universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Enhancing Authorship Attribution: Analysing the Impact of Emotional Language

verfasst von | submitted by

Maximilian Berens B.A.

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 647
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Digital Humanities
Betreut von | Supervisor:	Ass.-Prof. Mag. Mag. Dr. Andreas Baumann

Abstract

Authorship attribution is a long-standing method of determining the author of a text of unknown origin. It can be accomplished using computational methods and is classified as a text categorisation problem. The accuracy of its results generally measures the success of an authorship attribution problem.

This master's thesis presents methods of verifying authorship and aims to give an overview of their theoretical foundations. Further, it explicitly analyses the possibilities considering emotion in text, how these parameters can be measured, and their potential implications regarding the overall work process of predicting possible text authors. The master thesis employs methodologies to enhance the results of authorship attribution accuracy while considering emotional language utilising machine learning. To accomplish this, a diachronic text database is used to determine the authorship of texts. This database, compiled by scraping blogs from over 100 authors, contains over 80 million word tokens from active blogs written over the last 20 years. This extensive text corpus provides a vast dataset of texts and content linked explicitly to individual authors, paving the way for extracting appropriate metrics and potential results for correctly classifying authorship.

Authorship attribution is a complex task that comes with its own set of challenges. Ideally, it means having a small set of candidate authors with nearly unlimited training data on the content that these authors produced. However, since that is hardly the case, authorship attribution has its own set of problems and best practices in terms of methodology. As for a few examples to illustrate the method, the data can be analysed using stylistic and linguistic features, such as usage patterns, different usage of words, their frequencies, and sentence structure and punctuation. These features provide a baseline for verifying authorship and achieving a robust accuracy score.

While authorship attribution has traditionally focused on linguistic and stylistic features, this thesis introduces a novel approach by supplementing these methods with the emotional content in texts and its potential impact on the accuracy of predicting authorship. The overall goal of employing this new analysis layer is to analyse how emotions affect authorship attribution and implement this innovative approach to introduce new avenues for research in the field potentially.

Kurzfassung

Authorship attribution ist eine seit langem angewandte Methode zur Bestimmung des Autors eines Textes unbekannter Herkunft. Um dieses Problem der Textkategorisierung zu lösen, können Forschungsmethoden implementiert werden, die eine Textdatenbank und den Schreibstil der dort vorhandenen Autoren analysieren. Üblicherweise werden die Ergebnisse anhand des Erfolgs der korrekten Attribution gemessen.

Diese Masterarbeit stellt Methoden zur Autorenverifizierung vor und gibt einen Überblick über die theoretischen Grundlagen. Darüber hinaus wird insbesondere der Einfluss von im Text vorhandenen Emotionen untersucht und ob diese als zusätzliche Parameter gemessen werden können, und, im Idealfall, zu einer Verbesserung der Genauigkeit beitragen können, beziehungsweise, welche Auswirkungen diese emotionalen Marker auf den gesamten Arbeitsprozess der Vorhersage bestimmter Autoren haben können. Somit stellt diese Masterarbeit einen beispielhaften Arbeitsprozess vor, um die Ergebnisse von *authorship attribution* unter Rücksichtnahme auf emotionale Textattribute potenziell zu verbessern mit Hilfe von maschinellem Lernen. Zu diesem Zweck wird eine Textdatenbank verwendet, welche gesammelte Internetblogs der letzten 20 Jahre enthält. Insgesamt beinhaltet diese, durch *webscraping* erstellte, Datenbank über 100 Autoren mit über 80 Millionen Wörtern. Dieser umfangreiche Textkorpus bietet eine adäquate Größe von Texten, die explizit mit ihren jeweiligen Autoren verknüpft sind, und stellt damit eine gute Grundlage dar, um für die Forschungsanforderungen geeignet zu sein.

Authorship attribution ist eine vielschichtige Aufgabe, die verschiedene Herausforderungen beinhaltet. Im Idealfall gibt es eine kleine Gruppe von in Frage kommenden Autoren, die jeweils mit einer nahezu unbegrenzten Anzahl an Texten als Trainingsdaten zur Verfügung stehen. Da dies jedoch selten der Fall ist, haben sich über die Jahre eine Anzahl von Methoden zur Analyse durchgesetzt. Beispielsweise können die Texte auf stilistische und linguistische Merkmale untersucht werden, wie zum Beispiel unterschiedliche Wortverwendungen, die Häufigkeit bestimmter Worte oder die Satzstruktur, Fehler und die Zeichensetzung. Durch diese Untersuchungen kann Autorenschaft potenziell erkannt werden. Während die Autorenschaftsattribution sich traditionell auf diese Felder beschränkt hat, wird in dieser Masterarbeit ein neuer Ansatz eingeführt, der diese Methoden durch die Analyse emotionaler Sprache in den Texten ergänzt, um potenziell die Ergebnisse zu verbessern. Das übergeordnete Ziel ist es, diese neuen Methoden zu analysieren und die Auswirkungen darzulegen, um neue Wege für die Forschung auf diesem Gebiet zu eröffnen.

Table of Contents

1. Introduction

Authorship attribution means determining the unknown author of a given text. Historically, there have been many ways in which this has been accomplished. Nowadays, it can be done by utilising computational and statistical methods, while the field itself has a long history, starting in the 19th century (Juola, 2008). More and more methods were added to the repertoire over the decades. Analysing word frequencies or utilising linguistic stylometry were popular methods to determine authorship, and many researchers are still working on these topics today. Since then, and especially over the last decades, the options and capabilities of computational methods have drastically increased (Stamatatos, 2009). Nowadays, they offer a wide range of possible tools, enhancing traditional methodologies while offering unique new research methods. These range from linguistic text features analysis methods to natural language processing, machine learning techniques, and emerging artificial intelligence approaches for text analysis (Stamatatos, 2009).

Historically, the field often encompassed analysing disputed literary texts to determine their authorship. While this is still a popular avenue of research, the field has expanded to other directions, such as forensics or attributing software code (Chaski, 2001; Juola, 2008; Kalgutkar et al., 2020). Each avenue features unique challenges, although nowadays, most of these experiments are software-driven, utilising digital tools and programming to achieve their goal.

One potential research aspect of authorship attribution and its possibilities to improve is the field of emotional attribution and the different methods to analyse emotional language in texts. Successful implementations of these research methods could positively impact the classification results and emphasise how authors incorporate emotion into their texts. Considering existing traditional methods in the field, it seems natural that this might be one of the subsequent research steps to add new features to enhance the accuracy and to explore how different authors utilise emotional language, subsequently asking the question of individual patterns emerging from author to author, which can be differentiated. This means examining the emotional content of the text and isolating an author’s unique writing style. To make the results of this process comparable to traditional methods, it is essential to incorporate new variables that include the emotional tone of an author or a given text, which could then be compared to complement existing variables used for authorship attribution. This concept could add a new layer of depth to the analysis, and researchers in the field might develop a different viewpoint to analyse nuances in an author's style on a different level.

This thesis topic was mainly inspired by existing research, e.g. focusing on analysing speeches and written text by popular figures of public life and politicians (Martins et al., 2018). The research focused on utilising emotional models based on multiple emotional research theories in the field and analysing the texts based on these theories to increase the accuracy scores of authorship attribution. Other research regarding emotional tone for authorship attribution exists, but is relatively sparse. However, there are multiple promising directions to explore within this scope, which will be described in this thesis, such as using sentiment analysis or other emotional theories to pinpoint an author's style based on emotional metrics or emotion intensity.

This thesis aims to conduct a new experiment utilising these tools and focusing on written text over the last 20 years collected in a combined text corpus. Specifically, the easier distribution of written text via the internet (Abbasi et al., 2022) contributes to this by making it easier to collect text samples from many different authors, while not only limiting research on historical text or literary books from a few popular authors. Specifically, web platforms provide an avenue to quickly publish texts for anyone on social media, blogs or through more straightforward digital publication methods.

However, analysing these texts is a task with some challenges to consider. Writing styles evolve, and the tone of online texts and communication differs from traditional writing in books (Azarbonyad et al., 2015). Also, emotional tone is not always easy to analyse. For example, analysing sarcastic text with the help of a program might be a challenge since it can be difficult to detect sarcasm in texts correctly. Other stylistic elements in the digital world also contribute to these difficulties (e.g. the use of emoticons, which convey emotions without necessarily having to use written characters). Still, analysing emotions can be beneficial since it can add nuances to texts, which, otherwise, in traditional authorship attribution, might be overlooked since it focuses more on text structure or lexical features. Further, linguistic features can signify specific emotional attitudes (Rude et al., 2004). So, combining all these metrics can potentially improve the accuracy further.

One of those emotional theories is the dimensional emotional theory, which focuses on emotional dimensions and the intensity of emotions felt when expressing them (Scherer, 2001). Usually, these scores are defined as *valence*, *arousal*, and *dominance* scores. These scores can be calculated on a word-by-word basis, and the words are then defined by their scores and saved in word dictionaries from which the metrics for the analysis can be derived. For the longest time, there have been a few collected dictionaries from which researchers pulled their scores

(Warriner et al., 2013). In 2013, Warriner et al. computed these scores for almost 14.000 words in English and thus extended existing dictionaries significantly. This was also possible because, through newly emerging digital tools, collecting these word scores has become easier, for example, through crowdsourcing and mass-surveying people about the emotions they felt when confronted with a specific word. The obvious advantage is that it becomes easier for researchers to compare a larger word corpus more accurately based on more emotional word scores, which allows more words to be better defined by their emotional impact.

In this sense, the problem worth investigating revolves around emotional tone, which offers additional stylistic features that may contribute to authorship analysis. Following that thought, the main research question is: Can the implementation of emotional language features improve the accuracy of authorship attribution methods?

This thesis focuses on this question while considering additional questions surrounding this thematic complex. Of course, the question is whether accurate authorship attribution can be conducted on a given dataset and whether incorporating emotional tone further improves the results. Coming from that assumption means that the texts themselves must contain underlying evidence of authorship that is reliable enough to form an educated guess with a certain amount of potential success. Further, the texts must give enough information to convey an author’s emotional style so that it can be compared to other authors with notable differences.

