the main features were encoded by style markers such as the number of words per sentence, the number of conjunctions per sentence, the verb-noun ratio, and so on. Like the experiment in this thesis, quantifiable features were chosen as theirs was also a study employing stylometrics. It should be noted that even though the feature selection is similar to that employed in this thesis, the nature of the documents are not. While they (the documents in the experiment here and the one in this thesis) fall under genres, their form and content (as well as purpose) differs in the texts themselves, however stylometry is concerned with what it can count, regardless of the nature of the genre. The issue of differing genres is discussed more below with the Karlgren and Cutting (1994) study.

While the authors’ research dealt with text genre classification with web documents, their argument that the use of genre information can aid in helping users judge the relevance of a document is interesting, particularly when taking into account how users can also determine genre when encountering it (Rosso, 2008), making genre a two-way street perhaps from the user-perspective. In the case of fiction, genre is possibly the main determiner of relevance from a user-based perspective, and something incorrectly classified under a certain genre would no doubt prove problematic when not meeting user expectations upon encountering it (and determining it is another genre). Taking this into account, the machine – that is, the computer – is viewed as the user in a sense in the experiment in this thesis whereby it is trained through human supervision to attempt to determine genre when encountering it.

Lim, Lee, and Kim (2005) suggested features for the classification of web genres. While they did not mention stylometry specifically, their methods proved influential for this thesis. Their evaluation of the features found that punctuation mark counts and function words were appropriate for automatic web genre classification when using a corpus of Korean texts. These feature counts are in line with the stylometric method employed in this thesis, as the nature of this method is exactly what stylometry is concerned with (quantifiable features). The authors stressed the importance of the features used in text genre classification studies, and thus incorporated a wide variety of them, with a special distinction made for lexical features whereby frequencies such as the 50 most frequently used words and the 32 most frequently used punctuation marks were taken into account. A similar approach was used (albeit differently with the 300 most frequent terms) in this thesis for the lexical classification, distinguishing it from the stylometric classification where counts were used.

Lex, Juffinger, and Granitzer (2010) compared stylometric and lexical features for both web genre and emotional classification in blogs. Some stylometric features selected were punctuation, the distribution of sentences with specific word lengths, the average number of words and characters in sentences, and adjective and adverb rate, while some lexical features selected were unigrams, bigrams, trigrams, stems and feature spaces with nouns, verbs, adjectives, and adverbs. For the identification of the most valuable stylometric features the authors utilized the mutual information, a measurement that ascertains how much information the presence or absence of a term contributes to making the correct classification decision on. The Naive Bayes classifier and a C45 decision tree, both with and without boosting, were used to compare the stylometric and the lexical features. The study found that classifiers trained on lexical features performed significantly better than classifiers trained on stylometric ones.

This Lex *et al*. study was used as the basis for the preliminary structure of the research conducted in this thesis. The focus was originally on the stylometric features, and this study provided the inspiration and impetus to not only include lexical features but to compare them to the stylometric features. While the study made note that genre should be differentiated from subject and topic, it did not explore or provide a definition for the web genres they used. The study did, however, highlight that stylometric feature selection is topic independent, an important factor for the use of stylometric classification for genre, where topic and genre are closely correlated (discussed in the Concepts chapter below).

4.2 Classification studies with Gutenberg corpora

The studies here were of particular interest due to their use of texts extracted from the Gutenberg site and for being recent. They showed that the texts freely available on the Gutenberg site are suitable for experiments of this nature, and while not many text classification studies have taken advantage of this the opportunity to do so is clearly there, as is evidenced by the use of the Gutenberg texts in this thesis.

The methodology in the following studies such as the use of 10-fold cross-validation, the extraction of 300 features, the use of the Naive Bayes classification algorithm, and the distinction made between genre proved influential for the experiment performed in this thesis. The use of 10-fold cross-validation, for example, was chosen over 5-fold due to its common use in machine learning experiments besides those listed below (Fuller, Biros, and Delen, 2011; Liu, Bi, and Fan, 2017), as well the number 300 for terms

(Kumaran and Allan, 2004; Chang, Chen, and Liau, 2008; Wang, Hong, and Lau, 2018).

Samothrakis and Fasli (2015) also made a distinction between genre, but more in line with the work presented in this thesis. The authors conducted a study to not only predict fiction genre, but by also using texts and consequently genres from the Project Gutenberg website. In addition, they too used 10-fold cross-validation for their classification and extracted 300 features for use in their study, the same amount of features extracted in this thesis for the lexical classification.

Where the Samothrakis and Fasli (2015) study differed from the work in this thesis, however, is that they used the emotional content within the genres to train randomized trees (instead of feature selection used with the Naive Bayes classification algorithm), and did this with six genres selected from the Gutenberg site instead of three. Their focus was not on what genre is, but rather that genre conveys emotion, and this emotion can be used to distinguish genres from one another. Instead of emotion this thesis aims to present a case that quantifiable writing style, as well as lexical features, can distinguish genres from one another, resulting in a different feature selection. The studies are similar enough through the use of other methods (corpora used, amount of features selection, 10-fold cross-validation, genre determination) that they can be seen as being complementary to one another, or rather that this thesis hopes to add to such studies.

To continue with the Gutenberg website theme, Boran, Voss, and Hossain (2016) conducted an authorship attribution study using text files of three books each from five different authors extracted from the Project Gutenberg website. Accuracy was used to measure the experiment's success. They compared the performance of three machine learning algorithms (Artificial Neural Network, Naive Bayes Classifier, and Support Vector Machine) and found that the

Support Vector Machine, as well as punctuation and various suffixes, resulted in the highest accuracy.

This thesis' experiment is similar to the study in that it also uses texts (albeit a far larger amount) from the Gutenberg website, as well as also using the Naive Bayes classification algorithm (but not any others). The study used a variety of number counts for their features (pronouns, adverbs, hyphens, commas and so on), with the punctuation especially used to showcase the stylistic differences of the authors in their texts. While this thesis used punctuation marks as one feature on its own without distinguishing between the types of punctuation marks used in the texts, it should be noted the authors of the study were using far less texts and had to rely on extracting as many unique (on a smaller scale) features as possible. Nevertheless, the study adds to those using Gutenberg texts as corpora and the success of their experiment (albeit a success where the SVM classifier outperformed the Naive Bayes one) allows this thesis to continue studies with Gutenberg fiction corpora and quantifiable stylistic features.

4.3 Stylometry/genre classification studies

The studies below are concerned with either stylometry or genre. Two of them are authorship attribution studies, and the studies concerned with genre classification are not limited to fiction genres. Of interest however is their feature selection, and how despite being related to the study in this thesis there is still a lack of genre classification studies using stylometry, despite numerous articles on stylometry making mention that it is used, among other things, for genre classification.

Karlgren and Cutting (1994) used the Brown corpus to attempt to classify documents into one of four genres ("press," "miscellaneous," "fiction," and "non-fiction") with varying success (as the "miscellaneous" category proved problematic). A wide variety of features were used, including parts-of-speech count, character count, long word count, characters per word, and words per sentence, as well as a lexical count featuring specific words (“therefore,” “me,” “that,” “which,” “I,” “it”).

The authors identified two problems of research involving genres and text retrieval, namely the identification of genres and then subsequently the choice of criteria to cluster texts in the same genre with predictable recall and precision. They also took care to emphasize the difference between topic and genre, a problematic issue solved by the topic independent features used in stylometry (Lex *et al*., 2010). In addition to topic and genre, they mention that fiction types are “naturally defined in terms of their content,” a statement that takes into account how the genre of a text is perhaps defined by its form, purpose, and of course its content. This differentiation between topic and genre, as well as how fiction types are defined, aided in the theory (see Concept chapter below) used in this thesis, a problematic issue as to how genre – or genre fiction as the case may be – is defined, and how it is defined in terms of a stylometric classification.

Despite the length of time that has transpired since the Karlgren and Cutting classification of genres, their feature selection of character counts, character per words average, sentence counts, and parts-of-speech counts among others were influential for this experiment. The use of these quantifiable stylometric features were taken into account and utilized in this thesis and while many subsequent studies have added to the work of Karlgren and Cutting it was hoped that the work in this thesis could be counted among them.

Stamatatos, Fakotakis, and Kokkinakis (2000) used the British National Corpus to collect the frequencies of the 50 most frequent words (MFWs) in English and evaluated them on a Wall Street Journal corpus. They reported that the most frequently used punctuation marks also play an important role in the discrimination of a text genre, a point this thesis once more took into consideration and consequently applied to the experiment. While MFWs have been employed in numerous stylometry studies, they are most often used in stylometric studies dealing with authorship attribution, and their use is rare in stylometric genre studies, with the implication being that they possibly show better results when used for authorship attribution and consequently were not used for the experiment conducted for this thesis. Nevertheless, the study was considered worth noting as it both uses stylometric methods and is concerned with genre detection, a somewhat uncommon (compared to authorship attribution studies using stylometry) experiment that highlights the need for more in the area as this thesis hopes to do. In addition to this, despite the different methods, this study serves as an example of techniques that could possibly be used in conjunction with, for example, the ones used in this thesis in future studies, provided the techniques used in this thesis bear repeating.

