# Author Gender Classification from Blogs: A Comparative Analysis of N-Grams-Based Authorship Attribution Techniques

Sujin Kay
Haverford College
Department of Computer Science

Faculty Advisor
Deepak Kumar
Bryn Mawr College

April 25, 2019

Submitted in partial fulfillment for the requirements
of a Bachelor of Science degree in Computer Science.

## Abstract

This thesis explores n-grams-based gender classification analyses using various n-grams types, sizes, and feature sets. This study expanded on previous research by including a non-binary gender category. Non-binary data was scraped from a website and the collection has been made available as a resource for future research. First, a state-of-the-art n-grams analysis using a simple dissimilarity measure was replicated, and peak gender classification accuracy was reached at 71%. Seeking to improve this result, a formal feature selection process was performed. This secondary analysis yielded a lower overall peak accuracy of 61%, but non-binary and female-specific accuracy reached 99–100%. Results from both analyses are comparable to findings from previous research.

**Acknowledgement**: This work was partially supported by the Undergraduate Fellowship in the Center for Science of Information (an NSF Science & Technology Center) by award NSF CCF-0939370.

# Table of Contents

# 1 Introduction

Automated Authorship Identification is a field of study in computational linguistics which strives to identify characteristics of an author's writing style based on analysis of various linguistic features. The underlying idea behind this procedure is that each person has distinct and measurable traits that comprise their individual writing style and act as an "authorial fingerprint" [Joula, 2006]. The presence or absence of various lexical, character, syntactic and semantic stylometric features (or "style markers") are what differentiates one's writing style, and can thus be used to unambiguously classify texts [Doyle and Keselj, 2005, Stamatatos, 2009]. Major applications of this discipline are *authorship attribution*, which aims to attribute a text to one author in a given set; *author verification*, which seeks to determine whether a text was indeed written by a given author; and *author profiling*, which aims to extract certain demographic characteristics of an author based on their written texts [Mikros, 2012]. This thesis focused on the third application: given a set of blogs and articles written by an individual, I investigated whether it is possible – and with what accuracy – to algorithmically classify that author's gender (male, female, or non-binary).

This thesis explored the gender classification accuracy of various n-grams-based attribution analyses on the character, word, and Part-of-Speech tag levels. Further, classification performance was evaluated across different n-grams sizes and various *feature sets*. Male and female gender profiles were built using more than 680,000 blog posts obtained from a commonly-used corpus, and the non-binary gender profile was built using over 200 online articles from a website. A *simple dissimilarity measure* was used to compare a given author profile to each of the gender profiles, and to subsequently classify the author's gender among the three possible categories. This study extended the knowledge from previous empirical research by including the third non-binary gender category; to my knowledge, no previous study had attempted or considered non-binary gender classification, and this dearth has been criticized as an ethical issue due to its dangers of oversimplification/disrespect or lack of transparency in how gender labels are ascribed [Larson, 2017].

Within author profiling, *author gender identification* has two particularly interesting real-world applications. First, it is related to information retrieval and sentiment analysis, especially in the context of product development, targeted advertising and marketing strategies [Doyle and Keselj, 2005, Mukherjee and Liu, 2012]. For example, knowing the gender of online reviewers could help a company determine which products or services are preferred by men versus women, and to then utilize that information to prioritize a specific product's expansion or to implement gender-based targeted advertising [Mukherjee and Liu, 2012]. Second, gender identification is useful in social-media-related forensics, especially with the growing prevalence of social media platforms and their use by young people today. Since it is relatively easy for criminals or predators to create fake profiles online using fabricated personal information, law enforcement agencies and social network moderators would benefit from gender identification methods that detect and flag these deceptive profiles [Peersman et al., 2011].

Given the relevance of these two applications, blog posts and online magazine articles are relevant sources for online audience gender identification and have been widely used as data for this type of attribution analysis. Blogs and online media posts tend to be informal, personal, and publicly shared texts; post topics range from product reviews to baking recipes to financial advice [Zhang and Zhang, 2010]. Further, blogging trends have been on the rise in recent years: according to a survey done by Orbit Media Studios, in 2018 the average time for a blogger to write a post is nearly 3.5 hours (versus 2.5 hours in 2014), and the average post length was 1,151 words in (versus 808 in 2014)[1]. The percentage of bloggers who use professional editors has also risen, from 12% in 2014 to 24% in 2018[1]. These trends may suggest that the content of blog posts have improved in quality, which could make them even more valuable data sources for the commercial applications mentioned previously. However, many bloggers and online article authors do not explicitly make their gender nor age information public, which makes author gender classification on blogs an increasingly relevant and useful task [Zhang and Zhang, 2010].

The following sections of this paper are organized as follows:

- Section 2 discusses gender style differences in writing, and introduces n-grams, author profiles, feature sets, and the simple dissimilarity measure used for classification.
- Section 3 addresses previous research and related work.
- Section 4 summarizes the blog corpus data and explains the non-binary data acquisition process.
- Section 5 presents the methodology of n-gram feature selection, author / gender profile generation, and gender classification. Accuracy results for both the preliminary analysis and the formal feature selection process are presented and analyzed. Lastly, ethical considerations are discussed.
- Section 6 gives suggestions for future work.
- Section 7 summarizes the main findings and conclusions.

[1] https://www.orbitmedia.com/blog/blogging-statistics/

# 2 Background

This section first explains evidence of gender differences in writing, and then introduces n-grams, author and gender profiles, and feature set selection in the context of the similarity-based author gender classification approach.

## 2.1 Gender Differences in Writing

Previous research has indeed found that there exist quantifiable differences between male and female writing, across genres and types of formal written texts. Argamon et al. studied 246 English documents from the British National Corpus and identified around 50 linguistic features that were most useful for distinguishing author gender [Argamon et al., 2003]. These included substantially higher use of certain determiners (*a, the, that, these*) and quantifiers (*one, two, more, some*) by male authors, while the use of personal pronouns (*I, you, she, her, their, myself, yourself, herself*) represented strong female indicators [Argamon et al., 2003]. Their findings also characterized female writing as presenting things in a "relational" way (i.e., relaying information as if the reader is familiar with and understands what the author is referring to); in contrast to this "involved" style, male authors were characterized as having a more "informative" style (i.e., presenting details about the things being mentioned) [Argamon et al., 2003]. Schler et al. also found certain style-related and content-related features to be substantial in distinguishing gender: female bloggers write more personal posts and use more pronouns and words that express assent/negation, while male bloggers write more political / technological / financial posts and make more frequent use of articles and prepositions [Schler et al., 2006]. These blog-based findings support the assertions made by Argamon et al. on formal written texts.

However, Larson argued that the use of gender as a variable in natural language processing studies is an ethical issue when researchers are discriminatory or disrespectful in ignoring the nuances in gender identities, or fail to thoroughly explain how gender labels are defined and assigned to authors [Larson, 2017]. Bamman et al. emphasized that previous studies of the relationship between gender and language in social media are limited in their oversimplification of gender as a male / female binary variable [Bamman et al., 2014]. They further argued that prior research fails to address cases where authors' linguistic styles contradict population-level gender patterns, and that assuming a binary stylistic opposition by design only perpetuates this assumption and reinforces a binary status quo [Bamman et al., 2014]. To my knowledge, no previous research has attempted to determine the (if any) distinguishable stylistic traits of non-binary writing; thus, my thesis aimed to do so via an extensive feature selection process.

Table 1: Python NLTK Tagset

| Tag | Meaning | Examples | Tag | Meaning | Examples |
|---|---|---|---|---|---|
| **ADJ** | adjective | *new, good* | **NUM** | numeral | *first, 1997* |
| **ADP** | adposition | *on, of, with* | **PART** | particle | *at, on, over* |
| **ADV** | adverb | *really, still* | **VERB** | verb | *is, going* |
| **INTJ** | conjunction | *and, or, but* | **.** | punctuation | *. , ! ? %* |
| **NOUN** | noun | *college, tree* | **X** | other | *lol, umm* |

## 2.2 N-Grams

N-grams-based attribution techniques are widely used and considered state-of-the-art in modern natural language processing analyses. Beyond authorship attribution, applications of n-grams include language modeling, music or DNA representation, and text compression.

An **n-gram** is defined as a sequence of n elements that occur in a given text; these elements can be bytes, characters, words, Part-of-Speech tags, or any other unit of information [Leahy, 2009, Sidorov et al., 2014, Jurafsky and Martin, 2008].

As an example, consider the following phrase on the word-level:

*is it warm today or is it cold*?

Here, the **unigrams** (1-word sequences) are "*is*", "*it*", "*warm*", "*today*", "*or*", "*cold*", "?"; the **bigrams** are "*is it*", "*it warm*", etc.; the **trigrams** are "*is it warm*", "*it warm today*", etc. For unigrams, the frequency counts are simply the word frequencies themselves. In this example, note that the bigram "*is it*" occurs twice.

Beyond characters and words, **Part-of-Speech (POS)** tags are another type of n-gram. These tags distinguish the core categories of lexical and grammatical properties of words. This study utilizes Python's built-in Natural Language Toolkit (NLTK) tagset, shown in Table 1 [Bird et al., 2009]. The other most prominent tagsets include the Universal POS tagset (17 tags; includes determiners, pronouns, proper nouns, etc.) and the extensive Penn TreeBank tagset (45 tags; distinguishes between singular and plural nouns, personal and possessive pronouns, comparative and superlative adjectives, etc.) [Universal Dependencies, 2014, Jurafsky and Martin, 2008].

Now, consider the earlier phrase but on the Part-of-Speech tag level (shown in bold below):

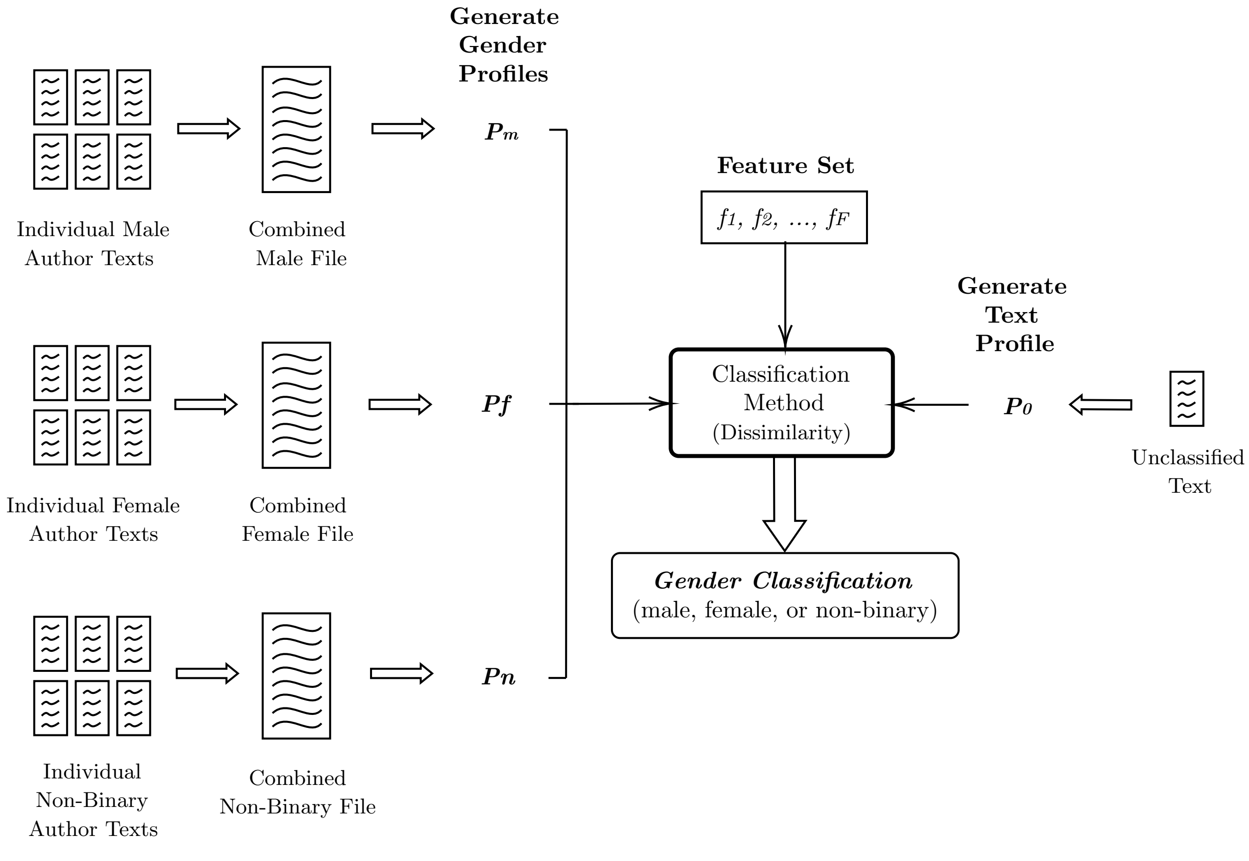
|  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| *is* | *it* | *warm* | *today* | *or* | *is* | *it* | *cold* | *?* |
| **VERB** | **NOUN** | **ADJ** | **NOUN** | **INTJ** | **VERB** | **NOUN** | **ADJ** | **.** |

Here, the unigrams are the **VERB**, **NOUN**, **ADJ**, **INTJ** and **.** tags, the bigrams are **VERB NOUN**, **NOUN ADJ**, etc. Byte and character n-grams follow the same logic. Due to their ability to capture a wide variety of stylometric patterns, n-grams are widely used as linguistic features in author profiling analyses, as discussed in the following section.