Other questions are worth considering, such as which emotional feature contributes most to the classification or how these emotion scores can be effectively computed and integrated into an efficient programming pipeline for authorship attribution. This thesis aims to answer these questions for a given dataset from which the authorship of different authors will be determined.

This will be done with computational methods while first implementing a purely traditional approach building on stylometry and lexical feature analysis of the texts before classifying them with machine learning models and, afterwards, incorporating emotional scores with a second pipeline which computes these scores based on the dictionary developed by Warriner et al. (2013) and integrating them into the traditional authorship attribution pipeline to compare the classification results. The dataset used for this master's thesis consists of online blogs written by over a hundred authors over the last 20 years, thus providing a significant dataset size on which to base this study. By accomplishing these steps, the thesis aims to

contribute to the niche of authorship attribution research, which focuses on emotional analysis in an author's style.

Structurally, the first chapter of the thesis gives an extended background on authorship attribution, sketching the historical developments of traditional methods and the start of incorporating computational methods. This section also introduces research and theories focusing on emotional features in text and how they can be utilised in the scope of this thesis. The second chapter focuses on the methodological approach, describing the dataset, the preprocessing steps undertaken, and the programming pipeline used to subsequently train the machine learning models.

Afterwards, the results are described in detail. The results show that a slight increase in accuracy is possible by utilising the proposed methods. Both machine learning models, Logistic Regression and a Support Vector Machine, benefit from the incorporated emotional scores, and the overall accuracies for both models increase by one per cent. After the results section, the subsequent chapter discusses the results and puts them into perspective with existing research. After examining these points, the results are tied to the original research questions described in this introductory section. Lastly, potential limitations are briefly discussed before the conclusive part of the thesis.

2. Literature Review

This chapter aims to give an overview of the broad avenues of research about authorship attribution, exploring existing research and, ideally, tailoring it to the abovementioned research questions and how these approaches might lead to evolving new methodologies and current best practices.

First, the review will give a short historical perspective on the field's foundations and development. It will then examine traditional approaches and their modern counterparts before delving into emotional language analysis and the specific literature in that subsection of the field. This review aims to give an overview of the topic and the current ongoing discussions derived from a broad history of methodologies, some of which are well over a century old.

2.1 Authorship Attribution

The idea of determining the authorship of texts is old and goes back centuries (Koppel et al., 2009). However, in recent years, it has become more prominent in several fields with practical implications, from humanities scholarship to forensic analysis and plagiarism detection (Holmes 1994; Kalgutkar et al., 2020; Somers and Tweedie, 2003; Stein and Meyer zu Eissen, 2007). Its capabilities have increased with the development of robust computational methods. As such, the field has evolved significantly over the years.

Initially, the practice relied heavily on stylometric methods, categorising unique features to objectify writing styles (Stamatatos, 2009). This builds on the assumption that different authors have varying literary styles, which can be analysed. The statistical analysis of these variations is known as stylometry. It is based on measurable text features that form an author's linguistic style. For these methods, text is considered a word sequence where each word is represented as a singular token. These tokens are then grouped into sentences. Holmes notes: "At its heart lies an assumption that authors have an unconscious aspect to their style, an aspect which cannot consciously be manipulated but which possesses features which are quantifiable and which may be distinctive" (Holmes 1998, 1). This includes various measures for evaluating texts, such as sentence- and word length, sentence structure, word frequencies, and vocabulary richness, as the backbone of stylometry in modern authorship attribution methods to differentiate between authors (Holmes, 1994).

Features can be roughly differentiated into four categories to characterise writing styles: lexical features, syntactic features, styling features, and idiosyncratic features (Hurtado et al.,

2014). Lexical features include word- and character frequencies, average word length, and the use of case. Syntactic features include, for example, function words and special characters. Those are especially helpful since they are context-independent and can be analysed without focusing on the text content. Styling features focus on document layout, or visual specifications that the author chose for their text. This could, for example, mean to analyse the use of code comments. Lastly, idiosyncratic features analyse incorrect text sequences in texts, such as grammar mistakes. They could also entail cultural markers which the author might have subconsciously integrated into their text based on their origin (cf. Hurtado et al., 2014; Juola, 2008; Stamatatos, 2009). Stamatatos (2009) mentions in his authorship attribution survey more potential stylometric identifiers and methodologies, for example, n-grams and other style markers, together with machine learning algorithms, to showcase how stylometry plays a crucial role in research models attributing authorship. All these metrics do not exist in a vacuum, though, as the context also influences the writing style when texts are written in formal or informal settings (Overdorf and Greenstadt, 2016).

One straightforward feature in that category is measuring word frequencies, meaning the frequency of different words occurring in a text. This could mean the relative frequency of function words (e.g. ‘and,’ ‘the,’ ‘of’) as these often are unconscious writing decisions by an author (Stamatatos, 2009; Juola, 2018). Juola (2018) remarks that those function words are characteristics of a specific author. The assumption is that each author has a linguistic fingerprint that can be captured by these frequency metrics of their word usage. The same can be said of words used as sentence openers, another function word category. The reliable outcomes of analysing these metrics and their consistency led to robust research models (Yang and Chow, 2014). Further, those word structures are frequent and often independent from the content of a specific text, and as such, they are not influenced by the topic of a text. As such, they are often excluded from topic-based approaches since they do not carry semantic information. Still, depending on the research goal, the frequency of content words can also play a role, for example, when analysing synonyms or comparing texts about different topics.

In the usual setting, stylometry has focused on datasets comprised of small, closed-world models. In the usual setup, there are only a small number of candidate authors. The task then assumes a predefined set of possible authors. A substantial amount of training data represents each author; ideally, all data belong to the same genre (Stolerman et al., 2014). However, real-world scenarios are often different. It is a challenge of authorship attribution that

an author's writing style depends on a multitude of factors, such as the topic and its evolution over time.

Initially, this method was first used in the 19th century in a foundational study by Mendenhall, who analysed works of literature, such as Shakespeare's plays, based on word- and sentence length (Mendenhall, 1887). The idea was that an author's expression could be categorised by a curve consisting of the word length and the frequency of occurrence. Two further works on this notion were done by Mascol at the end of the 19th century in the New Testament (Mascol, 1888). Through these works, a baseline for statistical analysis of written text emerged, built upon by future scholars with the idea that quantifiable text features grew over the years, adding further statistical features (e.g. word frequencies, part-of-speech, character n-grams) to quantify texts, especially over the advent of digital technologies. By the end of the 1990s, it was estimated that over 1000 potential variables for stylometric analysis had been proposed (Rudman, 1998).

In the mid-20th century, one of the most seminal studies in the field was conducted by Mosteller and Wallace, covering the authorship of 'The Federalist Papers,' a series of political essays written by Alexander Hamilton, James Madison, and John Jay under a pseudonym. The study applied statistical methods to determine the authorship. This was done by collecting a small set of common function words (such as e.g. 'the,' 'and,' 'about,' etc.) and analysing the frequencies of those words by essentially applying a Bayesian classification to the paper's stylistic features (basically what is now known as the 'Naïve Bayes' algorithm). They produced significant differences between the candidate authors (Mosteller and Wallace, 1964).

The main idea was that the used algorithm extracted the word frequencies of function words to represent the grammatical style of the text adequately. Each document was then numerically represented by a set of word frequency values to form a high-dimensional feature vector. This allowed texts to be treated as points in a vector space to compare similarities by comparing the function word distributions. The advantages here were that all texts were of the same genre and thematic area and had a fixed set of possible candidate authors. From that point onward, new types and modelling methods of text features emerged to enhance the results of that original idea. With this study, the modern work in authorship attribution came to be referred to as nontraditional authorship attribution (Madigan et al., 2005; Stamatatos, 2009), compared to the more traditional human-expert methods, which relied on people manually examining the given texts and deriving metrics directly from them.

Following these advancements, until the number of potential variables suggested by Rudman (1998), most of these methods were computer-assisted by the end of the 1990s. The goal was rarely to develop a fully functional system on its own, and instead, the analysis itself was supported by computer systems. Further, one central problem of that era was the lack of objective evaluation of the methods, and the accuracy of the results was hard to estimate because many of those early tests were done on literary works of disputed authorship (Stamatatos, 2009). There were several limitations regarding methodology during this era, some of which still need to be considered today. Some of these include, for example, highly edited work either by a second party or by the author, based on the length of the text. If whole books and longer works were to be analysed, it could lead to an inhomogeneous style. Since this was before the computational era, it was also hard to objectively judge the results, and as such, the evaluation was often even more subjective. So, this method was prone to bias and the researcher's interpretation rather than being a standardised process across different experiments. By only choosing a limited number of candidate authors, the scope was restricted and did not reflect scenarios where many authors could be potential candidates. Following that, the developed methods were tailored too closely to the potential authors and could not be generalised well (Stamatatos, 2009).

Since then, machine learning advances have contributed significantly to successful authorship attribution (Juola, 2008), for example, through increasing the possible size of feature sets (Stamatatos, 2009) or by being able to classify sentence structures efficiently by analysing syntactic n-grams with machine learning (Sidorov et al., 2013). Another factor is that more text is available in electronic forms, such as emails, blogs, or forum postings. Those can be scraped with digital tools to build a large corpus for analysis and expand the scope of research. However, the need to handle this data efficiently has significantly increased with so much data available (Stamatatos, 2009). This need influenced machine learning, information retrieval, and natural language processing (NLP) methods to tackle large-scale datasets effectively (Madigan et al., 2005). So, Information retrieval was needed to efficiently process large amounts of data, which was then further analysed by machine learning algorithms. Further, NLP was used to evaluate new measures to reflect the author's style.

The early 2000s introduced a new era of authorship attribution and shifted the research focus from analysing disputed literary texts to evaluating proposed methods and comparing those (Juola, 2004). In that spirit, new applications were developed to analyse real-world texts, such as emails and blogs, in addition to books and older manuscripts. It became clear that the

training text size and the number of candidate authors play crucial roles (Abazari et al., 2023; Stamatatos, 2016). Apart from problems in authorship attribution, the methodologies were also being tested for more general text-categorisation tasks (Marton et al., 2005).