Neal *et al*. (2018) performed a large scale stylometric study to determine authorship attribution. Their feature selection was of interest for this thesis, where a distinction was made for character, word, and syntactic features. Although this thesis did not employ syntactic features such as parts-of-speech bi- or trigrams, parts-of-speech tags for nouns, adjectives, verbs were employed instead to see how similar they would be across genres instead of attempting to attribute sets of words/phrases to specific authors. While the authors of the study found that lower-level representations such as character-level features generally outperformed higher-level features such as word-level representations, their use of word and sentence lengths, along with character-level level features such as upper case words, were found to be suitable stylistic markers to be used as part of the feature selection in this thesis in order to capture stylistic traits at various levels.

Onan (2018) compared five feature sets (features used in authorship attribution, linguistic features, character n-grams, part of speech n-grams, and the frequencies of discriminative words) using five different base learners (Naive Bayes, SVMs, logistic regression, k-nearest neighbor, and Random Forest) with Boosting, Bagging, and Random Subspace to ascertain which methods are most effective with the text genre classification of a corpus of book and camera reviews. The study took into account three genres, or rather three abstract classes, namely expressive, appellative, and informative. This choosing of three classes was of particular interest as three genres were used in the experiment of this thesis, especially considering that the aforementioned experiments in this review tended to use more.

To summarize, the studies presented above vary greatly in method and scope. There is no, to the author’s knowledge, any previous study that uses the same corpus and methods as the ones in this thesis, however that does not mean by any means that what this thesis is doing is unique; rather it showcases the variation in method and technique when performing text classification experiments. The potential issues are many; the feature selection might not be ideal, the classification could perform better with a different classification algorithm, the classes could be imbalanced, and so on. Like the aforementioned research, this thesis can only aim to see how successful the classification is at the end. This literature review was concerned with presenting text classification studies done using genre and/or stylometry, and the following section looks at the concepts behind the two and how they can possibly be used together.

5 Concepts

As this thesis is concerned with the classification of genre fiction it is worthwhile to explore what is meant by genre, and how stylometry relates – or can relate – to it. Is genre fiction, for example, the same as other literary texts, such as non-fictional texts? Previously web genres were mentioned in related studies, which brings about the question of how web genres are related to genre fiction genres, if at all. To add to this the opening paragraph of this very thesis mentioned that there is no consensus on what genre is (Kwasnik and Crowston, 2005). How, then, is genre defined, and how is it used in this thesis, and how does it relate to other “genre” studies that are using genres that are clearly not genre fiction? Does it differ from topic, and where does stylometry fit into all this? What sort of style does stylometry look for when it comes to discerning texts in different genres from one another? These questions will be looked at below with the aim of providing some clarification.

5.1 Genre

Genre has a range of definitions. The Oxford dictionary (2018) defines ‘genre’ as “(a) style or category of art, music, or literature” while Merriam-Webster (2018) defines it as “a category of artistic, musical, or literary composition characterized by a particular style, form, or content.” Both dictionaries mention style, which is interesting, as style (in genre fiction) is a clear part of this thesis, which would be one use of style - payslips, for example, would have their own style, as would restaurant menus, or rather documents of the payslip genre would have their own style, and likewise documents of the restaurant menu genre. This implies that almost anything that is a document, such as the random examples of payslips and menus, can fall under a genre.

When referring to documents, genre is often seen as a document type with form, content, and - phrased one way or another - purpose (Miller, 1984;

Swales, 1990; Yates and Orlikowski, 1992; Biber, 1994, Rosso, 2008). What is essentially *purpose* has been phrased in various forms, such as rhetorical action (Miller, 1984), communicative events with common communicative purposes (Swales, 1990), and communicative action (Yates and Orlikowski, 1992). This would mean that the document of the restaurant menu genre would have the purpose of informing potential eaters what the restaurant has available, its content would be lists of dishes, and its form would be sections on a physical piece of paper/booklet or in an online digital format. An online web genre such as Lex *et al*.'s (2010) news related blogs would have the purpose of informing the reader about the news/giving the reader an opinion about the news with news-related content in the form of blog posts. By this definition, that genre documents have form, content, and purpose, we can see the relation of genre fiction to other genres. The content is a fictional story, the purpose is to entertain the reader and meet their expectations with what they expect from that genre (Rosso, 2008), and the form is most often in chapters comprised of paragraphs following sequential events.

According to Petrenz and Webber (2011) all texts have one or more topics as well as one or more genres. The topic and genre can be found from several of the text's features, including both its lexical and its syntactic features. Topic is what the text is about, and while in theory a text from any genre can be about any topic (Finn and Kushmerick, 2006), relations between genre and topic occur frequently, for example in science fiction (genre) novels and first encounters with an alien race (topic). While conceivably the topic of a first encounter with an alien race could occur in other genres (historical fiction, for example) the likelihood of that taking place is unlikely. A problem arises when there is, for instance, a science fiction novel and a book about science fiction novels. Should they both be classed as science fiction, as the genre of one (the novel) is science fiction and the topic of the other is about science fiction?

Fiction is most commonly classed according to genre (Ward and Saarti, 2018), while distinctions are sometimes made for short stories, novellas, and novels.

Sections for mystery/detective novels, horror, fantasy and so on are commonplace across physical bookstores and libraries as well as digital ones, whereas with non-fiction the texts are often grouped according to topic, for example it would be expected that books about the Vietnam War will be grouped together, even though the range of topics within that specific topic would most likely vary significantly. Short stories, novellas, and novels will more often than not be classified according to genre and not the length of the story.

Both of the aforementioned dictionary definitions related genre to style. Karlgren (2004) pointed out that style is distinguishable from genre, where style is a consistent tendency to make certain linguistic choices, and genre is to be understood as groupings of documents that are stylistically consistent and intuitive to accomplished readers of the communication channel in question. For Finn and Kushmerick (2006), genre is an abstraction that is based on a natural grouping of documents written in a similar style and is orthogonal to topic. They view genre as referring to the style of text used in a document, and a genre class as a set of documents written in a similar style that serves a useful function for the users of the document collection.

This thesis takes into account form, content, and purpose, as well as reader expectations when defining text genres. Detective stories, fantasy, and science fiction were chosen as they are well-defined in the sense that the average reader should have expectations of their content (what they would be about), realize what the purpose of the texts are, and would thus be able to place them in their respective genres. Where this human reader would most likely be able to look at a work of fiction and identify what genre it falls under by title, cover, content, form, and what they would expect when reading it, the idea in this thesis is to see how well the computer can do so. While a person's definition of fiction genre can be subjective, a computer's can not. The person would know that a word such as 'I' or 'me' would generally not be used in academic genres, but the computer cannot make this distinction on its own, unless it has been programmed in some way to do so, or in the instance of this thesis, aided through supervised learning in a stylometric classification task.

5.2 Stylometry

Computational styolometry is concerned with the style of writing in a quantifiable manner. Certain features are measured, such as the number of words used per sentence, as it is expected that the author of a text uses a certain style of writing which in turn can be used to identify them as the author of the text. This measurement of writing style by individual authors can be extended to groups of authors, such as stylometric studies examining the differences in writing style between men and women (Rybicki, 2016). In turn, then, groups of authors writing genre fiction should, conceivably, have a certain style of writing that identifies their work as being written under their respective genre.

To do this, text categorization using stylometry is generally not concerned with categorization by topic, but rather categorization by writing style (Koppel, Argamon, and Shimoni, 2002). This is accomplished, or to be accomplished in the case of this thesis, through the use of topic-independent features such as part-of-speech tags and punctuation marks. The advantage of topic-independent features used in stylometry is that, without them, text classifiers can easily overfit to topics because of the correlation (as mentioned in the preceding section) between topic and genre (Kessler, Numberg, and Schütze, 1997). As computational stylometry is involved in counting things that the author/s are not aware of, it can be said to be focused on subconscious elements of style.

Stylometry, then, hopes to capture the essence of writing style through quantifiable criteria, which makes it suitable for large quantities of data. Instead of individual authorial style, a study looking at a significant amount of text documents such as the one in this thesis is apt for a quantifiable – a stylometric – analysis.

In the case of stylometric classification and genre, user expectations and purpose are not of concern in what makes the genre; rather countable measures are. If the algorithm shows significant success in such a classification then fiction genre has been determined on the quantifiable level where stylistic markers such as punctuation marks, average word and sentence lengths, parts-of-speech tags and so on are what differentiate one genre from another. As previously mentioned, these stylistic markers are independent of topic (Lex *et al.*, 2010), meaning that machine is not looking for quantifiable similarities between topics but rather between genres as intended. The idea of whether or not the text is a novel or short story is of no concern, not only because the texts in the corpus were extracted from the Gutenberg website regardless of their classification according to length, but because a different type of style pertaining to genre is sought after: a quantifiable one known as stylometry.

Stylometry is not without its shortcomings, however. Elements of style such as an author's use of metaphor and simile, for example, is something that sylometry falls short of identifying with the author/gender/genre. Another shortcoming is that an author's style can change over time, or they can change their style depending on the topic they are writing about. With a corpus of texts from fiction genres, two authors could have different styles despite writing under the same genre, and with more authors under that genre such as in this thesis' corpus the likelihood of that increases.