## 2.3 Author / Gender Profiles & Feature Sets

The author attribution task is to attribute an anonymous text to a known author in a candidate set. For this task, the *profile-based* (or *similarity-based*) *approach* computes a distance measurement between the anonymous text and each of the candidate authors, and attributes the unknown text to the author with the lowest distance (i.e., the least dissimilar).

Figure 1: Profile-based n-grams gender classification approach



Similarly, the gender classification task is to attribute an anonymous text to a gender. Thus, the profile-based approach can be used for gender profiling, and is also specifically applicable to n-grams-based analyses. As depicted in Figure 1, this *profile-based n-grams approach* is organized as follows:

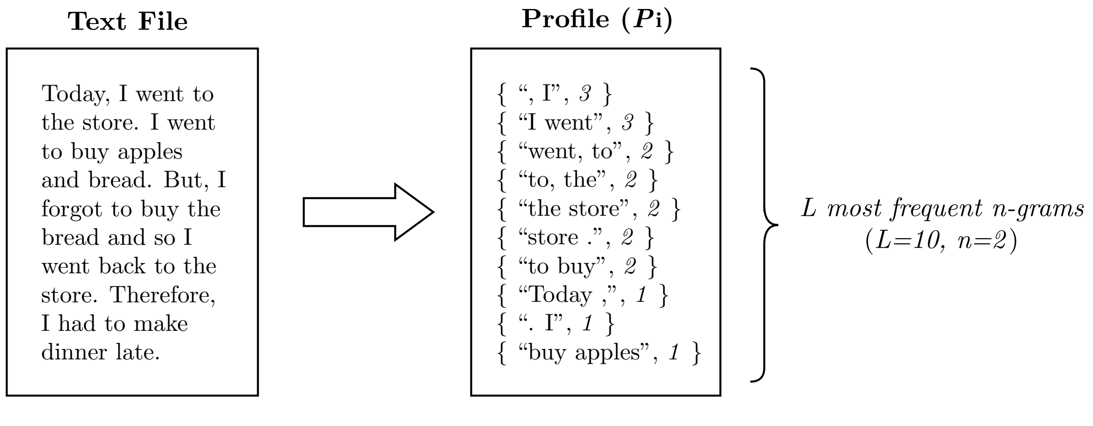
First, all individual text files are combined into their respective *gender textfiles* (e.g., the Female file contains all text data written by all the female authors). Next, each gender textfile is used to generate a corresponding *gender profile*; this profile contains the ***L* most frequent n-grams** (and their respective normalized frequency counts) in that textfile. Given an unclassified-gender author textfile, an *author profile* containing the $L$ most frequent n-grams (and their respective normalized frequency counts) in the author textfile is generated in the same manner.

Essentially,

$$\textit{profile} = \{(n_1, C(n_1)),\ (n_2, C(n_2)),\ ...,\ (n_L, C(n_L))\}, \tag{1}$$

where $n_i$ is a given n-gram and $C(n_i)$ is its normalized frequency count.

As an example, consider the document below. From the full textfile, the 10 most common bigrams and their frequencies are extracted to produce the document *profile*.



Today, I went to the store. I went to buy apples and bread. But, I forgot to buy the bread and so I went back to the store. Therefore, I had to make dinner late.

{ “, I”, *3* }
{ “I went”, *3* }
{ “went, to”, *2* }
{ “to, the”, *2* }
{ “the store”, *2* }
{ “store .”, *2* }
{ “to buy”, *2* }
{ “Today ,”, *1* }
{ “. I”, *1* }
{ “buy apples”, *1* }

Once the gender profiles and author profile are formed, a *feature set*, comprised of any number $F$ of stylometric features, is chosen. In the previous example, selected features could include average word length, the frequency of the character ‘a’, etc. The chosen feature set aims to distinguish between authors by capturing their individual writing styles, and typically combines a multitude of feature types. Table 2 shows the basic categories and their main advantages and disadvantages [Stamatatos, 2009, Leahy, 2009].

Table 2: Types of Stylometric Features and Their Advantages & Disadvantages

| Features | | Advantages | Disadvantages |
|---|---|---|---|
| **Lexical** | - Token-based (word length, etc.)<br>- Vocab. richness<br>- Word frequencies<br>- Word n-grams | - can be applied to any language using a tokenizer<br>- can disregard or capture contextual information | - depends on text length<br>- may capture content-specific rather than stylistic information |
| **Character** | - Character types (letters, digits, etc.)<br>- Character n-grams | - can capture lexical and contextual information, use of punctuation, etc.<br>- tolerant to noise (grammatical errors, etc.)<br>- language independent | - n-grams trade-off between capturing sub-word vs. contextual information<br>- higher n-grams dimensionality vs. lexical approach |
| **Syntactic** | - Part-of-Speech<br>- Sentence / phrase structure | - captures unconscious syntactic patterns and allows for structural analysis<br>- considered more reliable than lexical information | - requires robust and accurate NLP tools<br>- language and tool-specific<br>- noisy datasets due to inevitable parsing errors |

In n-grams-based analyses, the feature set selection process typically follows a straightforward bag-of-words approach. This **preliminary/baseline feature extraction** approach simply uses the *union* of the $L$ most frequent n-grams in each profile as the feature set. Alternatively, a more **formal feature extraction** process requires careful consideration of and experimentation with what stylometric features are most indicative of author gender. This thesis implemented the bag-of-words approach as a baseline, then implemented a formalized feature extraction process. Table 3 below illustrates the differences between these analyses.

Table 3: Differences between Analyses using Preliminary vs. Formal Feature Selection

| | **Author / Gender Profile** | **Feature Set** |
|---|---|---|
| **Preliminary Feature Selection** | $L$ most frequent n-grams | $L$ most frequent n-grams (author)<br>$\cup$<br>$L$ most frequent n-grams (gender) |
| **Formal Feature Selection** | $L$ most frequent n-grams | **Various n-grams-based features** (e.g., bigrams with the pronoun “her”, trigrams with the determiner “the”) |

The last step is the classification task (see Figure 1). For this similarity-based approach, the classification method shown in Figure 1 is a distance function, which computes how different the author profile and a given gender profile are, based on the relative frequency that each feature $f_i$ appears in each of the profiles. This calculation, called a *simple dissimilarity measure*, is outlined in the following section.

## 2.4 Simple Dissimilarity Measure

There exist two primary categories of authorship attribution methods. *Supervised methods* utilize machine learning algorithms to build and improve a classifier using a subset of the relevant documents (the "training data"), which is then used on the "test data" to attribute a document to an author [Koppel et al., 2013]. In contrast, in the *profile-based, unsupervised approach*, a distance function is used for classification. This thesis replicates the *simple dissimilarity measure* proposed by Keselj et al. for the authorship attribution task, which is directly applicable to the gender identification problem [Keselj et al., 2003].

$$\boldsymbol{dissimilarity}(author,\ gender) = \sum_{f \,\epsilon\, FeatureSet} \left( \frac{C_a(f) - C_g(f)}{\frac{C_a(f)\ +\ C_g(f)}{2}} \right)^2 \tag{2}$$

Given an unclassified author profile and a gender profile, each comprised of their respective $L$ most common n-grams, Equation 2 computes the dissimilarity score between the two profiles.

The specific process is depicted in Figure 2 and is explained as follows: For each n-gram $f$ in the feature set, the frequency count of $f$ in the author and gender profiles is computed ($C_a(f)$ and $C_g(f)$, respectively). Importantly, $C_a(f)$ and $C_g(f)$ are *normalized* frequency counts (i.e., the number of occurrences of the n-gram feature $f$ in the profile, divided by the total number of n-grams in the profile). Next, the square of the difference between these counts is calculated; again, this value is normalized by dividing by the *average frequency* of the feature $f$. By doing so, we avoid placing too little emphasis on less frequent n-grams; this is especially relevant as the frequency varies increasingly for larger n-grams sizes [Keselj et al., 2003].

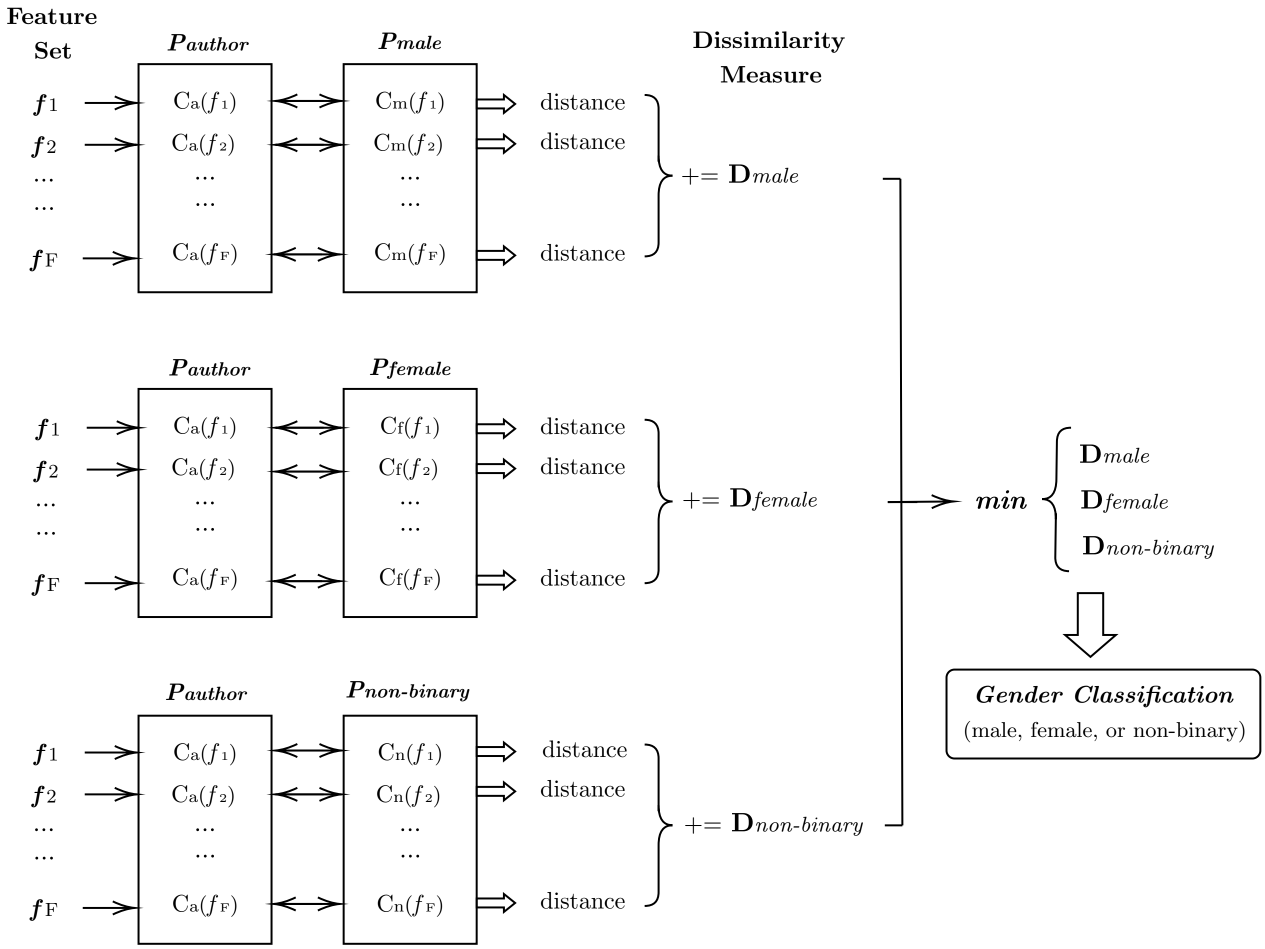
The dissimilarity calculation returns a positive number in the range [0, 1]. If an author profile is nearly identical to a given gender profile, then their dissimilarity score will be close to 0. If the two profiles share very few of the same n-gram features, then the score will be close to 1.