Nowadays, besides the often research-based question of whether a specific author wrote a text, authorship attribution has various real-world applications (Stamatatos, 2009). Naturally, it can be applied to all languages, although some, such as Chinese (Zhang et al., 2024), might need additional tools for detecting sentence boundaries; other texts might pose difficulties because of abbreviation usage and acronyms (e.g., text messages), which might lead to noise in the data (Stamatatos, 2009). For example, using Chinese texts can be challenging due to the characteristic structure of Chinese words with limited characters per word, which influences the effectiveness of some features, leading researchers to focus more on function words for stylistic analysis (Zhang et al., 2024).

As stylometric methods continue to develop, they have the potential to address the challenges posed by these issues deriving from modern digital communication. The impact that stylometry can have on accuracy is also dependent on the approach itself. The standard approach is the closed-world assumption of authorship attribution, where the author is within a suspect set of potential authors. As Stolerman et al. note: “Classic stylometric analysis requires an exact set of suspects in order to perform reliable authorship attribution” (Stolerman et al. 2014, 184). Nevertheless, this approach limits methodical options in open-world scenarios where the author might not be known at all. Addressing these limitations requires innovative approaches incorporating broader datasets and new problem statements diverging from the artificial situation of a test-based research data environment (Luyckx and Daelemans, 2008). Further, the size of the analysed corpus can play a significant role in correctly attributing potential authors. Smaller text corpora can be problematic and lead to incorrect classification (Eder, 2015). While exact numbers are hard to find, Eder (2015) notes that attribution tends to be untrustworthy with samples being shorter than 2.500 words (for Latin) and 5.000 words for most other languages.

To list a few real-world applications, authorship attribution can be used to determine plagiarism by finding stylistic similarities between two texts (Stein and Meyer zu Eissen, 2007). This can be used in academia or other branches where the question of authorship of a specific text is significant. It has been shown that authorship attribution is even effective in scenarios where the texts are specially reworked to make plagiarism harder to detect. This is done with the help of supervised machine learning models (Mahmood et al., 2020). There is also the

possibility that a text may have minor stylistic inconsistencies. Those can often happen in collaborative writing, where multiple authors with different writing styles work on a text together. For example, Graham et al. (2005) have developed methods to tackle this issue by analysing stylistic inconsistencies with the potential to determine if multiple authors have written a text.

Furthermore, to mention another example, text categorisation techniques can extract information about specific authors by exploiting combinations of lexical and syntactic features in texts, allowing for the inference of demographic information such as an author's age or level of education (Koppel et al., 2002).

Lastly, there are two branches related to security where authorship attribution can play a significant role. The first is in the context of cybersecurity to identify the author of malicious content, such as malware, through attributing written software code (Koppel et al., 2010). The challenge here lies in the enormous number of potential authors. This needs a robust model as there are problems, such as the potential of authors wanting to hide their authorship by obscuring their texts. Further, identifying the author of written code for programming is an extra challenge since code uses a different written text than literary works. In contrast, some code components recur despite authorship and instead because of code restrictions or specifications. Still, there have been propositions for effective identification when dealing with the authorship of a program (Frantzeskou et al., 2006; Abazari et al., 2023).

One of the more prominent examples in the security field is applications in forensic linguistics. These also deal with disputed or anonymously written texts, such as those of authors of threatening emails or criminal online forums (Stamatatos, 2017).

Since the history of authorship attribution goes back such a long time, a diverse array of methods has been developed and tested over the decades, from manual human-expert analysis to far more advanced computational methods involving large datasets, algorithms and machine learning-based methods. As such, the field has evolved significantly. Traditional approaches such as stylometry are still being utilised with great success today, and the other methodologies are being used to complement already existing methods for research. What is apparent is that computational technologies have played a significant role in bringing the field forward. As mentioned, there are many real-world scenarios where knowing the author of a text is important, and with the world shifting increasingly into the digital realm, this question will be ever more relevant.

The next chapter focuses on emotion theory models before delving into the impact those theories can play in authorship attribution when categorising texts based on their emotional tone. This will be done by giving an overview of the current research, which combines authorship attribution with emotion recognition.

2.2 Emotion Analysis

When discussing 'emotion,' a problem already arises when defining the term. The American Psychological Association defines 'emotion' as: "A complex reaction pattern, involving experimental, behavioural, and physiological elements, by which an individual attempts to deal with a personally significant matter or event. The specific quality of the emotion (e.g., fear, shame) is determined by the specific significance of the event […] Emotion typically involves feeling but differs from feeling in having an overt or implicit engagement with the world." Other researchers define the term differently while also considering the general, broader meaning of the word. Scherer notes: "While laypersons use 'emotion' interchangeably with terms such as affect, feeling, sentiment, or mood, psychologists define the construct as a process of changes in different components rather than a homogeneous state. Furthermore, it is assumed that the differentiation of the emotions (e.g., into fear, anger, or joy) is based on specific configurations of changes in the components" (Scherer 2001, 4472). Emotion detection then means to identify distinct human emotion types (Nandwani and Verma, 2021). When detecting emotion, various phrases are used, such as 'emotion detection,' 'affective computing,' 'emotion analysis,' and 'emotion identification,' which are often used synonymously and interchangeably (Munezero et al., 2014; Nandwani and Verma, 2021).

Even though these definitions provide a structure, another problem arises when trying to differentiate between emotions and trying to measure them. This question is heavily influenced by the theoretical stance adopted (Scherer, 2001). So, the concept of emotion must theoretically be examined to represent and distinguish different emotional categories and analyse them accurately. Some of these multiple theoretical positions are discussed in this chapter, as several models have been tested to systematise emotion and the associated behaviours. The current work describing and quantifying emotion can be divided into three main categories, discrete, dimensional and appraisal theories:

Discrete emotional theories describe emotion as an existing set of basic emotions, such as happiness, anger, sadness, surprise, and fear. Emotions are then grouped into these discrete

categories without an interconnection. Theorists such as Ekman, who proposed six basic emotions (happiness, sadness, fear, anger, surprise, and disgust), have made these models popular (Ekman, 1992). The assumption is that those basic emotions are universal across cultural backgrounds and human ethnicity (Pan et al., 2023).

If emotions can be categorised apart from each other and identified accordingly, texts can be categorised into a more accurate state of an author's emotions. So, the advantage of using the discrete emotional model is that basic emotions can be distinctly categorised based on their unique features. Another benefit is that "[…] it is intuitive to describe emotion based on the six emotion labels" (Pan et al. 2023, 2). However, researchers have also suggested that these basic emotions can be blended, "[…] explaining the large variety of emotion-descriptive verbal labels in many languages" (Scherer 2001, 4475). Ekman noted later that other universal emotions may exist beyond the original ones (Ekman, 2011). Using only the original six emotions may not fully consider the nature of more complex emotional states. Thus, the exact number of these basic emotions is controversially discussed (Geetha et al., 2024).

Dimensional emotional theories place emotions in three (sometimes two) dimensions (Rubin and Talarico, 2009; Osgood, 1966). The two primary dimensions are arousal and valence (Scherer, 2001), where valence indicates a positive or negative emotional response, measured as pleasure vs. displeasure. Arousal represents the intensity of the felt emotion. Low arousal, for example, is connected to less energy and vice versa. However, some emotional categories are more complex than those categorised by only two dimensions, as this approach would lack the required depth. For this reason, a third dimension, namely dominance, is often added to distinguish between closely related emotions more precisely.

Albert Mehrabian and James A. Russell first introduced the model, which is also referred to as the PAD (pleasure, arousal and dominance) emotional state model, to represent the general emotional dimensions (Mehrabian and Russell, 1974; Mehrabian, 1980). This dimensional emotion theory, with the valence, arousal, and dominance model, will be used in this thesis. The model structure with three evaluated states allows for a nuanced representation of emotional states. Since the dimensions influence each other, they can describe emotional states in a detailed form, and different fields of research have built their analysis on this framework. Integrating dimensional theories into a model allows for this more nuanced way of understanding emotional expressions linked, for example, to a writing style. This also represents the research approach that emotions are simultaneously distinct and dimensional, thus having a chance of overlapping (Lange and Zickfeld, 2021).

Lastly, appraisal emotional theories consider events and real-world situations with their interconnection to an emotional response (Scherer, 1999). This approach does not consider emotions to exist in a vacuum without reason. The question is how an individual responds emotionally to a specific event (Roseman, 2013). An event could be anything with which the individual interacts, such as a person or an object, or events like a sports match. Appraisal, then, is how the individual assesses and evaluates this situation. The idea is that emotion and reason are not disconnected and are interpreted by an individual (Martins et al., 2018). This also means that the same event can trigger different emotions in different individuals. Emotions, then, are caused and differentiated by the different levels of appraisals, which are individual responses to an event (Moors, 2017).

Generally, the commonality between these three models is a sense of positivity or negativity regarding emotion and their categorisation. The difference between these theories is then shown in different uses of values to distinguish emotion, such as a concrete event triggering an emotional response or the intensity felt when describing the impact of emotional perception. Still, these models share the 'polarity detection' (Cambria, 2010), which is the scores measuring how positive, negative, or neutral parts of a sentence are perceived on an emotional basis when talking about texts.

When considering the different models, the American Psychological Association's introductory definition of emotion seems more in line with the abovementioned appraisal theories. These theories connect emotions to a response to a specific event and the individual perceiving a particular event, which then triggers emotions.

2.3 Emotion Analysis Applications in Authorship Attribution

Traditionally, authorship attribution has been conducted with an approach based on quantitative textual features, such as word frequency or stylistic markers known as stylometry, and, e.g., machine learning approaches utilising these metrics as training data. However, those approaches do not always consider the qualitative aspects of writing, such as the emotional tone or the sentiment conveyed by the author in their texts. Without these considerations, traditional methods can overlook emotional nuances. For instance, an author might consistently write in a more positive tone, making their texts more recognisable and tailored to their writing style. The same thing might be apparent if an author employs a certain tone when talking about a specific subject, and, as such, it might be a distinctive feature of their writing style if it aligns with a

particular topic they are writing about. If they write about a topic they are enthusiastic about, this could be a marker which can be recognised, especially if the tone of writing is different when they write about topics they feel more neutral about. These emotional attributions add a new layer to the analysis by examining how the authors express their emotions. As such, emotional attribution is a potential additional field of authorship attribution for researchers to explore.