Despite this, the use of quantifiable stylistic features to determine genre is a prospect that this thesis feels can be investigated more, if only to see how to improve upon the use of such a method but ideally to see if quantifiable writing style can be used to distinguish between fiction genres.

6 Methodology

In this study a corpus containing texts extracted from the Project Gutenberg website was used to train a classifier for genre identification. All the available texts under the following genres from http://www.gutenberg.org were used: detective fiction (521 texts), fantasy (149 texts), and science fiction (1308 texts).

Following preprocessing, the experiment was run for the lexical features classification. The experiment was then run with all the features for the stylometric classification, and run again after being altered by the removal of certain feature sets to evaluate the performance. For the lexical experiment, feature selection was based on the chi-square measure using 300 terms subsequently weighted by tf-idf. A 10-fold cross-validation procedure was followed for the lexical and stylometric features whereby the corpus was divided into ten equal parts and in each fold a part was used as the test set and the remaining parts were used as the training set. The average results of each fold then provides the final results. The Naive Bayes classifier with Laplace 1 (see Naive Bayes section) was used as the learning algorithm in all instances.

This experiment is concerned with the machine’s ability, via supervised learning, to differentiate between genres, or rather how successful the machine will be at doing so through the use of a classifier. There are numerous options available through which to gauge this success, namely accuracy, precision, recall, or precision and recall together in what is known as the F_1 value or score. The minimization of false positives was deemed most desirable as the classifier correctly assigning the appropriate texts to their respective genres would result in a “most-correct” classification. As such, precision was used as the main determiner of the classifier’s success, with recall and the F_1 value also taken into consideration.

In the following sections we begin with an overview of the corpus used in the experiment before introducing and elaborating on the preprocessing steps that were applied. In the next section the algorithms behind the experiment are looked at in more detail, followed by the feature selection that was used. This chapter ends with a section on how the results were evaluated. The aim is to provide explanations and elaborate on why these methods were chosen.

6.1 The corpus

All the texts were extracted using R from the Gutenberg website. The reason for this is that R has a Gutenberg package (gutenbergr) that allows users to search, download, and subsequently use texts from the Gutenberg website. The switch from R to Python occurred because the text mining/classification packages in R were struggling to handle the amount of content. Both R and Python are suitable for this type of experiment so nothing was neither gained nor lost.

The texts were found under the Fiction Bookshelf on the Gutenberg website. Because all the texts from the detective fiction, fantasy, and science fiction genres were extracted, it is natural that there was some discrepancy between the amounts of texts (521, 149, and 1308, respectively). Related to this would be the actual sizes of the texts, which would naturally vary. To account for this, term frequency normalization and feature scaling was used (mentioned below under Algorithms).

The Project Gutenberg site is one of the largest and most successful digital libraries (Kresh, 2007) with over 50 000 books available online at www.gutenberg.org. Its mission statement outlines its goal, which is "(t)o encourage the creation and distribution of eBooks" (Hart, 2004), and also goes on to state that the Project aims to "provide as many eBooks in as many formats as possible for the entire world to read in as many languages as possible". This is possible as the books have fallen out of copyright and can be freely accessed by anyone, making it a suitable candidate for the nature of this experiment, especially considering the stringent EU laws regarding such things. For example, even if the material is used just for research with no intention of distribution (indeed, the corpus used for this experiment is not included in any shape or form with this thesis), the law would prevent such a thing from occuring in the first place. This no doubt hinders research as it limits the available material, however Gutenberg was both suitable and legal for this experiment.

It should be mentioned that under the fiction bookshelf, there are only 18 categories, three of which are detective, mystery, and crime. These three genres could most likely be placed under one all-encompassing detective/mystery/crime genre. Already there is an overlap of books on the site - for example, under gothic fiction the site says that the genre (gothic fiction) contains elements of both horror and romance, meaning the same text could appear twice or even three times under different genres. The same text by Edgar Allan Poe illustrates this as it has been placed under both the gothic bookshelf and the mystery one. According to Deng, Li, Weng, and Zhang (2018), in library science a book is classified to a class or category (presumably genre) if at least 20% of the content of that book has content about/pertaining to that class or category. This could explain texts on Gutenberg being under different categories at the same time, and could possibly provide a compelling argument about the need for a different kind of classification, such as the kind of text classification presented in this thesis.

So while the texts on Gutenberg were suitable for this experiment, their categorization meant that special care had to be taken when selecting which genres to use. It would not have been feasible to go through the genres with large amounts of texts one by one to remove the duplicate texts, nor would it have been feasible to use a genre that contained too few texts (such as the erotic fiction genre on Gutenberg which contains about ten texts). As far as could be seen, however, there was no overlap (at least not a significant) of the texts under the detective fiction, fantasy, and science fiction categories on the Gutenberg site.

6.2 Preprocessing

Preprocessing is a vital step that cleans the text in the documents and can noticeably influence the success of the classification (Uysal and Gunal, 2014). An important objective is to normalize the text, for example by converting upper case words to lower case ones and performing stemming. Words such as conjunctions and prepositions are often removed as they are generally unnecessary for the classification (Kobayashi, Mol, Berkers, Kismihok, and Den Hartog, 2018).

In this experiment, the following preprocessing steps were taken:

For all texts using both the lexical and stylometric features the licensing information at the top and bottom of the texts were automatically removed. The same preprocessing step was done in the Boran *et al*. study (2016) and is done to remove the influence the same licensing information would have on the classification.

For the lexical features the corpus was normalized. Punctuation, numbers, and stop words were removed, unnecessary white space was stripped away while single spaces were retained, Porter stemming was performed, and the text was converted to lower case. These preprocessing steps are regularly utilized in text classification studies (Syamala, Nalini, Maguluri, and Ragupathy, 2017, Oevermann and Ziegler, 2018), and allows the classification of the document to focus more on the topics (through words/terms) within the texts as opposed to a stylometric classification.

Stop words are the words that are filtered out, and are generally the most common words, or words without semantic value on their own, in the given language. In the majority of documents stop words occur at a high frequency and need to be accounted for, most often by removal (Alam and Yao, 2018). Stop word removal should not have a negative impact on the model's performance, and is beneficial in general (Toman, Tesar, and Jezek, 2006). For this experiment, the stop words provided by the NLTK module for Python was used. The module removes a total of 127 stop words such as possessive pronouns and conjunctions, presented in Table 1 below.

'ourselves', 'hers', 'between', 'yourself', 'but', 'again', 'there', 'about', 'once', 'during', 'out', 'very', 'having', 'with', 'they', 'own', 'an', 'be', 'some', 'for', 'do', 'its', 'yours', 'such', 'into', 'of', 'most', 'itself', 'other', 'off', 'is', 's', 'am', 'or', 'who', 'as', 'from', 'him', 'each', 'the', 'themselves', 'until', 'below', 'are', 'we', 'these', 'your', 'his', 'through', 'don', 'nor', 'me', 'were', 'her', 'more', 'himself', 'this', 'down', 'should', 'our', 'their', 'while', 'above', 'both', 'up', 'to', 'ours', 'had', 'she', 'all', 'no', 'when', 'at', 'any', 'before', 'them', 'same', 'and', 'been', 'have', 'in', 'will', 'on', 'does', 'yourselves', 'then', 'that', 'because', 'what', 'over', 'why', 'so', 'can', 'did', 'not', 'now', 'under', 'he', 'you', 'herself', 'has', 'just', 'where', 'too', 'only', 'myself', 'which', 'those', 'i', 'after', 'few', 'whom', 't', 'being', 'if', 'theirs', 'my', 'against', 'a', 'by', 'doing', 'it', 'how', 'further', 'was', 'here', 'than'

Table 1: Stop word list from the NLTK module for Python.

Stemming is the reduction of words to their word stem, base, or root form. This reduces the words so that they may unify across all the documents. For example, words ending in "-ed," "-ly," and "-ing" would have those parts removed. Words like "eating" and "eats" would thus be reduced to "eat." Lovins introduced the first stemmer in 1968, using a dictionary of 294 suffixes along with complex rules dictating conditions under which the suffixes could be removed. In 1980 Porter’s stemming algorithm came into being (Porter, 1980), which is not only a simpler version but a popular one as well, and is the stemmer used for this experiment. Porter stemming removes suffixes from words in English.

White space results from the spaces left over from the words that were deleted as well as the spaces that were not removed. Sequences of the space character, the tab character, and the line break fall under white space. As part of the cleaning up process this white space is then removed.

For the stylometric classification, no preprocessing steps besides the removal of the Gutenberg licensing information at the top and bottom of the texts were taken. This is because features such as the average sentence length were part of the feature selection and are used to determine vocabulary richness, and the stop words are necessary and important to keep the documents topic-independent (López-Escobedo, Méndez-Cruz, Sierra, and Solórzano-Soto, 2013). While preprocessing is an important part of text classification, in some instances it is not beneficial for the classification but rather would impede its success and thus is not performed (Michos *et al*., 1996; Lex *et al*., 2010; Boran *et al*., 2016; Onan, 2018).

6.3 Algorithms

6.3.1 Normalization

One issue when working with a large amount of texts is that they will inevitably vary in length. As the features in the different-sized texts all need to contribute equally, normalization is a technique used to put all terms on equal footing and was used in this experiment.