In my work, three dissimilarity measures are calculated:

- **dissimilarity**$(author\ profile,\ male\ profile)$
- **dissimilarity**$(author\ profile,\ female\ profile)$
- **dissimilarity**$(author\ profile,$ *non-binary* $profile)$

Finally, the author is classified as the gender for which its dissimilarity score is the minimum.

Figure 2: Author Gender Classification via Dissimilarity Calculations



# 3 Related Work

This section discusses previous research relating to n-grams-based authorship attribution and gender classification analyses, stylometric feature selection, and unsupervised vs. supervised classification methods. For each approach, ranges of accuracy and conditions under which accuracy is maximized are presented. Influenced by the most promising findings from previous research, I will also specify which features are included in this thesis.

## 3.1 N-Grams-Based Unsupervised Author Attribution and Gender Classification Analyses

In general, traditional style marker extraction / feature selection techniques are predominantly language-dependent, non-trivial processes that require subtle decision-making based on chosen thresholds for what constitute sufficiently informative features [Keselj et al., 2003, Peng et al., 2003]. In contrast, n-grams-based models have substantial advantages over other stylometric features, including that they are straightforward to calculate, with low computational costs [Stamatatos, 2009]. N-grams are a particularly popular method of analysis in various authorship attribution tasks, and have been found to provide high performance accuracy. Further, the profile-based *unsupervised* method which computes a distance function (as defined in Section 2.4) is a simple and straightforward method of attribution. Accuracy comparisons across n-grams types and approaches (unsupervised vs. supervised) are shown in Table 4 (at the end of Section 3.2).

**Byte n-grams** Byte-level n-grams are language-independent and require no text preprocessing [Keselj et al., 2003]. Using byte n-grams and the simple dissimilarity measure to classify an English data set comprised of 9 books by 6 different authors, Keselj et al. achieved accuracy ranging from **50% to 100%**, for n-grams sizes between 2 and 10, and values of $L$ between 20 and 5,000 [Keselj et al., 2003]. Despite these promising results for the authorship attribution task, byte-level analysis is not relevant for gender classification as it is unable to capture gender-differentiated style markers.

**Character n-grams** Similarly to byte n-grams, character-level n-grams can successfully capture morphological features, require no text pre-processing, and can be applied to any language (and even non-written media such as images, DNA, and music)[Doyle and Keselj, 2005, Keselj et al., 2003, Peng et al., 2003, Stamatatos, 2009]. For the *authorship attribution* task using character-level n-grams, the $L$ most common n-grams as the feature set, and the simple dissimilarity measure as the classifier, Hassan et al.'s study yielded accuracy between **40% and 73%** for bigrams and trigrams, and values of $L$ between 100 and 430 [Hassan and

Chaurasia, 2012]. Using an unsupervised Bayes Classifier approach, Leahy achieved accuracy between **80% and 93%** for n-grams sizes between 2 and 7, and values of $L$ between 1 and 45,000 [Leahy, 2009].

For the *gender classification* task, Doyle et al. utilized the similarity-based approach to classify author gender from a corpus of British student essays. They created profiles using character, word, and POS n-grams, and also used the $L$ most frequent n-grams as the feature set. On the character-level, accuracy between **51% and 76%** was reached for n-grams sizes between 1 and 5, and values of $L$ from 100 to 20,000 [Doyle and Keselj, 2005]. Further, they found that male and female authors shared nearly all of the most frequent characters (*space, e, t, i, a*) and the profiles remained extremely similar thereafter; this explains the low performance results [Doyle and Keselj, 2005].

**Word n-grams** are advantageous in their ability to combine lexical and syntactic elements to capture contextual information (e.g. "a fly" versus "to fly") [Joula, 2006, Stamatatos, 2009]. Doyle et al. found that gender classification on the word-level had substantially better overall performance than on the character-level, with accuracy between **64% and 81%** for n-grams sizes between 1 and 5, and values of $L$ from 100 to 20,000 [Doyle and Keselj, 2005].

**POS n-grams** capture information on the syntactic level, but are both language-specific and dependent on the accuracy of the tagger used [Doyle and Keselj, 2005, Stamatatos, 2009]. For the gender classification task, Doyle et al.'s study produced POS-level performance results between **42% and 76%** accuracy, which was similar to their character-level n-grams analysis [Doyle and Keselj, 2005].

Thus, it is evident that n-grams-based techniques can yield high gender identification accuracy and offer a broad versatility. Further, allowing the size of the n-grams and the profile size $L$ to vary can give insight into an author's style on the lexical, syntactical, and structural levels [Stamatatos, 2009]. For example, a small n-grams size would better glean sub-word or syllable-level information, but may not be able to capture thematic, lexical, and contextual elements. Extracting these features of variable length would promote understanding of the conditions that maximize gender classification accuracy.

### 3.1.1 Relevant N-Grams-Based Features

Based on promising results from previous research, character, word, and POS n-grams were analyzed as features in this study. Byte n-grams were excluded. N-grams size varied from 1 to 5, with values of $L$ ranging from 1,000 to 10,000.

### 3.2 Supervised / Machine-Learning Approaches

In contrast to *unsupervised methods* such as the dissimilarity measure, *supervised methods* utilize machine learning algorithms to build, train, test, and improve the chosen classifier on a subset of the data [Koppel et al., 2013]. Two recent studies have focused on the gender profiling task using word and character n-grams for profiles and feature sets. Peersman et al. predicted gender in online social networks, while Mikros et al. performed author and gender identification using a corpus of same-topic Greek blogs; both utilized a **Support Vector Machine (SVM)** for supervised training and classification, which utilizes vectors to define a hyperplane that separates points from the training and testing data into two classes [Peersman et al., 2011, Mikros, 2012].

On the *character level*, accuracy ranged between **41% and 43%** for n-grams sizes between 1 and 3, and values of $L$ between 1,000 and 50,000 [Peersman et al., 2011]. Mikros et al.'s study reached author gender identification accuracy of **83%**, but failed to explicitly convey under what conditions (n-grams size and type, $L$, etc.) this was maximized [Mikros, 2012]. Both studies found that utilizing *word-level* n-grams features improved performance, with gender classification accuracy ranging from **59% to 64%** [Peersman et al., 2011]. Interestingly, Mikros et al. found that word n-grams have increased value in the gender attribution task, whereas character n-grams were predominantly useful for authorship attribution [Mikros, 2012]. As sequences of word/character n-grams increased (from unigrams to trigrams), gender identification accuracy improved (which had the exact opposite effect in the authorship attribution task) [Mikros, 2012].

Three recent studies have utilized various machine learning techniques for *POS-level* n-grams-based classification approaches. Mukherjee et al. and Zhang et al. both utilized the same weblog dataset and focused on training SVMs using various POS-tag n-grams, POS patterns, and/or common lexical features [Mukherjee and Liu, 2012, Zhang and Zhang, 2010]. Accuracy ranged from **50% to 86%** using the SVM classifier and from **70% to 89%** using an SVM regression (for n-grams sizes from 1 to 3 and values of $L$ from 25 to 56,024 [Mukherjee and Liu, 2012, Zhang and Zhang, 2010]. Zhang et al. also used a Linear Discriminant Analysis that yielded accuracy from **61% to 65%** for values of $L$ from 50 to 1,000 [Zhang and Zhang, 2010].

Koppel et al. also investigated the use of POS-tag uni/bi/trigrams and function words as features, with a vector-based Exponential Gradient supervised approach [Koppel et al., 2002]. Using data from the British National Corpus that included both fiction and nonfiction work across many sub-genres, they found that using POS-tags in conjunction with function words yielded higher overall (fiction and non-fiction) average accuracy (**77%**) than POS-tags alone (**71%**) or function words alone (**74%**) [Koppel et al., 2002].

Although much previous research has been done using n-grams-based models on the word/character/byte/POS levels, varying the size of the n-gram, varying the size of the feature set L, and using various algorithms and classifiers, to my knowledge no research has been done to analyze all of these variables at once, using the simple dissimilarity measure and focusing on *non-binary* gender identification in blog posts.

Table 4: Accuracy across N-Grams Types and Classification Approaches in Previous Work (for Male / Female Classification)

| | **Unsupervised** | **Supervised** |
|---|---|---|
| **Byte** | **50 - 100%**<br>$n = 2 - 10$<br>$L = 20 - 5{,}000$ | |
| **Char** | **40 - 93%**<br>$n = 1 - 7$<br>$L = 100 - 45{,}000$ | **41 - 43%**<br>$n = 1 - 3$<br>$L = 1{,}000 - 50{,}000$ |
| **Word** | **51 - 76%**<br>$n = 1 - 5$<br>$L = 100 - 20{,}000$ | **59 - 64%**<br>$n = 1 - 3$<br>$L = 1{,}000 - 50{,}000$ |
| **POS** | **42 - 76%**<br>$n = 1 - 5$<br>$L = 100 - 20{,}000$ | **50 - 89%**<br>$n = 1 - 3$<br>$L = 25 - 56{,}024$ |

## 3.3 Stylometric Feature Selection and Extraction

Rather than simply employing the $L$ most frequent n-grams as the basis for the feature set, a more methodical feature selection process often plays a principal role in more granular authorship attribution or gender classification techniques. Because lexical analyses reflect the way humans naturally see texts as collections of words, and because token-level analyses retain the flexibility of being able to capture a variety of lexical, semantic, style-based and topic-based elements, the majority of non-trivial feature selection processes focus on the word-level. Still, the majority of past research has explored gender identification from formal written texts, and studying the gender differences in informal or online texts – which include personal, conversational cues – is relatively new and increasingly relevant today [Koppel et al., 2002].

**Bag-of-Words** The simplest, most straightforward lexical analyses are "bag-of-words" approaches (i.e., computing basic word frequencies of function words, articles and prepositions, etc.). Analyzing 75,000 blog entries from the Xanga blog service and performing a bag-of-words unigram analysis using a Naïve Bayes Classifier, Yan et al. found that excluding *stop words* (very common words such as *a, the, and*) hurt gender classification performance [Yan and Yan, 2006]. Across feature set sizes ranging from 5,000 to 50,000, the calculated $F$-measure was statistically significantly lower when stop words were removed from the training data [Yan and Yan, 2006]. Their findings suggest that the frequency of these common, topic-independent words may be indicative of author gender. However, these conclusions are directly contradicted by Doyle et al.'s findings that using function word n-grams as features yielded poor gender classification results between 42% and 76% [Doyle and Keselj, 2005].