Defining and examining emotional attribution is an important part of this study. It can be referred to as identifying and categorising emotional content within a text and trying to understand how emotions manifest in language. This could include following the discrete emotion theory, which identifies and recognises basic emotions such as happiness, sadness, anger, fear, or more nuanced feelings (Ekman, 1992).

So far, this is a smaller niche in the field of authorship attribution, reflected by a limited number of studies focusing on this area. These studies use different theories regarding emotion recognition to attribute authorship based on these criteria.

For example, one approach is to utilise sentiment analysis by classifying words based on their emotional sentiments. 'Sentiment,' in this instance, refers to 'the expression of subjectivity as either a positive or negative opinion,' and the goal of sentiment analysis is to categorise whether a text is subjective and whether the expressed views are positive or negative with the help of computational methods (Taboada, 2016). From the limited research combining authorship attribution and emotion recognition, most focus on utilising sentiment analysis to quantify emotion, as it is often utilised to detect emotion from text (Nandwani and Verma, 2021). Ivanov and Perez (2024) used sentiment analysis for authorship attribution of poetry, adding it to a traditional-feature-based ensemble classifier and improving the attribution accuracy through this method, specifically noting that, while sentiment also works decently as a stand-alone stylistic feature, it is well-suited to be combined with traditional stylistic features. Hammoud et al. (2021) use sentiment analysis for author verification for short texts in the context of social media fraud. They examined 3000 authors who each wrote 1000 short text posts on the social media platform *Twitter (*now *X)*, saying that they were able to pinpoint the opinion of an author towards a specific topic while achieving a promising accuracy, even though falling short of accuracy scores compared to style-based methods (Hammoud et al., 2021). Another study also found that initially, the accuracy scores for sentiment analysis in authorship attribution were relatively low. However, additional features in combination with sentiment analysis increased the results further from 22% to 57% (Narayanan et al., 2018). Gaston et al.

(2018) also use sentiment analysis in adversarial authorship attribution. They use LIWC (Linguistic Inquiry and Word Count), a software to extract psychological insights from texts. This is done with an internal dictionary of ‘almost 6400 words that are placed in different hierarchical categories,’ to process each word individually (Gaston et al., 2018). Sentiment analysis then supplements these features. Again, sentiment analysis alone results in low accuracy scores across different models, ranging from 18-22%. In contrast, the accuracy increases when used in addition to the other computed features in a hybrid feature set. They conclude that “sentiment analysis features do not provide much information alone which might be due to the sparsity of the sentiment analysis feature vectors for each of the documents” (Gaston et al. 2018, 937). So, these studies suggest that sentiment analysis might not be effective on its own but can be a valuable asset to authorship attribution when combining it with other features (Narayanan et al., 2018).

So, even though limited, some research revolving around authorship attribution through emotional analysis has been done. Martins et al. (2018; 2021) discuss the impact of emotion analysis by predicting prominent authors of social media posts, politicians and non-politicians based on the emotion contained in their texts. The research utilised different methods to achieve its goal. Firstly, a polarity analysis was employed to determine the positive and negative words in the texts, independently of their intensity. Following that, they employed a lexicon-based approach, running all the collected words against words in a lexicon with labelled emotion scores connected to each word, using a discrete approach; lastly, to further identify the emotional patterns of each author, they ran a machine learning-based emotion analysis, showing a prediction rate of roughly 87%. This study also served as a primary inspiration for this master's thesis by using the impact of emotion scores directly to classify texts to specific authors better while utilising a lexicon-based approach.

In another study, the researchers analysed similarities between authors on *Twitter* (now *X*), exploring their emotional and grammatical writing styles. This was done to define an emotional profile for each author (Martins et al., 2019). They again used a lexicon-based approach to determine emotions contained in tweets.

Other notable research has been done by Sailunaz and Alhajj, who analysed emotion and sentiment from social media postings on *Twitter* (now *X*) (Sailunaz and Alhajj, 2019). After preprocessing, they trained a Naïve Bayes classifier to build a recommender system based on emotion scores. While not explicitly trying to predict authorship, they use emotion recognition as the main differentiator to build the recommender system. Specifically, their argumentation is

noteworthy as they note that, while successful, the representation of emotions following a discrete model can still carry different interpretations regarding the meaning. For instance, depending on the sentiment, the emotion 'surprise' can still be negative or positive, which poses a challenge for the correct classification. Occurrences like that further underline the challenges involving emotion analysis within text data, even when considering multiple approaches.

Boyd (2018) tries to assess authorship based on a psychological analysis of an author, for example, by looking at short phrases or spelling errors to identify a unique set of psychological attributes per author. The difference here is that these psychological features include social and cognitive tendencies in addition to emotional characteristics in their writing. The aim was to combine these methods by creating a unified approach named *Mental Profile Mapping*, noting that the approach was still in its early stages while highlighting potential benefits of this method with future research possibilities (Boyd, 2018).

Generally, while considering emotion in text, researchers must be wary of some common difficulties when extracting text data. Some of these issues are inherent to the standard way of using online platforms for communication, which might then be used as resources for extracting text corpus data. For example, users might not be restricted to just using text to express themselves. They can use photos, videos, hashtags, or emoticons, to name some options. All that can create noise in the data, and some of this apparent noise might also carry information. Emoticons, for example, can easily be used to convey emotion.

Taking all these considerations into account might require extensive preprocessing and decision-making steps, specifically depending on the research goal and what researchers try to accomplish when analysing the data.

3. Methodology

The methodology section focuses on the approach to answer the research question and the concrete steps to implement the necessary computational methods. First, the dataset will be introduced and discussed, and the necessary steps will be explained to ensure its representativeness before describing the preparation steps of the dataset. This entails taking measures to reduce the scope of the dataset to an appropriate size and the concrete preprocessing steps in which the dataset was altered, as well as why these steps are necessary to build a robust evaluation afterwards. Where appropriate, concrete examples of how the pipeline works on a computational level are discussed. Further, the extraction of stylometric markers and why these markers are important are explained, as well as why they are an important stage in the analysis, since they represent an author's unique writing style. Following that, another extraction process is undertaken, which specifically extracts emotional word features from the dataset, which is essential to the topic of this thesis, to compare those combined features with the stylometric features. The comparison happens after the model training, where those extracted features are used to train two machine learning models. Lastly, the results of the analysis will be shown and put into perspective after this section. The indications and potential explanations are investigated, bridging other fields and related studies in stylometry and authorship attribution. In general, the concrete analysis steps and their implementation will be looked at in detail to describe the necessary steps best and explain why they are needed to answer the research question accurately.

3.1 Dataset

When it comes to determining the ideal size of the dataset and which exact sizing to strive for, it depends on several factors. Generally, the more authors are looked at in terms of authorship attribution, the more challenging it gets (Stamatatos, 2016). These authors need to be differentiated based on their unique writing style, which is especially challenging when analysing these aspects in a large corpus with many authors. So, unsurprisingly, many authorship attribution studies have been done on small to medium-sized datasets with a relatively low fixed number of potential candidate authors. Abazari et al. note that: “Traditionally, author attribution techniques have been applied to problems where a small set of candidate authors is known in advance (e.g. plagiarism detection)” (Abazari et al. 2023, 516). According to their paper, most studies they examined (eleven out of 21 total studies) compare up to 20 authors, while some studies exceed that and settle between 29 and 70 authors. At the

same time, they note that only a few researchers consider larger numbers of authors to classify. In their own experiment focusing on authorship attribution in code, they also note that the accuracy decreased significantly as the number of authors increased. Furthermore, they explore how many text samples are needed for classification. According to their research, attributing a small number of samples is challenging for the experiment. In contrast, the performance plateaus once a sufficient threshold of sample sets is reached (Abazari et al., 2023). However, it should be noted that this paper specifically focused on code authorship attribution instead of text.

Other researchers note that the standard way to approach an authorship attribution problem is by assuming that an unknown text was always written by an author from the training set, compared to considering that an entirely separate author from outside the dataset might have written the text. Through this, the problem changes completely (Puig et al., 2016). They also point out the difficulties of machine learning algorithms to detect authorship correctly if the author text samples are not enough or too short. This means the longer the text, the better the performance of the models (Al-Sarem & Emara, 2019; Puig et al., 2016).

An ideal dataset for authorship attribution consists of an adequately large set of candidate authors, each with enough texts and documents, to have sufficient sample data to analyse. This also means ensuring that the author's texts are equally distributed to verify that the classes are balanced; otherwise, the results may be influenced accordingly, and the authors with more texts might be preferred in the predictions (Stamatatos, 2016). Considering this, the next part of this chapter will focus on how the dataset used for this master's thesis was created.

The dataset used was initially created for a separate project focusing on data analysis of text corpora at the University of Vienna. The methodology involved scraping openly accessible blog data from the web spanning roughly two decades[1]. The database collects blog posts of users' writings across various thematically different fields of interest. As such, the dataset provides sufficient data for further analysis based on the current research question, particularly by ensuring representativeness at the individual level by providing an adequate sample size of data per author.

The original problem of the dataset, which is also crucial for this thesis, was to build a database with large amounts of individually written data to have at least 100.000 tokens per

[1] The earliest blog entries were published in 2004, with the whole corpus spanning across 20 years from 2004 to 2024.

participant to ensure an appropriate sizing for data analysis and so that each author is represented with a specific minimum word count. To solve this problem, web-scraping blog data seemed a sufficient way to accomplish this task. Further, with this generous baseline of tokens per author, eventual problems arising from not having enough samples per author (cf. Abazari et al., 2023) are circumvented, even after performing extensive preprocessing on the dataset.