A distinction can be made between term normalization and term frequency normalization. Term normalization removes variants and standardizes the terms. The removal of punctuation, stemming, conversion of upper case letters to lower case, and stop word removal are examples of term normalization. A word such as U.S.S.R. would find itself equivalent to USSR.
Term frequency/document length normalization, on the other hand, takes each term’s weight and makes it proportional to the length of the document (Manning *et al.*, 2008). This normalization looks at the row of the document vector and transforms it to have a unit length of 1.

For this thesis, the ℓ^2 (Euclidean) norm was used as the normalization method for the stylometric features. The ℓ^2 norm is a distance based measurement used to normalize a document’s vector by measuring its length (while the Cosine similarity, for example, measures the similarity between the angles of the vectors). Every value in the vector was divided by the ℓ^2 norm (the geometric length of the vector), which then made the document vector’s length exactly 1. The ℓ^2 norm is a commonly used norm in machine learning (Jia, Miratrix, Yu, Gawalt, Ghaoui, Barnesmoore, and Clavier, 2014; Miratix and Ackerman, 2016) and keeps the model’s coefficients small as the length of the document vector is reduced to 1. This action was performed as compensation for the unequal lengths of the documents as longer documents are likely to have higher frequencies.

6.3.2 Feature scaling

While the length of the documents in an experiment such as this can be accounted for with term frequency normalization, another step would be needed to account for the magnitude of the features. Without such a step features with high magnitudes, or large features, would weigh more in the distance calculations than low-magnitude (or small) features. For this step feature scaling was applied to account for optimization problems whereby each feature is scaled in the column so that the highest value in the column is 1 and lowest value is 0. The idea, essentially, is that all features are weighted equally so as not to affect the result of the classifier.

When the features are too large, for example numbers in a large range such as 320, 340, 360, 400, 420, 440 and so on, the large features will be given more weight which can result in an incorrect classification. By applying feature scaling to [0, 1] for example, the range of numbers can be transformed into something like 0.2, 0.4, 0.6, 0.8, 1.0, resulting in the features being brought to the same magnitude. This ensures that just because some features are large they will not end up being used as the main predictor, and the features will then have an equal opportunity to influence the weight.

For the stylometric experiment Z-score normalization (standardization) was implemented whereby each feature value was subtracted by the column mean and then divided by the standard deviation of that feature (min-max scaling).

While feature scaling is generally not necessary with the Naive Bayes algorithm as it does it by design, it was still performed as not only is it better to err on the side of caution, but as it was assumed that the features have a Gaussian distribution (elaborated on in the Naive Bayes section below), the Gaussian Naive Bayes algorithm was used for the stylometric classification. Feature scaling has also been shown to perform well with the algorithm (Cuzzola, Jovanovic, Bagheri, and Gasevic, 2015).

6.3.3 Cross-validation

The aim of cross-validation is to estimate how accurately a predictive model will perform in practice and the way this is done entails the data set being partitioned into *K* bins. Out of these *K* bins, *K*-1 are used for training and 1 for testing, performed in such a manner that each bin is used for testing exactly once. It is most often used when the goal is prediction, and where there is prediction there is a model, which in turn brings about errors, or rather error concepts.

The two error concepts are the training error, and the test error. The training error is the average loss over the training sample and occurs when applying the model to the same data that has been trained. The test error is the prediction error that is incurred over an independent test sample.

With the model, a validation set measures the performance of it after it has been trained. After training an error estimation for the model is made, namely an evaluation of residuals that results in a training error (the error that occurs when the trained model is run back on the training data). The training error provides an estimate of the difference in predicted and original responses. The trade-off between high bias (which causes underfitting) and high variance (which causes overfitting) are a problem. An underfit model will have both a high training and a high test error. An overfit model will also have a high testing error as well as an incredibly low training error. An issue, then, is that no indication is given as to how well the learner will generalize with an unseen or independent set of data.

A simple solution for this is the holdout method, which removes part of the training data and uses it to get predictions from the model trained on the rest of the data. The first problem with this is the removal of data, which can lead to underfitting. The second problem is that the holdout method has problems with high variance as it is not sure which data points will be in the validation set. This means that there could be different results for different sets of data.

One such solution to this is *k*-fold cross-validation. This method leaves enough data for both the training model and for validation.

K-fold cross-validation takes the samples and randomly partitions them into *k* subsets (known as folds) that are more-or-less equally sized. *K* is a single parameter, and is assigned to the number of groups that the data is sample is going to be split into. A value for *k* can be chosen, such as *k* = 10, hence 10-fold cross-validation. During *k*-fold cross-validation the aforementioned holdout method is repeated *k* times. A random shuffling of the dataset takes place. When this occurs, one of the *k* subsets gets used as the test/validation test while the other *k* subsets are put together to make a training set (Onan, 2018). A model is fitted on the training set and evaluated on the test set. This occurs for each group of *k* subsets. The error estimation is averaged, the score retained, and the model discarded. Both bias and variance are reduced as every point of data is in a validation set once and in a training set *k*-1 times.

For this thesis the value of *k* = 10 was chosen for the lexical and stylometric features as it is not only common in machine learning tasks (Lex *et al.*, 2010; James, Witten, Hastie and Tibshirani, 2013; Onan, 2018), but is also the usual norm in classification data training (Pennacchiotti and Popescu, 2011). *K* = 10 generally results in a model skill estimate with both low bias and a moderate variance.

6.3.4 Naive Bayes

Besides the data set used in the experiment(s), the features (detailed in the section after this one) also have to be considered when choosing the classification algorithm. When working with short feature vectors, for example, *K*-nearest neighbor is an appropriate classification algorithm, while support vector machines would be near-insensitive to the feature vector length (Makrehchi, 2015). As this experiment uses numerous features, the Naive Bayes classification algorithm is appropriate. What follows is an explanation of the algorithm and how it was utilized for the lexical and the stylometric classifications.

The Naive Bayes algorithm is a probabilistic classification technique stemming from the Naive Bayesian theorem. It is not only a common classification technique (Lex *et al*., 2010; Boran *et a*l., 2016; Onan, 2018) but also a convenient one. It is powerful and not only works well with large datasets but often outperforms more complex algorithms in such instances.

With Bayesian probability, a key concept is that of conditional probability. This is the probability of event B seeing that event A has occurred, or the likelihood that another event (B) will happen after something has already happened (A), and can be denoted as P(B|A).

In the formula below, P(A|B) is the posterior probability of the class, A (the target) given the predictor, and B (the attributes). P(A) is the prior probability of the class, or what was believed prior to seeing the evidence, while P(B|A) is the likelihood of which is the probability of predictor given class, that is, the likelihood of seeing the aforementioned evidence if the hypothesis turns out to be correct, and P(B) is the prior probability of predictor, or the likelihood of the evidence no matter the circumstance.

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

In the case of this thesis, the experiment is looking for the probability of a text belonging to genre class A_k. A_k denotes the kth class where k is the numeric index given that its predictor values are x_1, x_2, x_3, ...,…,x_p. This can be written as $P(A_k|x_1,\ldots,x_p)$.

Naive Bayes is often fast in relation to other classification algorithms and works well in multi-class prediction and despite the conditional independence assumption, and requires less training data. It also works well with categorical input variables compared to numeric ones, and as such is recommended if the input features are categorical. Sometimes, however, an issue known as Zero Frequency occurs. This takes place when a categorical variable has a category in the test data set that was not observed in the training data set so the model is then unable to make a prediction as it assigns a zero probability to it. The solution to this is to use a smoothing technique, the most common and frequently-used in text classification being Laplace estimation (Diab and El Hindi, 2017).

The Laplace smoother adds a small number, most often 1, to each of the counts in the frequencies for each feature. Thus there is a certainty of each feature having a nonzero probability of transpiring for each class.

While the classifier is fast and easy to implement, the assumption of independent predictors is a limitation as it is highly unlikely that one gets predictors which are entirely independent.

Two variations of the Naive Bayes classification algorithm were used in this thesis. When choosing the type of algorithm in an experiment such as this one, where the probabilities of the occurrence of the different possible outcomes is of concern, the probability distribution is the function that provides the aforementioned probabilities.

In the case of discrete data, a discrete probability distribution is apt for the sample space (set of all possible outcomes) while in the case of continuous data, a continuous probability distribution is appropriate for the sample space.

In this thesis, the lexical features experiment was run using the multinomial Naive Bayes and the stylometric features experiment was run using the Gaussian Naive Bayes. The Gaussian Naive Bayes is used for features that are continuous, that is, features that can have different values in the data set. The Gaussian, or normal distribution, is a probability distribution with a continuous cumulative distribution function. The multinomial Naive Bayes classifies a document based on the counts it finds of multiple keywords and is used when the data is discrete (non-binary data that is based on counts and can only take certain values). As multinomial Naive Bayes takes into account term frequencies in the document (Deng *et al*., 2018) it was suitable for the lexical classification. In both instances the algorithm assumes that the features (words/terms) occur independently of each other in the documents.