**Personal & Possessive Pronouns** Previous findings by Koppel et al. suggest that male/female pronouns may be particularly useful distinguishing features [Koppel et al., 2002]. In their gender classification study using 920 English cross-genre and fiction/non-fiction documents from the British National Corpus, they found that in conjunction with certain POS tags, including certain gender pronouns in the feature set yielded considerably improved accuracy of around 80% [Koppel et al., 2002]. They found that men used *he* as often as women, but that female authors used all other pronouns much more frequently [Koppel et al., 2002]. A particularly notable finding was that of the 58 test documents in which the gendered word *herself* appeared more than 5 times, only two were written by male authors [Koppel et al., 2002]. However, a later study of blog data by Herring et al. found that only *he* and *we* were statistically significant female markers, and *you* was preferred by male authors [Koppel et al., 2002, Herring and Paolillo, 2006].

**Other Gender Markers** In their study, Koppel et al. found that women used the prepositions *for* and *with* at significantly higher rates, but men use *of* more [Koppel et al., 2002]. In general, men employ determiners in their writing at much higher rates [Koppel et al., 2002]. In both fiction and non-fiction writing, female authors were found to have higher usage of negations, which again supports previous conclusions.

Bamman et al. specifically aimed to investigate the relationship between gender identity and stylistic differences in language used in social media networks [Bamman et al., 2014]. Using a corpus of 14,000+ Twitter user data to analyze the impact of gender on word-level stylistic choices, they employed a "clustering" method rather than grouping solely based on a binary gender variable. In order to determine the most strongly-gendered words, for each word they computed the ratio of men and women who use that word and noted the ones with the highest dissimilarity or imbalance compared to the overall usage [Bamman et al., 2014]. Although some features reversed previously assumed gender style trends, the majority of clusters displayed a strong gender-language relationship.

Of the words ascribed as "gender markers", the features strongly associated with female authorship include the following: pronouns (including casual / alternative spellings such as *u* and *ur*), emotion terms and gendered-emoticons (*sad, love, :), ;)*), kinship terms (*mom, child, bff*), several abbreviations (*lol, omg, ...*) and "excessive lengthenings" (*coooool*), exclamation and question marks, verbalized sounds (*grr, ugh*), and hesitation / assent / negation terms (*um, yes, noooo*) [Bamman et al., 2014, Schler et al., 2006]. Only a few elements in the majority-female marker categories were associated with male authors: certain kinship terms (*wife, bro, bruh*), assent / negation terms (*yessir, nah, ain't*), and abbreviations (*2* rather than *to*) [Bamman et al., 2014]. In previous research, numbers and quantifiers were largely found to be male-associated. Another predominantly male-associated lexical category was swear and taboo words. Further, in Yan et al's study of male and female blog data, many of the most "gender-discriminant" words were swear or derogatory words [Yan and Yan, 2006].

These social-media-based findings largely support the formal-text-based assertions of Argamon et al. and Schler et al. (see Section 1) in characterizing women as having a relational, personal style in contrast to men's informative style [Argamon et al., 2003, Schler et al., 2006]. However, Bamman et al. found no statistically significant evidence that articles or determiners act as male gender style markers; this contradicts the findings of Argamon et al. and Schler et al. [Argamon et al., 2003].

While the primary focus of this thesis is on n-grams-based analysis, it also employed a more deliberate approach to feature selection and extraction. Preliminary experiments were carried out using the $L$ most common n-grams of varying types and sizes of n. Next, more granular empirical testing was carried out using combinations of n-grams-based features influenced by the most promising features highlighted in the previous work mentioned above (e.g., bi / trigrams that include the pronoun *she*, etc.).

### 3.3.1 Relevant Stylometric Features

This study analyzed various categories of words as features, based on promising findings identified in the previous research detailed above. No previous research has been dedicated to analyzing writing styles or markers for non-binary gendered authors, so this study extrapolated the binary findings from Section 3.3. Table 5 on the following page summarizes these categories and their predominant gender associations from prior work (though in several categories, previous findings are mixed). The specific chosen words for this study can be found in Table A1 of the Appendix (also discussed in Section 5.5).

Table 5: Gender Associations of Gender Markers from Previous Research

| **Gender Markers** | **Previous Research** |
|---|---|
| Personal & Possessive Pronouns | Female |
| Articles | Male |
| Prepositions | Mixed |
| Assent | Female |
| Dissent | Female |
| Negative Emotions | Female |
| Positive Emotions | Female |
| Punctuation | Female |
| Abbreviations | Female |
| Numerals | Male |
| Ordinals | Male |
| Quantifiers | Male |
| Kinship | Female |
| Swear | Male |
| **Non-binary** | **Not previously studied** |

# 4 Data

There are several large, public blog data sets free to download online; the one selected for this study has pre-labled (male and female only) gender indicators. However, the lack of a readily available non-binary gender data set necessitated a separate data acquisition and web scraping process.

## 4.1 Male and Female Blog Data

Two datasets were used in this study. Male and female blogger data were obtained from the Blog Authorship Corpus, which is a collection of blog posts from blogger.com in August 2004 [Koppel et al., 2002]. This corpus is comprised of 681,288 blog posts in English from 19,320 bloggers, representing a collection of over 140 million words. There is an equal number of male and female authors (9,660 each), who have self-identified their gender. All bloggers were categorized into one of three age groups:

· 8,240 "10s" bloggers (ages 13-17)
· 8,086 "20s" bloggers (ages 23-27)
· 2,994 "30s" bloggers (ages 33-47)

## 4.2 Non-Binary Online Article Data

To my knowledge, there is no corpus or readily downloadable dataset for pre-labeled non-binary blog data. Instead, I used Python to scrape non-binary gender text data from over 200 articles in English, which are publicly available on the Beyond the Binary submissions-based U.K. magazine website [Beyond the Binary]. This organization's mission is to promote visibility and representation in the media for people who identify as non-binary, and posts submissions from authors who identify as such.

Articles on the website date from May 2014 through January 2019. Posts are typically 500 to 700 words long, and cover a broad range of topics including sports, relationships, politics, and art. This dataset is described below:

· 189 total authors (all contributors to the site self-identify as non-binary)
· 222 total articles/posts
· 187,791 total words

Throughout the data acquisition process, Python's BeautifulSoup module was used to extract data from webpages [Richardson, 2004]. The methodology is described in Algorithm 1 on the following page.

**Algorithm 1** Non-Binary Data Acquisition Methodology using Python's BeautifulSoup

```
procedure NON-BINARY DATA ACQUISITION
    website ← Beyond the Binary website
    for webpage in website do
        hyperlinks ← extract from webpage (exclude invalid / duplicate links)
    for link in hyperlinks do
        post data ← extract from relevant HTML class / tags
        write post data into separate text file
        filename in format [author name] + .nonbin. + [counter] + .txt
```

This data set, along with a detailed README file, is available to download on GitHub at: https://github.com/sujinkay02/Thesis

There are 222 individual article files, as well as another version of the data where all texts from a given author are combined into a single text file (189 total files). This study used the 189 author-combined files as non-binary data.

# 5 Methodology

This thesis first replicated a state-of-the-art n-gram gender classification method, using the simple dissimilarity calculation and a basic feature set of the $L$ most common n-grams. Baseline gender classification accuracy was recorded across different types of n-grams, and various sizes of the n-grams and $L$. Then, a formal feature selection process was performed in an attempt to improve upon the baseline gender identification performance. All programs were written in Python 3.7.

## 5.1 Types of N-Grams

Three primary experiments were run using the following types of n-grams: characters, words, and Part-of-Speech (POS) tags. The Python 3.0 Natural Language ToolKit (NLTK) package was used to perform word segmentation and POS tagging [Bird et al., 2009]. No other data pre-processing was done (e.g., normalizing text to lowercase, removing punctuation or HTML metadata, etc.). Character analyses did not require any text pre-processing.

## 5.2 Feature Set and Size of N-Grams

Within each of these three experiments, this study evaluated how accuracy is maximized across various sizes of the n-grams, as well as the profile size ($L$). N-gram sizes included values of n between 1 (unigrams) and 5, while $L$ varied among $L = \{10, 500, 1000, 5000$ and $10{,}000\}$.

## 5.3 Profile Generation & Gender Classification

Given a chosen feature set, the author or gender profile was built using NLTK's built-in *ngrams* module [Bird et al., 2009]. This module counted the frequencies of all n-gram occurrences in a given text file, and the program extracted the specified $L$ most common n-grams-based features and their respective frequency counts. The resulting output was thus the n-grams frequency profile for the given anonymous document or gender text file, which was subsequently normalized and used to calculate dissimilarity in the classification step. This process is described in Algorithm 2.

In this study, 85% of the total 19,542 male, female, and non-binary data files were concatenated into their respective *gender profiles* and used as “training” data (16,611 files). The other 15% (2,931 randomly-selected files) were left as individual files and used as “test” data to measure the accuracy of the chosen gender classification approach.

**Algorithm 2** Compute Dissimilarity Measure using Bag-of-Words Approach

```
procedure PROFILE DISSIMILARITY(author profile, gender profile)
    $feature\ set \leftarrow$ (author profile $\cup$ gender profile)
    for n-gram $f$ in $feature\ set$ do
        $C_a(f) \leftarrow$ normalized frequency of $f$ in author profile
        $C_g(f) \leftarrow$ normalized frequency of $f$ in gender profile
        $sum \leftarrow sum + [2 * (C_a(f) - C_g(f)) \ / \ (C_a(f) + C_g(f))]^2$
    Return $sum$
```

## 5.4 Data Summary & Preliminary Results

Tables A2 and A3 in the Appendix show the basic summary data, which includes the 10 most frequent characters, all POS tags, and the 20 most frequent words for each gender's data, along with their normalized frequencies.

For this preliminary analysis, no specific n-grams-based features were selected in particular, and simply the ***L* most frequent n-grams** were used. Figure 3 shows the baseline results of non-binary gender classification, for $L$ = 5,000.

Figure 3: Baseline Non-Binary Results, using $L$=5,000 most frequent Word n-grams

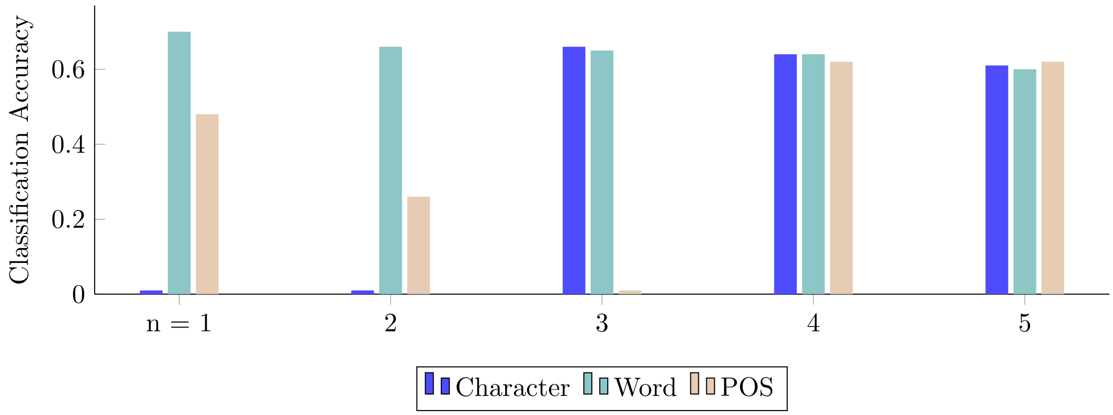


Figure 3 (and Figures A1–A3 in the Appendix) shows the following distinctive findings:

- Character unigrams and bigrams, and POS bigrams and trigrams were not useful features (accuracy near 0 was due to *all* texts being classified as Non-Binary).
- In general, accuracy increased from unigrams to bigrams, but degraded beyond n-grams of size n = 2.

- $L$ = 10 had lower maximum accuracy, and there was little gain in accuracy beyond $L$ = 5,000 (see Appendix for $L$ = 10).
- Otherwise, binary vs. non-binary classification accuracy was similar (see Appendix).