As a baseline, the general approach focused on searching long-running blogs managed by native English speakers who have written close to or above 100.000 words. Further, the collected blogs should represent various interests spanning different fields to diversify the topics and construct a well-rounded corpus. This is an inherent feature of this dataset, although in terms of attributing authorship, a dataset bridging multiple topics across authors can be a complicating factor. Stylometric features should ideally be immune to topic shifts or different genres and only reflect an author's personal style (Stamatatos, 2017). However, Stamatatos (2017) also notes that this is not fully proven. So, this is a complicated problem in authorship attribution, as writing styles are often influenced by subject matter or genre, and both aspects can mix (Song et al., 2019; Stamatatos, 2017).

The potential blogs were researched and collected in an *Excel* file after looking for them manually via online searches and blog search engines. Regarding personal blogs, it can be an issue to ensure whether the author is a native speaker. Each blog was manually looked at to ensure this would be the case[2]; although there is an element of uncertainty, it is hard to say that this is always true. Rudman (1998) mentions potential problems with corrupted texts as a 'major data problem' in authorship attribution studies. Aspects like plagiarism, imitation, translation, or lengthy quotations can cause problems with the analysis if the dataset itself is problematic due to not considering these problems. Those are just the aspects of 'authorial corruption.' Other aspects to consider are also experimental or editorial, such as, e.g., typesetting mistakes, or supplementing words to fill damaged text (Rudman, 1998). Some of these problems are circumvented by using a dataset comprised of blogs. It is relatively safe to assume that these private blog entries from individuals did not run through a prolonged editorial process from external editors, or that all blog posts have been translated into English from another language. So, an extensive blog dataset seems well-suited to avoid some potential pitfalls for authorship attribution that other datasets might have.

[2] For example, some blogs have an 'About' section where the author could have specified that they were from an English-speaking Country. However, these aspects can hardly be verified. Thus, some uncertainty remains.

Another predetermined selection factor was that all blogs were hosted on the platform *Blogspot* (a domain of *Blogger.com*), which was done for technical reasons. *Blogspot*, one of the oldest and most widely used blogging platforms, has hosted enormous content over the last decades. The technical advantage of scraping data from *Blogspot* lies in its framework, which consists of a straightforward HTML structure, making scraping uncomplicated through HTML element naming conventions in the backend. With this, the scraping process is simplified by employing automated scripts to extract content from multiple blogs within the platform efficiently. This was done with Python and the *Scrapy*-library[3].

At the end of the scraping process, the extracted contents for each blog were compiled into a structured JSON file consisting of the title, publication date, and textual content for every post. Thus, every blog has a JSON file consisting of all blog posts written by the author of this blog. With that step, each blog file is ready for further preprocessing.

All in all, the collected *Blogspot* text corpus contains approximately 81.000.000 words divided across 104 blogs (meaning each author is represented with their respective blog) written over 20 years. All blogs reach the threshold of roughly 100.000-word tokens per author, forming an overall robust corpus for further processing. Some authors extend that number significantly as they have written millions of words on their blogs over the years.

For this thesis, 25 authors were manually picked from the dataset to ensure that various writing styles were analysed while not oversaturating the data and the number of authors, partly due to the mentioned restrictions many authorship attribution studies face. All these authors have written thousands of texts on their blogs. To provide an equal number of texts for each author, the JSON files were capped at 1.000 texts per author so that every author has an equal amount of text with which they are represented in the dataset for this thesis. This does not necessarily mean that the word count is identical for each author, but rather that the number of blog posts that go into the next step is the same across authors. So, in total, the final dataset used for this thesis consists of 25 authors with 25.000 blog posts. The complete overview is given in the table below, including the authors, represented by their blog name, and the total word count they wrote in the 1.000 texts included in the dataset (Table 1).

[3] https://scrapy.org/

Author (Blog)	Word Count
baileysbeerblog	1127024
bluehousejournal	213636
boomergirlsguide	296088
cashonlyliving	383156
christasrandomthoughts	155049
crpgaddict	2921428
culturalsnow	354053
cuponthebus	449972
danabugseyeview	378947
dorcassmucker	817842
fatroland	455096
frikosmusings	537407
frugalistanbul	309226
glassincarnate	520200
interimarrangements	331395
kimmy-cookingpleasure	324885
lifeatgoldenpines	364858
lindahoof	446410
momsscribbles	401176
mumssimplylivingblogat	151798
myheartisalwayshome	280147
newamusements	299685
stitchesandseams	545731
thriftathome	285688
understandingsociety	1458333

***Table 1**. Dataset Overview (each Blog is represented with 1.000 Texts (Blog Posts)*

3.2 Preprocessing Steps

This section describes the text preprocessing steps in detail. A practical preprocessing pipeline is a crucial step in the workflow, as it ensures that the raw data is in a structured, cleaned format before starting the following analysis steps and extracting metrics as a numerical representation (Singh, 2023). As such, the data is structured and transformed appropriately for the stylometric analysis and the model training. This is an essential step to enhance the outcome of natural language processing tasks (Nazir et al., 2024). Thus, accurate preprocessing directly influences the potential outcome of later analysis steps.

The approach follows a two-part structure in analysing and processing the dataset. Firstly, the authorship attribution pipeline follows an approach without the context of emotional

language analysis; thereafter, the pipeline will be restarted with incorporated emotional markers. This is done to provide a basis for comparing both approaches later and to see if the incorporation of emotional features impacts the overall accuracy of the analysis. To start this process, both approaches must be handled by preprocessing the text files appropriately and, therefore, differently based on the needs of both pipelines. However, the different preprocessing steps are relatively nuanced.

The dataset consists of 25 JSON files, one for each author, which are loaded into one file. The popular Python library *pandas*[4] (McKinney, 2010) is used for this step to aggregate all files into a *pandas* DataFrame format. This data structure is included in the *pandas* library, which handles large amounts of data efficiently and quickly for further analysis. *Pandas* is one of the most popular and widely used libraries for efficiently handling large amounts of data (Petersohn et al., 2020). In this format, all information about the authors and their texts can be stored in a single large table format.

After loading the dataset into a DataFrame, another function is designed to prepare the text for stylometric analysis. The linguistic features of the text have to be preserved since they are needed for the stylometry analysis, while they would be discarded for most other text analysis tasks. For example, sentence structures and lengths are still needed for later steps. Thus, in this case, punctuation remains in the cleaned texts, as well as capitalised letters. Since the dataset has a significant size, the steps undertaken to transform the text have been manually verified on subsamples of the function outputs to inspect if the texts were accurately transformed. However, there is some uncertainty about whether this task will always be performed entirely accurately. Still, manually checking samples of the output after each step was done to ensure that the target was achieved with a high probability.

Some text elements are excluded during this preprocessing step to standardise the text and to reduce the noise. Certain text elements, such as special characters, website URLs, email addresses, or social media tags, are unnecessary and were removed using regular expressions in the appropriate Python preprocessing function so that only relevant text content remains. The same is done with excessive whitespace so that multiple lines of consecutive spaces are replaced with a single whitespace. With that, word boundaries remain while unnecessary characters, which could lead to frequent noise in the data, get removed.

[4] https://pandas.pydata.org/docs/

Another step in the preprocessing pipeline is tokenising the text, which can now be done after ensuring no extended whitespace or non-printable characters remain in the text. This involves splitting the text into individual chunks of words (character strings) so that each word can later be analysed. The goal of this preprocessing task is 'to convert a stream of characters into a stream of processing units called tokens' (Hassler and Fliedl, 2006). The Python library *NLTK* (Natural Language Toolkit) is used with its included tokeniser function. After the abovementioned steps, a tokenised copy of the input text is returned.

Lastly, the text is lemmatised to standardise the word forms. Lemmatisation means reducing words to their base form so that they are represented equally. This simplifies words and word counts by removing different context-based word forms they otherwise might have based on, e.g., tense or case (Balakrishnan and Ethel, 2014). In this case, this process has, for example, the advantage that word frequencies can be more accurately measured later, when the words are represented by their lemmatised form only. This step is also done by using the feature provided by *NLTK* (Bird et al., 2009). The *NLTK* documentation states that POS tags are valid options to use with the lemmatiser (e.g. 'n' for nouns, 'v' for verbs, 'a' for adjectives, 'r' for adverbs). It then returns the 'shortest lemma of words for the given pos.'[5] Part-of-speech (POS) tagging is a natural language processing task which assigns grammatical tags to each word token throughout the dataset, indicating if the given word is, for example, a noun or an adjective. This carries information regarding the grammatical structure of a sentence (Zhou et al., 2018). Here, it is explicitly implemented as part of the lemmatisation preprocessing step so that the lemmatisation is more accurate since the function knows which grammatical cue the given word has and, as a result, lemmatises the word with better accuracy. Later, POS tagging is also implemented separately in the function to extract stylometric features.

Having tested both variants, the lemmatisation performs better when including POS tags, on the basis of which the tokens get lemmatised. So, this variant has made it into the final preprocessing function. With these steps, the general preprocessing steps are presented in Figure 1.