6.4 Feature selection

Feature selection is a crucial part of any text classification task. The choice of features were decided upon through how common they were in other studies of a similar ilk as there is no overall consensus as to which features are the best. Rudman (1998), for example, proposes that over 1000 style markers exist in stylometric research. Daunting as the amount may seem, it offers a wealth of different options to use when conducting experiments of this nature.

No matter what classification algorithm is used, a document containing thousands of words where many of them are “noisy” or redundant, or contain any other such information that is less informative to the class label, can negatively affect the performance of the classifier (Sebastiani, 2002). In order to help the classification algorithm, to supervise it and to provide it with what would be the best (or most suitable) conditions under which to work with the documents, feature selection is performed.

Feature selection can be viewed as preprocessing. Where it differs from preprocessing steps such as white space removal, stop word removal, stemming and so on is that the aforementioned steps are transformations to the texts, while feature selection is the filtering of the texts (Feinerer, Hornik, and Meyer, 2008). While difficult to know for certain what features should be selected, what was hoped to be relevant features were selected based on previous research (Lim *et al*, 2005; Lex *et al.*, 2010; Boran *et al.*, 2016; Neal *et al.*, 2018) for the machine learning model used in this thesis. Both topic independent stylometric features as well as common lexical features were chosen and utilized.

Average sentence length, average word length, number of characters, number of punctuation marks, number of uppercase words, number of title case words, and parts-of-speech tags for nouns, verbs, and adjectives were generated as the feature sets for the topic independent stylometric features. For the lexical features, the univariate filtering method was used whereby terms were selected by means of ranking according to the chi-square measure, which was then followed by the selection of the top K terms ($K = 300$). Both tf-idf and chi-square are unsupervised term weights metrics, meaning that they assign weights to terms to allow for the discrimination of the important terms from the less important ones. They are both frequently used in text classification experiments (Haddoud, Mokhtari, Lecroq, and Abdeddaim, 2016).

One issue that arises from the feature sets above is the inconsistency between text sizes and word counts for example. If frequencies on their own were counted, it would result in discrepancies between data points in texts of different sizes. This was taken into account through the normalization/standardization of the feature sets for the stylometric classification, as well as the use of the Naive Bayes classification algorithm to minimize such inconsistencies (discussed in a preceding section but mentioned again as it is a potentially pertinent issue). What follows below are explanations of the features and justification for their use in this experiment.

6.4.1 Lexical features

Univariate filtering (where each feature is evaluated independently) for the lexical experiment was used whereby the texts were vectorized using tf-idf weighting and chi-square based feature selection in order to keep the 300 best terms.

Term frequency-inverse document frequency (tf-idf) aims to statistically represent how important a word is to a document in a corpus. Local term frequency refers to the number of times a term occurs in a document and global term frequency refers to the total number of occurrences of a certain term. Term weighting provides a value indicating how important a term is in relation to the document. The higher the weight of the term, the more the term indicates about the document.

When documents are emphasized because common words are used more frequently, other meaningful terms present are not given enough weight and their importance is thus diminished. Inverse document frequency is a factor that lessens the weight of the common words in relation to the entire collection and increases the weight of the less common terms. The terms are measured by how much information they provide regardless of their rarity and is obtained by dividing the number of documents by the number of documents containing the term.

To account for a term appearing more times in longer documents than shorter ones, term frequency is usually divided by the length of the document for normalization. Thus, TF(a) is the number of times the term ‘a’ appears in a document divided by the total number of terms in said document. Inverse document frequency then measures the importance of the term. Here the frequency (IDF, or rather, log {DF}) is attained by the total number of documents divided by the number of documents containing term ‘a.’ If the TF is, for example, 0.14 and the IDF is 1.64, then the two are multiplied to get the tf-idf for the term ‘a,’ resulting in 0.229 (from 0.14 x 1.64).

What td-idf means, then, is that words which contain the greatest information about a document will then be the words that appear many times in that document, while appearing less or not at all in others. Tf-idf is commonly used, popular, and has many variants (Peng, Liu, and Zuo, 2014). By giving each feature vector component related to a word of the vocabulary in the texts a weight to estimate its importance, it was deemed an apt weighting method for the lexical classification as it hopes to capture more of what the text is about (the topic) than the stylometric classification (which is topic-independent). As tf-idf weighting is based on the bag-of-words model (discarding the position and structure of words in texts), it is thus is unable to capture semantics. The focus, then, is not on the meaning of what the texts contain but rather the terms of importance in them, while the stylometric feature selection is concerned with features of a quantifiable nature.

Widely used in text categorization (Deng *et al.*, 2018), the chi-square test is a statistical test of independence used to to determine the dependency of two variables and is performed on the features of a dataset to ascertain how a feature is correlated to another feature. The chi-square test does exactly that - it tests whether or not a dependent feature variable is related to an independent feature variable. If the class label feature variable is the target variable, then chi-square calculates the relationship between that target variable (the class label) and the other feature variables to see if there is a relationship between the two. If they are independent of one another, the relationship is not of significance and the feature variable can be discarded, while if they are dependent of one another then the relationship between the two is important. This dependency can be seen as a feature being important enough to belong to the class, for example the detective class. Thus, if the word is frequent in many categories, the chi-square value is low, while fs the word is infrequent the value is high.

The univariate chi-square method is commonly used in machine learning (Haddi, Liu, and Shi, 2013) when it is assumed that some feature variables are independent of the class variable (Forman, 2003). Because of the frequency of its use and the nature of the lexical experiment (words/terms favored over semantics), it was used for this thesis as well.

6.4.2 Stylometric features

The average sentence length, the average word length, the number of characters, the number of punctuation marks, the number of uppercase words, the number of title case words and PoS tags for nouns, verbs, and adjectives were generated as the stylometric features. These are mentioned below under document-based features, word-based features, character-based features, and parts-of-speech features.

6.4.2.1 Document-based features

The average sentence length of texts is often used in authorship attribution studies as it is a characteristic of particular authors. In this thesis it is used in the same manner but making the assumption that authors writing in a specific genre are prone to using similar sentence lengths. A genre such as fantasy for example may make use of longer sentences due to using an excessive amount of adjectives for the world-building process, while detective fiction might use less words in order to present the colder, harsher reality associated with crime the solving of it. Sentence length as a style marker in prose has long been used (Yule, 1938), and is used in text classification studies still (Duric and Song, 2012), and is a stylistic marker in that it is quantifiable and aiming to capture authorial style that was written on the subconscious level (as sentence length cannot be said to be something the author is constantly and consistently aware of when writing genre fiction).

6.4.2.2 Word-based features

The average word length is often used in conjunction with the average sentence length in stylometric studies (Zheng, Li, Chen, and Huang, 2006; Duric and Song, 2012). Along with character-based features such as upper case words and title case words, they are used to capture stylistic traits at each level and have the additional advantage of being able to be applied to any corpus (Neal *et al.*, 2018). In the Gutenberg fiction genre text documents, the word was considered of importance due to the nature of the documents. The ability to capture it (that is, utilize its presence) in both the lexical and stylometric classifications provided a good opportunity to showcase how it (the word/term) can be used differently in text classification experiments depending on the purpose. In this case that purpose was a comparison of lexical and stylometric features in a text classification experiment.

6.4.2.3 Character-based features

The number of characters, the number of punctuation marks, the number of upper case words, and the number of title case words were used as the character-based features. Punctuation mark counts have proved both popular and successful in previous studies (Stamatatos *et al*., 2000; Lim *et al*, 2005; Lex *et al*., 2010; Boran *et al*, 2016).

Chaski (2001) hypothesized that punctuation frequencies could differentiate between authors. Once more we find stylometric features applied to an authorship attribution task, but like before it suits the purpose of this thesis.

The author found punctuation mark frequencies to be mostly successful with promising results except for one disappointing result where the technique failed in clustering the document with the actual author. Despite this punctuation mark counts were deemed suitable not just because of their success in previous studies but because of their use in genre fiction and the ability of a stylometric classification to utilize their ability to be quantified.

6.4.2.4 Parts-of-speech features

PoS tagging is used in corpus linguistics to identify words in particular parts of speech as belonging to a specific word class. Essentially, PoS tags are another way labeling text data, in this case the words (tokens) in the text. For this experiment nouns, verbs, and adjectives were identified and appropriately labeled. PoS are apt for this experiment as works of fiction are written in full, complete sentences. In addition to this, the process of PoS tagging is quick (faster than parsing) and considered accurate (Dale, 2000).

Similar to Lex *et al*.'s (2010) use of feature spaces with nouns, verbs, adjectives, and adverbs, the PoS tags that were used in the feature selection for this thesis were nouns, verbs, and adjectives. When one considers the features of writing, that is, the writing present in works of fiction, grammar involving these three key features of any language naturally spring to mind, whereas for something such as a sentiment analysis classification, adverbs and adjectives have been reported to perform better as the PoS tags (Xia, Zong, and Li, 2011). Stamatatos, Fakotakis,and Kokkinakis (2001), for example, used a method to extract PoS features that performed better compared to lexical-based feature approaches. As such it was decided that these features, for a stylometric classification compared to a lexical one, were suitable.