Tables A4–A6 in the Appendix show the baseline results, for both binary and non-binary analyses (with corresponding Figures A1–A3). For each n-grams type, gender and document profiles (of size $L = \{10, 500, 1000, 5000, 10000\}$) were generated, using n-grams ranging from sizes n = 1 to n = 5. Accuracy results from this preliminary analysis are comparable to those from previous research; thus, this study was able to successfully replicate prior work.

Referring back to Tables A2 and A3, it is evident that across the n-grams types, the most frequent features for all three profiles were nearly identical. Thus, it is reasonable to hypothesize that simply using the $L$ most common features will not capture nuanced gender-based stylistic differences and thus will fail to yield maximal classification accuracy.

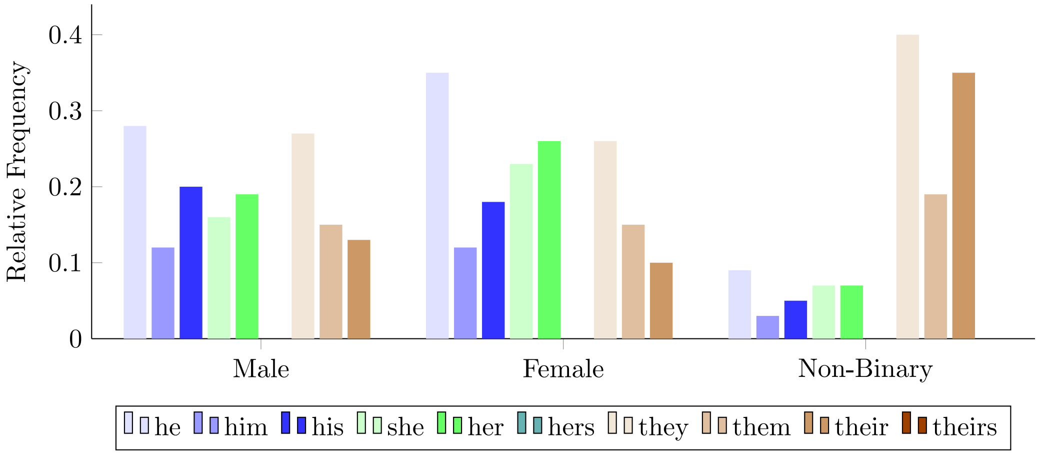
Accordingly, this research focused primarily on word-level n-grams in an attempt to improve this classification method, as token-level analyses are likely to act as the most prominent style markers and gender distinguishers. Further, as shown in Figure 3, it is evident that word-level n-grams yielded the most consistent and highest overall accuracy.

Thus, a formal, targeted n-grams-based feature selection process was undertaken, and specific features were extracted based on high-potential features in previous research, as well as independently-hypothesized style markers. Since binary results were similar, only non-binary classification was performed in the following analysis.

### 5.5 Formal N-Gram Feature Selection & Extraction

Previous research had only studied whether, or how, various gender markers could be attributed to male or female writing styles. This study hypothesized that, especially with the inclusion of the *non-binary* gender category, the use of *gendered personal pronouns* (e.g., *she, her* rather than *I, we*) would remain a key differentiating linguistic element. Figure 4 shows the relative frequency of these pronouns for each gender profile. In addition, this study included an additional style marker category that are words related to gender identity or sexual orientation (based on the hypothesis that these words may be used more frequently by non-binary authors).

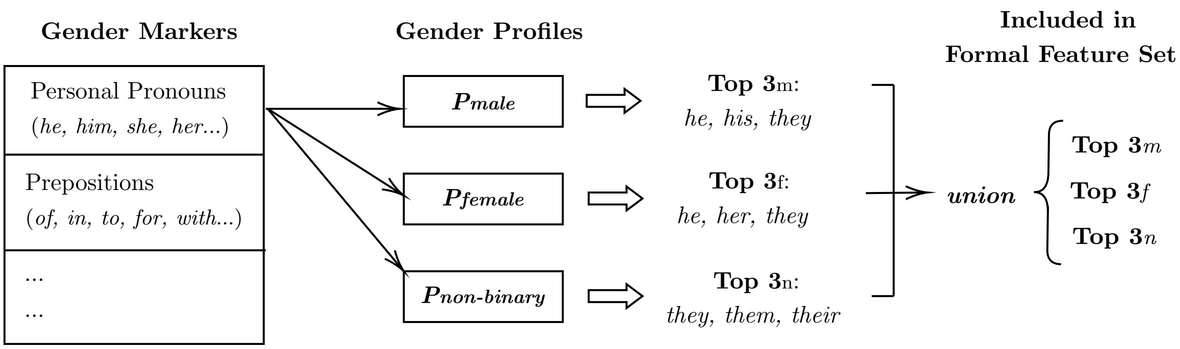
Figure 4: Relative Frequency of Personal Pronouns across Gender Textfiles



As shown in Figure 4, it is evident that personal pronouns could be a useful feature, and thus were included in the formalized token-level feature extraction process. Further, this feature selection process included consideration of stylometric features that have been found to be useful gender markers in previous research (see Section 3.3).

Based on the preliminary findings, formal feature selection was performed uniquely on the word-level. Further, only uni/bi/trigrams were considered, and the profile size excluded $L = 10$. The gender marker categories included in the formal feature selection process are specified in Table A1 of the Appendix (also see Section 3.3). Due to dimensionality and runtime constraints, only the *union* of the top 3 most frequent word n-grams in each gender profile was extracted and included in the feature set. Figure 5 illustrates this process.

Figure 5: Formal Feature Selection Process



Using the process outlined above, the full set of n-grams-based selected word features is listed below (with corresponding relative frequencies shown in Table A1 in the Appendix):

- Gendered personal pronouns (*he, his, her, they, them, their*)
- Other personal pronouns (*you, me, we*)
- Possessive pronouns (*my, your, our*)
- Determiners (*the, a, it*)
- Demonstratives (*that, this, This, these*)
- Coordinating conjunctions (*and, for, but, so, or*)
- Prepositions (*to, of, in*)
- Assent (*sure, yeah, yes, Yes*)
- Dissent (*no, never, No*)
- Negative emotions (*hate, sad, mad, angry*)
- Positive emotions (*love, hope, happy*)
- Punctuation (*..., !, ?*)
- Abbreviations (*u, lol, haha*)
- Numerals (*one, two, 2, One*)
- Ordinals (*first, second, third, First*)
- Quantifiers (*all, some, much, many*)
- Kinship terms (*friends, friend, family, mom*)
- Swear words (*shit, hell, stupid*)
- Gender / Sexuality terms (*man, girl, woman, gender, trans, non-binary*)

This classification procedure used the same dissimilarity measure as before; the only difference was the feature set used for comparing gender profiles with the unclassified text. Algorithm 3 outlines this procedure.

**Algorithm 3** Compute Dissimilarity Measure using Formally-Extracted Features

```
procedure PROFILE DISSIMILARITY(author profile, gender profile)
    $feature\ set$ ← formally-selected features
    for n-gram $f$ in $feature\ set$ do
        $C_a(f)$ ← normalized frequency of $f$ in author profile
        $C_g(f)$ ← normalized frequency of $f$ in gender profile
        $sum \leftarrow sum + [2 * (C_a(f) - C_g(f)) \ / \ (C_a(f) + C_g(f))]^2$
    Return $sum$
```

## 5.6 Overall Accuracy Results

Figure 6 shows the classification accuracy results using the formal feature selection process (see data in Table A7). Using **word unigrams**, peak accuracy of 56% was reached for $L =$ 10,000; peak accuracy was 61% for **bigrams** and 51% for **trigrams** (both for $L = 500$).

Figure 6: Gender Classification Accuracy using Formal Feature Selection

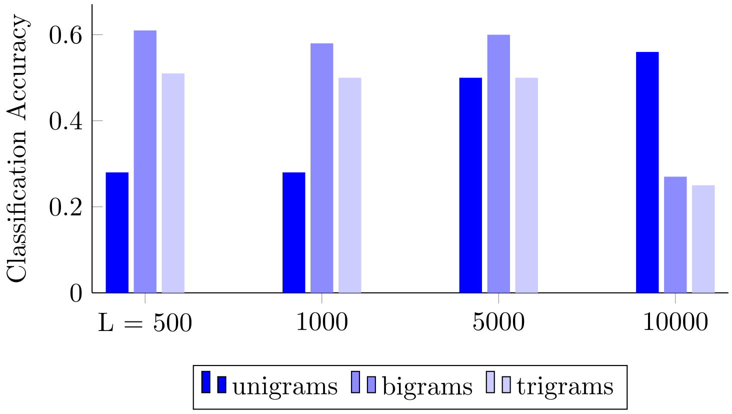
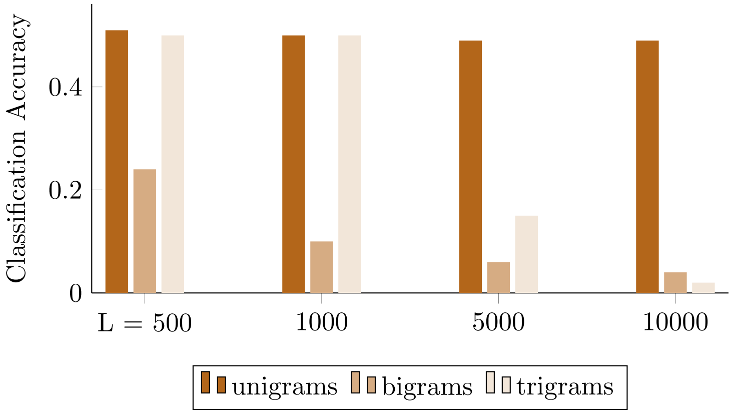


Figure 7 shows results utilizing only the gendered personal pronouns (*he, him, his, she, her, they, them, their*) as features (see data in Table A8). Peak accuracy was 51% for **word unigrams**, 24% for **bigrams** and 50% for **trigrams**, all for $L = 500$.

Figure 7: Gender Classification Accuracy using Gendered Pronouns



Based on these results, it is apparent that overall, this study's formal feature selection process did *not* improve upon the baseline gender classification accuracy obtained in Section 5.4.

Specifically, we recognize the following from the secondary analysis:

- Unigrams accuracy improved substantially as $L$ increased when using the full feature set, but decreased slightly when using only gendered pronouns
- Bigrams analysis yielded the highest accuracy values when using the full feature set, but yielded very low accuracy with gendered pronouns
- Trigrams accuracy decreased substantially as $L$ increased

Therefore, this study found that using the $L$ most common n-grams as the feature set yielded the highest gender classification accuracy (specifically, utilizing word unigrams with profile sizes of above $L$ = 1,000).

### 5.7 Gender-Specific Accuracy Results

Figures A4–A6 (and corresponding Tables A9–A11 and Figures 8–10 in the Appendix) show the *gender-specific* classification results using formal feature selection.

These results display significant differences in classification accuracy across the three gender categories. Most notably, female and non-binary identification reached peak accuracy of **99.1**% and **100**% accuracy, respectively. Peak accuracy for male authors was still high, at **90.9**%.