[5] https://www.nltk.org/api/nltk.stem.wordnet.html

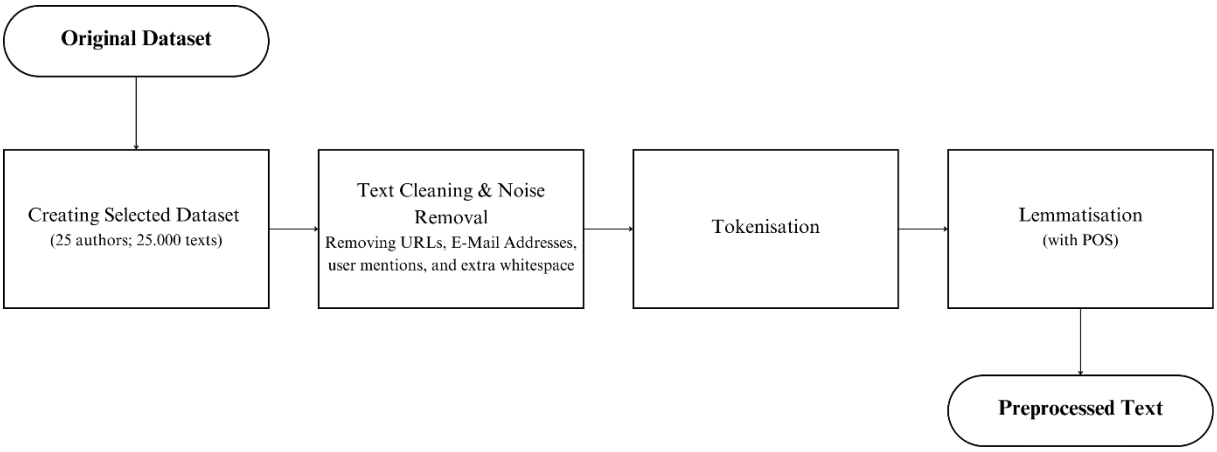

***Figure 1**. Preprocessing steps*

3.3 Building an Emotion Score Lexicon

The emotion scores are implemented in the second part of the preprocessing steps as a separate function. This builds on an emotion lexicon of valence, arousal, and dominance (VAD) scores collected by Warriner et al. (2013). Generally, several well-known emotion lexicons exist, such as the EmoLex lexicon (Mohammad and Turney, 2013), ANEW (Bradley and Lang, 1999), WordNet Affect (Strapparava and Valitutti, 2004), and SentiWordNet (Esuli and Sebastiani, 2006), with the ANEW norms collected by Bradley and Lang (1999) being the ones on which nearly all research has been based (Warriner et al., 2013). Warriner et al. (2013) extend the ANEW lexicon significantly, which is why their lexicon is used for this thesis. The data they created can be used through the additional files published in addition to the original research paper, which are downloadable as a CSV file. This file contains multiple scores, which the researchers aggregated as a general resource for other studies to use. The collection contains 13.915 words of the English language and their adjacent scores. For this thesis, the mean scores are specifically needed for the analysis. So, the original CSV file has been cleaned to only represent these scores per word (Figure 2). This is the basis for a word dictionary, which is built through a function in the preprocessing pipeline.

Word	Valence	Arousal	Dominance
aardvark	6.26	2.41	4.27
abalone	5.3	2.65	4.95
abandon	2.84	3.73	3.32
abandonment	2.63	4.95	2.64
abbey	5.85	2.2	5.0
abdomen	5.43	3.68	5.15
abdominal	4.48	3.5	5.32
abduct	2.42	5.9	2.75
abduction	2.05	5.33	3.02
abide	5.52	3.26	5.33
abiding	5.57	3.59	6.6
ability	7.0	4.85	6.55
abject	4.0	3.94	4.35
ablaze	5.15	6.75	4.58

Figure 2. *The first lines of the extracted VAD scores used for the dictionary.*

The reduced VAD lexicon is read from the CSV file and transformed into a DataFrame. The DataFrame contains the column names 'Word,' 'Valence,' 'Arousal,' and 'Dominance' based on the original names in the CSV file. Based on this DataFrame, a Python dictionary is then built. This dictionary maps each word to its corresponding valence, arousal, and dominance scores so that each entry contains: 'Word: (Valence score, Arousal score, Dominance score)'. The idea is that this dictionary is then used in later analysis steps, making it possible to quickly retrieve the word and its corresponding VAD scores from the dictionary for any given word in the original dataset.

The actual score calculations happen in a different function that calculates the average valence, arousal, and dominance scores for all words in a given text (blog post per author). Each word that matches a word in the VAD lexicon is identified, and its corresponding VAD scores are retrieved. This is also where the lemmatisation process of the original preprocessing becomes important. The VAD scores by Warriner et al. (2013) are computed and reported with word lemmas. So, to accurately match those scores against words in the dataset, those words also need to be lemmatised to be recognised by the matching algorithm. If it still happens that a word is not matched, or if a word is not present in the VAD lexicon, the function assigns

default values of (0.0, 0.0, 0.0) to prevent missing values. Then, the function computes the mean VAD scores across all matched words to represent the input text's overall emotional characteristics, which can be individually saved as features.

3.4 Feature Extraction

With all the preprocessing steps done, the necessary groundwork is established for enhancing the DataFrame with stylometric measures and focusing on the analysis and model training in the next step.

The pipeline's preprocessing steps add the processed text as a ‘clean_text’ column to the existing DataFrame from which the stylometric markers are extracted. All those extracted values are then individually added to the DataFrame as new columns based on the individual cleaned texts by the authors. As described, each author has 1.000 texts (blog posts) in the dataset. With 25 authors, the complete DataFrame consists of 25.000 texts; for each of these texts, all metrics are calculated.

The calculation of stylometric markers happens in a twofold fashion, with two main functions that handle the extraction. With the function stylometric_features, all cleaned texts in the clean_text column of the DataFrame are processed. The function then extracts the following stylometric features, with their selection based on the existing research described in the literature review:

- Word Count (word_count), to determine the number of words in a given text.
- Character Count (char_count), the total number of characters in a given text.
- Sentence Count (sentence_count), the number of sentences in a given text, differentiated by the use of punctuation, which is also, for example, why these punctuation markers had to be kept in the text files during the preprocessing steps.
- Average Sentence Length (avg_sentence_length), the average number of words per sentence.
- Type Token Ratio (type_token_ratio), the lexical diversity determined by the number of unique words in ratio to the total number of words in a given text.
- Function Word Frequency (function_word_ratio), the amount of function words relative to the word count.
- Part-of-speech (POS) Frequency (pos_features), analyses the frequency distribution of Part-of-Speech tags while utilising the appropriate function from the Python library

NLTK. Compared to the earlier computation of the POS tagging for the lemmatisation preprocessing step, the POS tags are now saved as features in the DataFrame based on the cleaned texts.

After running this function and computing all these features, the function returns an updated DataFrame with all features added as separate columns after the clean_text column for each author. This function has some potential efficiency limitations that have to be circumvented by the steps taken before. The function builds on the preprocessing steps, so the better the quality of the preprocessing is, the better the results of the extracted features are. For example, the POS tagging depends on the quality of the tokenisation and lemmatisation that happened before. By again manually inspecting random output samples, an attempt was made to ensure that the quality of the extracted features was sufficient. Table 2 summarises the characteristics of the stylometric features in the dataset and their variance distribution across the corpus.

Feature	Min	25%	Median	75%	Max
Word Count	0	174	341	610	8798
Char Count	0	820	1617	2891	43935
Sentence Count	1	10	20	35	508
Avg Sentence Length	0	13,84	17,2	21,57	404
Type-Token Ratio	0	0,45	0,53	0,62	1

Table 2. *Stylometry Feature Overview*

The second function (add_word_frequencies) computes word frequencies for the clean_text column of the DataFrame and adds those words with their frequencies to the existing DataFrame for further processing. The idea for computing these word frequency scores is that the function calculates the 100 most frequent words across the whole dataset and keeps them internally stored. Then, it looks through each individual text document to count the word occurrences of these 100 words in that document and adds them as values in these cells. In the end, each of these 100 words is a separate feature of the DataFrame. There is no strict consensus on the number of frequent words which should be used as features; generally, the numbers vary, with some older studies using at most the top 100 most frequent words, which were considered an adequate representation (Stamatatos, 2009). Newer studies used more words (Koppel et al., 2007; Stamatatos, 2006). However, those studies often use word frequencies as their sole

features (cf. Koppel et al., 2007). In the scope of this thesis, it seemed like a reasonable choice to limit the number of most frequent words to 100 to not overshadow all the other extracted features, especially since the primary focus should be on the impact of emotion features in addition to the other features, which are also considered apart from just word frequencies.

Computing the word frequencies can be accomplished by using CountVectorizer, which is a feature extraction method of the *scikit-learn* Python library (Pedregosa et al., 2011). The library documentation mentions that the module converts a collection of text documents to a matrix of token counts. Further, '[t]his implementation produces a sparse representation of the counts using scipy.sparse.csr_matrix' (scikit-learn developers, 2007). It creates a vocabulary out of all unique words from the text corpus. It converts each text into a numerical vector to count each individual word representation in the given document. So, for each document, it creates a row in the Document-Term Matrix for the occurrence of a given word. Each column then corresponds to a word from the built vocabulary. The number of features will equal the given vocabulary in the corpus if not otherwise specified. For this reason, the output will be limited to the most common 100 words in the text corpus. This is specified in the function call by limiting the maximum number of features to 100. Finally, it returns a sparse matrix representing all documents in the dataset, where each row represents a text, and each column represents a word. The values in the matrix are equivalent to the number of times the given word appears in the specific document.

To fit this matrix into the existing DataFrame, it is converted into an array before being stored in an intermediate DataFrame. Afterwards, it is merged with the existing DataFrame while adding new columns for each word (where the columns are named after a word). The aggregated DataFrame now has the existing stylometric features and the word frequencies computed with CountVectorizer.

Since the original dataset is quite large, computing all these different features is resource-intensive and can take considerable time. POS tagging is a good example since it tokenises and tags each word in the process individually, slowing the computing time for large amounts of data.

To circumvent reprocessing each of these steps for every program run, the DataFrame is saved after extracting all features using the Python library *Joblib*, an efficient storage library for handling large data. Then, it can be quickly loaded for the next part of the program or to test different pipelines without needing to compute every abovementioned step again for each run.

With these steps taken, the DataFrame is ready for the next steps. Overall, the computed DataFrame incorporates 25.000 entries with 154 columns each. Those columns first include the author, the unprocessed and the cleaned text, and then the stylometric features, as well as the 100 most common word frequencies. Since the DataFrame is too large to display, Table 3 shows a sample of the DataFrame, consisting of four entries from different blogs (entries 100, 1000, 10000, and 20000) by different authors, with some of the extracted stylometric values (word_count, char_count, sentence_count, avg_sentence_length) and word frequencies for five random words (which, you, work, who, what).

author (blog entry)	100 *baileysbeerblog*	1000 *bluehousejournal*	10000 *fatroland*	20000 *myheartisalways home*
word_count	1830	3352	381	957
char_count	8851	15020	1933	4314
sentence_count	58	204	34	67
avg_sentence_length	31,55	16,43	11,21	14,28
which	5	9	1	0
you	6	11	8	1
work	3	9	0	10
who	1	2	0	2
what	2	8	0	1

Table 3. *Exemplary data from blog posts across the dataset.*

3.5 Emotion Pipeline

With the main DataFrame established, the second, separate part of the pipeline can be added, namely the computing and addition of emotion scores for each text. This builds on the preprocessing steps as well by using the described functions regarding the VAD scores of words collected by Warriner et al. (2013).