6.5 Evaluation

When evaluating the performance of a classifier, one can distinguish between the measures of efficiency and effectiveness (Sebastiani, 2005). Efficiency takes into account the time required to build a text classifier (training efficiency) and the average time required to perform a classification of a document (classification efficiency), while the effectiveness measure regards the average correctness of the performance of the classifier. While both criteria are important, one is of particular concern in the field of text classification.

For text categorization purposes, effectiveness is considered to be more important and can be used as the general term for measures which evaluate the quality of classification decisions such as precision, recall, the F_1 score/measure (which combines precision and recall into one joint aspect), and accuracy (Gonçalves, 2011).

The measure of accuracy determines the percentage of the correct assessments of the document categorization done by the classifier or in other words, how often the classifier is actually correct. It does this by taking the the number of correct predictions made and dividing them by total number of predictions made. This is then multiplied by 100 to convert it into a percentage. Relying on the accuracy value alone can be problematic since the classes can contain a highly unequal number of members. In this case the classifier allocates all the documents to the most populated class, producing a very high accuracy value which is misleading (Sebastiani, 2005).

As such, the measures of precision and recall are therefore necessary to perform a more accurate and consistent evaluation of the model. Precision calculates the percentage of the documents assigned to a certain class by the classifier that actually belong to the class according to the training set. Recall calculates the ratio of the documents belonging to the target class according to the training set that were retrieved by the classifier to the total amount of the documents in the target class.

Recall is defined by the number of true positives divided by the number of true positives plus the number of true negatives, while precision is then defined as the number of true positives divided by the number of true positives plus the number of false positives. When the focus is on the minimization of false negatives, the closer recall is to 100% (with precision not suffering too much), the better. If the focus is on minimizing false positives, then the closer precision is to 100% the better.

The F-measure tests accuracy and let's us know the harmonic mean of precision and recall, where perfect precision and recall would result in a score of 1, while a score of 0 would be the worst value of an F-measure.

The experiment in this thesis is concerned with the classifier assigning relevant texts to the correct classes that actually are relevant, so the minimization of false positives is desirable and the main determiner of success will be precision, while still taking into account recall and consequently the F-measure. Accuracy, while reported, was not considered a suitable measure for assessing the model's performance as measures such as recall and precision simply work better for the assessment of imbalanced classification tasks and issues such as the accuracy paradox don't have to be taken into consideration.

6.5.1 Confusion matrix

A confusion matrix (or error matrix) is a contingency table that describes the performance of a classifier by listing the amount of true positives, true negatives, false positives (type I errors), and false negatives (type II errors).

To elaborate on the four types of result that are possible from each classification: a true positive would occur when the human classifier and the machine classifier agree on assigning a particular document to a particular class, a false positive result (Type I error) would occur when we predict the document as belonging to a specific class but in actuality it does not belong to that class, a true negative would occur when we predict the document as not belonging to a specific class and it does not belong to that class, and finally a false negative would occur when we predict the document as not belonging to a specific class when it actually does belong to that class.

The confusion matrix was used in this experiment in order to see where the classifier correctly and incorrectly assigned documents to the classes.

7 Results

7.1 Lexical features

Classification for the lexical features used multinomial Naive Bayes with 10-fold cross-validation. The results are presented in Table 2.

	Detective	Fantasy	Science fiction
Precision (%)	100	100	69
Recall (%)	7	15	100
F_1 value (%)	14	26	11

Table 2: Results for the lexical classification.

The high precision and low recall for the detective and fantasy classes indicates that the documents the classifier ascribed to the respective classes truly belonged to those classes, however there were many documents it did not assign to those classes that belonged in them. For the science fiction class, the high recall and somewhat low precision indicates that the classifier incorrectly labeled a significant amount of documents as belonging to that class. Accuracy for the experiment was 0.692, or 69%. The confusion matrix is presented in Table 3.

	Detective	Fantasy	Science fiction
Detective	39	0	482
Fantasy	0	22	127
Science fiction	0	0	1308

Table 3: Confusion matrix for lexical classification.

In the confusion matrix in Table 3 the actual classes are presented in the left column. Ideally, accuracy-wise for the classes, a diagonal line from the top left to the bottom right filled with non-zeros (while the rest of the spaces in the matrix would be filled with zeros) would be a result of 100% accuracy. With the detective class, the classifier was able to identify 39 of them correctly, and 482 of them of as science fiction, resulting in the poor recall score for the detective class in Table 2. We can see a similar recall score for the fantasy class, while the science fiction class has a high recall score (none were identified as falling under detective or fantasy during the classification; however when taking into account the number of false positives – the precision score – we can see that the classifier performed poorly with the science fiction class in comparison to the other two classes).

7.2 Stylometric features

Classification for the stylometric features used Gaussian Naive Bayes with 10-fold cross-validation. The results are reported in Table 4.

	Detective	Fantasy	Science fiction
Precision (%)	58	32	88
Recall (%)	80	12	80
F_1 value (%)	67	18	83

Table 4: Results for the stylometric classification involving all features.

The classifier was most successful with the science fiction class classification. The low fantasy class scores implies a class imbalance. Accuracy for the experiment was 0.748. The confusion matrix is presented in Table 5.

	Detective	Fantasy	Science fiction
Detective	420	11	90
Fantasy	69	18	62
Science fiction	239	27	1042

Table 5: Confusion matrix for the stylometric classification involving all features.

The confusion matrix in Table 5 shows that 420 documents were correctly identified as belonging to the detective class, and 1042 were correctly identified as belonging to the science fiction class. Only 18 documents were correctly identified as belonging to the fantasy class, showcasing the problem of a class imbalance.

Next, the experiment was run again with all the stylometric features except the parts-of-speech tags removed. The results are presented in Table 6.

	Detective	Fantasy	Science fiction
Precision (%)	56	0	86
Recall (%)	82	0	80
F_1 value (%)	66	0	83

Table 6: Stylometric analysis with parts-of-speech only tags only.

While the classifier performed similar with the detective and science fiction classes, the fantasy class fared worse with only the parts-of-speech tags used in the feature selection. Accuracy for the experiment was 0.745. The confusion matrix is presented in Table 7.

	Detective	Fantasy	Science fiction
Detective	426	0	95
Fantasy	78	0	71
Science fiction	260	0	1048

Table 7: Confusion matrix for stylometric analysis with part-of-speech tags only.

In Table 7 the confusion matrix reveals that the classifier incorrectly classified 78 fantasy documents as belonging to the detective class, and 71 fantasy documents as belonging to the science fiction class.

Next, all the stylometric features were retained and the parts-of-speech tags were removed. The results are shown in Table 8.

	Detective	Fantasy	Science fiction
Precision (%)	59	35	87
Recall (%)	80	13	81
F$_1$ value (%)	68	19	84

Table 8: Stylometric classification with part-of-speech tags removed.

The results indicate that without the parts-speech-tags the classifier performed similarly to the experiment run with them (see Tables 4 and 5). Accuracy for the experiment was 0.754. The confusion table follows in Table 9.

Detective	418	12	91
Fantasy	66	19	64
Science fiction	229	23	1056

Table 9: Confusion matrix for stylometric analysis with part-of-speech tags removed

7.3 Results discussion

Precision was chosen as the main factor to determine how successfully the classifier performed. This measure was chosen as the minimization of false positives was considered suitable in an experiment where the classifier would ideally assign relevant texts to the correct (and relevant) classes. The user in search of correctly assigned texts would be satisfied knowing that the retrieved documents were assigned correctly with a minimal need to worry over retrieving an incorrectly assigned document.

The lexical features experiment produced precision scores of 100% for both the detective and the science fiction class. The recall scores for both classes were low (7% for the detective class and 15% for the fantasy class) meaning that while the classifier minimized false positives, it also resulted in a large amount of false negatives. What the classifier predicted as being true positives, however, it did so with certainty. While few results were returned, most of its predicted labels were correct.

The science fiction class in the lexical features experiment had high recall (100%) and a precision score of 68%. The classifier predicted 482 detective class texts as belonging to the science fiction class, and 127 fantasy class texts as belonging to the science fiction class as well. Taking the high recall and low precision of the science fiction class into account, the classifier’s net was cast quite wide indeed, meaning that while many results were returned a large amount of the predicted labels were incorrect.

While various combinations were performed regarding the removal of the stylometric features, none were found to have an overly significant effect compared to the stylometric analysis using all the features. Removal of the part-of-speech tags increased recall slightly for the fantasy and science fiction genres, but again nowhere near enough to be considered significant resulting in a near-replication (results-wise) of the experiment with all the stylometric features. Running the classifier using only part-of-speech tags resulted in poor precision (56%) and relatively high recall (82%) for the detective class and a 0% precision, 0% recall, and consequently a 0% F_1 score for the fantasy class, meaning the number of true positives was most likely zero while the false positives and false negatives were higher than zero. It might have been that the part-of-speech speech features had no effect on the fantasy class and thus the classifier in turn returned returned a true positive value of nothing (zero).

The experiment with all the stylometric features added had good precision (88%) and recall (80%) scores for the science fiction class. With the somewhat high precision and recall, it would seem that the classifier performed best on the science fiction class, however the low recall and low precision of the fantasy class (many false positives as well as false negatives), as well as the low precision of the classifier when predicting the detective class implies a class imbalance in the experiment, something that should have been addressed.