Figure 8: Male Classification Results, using Formal Feature Selection

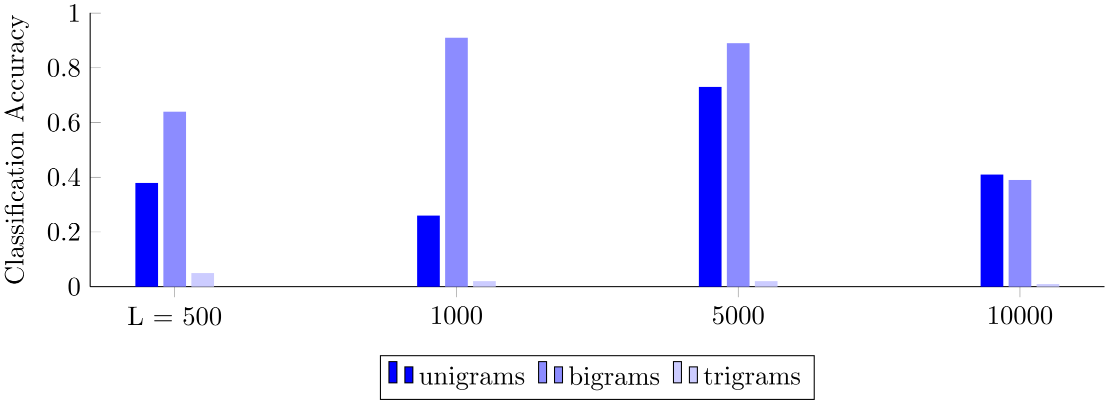


Figure 9: Female Classification Results, using Formal Feature Selection

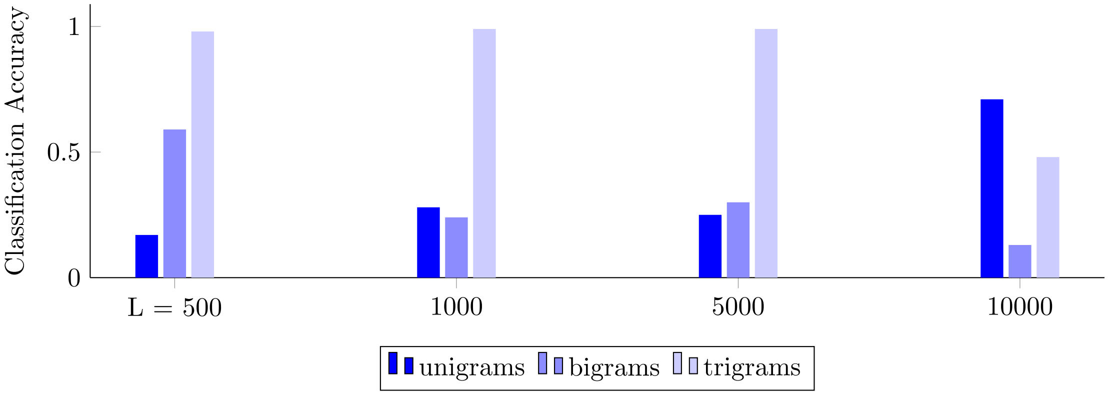
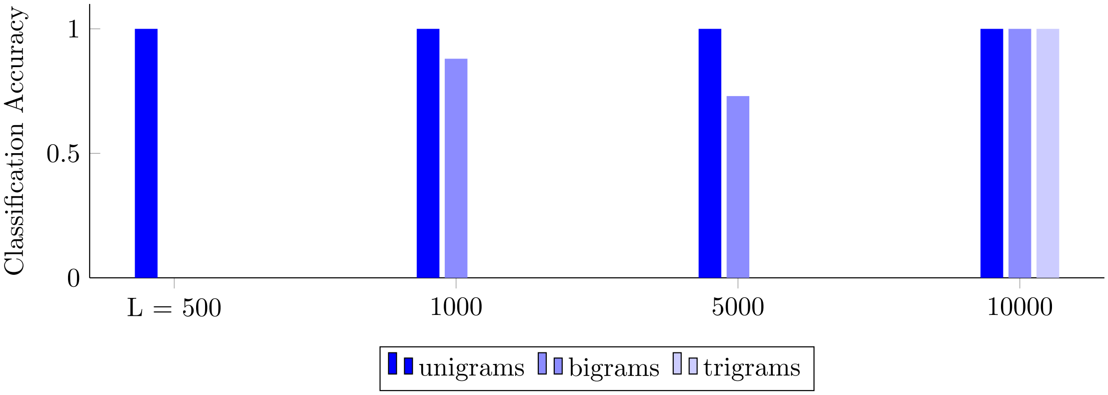


Figure 10: Non-Binary Classification Results, using Formal Feature Selection



The following main conclusions can be drawn from these breakdowns:

- **Male classification:** bigrams yielded highest accuracy for $L$ = 1,000 and 5,000; unigrams and especially trigrams yielded significantly lower accuracy.
- **Female classification:** trigrams yielded significantly accurate results until $L$ = 10,000; unigram results improved as $L$ increased and trigram results had the opposite effect (but both yielded much lower accuracy).
- **Non-binary classification:** unigrams yielded 100% accuracy across all values of $L$; bigrams had high performance except for $L$ = 500, and trigrams had 0% accuracy for all values of $L$ (except for $L$ = 10,000, which interestingly had 100% accuracy).

Table 6 on the following page re-illustrates the gender classification accuracy results from previous research, with the addition of this study's findings. In the preliminary analysis, binary and non-binary classification performances were very similar on the word-level. This is notable, since baseline performance (i.e., guessing) for non-binary classification is 33.33%, compared to 50% for binary gender classification. Results from this work were comparable and peak non-binary gender classification performance fell within the mid-upper range of previous binary gender classification research.

Although the *gender-specific* results showed improved accuracy, the overall results indicated that ultimately, this study did not find evidence to support previous research on distinctive gender markers. This is likely to be explained by the nature of blog data and issues of asymptotic convergence. Much of the previous research analyzed large corpora of texts, which in turn yield a large sample size. The larger the corpus, the greater the probability that the texts capture a high degree of vocabulary richness. With a substantive sample size, it is possible to reasonably deduce estimated empirical probabilities, regarding the validity of certain words as gender markers / features.

However, this study's data set was comprised of blogs and short online media articles (typically less than 1,500 words). Further, due to a lack of previous research, the non-binary data needed to be manually scraped and was not substantial (fewer than 250 articles). Therefore, it is reasonable to conclude that this was too small a sample size to see a statistical distribution of gender-specific vocabulary usage that matches the expected asymptotic convergence. This may explain the drop in gender classification accuracy using features suggested by previous research. This in turn implies that the feature selection process for non-asymptotic data needs to be even more deliberate and cannot simply utilize the gender markers that were found to be reliable in previous research that used large corpora.

Table 6: Accuracy across N-Grams Types and Classification Approaches:
Previous Work (Binary) vs. This Study's Findings (Binary & Non-Binary)

| | **PREVIOUS RESEARCH** | | **THIS STUDY** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | **Preliminary Analysis** | | **Formal Feature Selection (Non-Binary)** | | | |
| | **Unsupervised** | **Supervised** | **Binary** | **Non-Binary** | **Overall** | **Male** | **Female** | **Non-Binary** |
| **Byte** | 50 - 100% | | | | | | | |
| **Char** | 40 - 93% | 41 - 43% | 49 - 67% | 45 - 68%* | | | | |
| **Word** | 51 - 76% | 59 - 64% | 52 - 70% | 45 - 71% | 25 - 61% | 26 - 91%* | 13 - 99% | 73 - 100%* |
| **POS** | 42 - 76% | 50 - 89% | 55 - 65% | 43 - 64%* | | | | |

*A few tests resulted in < 5% accuracy, due to all test files being classified as non-binary. These were excluded from the ranges presented in Table 6 above, in order to present a clearer picture of the results in comparison to previous research findings.

### 5.8 Ethical Considerations

The work presented in this thesis raises several ethical questions and considerations. Automated gender classification on publicly available data can be problematic, especially in sensitive discussion forums where certain demographics can be targeted for harassment. Females and non-binary people are most often subject to or at risk of online aggravation; thus, this study's high gender-specific classification accuracy for female and non-binary authors necessitates consideration of how these tools might be used in the real world.

It is important to question the motivation and premise of this study's work. A substantial part of the reasoning behind the inclusion of the third non-binary gender category is that utilizing a strict gender binary is itself problematic. Prior binary gender profiling research has been criticized for its ethical issues of discrimination or disrespect. More specifically, utilizing gender as a binary variable ignores the fluidity in gender identities and may artificially label data or participants into an overly-simplistic or restrictive category.

Regardless, the tools presented in this thesis could be used for harm, and the 99-100% gender identification accuracy for female and non-binary authors presented in this study could be cause for concern. However, it is important to also consider that *all* technology inevitably yields both useful and criminal intentions and outcomes. Although gender classification tools can be used for malicious targeting and harassment, gender profiling is widely utilized for targeted advertising or to aid in the detection of deceptive or fake social media profiles. Even within these latter two applications, automated gender identification can be used and manipulated for good or bad. Therefore, it is both a personal and social responsibility to understand the risks associated with these technologies, and to utilize them for the purpose of advancing knowledge in the discipline or to create a beneficial, lasting impact on others.

## 6 Future Work

For future research that seeks to expand on this thesis, there are several promising paths to explore. First, gathering a more robust non-binary dataset would be beneficial (in regards to both quantity and a variety of sources), especially as the non-binary data was substantially smaller than the male / female blog data. Second, a more sophisticated classifier could be utilized in lieu of the simple dissimilarity measure (e.g. Bayes classifier, Support Vector Machines (SVMs), etc.). Lastly, more time and exploration could be undertaken in the manual feature selection process. More feature set combinations could be tested, as well as investigating other types or mixes of features, such as combining Part-of-Speech and word n-grams.

## 7 Summary

In this study, three main contributions were made. First, a non-binary dataset was gathered, and this collection has been made available for future educational research. Second, state-of-the-art authorship attribution techniques were utilized for both the binary and non-binary gender classification tasks, using word, character, and POS n-grams. Accuracy results were presented using the $L$ most common n-grams (peak of 71%). Third, non-binary gender classification was performed using a formal word-based feature selection and extraction process, and results for both overall (peak of 61%) and gender-specific (peak of 100%) performance were presented. Both results were comparable to findings from previous research, and fell within the mid-upper range of both unsupervised and supervised classification approaches, as well as byte, character, word, and POS n-grams. For word-level analyses in particular, this study's results were similar to the highest accuracy results from previous work.

All programs and the collection of non-binary data are publicly available on GitHub, along with a detailed README file: https://github.com/sujinkay02/Thesis

# Appendix

Table A1: Selected Word Features

| | | Relative Frequencies (in %) | | |
|---|---|---|---|---|
| **Gender Markers** | **Chosen Features** | **Male** | **Female** | **Non-Binary** |
| Gendered Personal Pronouns | *he* | 0.28 | 0.35 | 0.09 |
| | *his* | 0.20 | 0.18 | 0.05 |
| | *her* | 0.19 | 0.26 | 0.07 |
| | *they* | 0.26 | 0.26 | 0.41 |
| | *them* | 0.15 | 0.15 | 0.19 |
| | *their* | 0.13 | 0.10 | 0.33 |
| Other Personal Pronouns | *you* | 0.70 | 0.75 | 0.72 |
| | *me* | 0.46 | 0.64 | 0.48 |
| | *we* | 0.35 | 0.41 | 0.36 |
| Possessive Pronouns | *my* | 0.69 | 0.92 | 0.79 |
| | *your* | 0.17 | 0.18 | 0.27 |
| | *our* | 0.12 | 0.12 | 0.22 |
| Determiners | *the* | 3.23 | 2.68 | 2.91 |
| | *a* | 1.76 | 1.62 | 2.06 |
| | *it* | 0.96 | 1.02 | 0.83 |
| Demonstratives | *that* | 1.06 | 1.07 | 1.39 |
| | *this* | 0.46 | 0.43 | 0.41 |
| | *This* | 0.09 | 0.07 | 0.10 |
| | *these* | 0.07 | 0.05 | 0.11 |
| Coordinating Conjunctions | *and* | 2.00 | 2.11 | 2.66 |
| | *for* | 0.72 | 0 .70 | 0.72 |
| | *but* | 0.42 | 0.46 | 0.38 |
| | *so* | 0.35 | 0.48 | 0.23 |
| | *or* | 0.27 | 0.25 | 0.48 |