Now, these scores are implemented in the DataFrame based on the cleaned texts from each author. The function is applied to each element from the clean_text column of the DataFrame. Then, it calls the function that computes the VAD scores depending on the dictionary built from the original word scores from the CSV file. These scores get assigned to three new columns in the DataFrame. Lastly, this DataFrame gets returned by the function. With

these steps, the original DataFrame gets extended with three new features: valence, arousal, and dominance scores per text.

To compare both pipelines efficiently, the DataFrame is saved separately via *Joblib*. That way, one DataFrame, including the stylometric extractions and the word frequencies, is available, and a second DataFrame builds on the first one with included VAD scores. The latter then includes 157 features.

3.6 Exploratory Data Analysis

Now that the DataFrame contains all the necessary information, it also reveals some insights about the distribution of the different features, which can be examined in detail to compare the stylometric extractions from each author. This is mainly practical to get a feel for the data or the extracted features, to look and discover patterns and to visualise that the impact of the different extracted features can be significant since the differences across authors can be vastly different. This "can be loosely characterized by (a) an emphasis on the substantive understanding of data that address the broad question of 'what is going on here?' […]" (Behrens 1997, 131-132). This also includes possible visual representations to represent the patterns in the data.

For example, the average sentence length across the dataset can be examined. This data can be visualised using the Python libraries *Seaborn*[6] and *Matplotlib*[7] (Hunter, 2007; Waskom, 2021). The histogram in Figure 3 shows the distribution of average sentence lengths across the entire dataset of 25.000 texts. According to the data, the distribution peak contains between ten and twenty words per sentence, even though longer sentences exist (for better readability, the maximum sentence length for the graphic has been capped at 50 words). A small peak exists right at the beginning of the curve. Since the dataset contains blog posts, this could indicate that some blog posts contain very few words, leading to a slightly higher distribution of very short sentences right at the beginning of the curve. This might also explain the outliers in Table 2, where some blog posts had one sentence with zero words and characters. This might be a blog-specific idiosyncrasy where authors left the blog post empty and posted it with just a headline.

[6] https://seaborn.pydata.org/
[7] https://matplotlib.org/

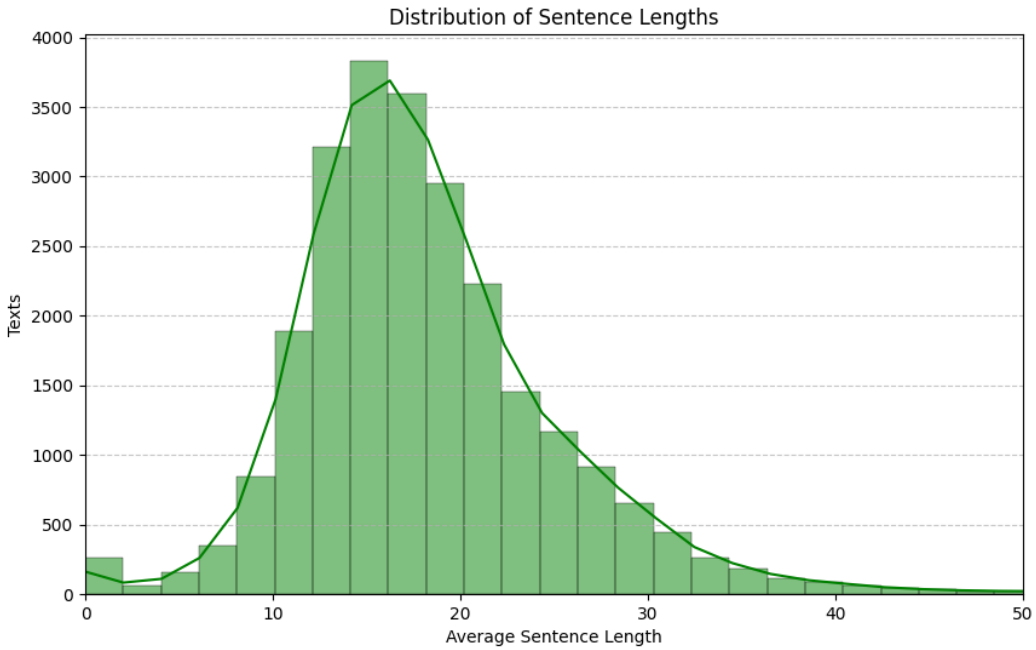

Figure 3. *Average Distribution of Sentence Lengths across the dataset.*

This first visual representation of the average sentence length can be more granularly defined by examining the average sentence length distribution per author, as presented in Figure 4.

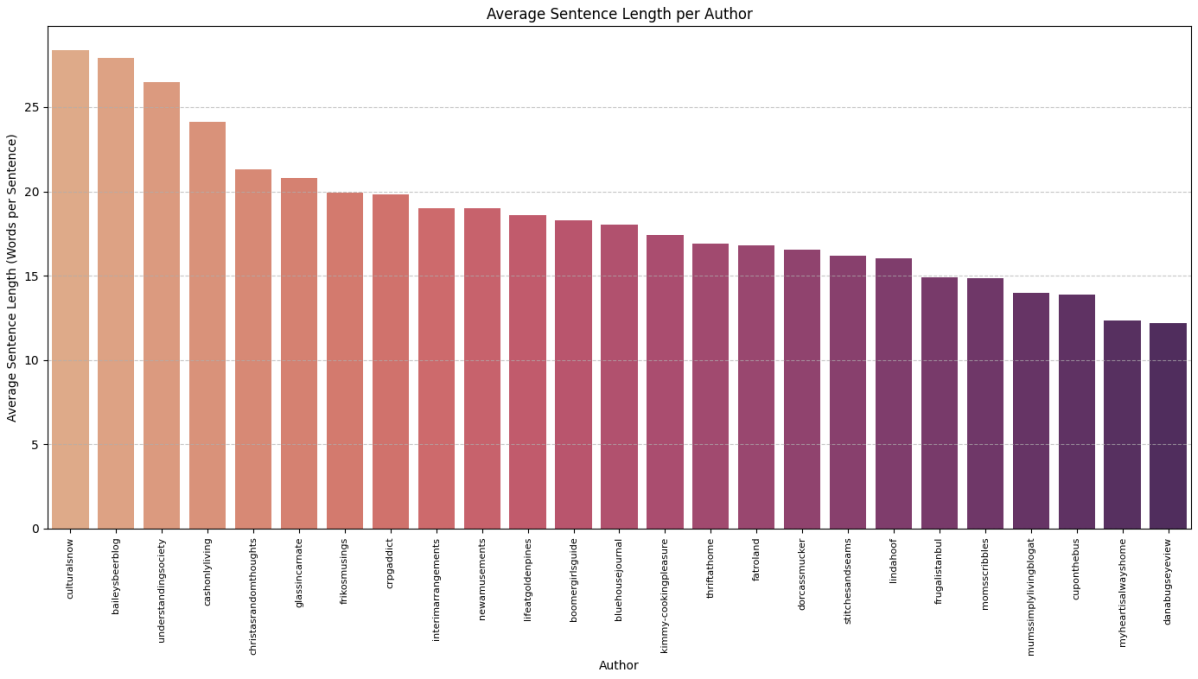

Figure 4*. Average Sentence Length per Author*

Examining this visualisation, it can be seen that although some authors use longer sentences, many authors tend to write shorter to medium-long sentences, with 15 to 20 words per sentence, which aligns with the distribution of Figure 3. Nevertheless, the distribution of sentence lengths diverges vastly across authors in the dataset, specifically when comparing the top-end and the low-end.

Thus, it becomes clear that there are distinctive differences between the authors and their writing styles, even just by looking at one of the stylometric features from the DataFrame. In the most extreme comparison, the longest average sentences (by *culturalsnow*) are twice as long as those by the author with the shortest average sentence lengths (by *danaburgersreview*).

Another metric that can be examined is the distribution of the most common words across the dataset. Since stopwords are included in the dataset and did not get extracted during the preprocessing steps, it seems likely that they are most represented in the dataset. This is confirmed by looking at the visual representation in Figure 5. Still, since stopwords are vital for examining an author's writing style, leaving them in the corpus is necessary, and their distribution varies greatly.

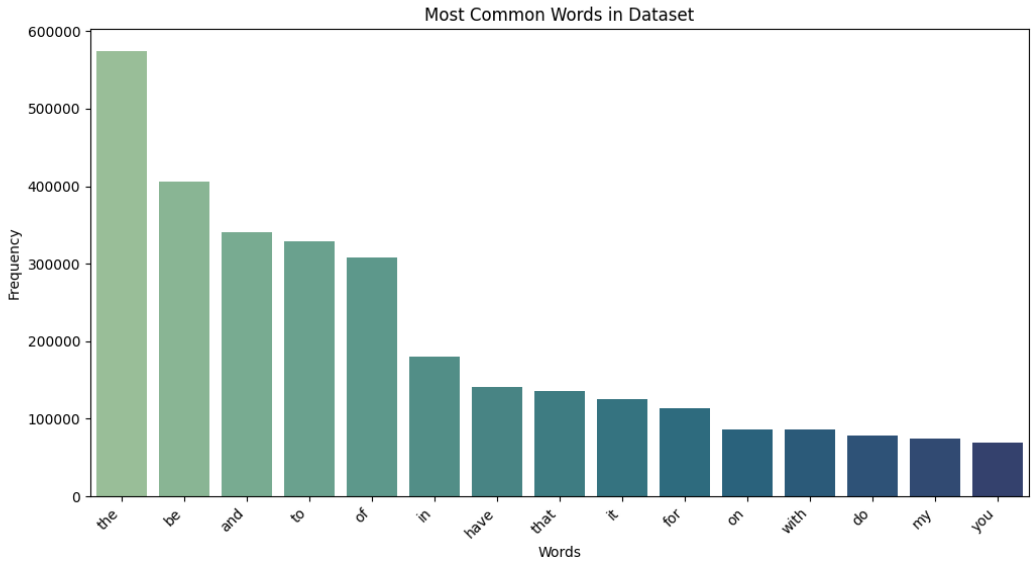

***Figure 5**. Distribution of the most common words in the dataset.*

For comparison, if stopwords were to be removed during the preprocessing phase, the most common words almost entirely differed, with the most common words having a much lesser frequency than before comparing it with their counterpart stopword frequencies, e.g., ‘the’ having nearly 600.000 occurrences as the most common word, while the most common word with stopwords removed (‘get’) has only slightly above 40.000 occurrences in the dataset (Figure 6). As a sidenote, stopwords were only excluded in this step for the exploratory data analysis to visualise and compare the frequency differences.