Overall, when taking into account precision as the main determiner of success, the classifier performed better when predicting labels for the lexical features. The poor performance of the classifier on the fantasy class suggests a class imbalance and downsampling (where a balanced dataset is created through the matching of a number of samples in the minority class with a random sample from the majority class) is needed.

8 Discussion

This thesis performed a text classification experiment for three genres of fiction extracted from the Project Gutenberg website (http://www.gutenberg.org/). Previous research has conducted investigations into web genre classification (Michos *et al.*, 1996; Lim *et al.*, 2005; Mason *et al.*, 2009; Lex *et al.*, 2010) and classification for authorship attribution using text from Gutenberg (Boran *et al.*, 2016) have been conducted, as well as genre classification via emotional content using Gutenberg texts (Samothrakis and Fasli, 2015). The overall aim was to see how well the classifier performed when discerning between the genres of detective fiction, fantasy, and science fiction, as well seeing how lexical features performed when compared to stylometric features. In that sense, it was assumed that instead of individual authorial style providing a unique voice to a text, entire genres encompassed a certain way of writing that allowed them (the genres) to be differentiated from others.

The differentiation of genres from others based on styles of writing through quantifiable stylometric features was reliant on whether or not the genres were discernible enough from one another (from the classification algorithm's perspective so to speak). The genres, however, were considered carefully and it was thought they would most likely use terms specific to them. Fantasy from that era would, again assumed, have more whimsical and florid prose, while pre-Golden Age science fiction (as well as the Golden Age science fiction that had crept into the out of copyright cracks) would no doubt be heavily focused on rocket ships, space, exotic planets, and so forth. These three genres between themselves held enough variety in their prose (in their respectful genres) to attempt to show that perhaps authors are bound to the field they were writing in, that their imaginations were constrained by the genres they themselves had chosen. The very idea of the author, regardless of voice, gender, age, and nationality had been reduced to that of the genre their works had been categorized in. The lexical classification, then, took advantage of what lay within the texts, and focused on term-weighting. The stylometric classification on the other hand was not concerned about the topics within the texts, but rather subconscious decisions made by the authors through quantifiable features such as word and sentence lengths, punctuation mark counts, and PoS tagging for nouns, verbs, and adjectives.

This careful consideration of which genres to include proved to be an ill-informed decision. Only once the study was well under way did the realization set in that it would have been better to include to more genres – more classes – in the experiment. While there is no perfectly valid excuse for not doing so, the source needs to be considered. The Project Gutenberg website's bookshelves (where the texts are arranged according to genre) is problematic. The fiction section consists of 18 genres, which includes crime fiction, detective fiction, and mystery fiction. What is the difference, and why is the distinction made when there are only 15 other options left? Gutenberg does not provide information on how books are assigned to categories. The works of Franz Kafka, for example, are listed under horror. Mary Shelley's Frankenstein, while on Gutenberg, is not under the horror nor the science fiction bookshelf. This possibly highlights the need for text classification in such matters, to see how the machine fares when assigning texts to classes, or as in this case, genres.

With genres in mind, another issue was the very definition of it – what is genre? While it is generally clear what means when the words *science fiction* are said, an issue arises when defining genre in light of previous studies, such as the web genre studies (Michos *et al.*, 1996; Lim *et al.*, 2005; Mason *et al.*, 2009; Lex *et al.*, 2010) or with studies featuring more abstract genre types (Onan, 2018). Their genres, while still genres, were different in form, content, purpose, and if one agrees with people like Rosso (2008) – which this thesis has tended to do – reader evaluations of what the genre is when encountering it. Because of this the experiment in this thesis, while building on previous research, had to select/use parts here and there from previous research, as the nature of the corpus, that is, the type of genre (genre fiction) used differed from that in other studies, or when similar, the intent differed (genre fiction classification) to authorship attribution (Boran *et al.*, 2016) or genre classification according to emotional content (Samothrakis and Fasli, 2015). This meant that the feature selection and algorithms used, while deemed appropriate, had no one specific study near-similar in nature. While it could be argued that the work in this thesis is unique, it is not (and it would be a weak argument).

Perhaps, though, the work here is unique in the sense that in 2018, more could have been done. Recent research would, for example, use more than one classification algorithm to compare them (Onan, 2018), which would have added another dimension to the results, however the overall aim was simply genre fiction classification, with the hope that stylometric analysis would perform well enough to encourage further attempts at trying to discern sub-genres of texts with a classifier from one another. As the lexical features classification was more successful (that is, had less false positives), some work is clearly still needed with what stylometric features should be used.

Another issue was the class imbalance. The poor performance of the classifier on the fantasy class suggests that downsampling is needed, as well as more features added to the stylometric features experiment, such as n-gram collocations for example. The idea, essentially, is not to see how genres differ, but rather what features to use/not use. As such, the experiments with the lexical and stylometric features can be seen as humble albeit flawed beginning that, with the necessary modifications, could see text classification transcending genres and delving into sub-genres, an enticing prospect riddled with opportunities for interesting experiments based on authorial style connected to genre.

9 Conclusion

While studies have used stylometry for authorship attribution machine learning learning tasks, few have used stylometry for genre fiction classification, especially to compare such an experiment with one using lexical features for the classification. The thesis looked at various interpretations of what genre is, and how it can be viewed in a quantifiable stylistic manner. Three fiction genres classes (detective fiction, fantasy, and science fiction) were chosen and extracted from the Project Gutenberg website to comprise the corpus used in the experiment.

The Naive Bayes classifier in conjunction with 10-fold cross validation was used for both experiments. The performance of the classifier with lexical features and stylometric features was compared. The lexical features provided a better recall once the experiments were run, implying that the classifier performed more successfully on them. The fantasy class provided problems in the stylometric classification, however, an issue that downsampling could resolve.

In terms of stylometric features performing well in this text classification experiment, the results (lexical features performing better) indicate that improvements to the design are needed. The class imbalance as well as the feature selection could be improved upon/changed, however the results were suitable enough to provide answers to the research questions, as well as address (besides the class imbalance and feature selection) other issues such as including more classes in the experiment, and possibly using more than one classification algorithm.

In spite of this, the stylometric classification performed reasonably well with the detective class and well with the science fiction class, so the potential to improve upon the experiment is there. The use of more features in the feature selection, comparing different classifiers, and a corpus of more classes would not only aid the experiment but bring it in line with current experimentation in the area that utilizes a wide variety of the aforementioned methods.

The lack of studies using stylometry specifically for genre detection, however, indicate that this is an area with a lot of exploration left in it. The use of subconscious quantifiable features used by authors in their writings to determine genre is an interesting if not fascinating one, and something well worth endeavoring to supervise a machine over.

References

Abbasi, A. & Chen, H. (2008). Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, **26**(2), 1-29.

AbeBooks: collections. (2018). https://www.abebooks.com/collections/?cm_sp=TopNav-_-Home-_-Collections (accessed on 14-11-2018).

Alam, S. & Yao, N. (2018). The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. *Computational and Mathematical Organizational Theory*, 1-17.

Amazon: literature & fiction. (2018). https://www.amazon.com/b/ref=gbpp_itr_m-3_01bf_Literatu?node=17&ie=UTF8 (accessed on 14-11-2018).

Biber, D. (1994). An analytical framework for register studies. In D. Biber & E. Finnegan (Eds.) *Sociolinguistic Perspectives on Register* (pp. 31-56). New York: Oxford University Press.

Boran, T., Voss, J., & Hossain, Md. S. (2016). Authorship categorization of public domain literature. In IEEE 7th Annual Ubiquitous Computing, *Electronics and Mobile Communication Conference*, 1-7.

Chang, Y., Chen, S., & Liau, C. (2008). Multilabel text categorization based on a new linear classifier learning method and a category-sensitive refinement method. *Expert Systems with Applications*, **34**(3), 1948-1953

Chaski, C. E. (2001). Empirical evaluation of language-based author identification techniques. *Forensic Linguistics*, 8, 1-65.

Cuzzola, J., Jovanovic, J., Bagheri, E., & Gasevic, E. (2015). Automated classification and localization of daily deal content from the web. *Applied Soft Computing*, 31, 241-256.

Dale, R. (2000). *Handbook of natural language processing*. New York, NY: Marcel Dekker.

Deng, X., Li, Y., Weng, J., & Zhang, J. (2018). Feature selection for text classification: a review. *Multimedia Tools and Applications*, 1-20. https://doi-org.ezproxy.its.uu.se/10.1007/s11042-018-6083-5

Dewe, J., Bretan, I., & Karlgren, J. (1998). Assembling a balanced corpus from the internet. In *Proceedings of 11th nordic computational linguistics conference, Copenhagen.*

Diag, D. M. & El Hindi, K. M. (2017). Using differential evolution for fine tuning naive Bayesian classifiers and its application for text classification. *Applied Soft Computing*, 54, 183-199.

Duric, A. & Song, F. (2012). Feature selection for sentiment analysis based on content and syntax models. *Decision Support Systems*, **53**(4), 704.

Eder, M., Kestemont, M., & Rybicki, J. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, **16**(1), 107-121.

Feinerer, I, Hornik, K, & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, **25**(5), 1-54.

Finn, A. & Kushmerick, N. (2006). Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, **57**(11), 1506-1518.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3, 1289-1305.