| | | **Relative Frequencies (in %)** | | |
|---|---|---|---|---|
| **Gender Markers** | **Chosen Features** | **Male** | **Female** | **Non-Binary** |
| Prepositions | *to* | 2.36 | 2.40 | 2.65 |
| | *of* | 1.60 | 1.31 | 2.15 |
| | *in* | 1.06 | 0.96 | 1.24 |
| Assent | *sure* | 0.05 | 0.05 | 0.04 |
| | *yeah* | 0.02 | 0.03 | 0 |
| | *yes* | 0.02 | 0.03 | 0.01 |
| | *Yes* | 0.02 | 0.02 | 0.01 |
| Dissent | *no* | 0.15 | 0.15 | 0.14 |
| | *never* | 0.07 | 0.09 | 0.08 |
| | *No* | 0.04 | 0.04 | 0.03 |
| Negative Emotions | *hate* | 0.02 | 0.03 | 0.02 |
| | *sad* | 0.01 | 0.02 | 0.01 |
| | *mad* | 0.01 | 0.01 | 0 |
| | *angry* | 0.01 | 0.01 | 0.01 |
| Positive Emotions | *love* | 0.08 | 0.13 | 0.06 |
| | *hope* | 0.04 | 0.04 | 0.03 |
| | *happy* | 0.03 | 0.04 | 0.02 |
| Punctuation | ... | 0.75 | 0.95 | 0 |
| | ! | 0.67 | 1.07 | 0.14 |
| | ? | 0.48 | 0.54 | 0.33 |
| Abbreviations | *u* | 0.03 | 0.05 | 0 |
| | *lol* | 0.02 | 0.03 | 0 |
| | *haha* | 0.01 | 0.02 | 0 |
| Numerals | *one* | 0.25 | 0.25 | 0.19 |
| | *two* | 0.07 | 0.07 | 0.06 |
| | *2* | 0.07 | 0.07 | 0.02 |
| | *One* | .03 | 0.02 | 0.03 |

|  |  | **Relative Frequencies (in %)** |  |  |
|---|---|---|---|---|
| **Gender Markers** | **Chosen Features** | **Male** | **Female** | **Non-Binary** |
| Ordinals | *first* | 0.09 | 0.08 | 0.08 |
|  | *second* | 0.02 | 0.02 | 0.01 |
|  | *third* | 0.01 | 0.01 | 0.01 |
|  | *First* | 0.01 | 0.01 | 0 |
| Quantifiers | *all* | 0.34 | 0.36 | 0.24 |
|  | *some* | 0.21 | 0.18 | 0.16 |
|  | *much* | 0.12 | 0.14 | 0.10 |
|  | *many* | 0.06 | 0.05 | 0.14 |
| Kinship | *friends* | 0.05 | 0.06 | 0.05 |
|  | *friend* | 0.04 | 0.05 | 0.03 |
|  | *family* | 0.03 | 0.03 | 0.03 |
|  | *mom* | 0.02 | 0.04 | 0 |
| Swear | *shit* | 0.02 | 0.04 | 0.01 |
|  | *hell* | 0.02 | 0.03 | 0.01 |
|  | *stupid* | 0.02 | 0.02 | 0 |
|  |  | 0.02 | 0.02 | 0.05 |
| Non-binary | *man* | 0.05 | 0.04 | 0.05 |
|  | *girl* | 0.03 | 0.03 | 0.02 |
|  | *woman* | 0.01 | 0.02 | 0.05 |
|  | *gender* | 0 | 0 | 0.43 |
|  | *trans* | 0 | 0 | 0.35 |
|  | *non-binary* | 0 | 0 | 0.35 |

Table A2: Data Summary of Male, Female, and Non-Binary Texts

| | **Male** | | **Female** | | **Non-Binary** | |
|---|---|---|---|---|---|---|
| | Item | Freq. | Item | Freq. | Item | Freq. |
| **Character** | ’ ’ | 20.4% | ’ ’ | 20.8% | ’ ’ | 17.1% |
| | e | 8.5% | e | 8.3 | e | 9.6% |
| | t | 6.6% | t | 6.5% | t | 7.0% |
| | o | 5.8% | a | 5.8% | a | 6.2% |
| | a | 5.8% | o | 5.8% | o | 6.1% |
| | n | 5.0% | n | 5.0% | n | 6.0% |
| | i | 4.8% | i | 4.7% | i | 5.5% |
| | s | 4.5% | s | 4.3% | s | 4.9% |
| | r | 3.9% | h | 3.8% | r | 4.7% |
| | h | 3.7% | r | 3.7% | h | 3.5% |
| **POS** | NOUN | 22.3% | NOUN | 20.8% | NOUN | 23.0% |
| | VERB | 17.6% | VERB | 18.5% | VERB | 18.1% |
| | . | 13.2% | . | 13.3% | ADP | 10.7% |
| | ADP | 9.0% | PRON | 10.2% | . | 9.4% |
| | PRON | 8.9% | ADP | 8.5% | PRON | 9.2% |
| | DET | 8.2% | DET | 7.2% | ADJ | 8.3% |
| | ADJ | 6.4% | ADV | 6.8% | DET | 8.1% |
| | ADV | 6.3% | ADJ | 6.3% | ADV | 5.7% |
| | PRT | 3.3% | CONJ | 3.6% | CONJ | 3.8% |
| | CONJ | 3.3% | PRT | 3.4% | PRT | 3.1% |
| | NUM | 1.3% | NUM | 1.2% | NUM | 0.7% |
| | X | 0.2% | X | 0.2% | X | 0.1% |

Table A3: Data Summary of Male, Female, and Non-Binary Texts

| | **Male** | | **Female** | | **Non-Binary** | |
|---|---|---|---|---|---|---|
| | Item | Freq. | Item | Freq. | Item | Freq. |
| **Word** | . | 4.4% | . | 4.4% | , | 4.3% |
| | , | 3.9% | , | 3.5% | . | 3.6% |
| | the | 3.2% | I | 2.9% | the | 2.9% |
| | I | 2.4% | the | 2.7% | to | 2.7% |
| | to | 2.4% | to | 2.4% | and | 2.7% |
| | and | 2.0% | and | 2.1% | I | 2.4% |
| | a | 1.8% | a | 1.6% | of | 2.2% |
| | of | 1.6% | of | 1.3% | ’ | 2.2% |
| | that | 1.1% | ! | 1.1% | a | 2.1% |
| | in | 1.1% | that | 1.1% | that | 1.3% |
| | it | 1.0% | it | 1.0% | in | 1.2% |
| | is | 0.9% | in | 1.0% | is | 1.0% |
| | ... | 0.7% | ... | 0.9% | it | 0.8% |
| | for | 0.7% | my | 0.9% | my | 0.8% |
| | you | 0.7% | i | 0.8% | as | 0.7% |
| | my | 0.7% | is | 0.8% | for | 0.7% |
| | was | 0.7% | was | 0.8% | you | 0.7% |
| | ! | 0.7% | you | 0.8% | people | 0.7% |
| | ’s | 0.7% | for | 0.7% | with | 0.7% |
| | on | 0.6% | ’s | 0.7% | s | 0.6% |

In Tables A4–A6, accuracy for binary classification is listed first, with non-binary classification accuracy shown in parentheses. Peak accuracy is highlighted in bold.

Gender Classification Accuracy (%) – Binary and Non-Binary Profiling Results

Table A4: Results Using Character N-Grams

| **Profile** | **N-Gram Size** | | | | |
|---|---|---|---|---|---|
| **Size** | 1 | 2 | 3 | 4 | 5 |
| 10 | 57 (*54*) | 59 (*53*) | 59 (*45*) | 62 (*50*) | 57 (*53*) |
| 500 | 62 (*1*) | 64 (*65*) | 63 (*62*) | 63 (*62*) | 64 (*63*) |
| 1000 | 62 (*1*) | 66 (*67*) | 63 (*63*) | 64 (*63*) | 64 (*62*) |
| 5000 | 62 (*1*) | **67** (*1*) | 65 (*66*) | 65 (*64*) | 64 (*61*) |
| 10000 | 49 (*1*) | 66 (*1*) | 66 (*67*) | **67** (***68***) | 64 (*65*) |

Table A5: Results Using Word N-Grams

| **Profile** | **N-Gram Size** | | | | |
|---|---|---|---|---|---|
| **Size** | 1 | 2 | 3 | 4 | 5 |
| 10 | 61 (*53*) | 57 (*45*) | 54 (*52*) | 54 (*52*) | 52 (*51*) |
| 500 | 65 (*65*) | 62 (*61*) | 61 (*60*) | 58 (*56*) | 55 (*54*) |
| 1000 | 66 (*67*) | 64 (*63*) | 63 (*62*) | 60 (*59*) | 56 (*54*) |
| 5000 | 69 (*70*) | 65 (*66*) | 65 (*65*) | 63 (*64*) | 59 (*60*) |
| 10000 | **70** (***71***) | 66 (*67*) | 65 (*66*) | 65 (*66*) | 63 (*64*) |

Table A6: Results Using POS N-Grams

| **Profile** | **N-Gram Size** | | | | |
|---|---|---|---|---|---|
| **Size** | 1 | 2 | 3 | 4 | 5 |
| 10 | 63 (*56*) | 59 (*48*) | 60 (*46*) | 57 (*43*) | 55 (*43*) |
| 500 | 64 (*48*) | 63 (*26*) | 63 (*59*) | 62 (*56*) | 62 (*57*) |
| 1000 | 64 (*48*) | 63 (*26*) | 62 (*62*) | 63 (*59*) | 61 (*56*) |
| 5000 | 64 (*48*) | 63 (*26*) | 64 (*1*) | 62 (*62*) | **65** (*62*) |
| 10000 | 64 (*48*) | 63 (*26*) | 64 (*1*) | 62 (*2*) | **65** (***64***) |

Figure A1: Baseline Non-Binary Results, using $L$ most frequent Character n-grams