Fuller, C. M., Biros, D. P., & Delen, D. (2011). An investigation of data and text mining methods for real world deception detection. *Expert Systems with Applications*, **38**(7), 8392-8398.

Gonçalves, M. (2011). Text Classification. In R. Baeza-Yates, & B. Ribeiro-Neto (Eds.) *Modern information retrieval : The concepts and technology behind search* (2nd ed.). Harlow: Addison Wesley/Pearson.

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17, 26-32.

Haddoud, M., Mokhtari, A., Lecroq, T., & Abdeddaim, S. (2016). Combining supervised term-weighting metrics for SVM text classification with extended term representation. *Knowledge and Information Systems*, **49**(3), 909-931.

Halvani, O., Winter, C., & Pflug, A. (2016). Authorship verification for different languages, genres, and topics. *Digital Investigation*, 16, 33-43.

Hart, M. S. (2004). In Gutenberg mission statement by Michael Hart. Retrieved 31 July, 2018 from http://www.gutenberg.org/wiki/Gutenberg:Project_Gutenberg_Mission_Statement_by_Michael_Hart

Hayes, S. (1992). Enhanced catalog access to fiction: a preliminary study. *Library Resources and Technical Services*, **36**(1), 441-459.

Hogenboom, A., Frasincar, F., de Jong, F., & Kaymak, U. (2015). Using rhetorical structure in sentiment analysis. *Communications of the ACM*, 58(7), 69-77.

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, **13**(3), 111-117.

James, R. G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Berlin:Springer.

Jia, J., Miratrix, L., Yu, B., Gawalt, B., Ghaoui, L. E., Barnesmoore, L., & Clavier, S. (2014). Concise comparative summaries (CCS) of large text corpora with a human experiment. *The Annals of Applied Statistics*, **8**(1), 499-529.

Karlgren, J. & Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th International Conference on Computational Linguistics*, 1071-1075.

Karlgren, J.(2004). The wheres and whyfores for studying text genre computationally. *Workshop on Style and Meaning in Language, Art, Music, and Design, 19th National Conference on Artificial Intelligence*, Washington, DC.

Kessler, B,. Number, G., & Schütze, H. (1997). Automatic detection of text genre. In *ACL-35: 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*, 32-38.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihok, G., & Den Hartog, D. N. (2018). Text classification for organizational researchers. *Organizational Research Methods*, **21**(3), 766-799.

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic*

Computing: Journal of the Association for Literary and Linguistic Computing, **17**(4), 401-412.

Koppel, M., Schler, J., & Argamom, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, **60(**1), 9-26.

Kresh, D. (Ed.). (2007). *The whole digital library handbook*. Chicago, IL: American Library Association.

Kumaran, G. & Allan, J. (2004). Text classification and named entities for new event detection. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 297-304.

Kwasnik, B. H.& Crowston, K. (2005). Introduction to the special issue: genres of digital documents. *Information Technology and People*, **18**(2), 76-88.

Lex, E., Juffinger, A., & Granitzer, M. (2010). A comparison of stylometric and lexical features for web genre classification and emotion classification in blogs. *2010 Workshops on Database and Expert Systems Classifications*, 10-14.

Li, Q., Shah, S., Liu, X., Nourbakhsh, A., & Fang, R. (2016). Tweet topic classification using distributed language representations. In 2016 *IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 81.88).

Lim, C., Lee, K., & Kim, G. (2005). Multiple sets of features for automatic genre classification of web documents. *Information Processing and Management*, **41**(5), 1263-1276.

Liu, Y., Bi, J., & Fan, Z. (2017). Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms. *Expert Systems with Applications,* **80**, 323-339.

López-Escobedo, F., Méndez-Cruz, C., Sierra, G., & Solórzano-Soto, J. (2013). Analysis of stylometric variables in long and short texts. *Procedia - Social and Behavioral Sciences,* **95**, 604-611.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Mason, J. E., Shepherd, M., & Duffy, J. (2009). Classifying web pages by genre: an n-gram approach. *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 1, 458-465.

Michos, S., Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (1996). An empirical text categorizing computational model based on stylistic aspects. In *Proceedings of the 8th International Conference on Tools with Artificial Intelligence*, 71-77.

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech*, **70**(2), 151-167.

Miratrix, L., & Ackerman, R. (2016). Conducting sparse feature selection on arbitrarily long phrases in text corpora with a focus on interpretability. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **9**(6), 435-460.

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. (2018). Surverying stylometry techniques and applications. *ACM Computing Surveys*, **50**(6), 1-36.

Oevermann, J. & Ziegler, W. (2018). Automated classification of content components in technical communication. *Computational Intelligence*, **34**(1), 30-48.

Onan, A. (2018). An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science*, **44**(1), 28-47.

Ortigosa, A., Martin, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*, 31, 527-541.

Oxford Dictionaries: genre. (2018). https://en.oxforddictionaries.com/definition/genre (accessed on 20-04-2018).

Oxford Dictionaries: stylometry. (2018). https://en.oxforddictionaries.com/definition/stylometry (accessed on 20-04-2018).

Oxford Dictionary of Literary Terms, The. (4th Ed.): genre fiction. (2015). http://www.oxfordreference.com.ezproxy.its.uu.se/view/10.1093/acref/9780198715443.001.0001/acref-9780198715443-e-494 (accessed on 14-11-2018).

Peng, T., Liu, L, & Zuo, W. (2014). PU text classification enhanced by term frequency-inverse document frequency-improved weighting. *Concurrency and Computation: Practice and Experience*, **26**(3), 728-741.

Petrenz, P. & Webber, B. (2011). Stable classification of text genres. *Computational Linguistics*, **37**(2), 385-393.

Porter M. F. (1980). An algorithm for suffix stripping. *Program*, **14**(3), 130–137.

Rafferty, P. & Hidderly, R. (2005). *Indexing multimedia and creative works: the problem of meaning and interpretation*. Hants, England: Ashgate.

Rosso, M. (2008). User-based identification of Web genres. Journal of the American Society for Information Science and Technology, 59(7), 1053-1072.

Rudman, J. (1998). The state of authorship attribution studies: some problems and solutions. *Computers and Humanities*, 31, 351-365.

Rybicki, J. (2016). Vive la différence: Tracing the (authorial) gender signal by multivariate analysis of word frequencies. *Digital Scholarship in the Humanities*, **31**(4), 746-761.

Saarti, J. (2000). *Fiction indexing by library professionals and users*. Scandinavian Public Library Quarterly, **33**(4), 6-9.

Samothrakis, S. & Fasli, M. (2015). Emotional sentence annotation helps predict fiction genre. Plos One, 10(11). doi:10.1371/journal.pone.0141922

Satija, M. P. & Martinez-Avila, D. (2015). Features, functions, and components of a library classification system in the LIS tradition for the e-environment. *Journal of Information Science Theory and Practice*, **3**(4), 62-77.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1), 1-47.

Sebastiani, F. (2005). Text categorisation. In A, Zanasi (Ed.) *Text mining and its applications (advances in management information series)*. Southampton: WIT.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Text genre detection using common word frequencies. In *Proceedings of the 18th conference on Computational linguistics*, 2, 808-814.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, **35**(2), 193-214.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology,* **60**(3), 538-556.

Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, **21**(2), 421-439.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press.

Syamala, M., Nalini, N. J., Maguluri, L., & Ragupathy, R. (2017). Comparative analysis of document level text classification algorithms using R. In *IOP Conference Series: Materials Science and Engineering,* **225**(1), 12076.

Toman M., Tesar R., & Jezek K. (2006). Influence of word normalization on text classification. In *Proceedings of InSciT*, 354–358.

Uysal, A. K. & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management,* **50**(1), 104-112.

Varga, A., Basave, A. E. C., Rowe, M., Ciravegna, F., & He, Y. (2014). Linked knowledge sources for topic classification of microposts: a semantic graph-based approach. *Web Semantics: Science, Services, and Agents on the World Wide Web*, 26, 36-57.

Wagers, R. (1981). Popular fiction selection in public libraries: implications of popular culture studies. *The Journal of Library History*, **16**(2), 343-344.

Wang, H., Hong, M., & Lau, R. Y. K. (2018). Utility-based feature selection for text classification. *Knowledge and Information Systems,* 1-30.

Ward, M. & Saarti, J. (2018). Reviewing, rebutting, and reimagining fiction classification. *Cataloging and Classification Quarterly*, **56**(4), 317-329.

Merriam-Webster: genre. (2018). https://www.merriam-webster.com/dictionary/genre (accessed on 20-04-2018).

Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, **181**(6), 1138-1152.

Yates, J. & Orlikowski, W. (1992). Genres of organizational communication: a structurational approach to studying communication and media. *Academy of Management Review*, **17**(2), 299-326.

Yule, G.U. (1938). On sentence length as a characteristic of style in prose. *Biometrika*, 30, 363-390.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, **57**(3), 378-393.

Zubiaga, A., Vicente, I., Gamallo, P., Pichel, J., Alegria, I., Aranberri, N., Ezeiza, A., & Fresno, V. (2016). TweetLID: a benchmark for tweet language identification. *Language Resources and Evaluation*, **50**(4), 729-766.