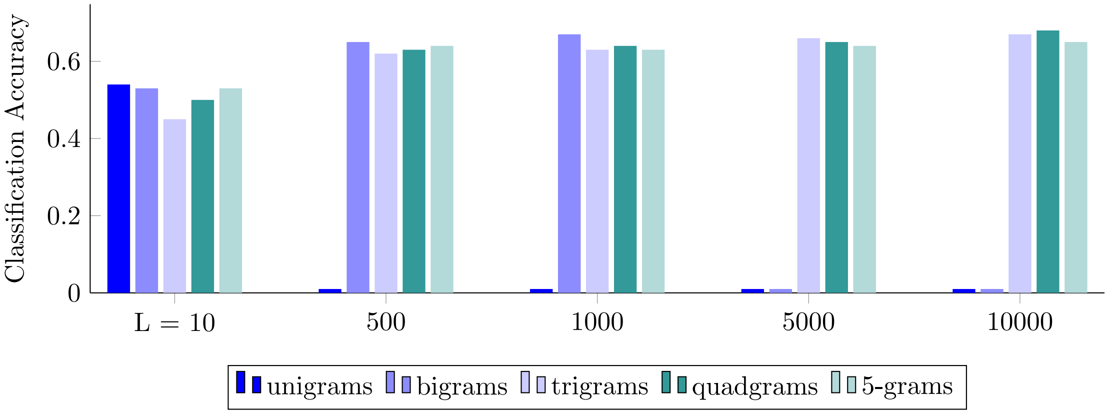


Figure A2: Baseline Non-Binary Results, using $L$ most frequent Word n-grams

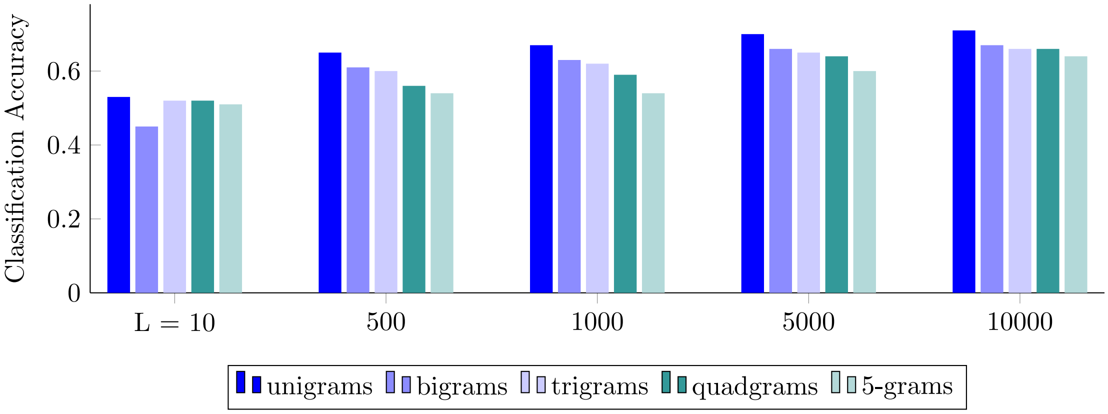
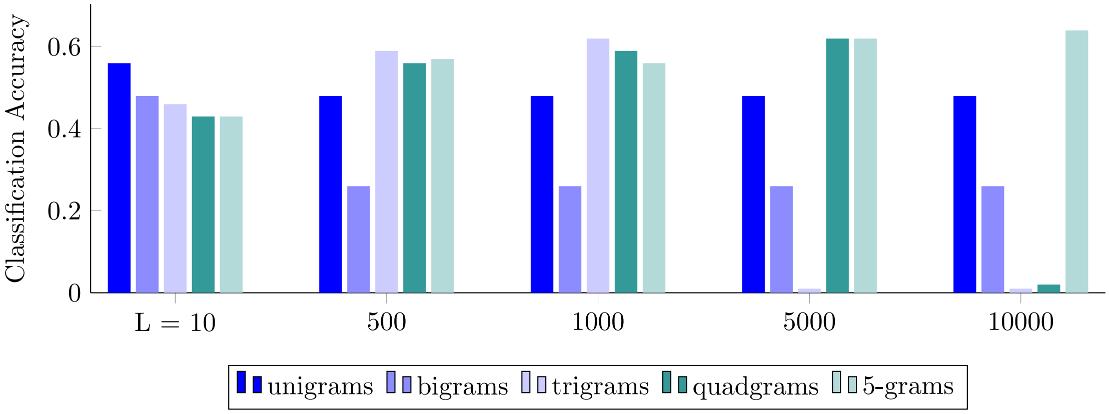


Figure A3: Baseline NonBinary Results, using $L$ most frequent POS n-grams



In Tables A7–A8, highest accuracy values are highlighted in bold.

Gender Classification Accuracy (%) using Formal Feature Selection

Table A7: Results Using Formally-Selected Word Features (Top 3)

| **Profile** | **N-Gram Size** | | |
|---|---|---|---|
| **Size** | 1 | 2 | 3 |
| 500 | 28 | **61** | 51 |
| 1000 | 28 | **58** | 50 |
| 5000 | 50 | **60** | 50 |
| 10000 | **56** | 27 | 25 |

Table A8: Results Using Gendered Pronouns

| **Profile** | **N-Gram Size** | | |
|---|---|---|---|
| **Size** | 1 | 2 | 3 |
| 500 | **51** | 24 | **50** |
| 1000 | **50** | 10 | **50** |
| 5000 | **49** | 6 | 15 |
| 10000 | **49** | 4 | 2 |

Table A9: Male Classification Accuracy

| | **Unigrams** | **Bigrams** | **Trigrams** |
|---|---|---|---|
| **L = 500** | 38.2 | 64.3 | 4.8 |
| **L = 1,000** | 26.2 | **90.9** | 1.9 |
| **L = 5,000** | 73.2 | **89.0** | 1.7 |
| **L = 10,000** | 40.6 | 38.6 | 1.0 |

Table A10: Female Classification Accuracy

| | **Unigrams** | **Bigrams** | **Trigrams** |
|---|---|---|---|
| **L = 500** | 17.0 | 59.4 | **98.4** |
| **L = 1,000** | 27.8 | 23.6 | **99.0** |
| **L = 5,000** | 25.0 | 30.0 | **99.1** |
| **L = 10,000** | 70.9 | 12.8 | 47.7 |

Table A11: Non-Binary Classification Accuracy

| | **Unigrams** | **Bigrams** | **Trigrams** |
|---|---|---|---|
| **L = 500** | **100** | 0 | 0 |
| **L = 1,000** | **100** | 87.9 | 0 |
| **L = 5,000** | **100** | 72.7 | 0 |
| **L = 10,000** | **100** | **100** | **100** |

Figure A4: Unigram Classification Results by Gender

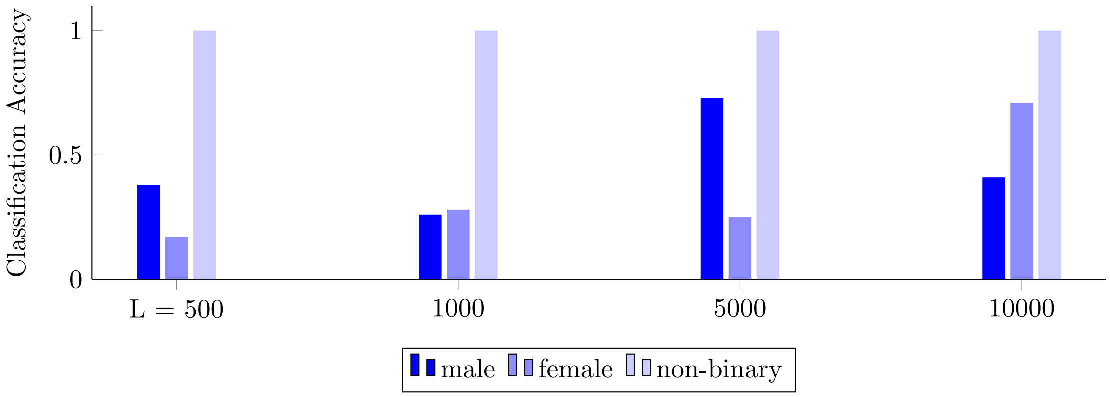


Figure A5: Bigram Classification Results by Gender

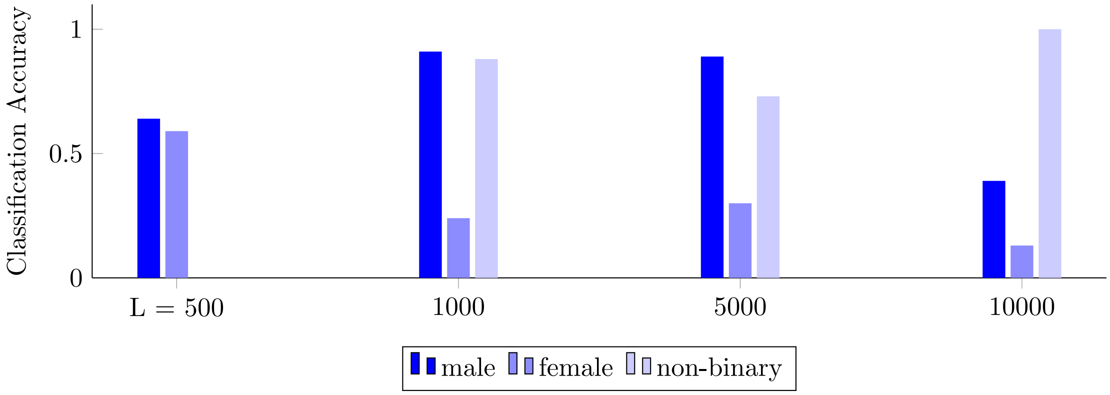
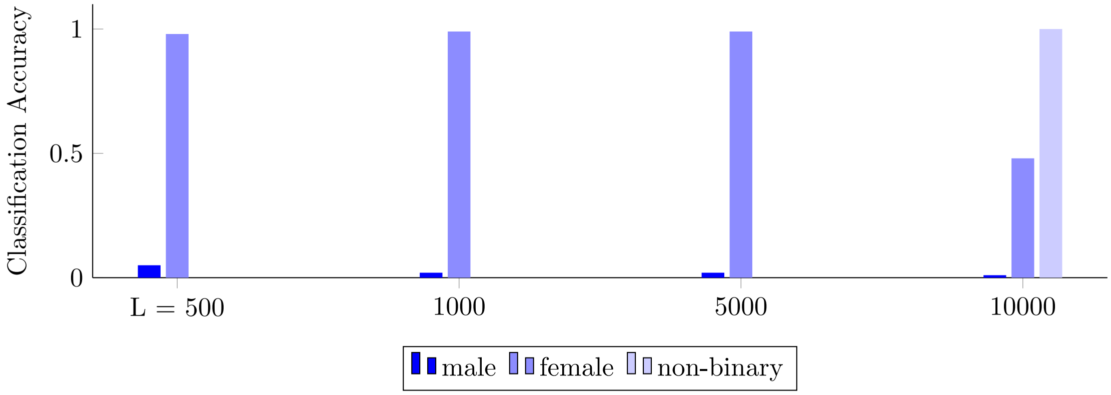


Figure A6: Trigram Classification Results by Gender



# References

[Argamon et al., 2003] Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). Gender, genre and writing style in formal written texts. *TEXT– Interdisciplinary Journal for the Study of Discourse*, 23:321–346.

[Bamman et al., 2014] Bamman, D., Eisenstein, J., and Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

[Beyond the Binary, ] Beyond the Binary. http://beyondthebinary.co.uk/.

[Bird et al., 2009] Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, 1 edition.

[Crestodina, 2018] Crestodina, A. (2018). Blogging statistics and trends: The 2018 survey of 1000+ bloggers. https://www.orbitmedia.com/blog/blogging-statistics/.

[Doyle and Keselj, 2005] Doyle, J. and Keselj, V. (2005). Automatic categorization of author gender via n-gram analysis.

[Hassan and Chaurasia, 2012] Hassan, F. I. H. and Chaurasia, M. A. (2012). N-gram based text author verification. In *2012 International Conference on Innovation and Information Management*, volume 36.

[Herring and Paolillo, 2006] Herring, S. C. and Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10:439–459.

[Joula, 2006] Joula, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieva*, 1(3):233–334.

[Jurafsky and Martin, 2008] Jurafsky, D. and Martin, J. H. (2008). *Speech & Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Prentice Hall, 2 edition.

[Keselj et al., 2003] Keselj, V., Peng, F., Cercone, N., and Thomas, C. (2003). N-gram-based author profiles for authorship attribution. *Pacific Association for Computational Linguistics*, pages 255–264.

[Koppel et al., 2002] Koppel, M., Argamon, S., and Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

[Koppel et al., 2013] Koppel, M., Schler, J., and Argamon, S. (2013). Authorship attribution: What's easy and what's hard? *Journal of Law and Policy*, 21(2):317–331.

[Larson, 2017] Larson, B. (2017). Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11. Association for Computational Linguistics.

[Leahy, 2009] Leahy, P. (2009). N-gram-based text attribution.

[Mikros, 2012] Mikros, G. (2012). Authorship attribution and gender identification in greek blogs. In *8th International Conference on Quantitative Linguistics (QUALICO)*.

[Mukherjee and Liu, 2012] Mukherjee, A. and Liu, B. (2012). Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 207–217.

[Peersman et al., 2011] Peersman, C., Daelemans, W., and Vaerenbergh, L. V. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents*, pages 37–44.

[Peng et al., 2003] Peng, F., Schuurmans, D., Keselj, V., and Wang, S. (2003). Language independent authorship attribution using character level language models. In *10th Conference of the European Chapter of the Association for Computational Linguistics, EACL*.

[Richardson, 2004] Richardson, L. (2004). Beautiful soup documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[Schler et al., 2006] Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. (2006). Effects of age and gender on blogging. In *AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pages 191–197.

[Sidorov et al., 2014] Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., and Chanona-Hernandez, L. (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3):853–860.

[Stamatatos, 2009] Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, pages 538–556.

[Universal Dependencies, 2014] Universal Dependencies (2014). Universal POS Tags. https://universaldependencies.org/u/pos/.

[Yan and Yan, 2006] Yan, X. and Yan, L. (2006). Gender classification of weblog authors. In *AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pages 228–230.

[Zhang and Zhang, 2010] Zhang, C. and Zhang, P. (2010). *Predicting Gender from Blog Posts*. University of Massachusetts Amherst.