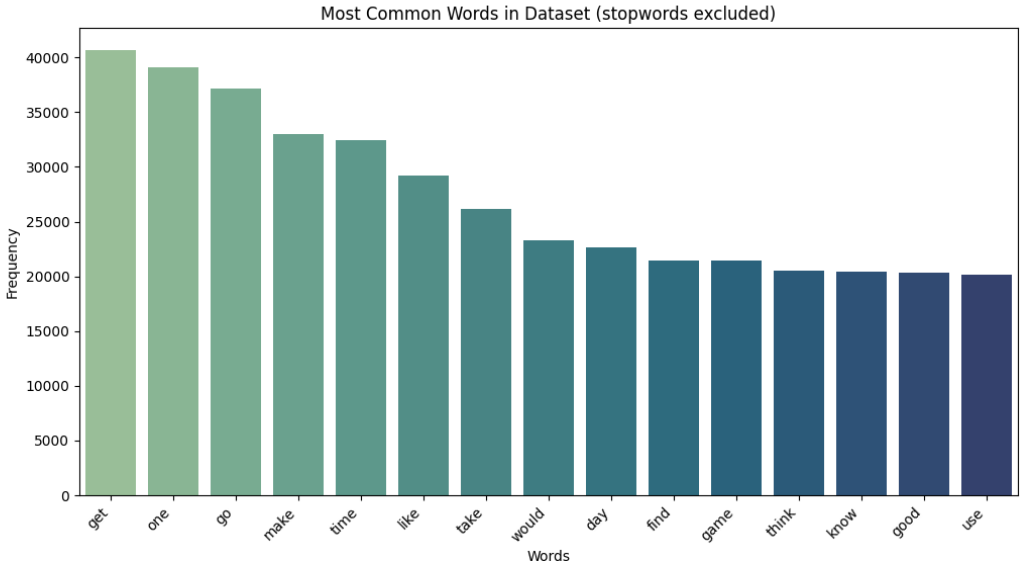

Figure 6. *Distribution of the most common words across the dataset excluding stopwords*

The extracted emotion features (VAD scores) can also be visualised for each author. It becomes apparent that most scores have a relatively similar distribution over the dataset, with slight differences per author, showing some outliers across the different valence, arousal, and dominance scores (Figure 7).

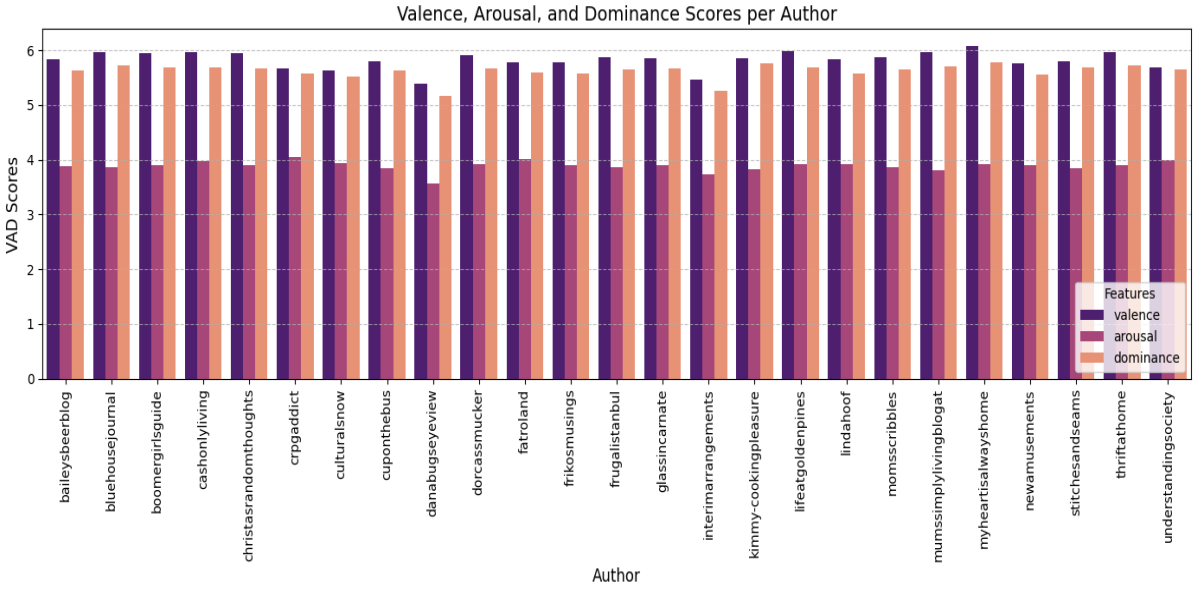

Figure 7. *Emotion Scores distribution per author*

As an exemplary combination of both extracted metrics, the VAD scores, or, in this case, the valence score, can be combined with the sentence length distribution in the dataset. This is done in Figure 8.

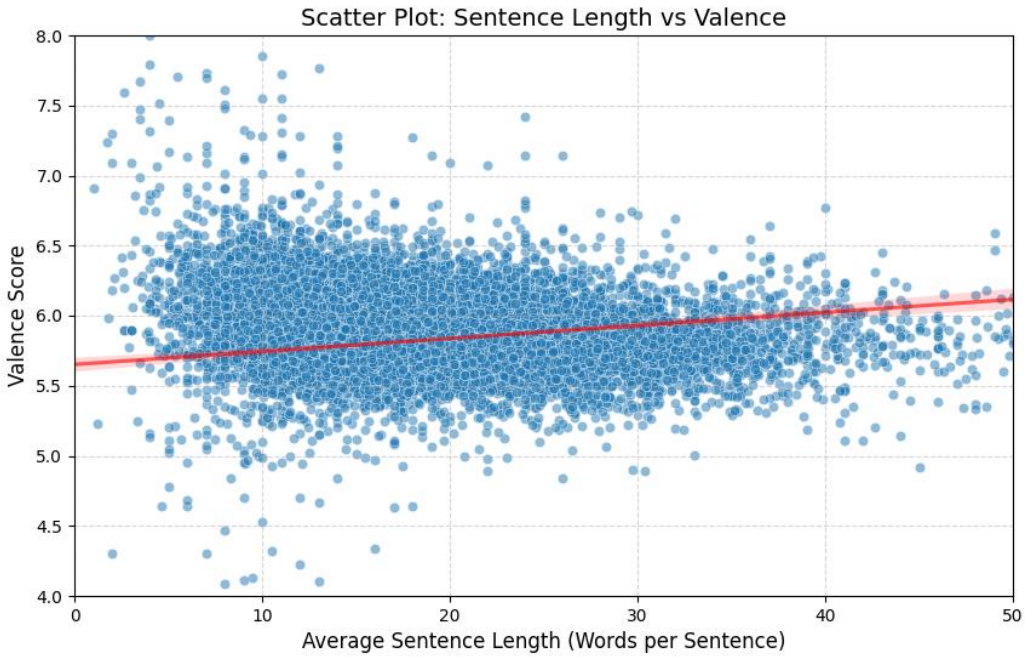

Figure 8*. Average Sentence Length distribution and Valence Scores across the dataset*

This figure shows the sentence length distribution in comparison with their valence scores to show if there might be a correlation between the two metrics. That would, for example, show that sentences with higher sentence length show a higher valence score than shorter sentences. In the visualisation, there seems to be a weak indicator that this is the case, judging from the regression line increasing slightly as the number of words in a sentence increases. However, there are many outliers across the dataset in both directions, meaning that even though the assumption might be valid, the correlation is somewhat weak. Further, the distribution appears to exhibit a U-shaped pattern, with shorter sentences also showing higher valence scores than their medium-length counterparts. However, these metrics alone might not provide sufficient data to confirm this assumption, and further examination of stylometric features and their interplay would be needed.

Visualising extracted features across the dataset shows that the distribution of stylometric features varies widely from author to author. This can be beneficial for classifying and attributing authorship. As mentioned before, these features define successful authorship attribution by ensuring that texts with unique features can eventually be attributed to the correct author with a similar writing style. The better the underlying data, the better the chances that unknown texts can be correctly assigned to an author.

3.7 Machine Learning Models

The next step in the analysis entails training the different machine learning models based on the created DataFrame. After testing different models, two of them are described here, which will then be compared directly. The implementation of those models happens after the DataFrame is created. Both models can then be compared to each other, and with the modular structure of the program, both can easily be trained on the two available DataFrames – the first DataFrame with only stylometric and word frequency features extracted, while the second DataFrame also encompasses the three emotion features.

This happens again using the available methods from the *scikit-learn* library tailored towards the input data. At the very start, two variables are created to prepare the data for model training. ‘X’ contains a copy of all computed numerical features of the original DataFrame while dropping the ‘author,’ ‘text,’ and ‘clean_text’ columns since these are not needed for the model training based solely on the extracted features. The variable ‘Y’ then saves the authors from the

dataset as labels for the model training to ensure that the authors are predicted correctly. Since the rows are still aligned consecutively, the model can still know which line of numerical features corresponds to which author from the labels afterwards.

As a last step before training a model, all features are standardised to ensure that all features are considered equally for model training. This is done with the standard scaler provided by *scikit-learn,* which is widely used for data scaling and feature normalisation. If this step were not implemented, some features might have much larger values than others, and these features might dominate the model. The standard scaler considers each feature used for model training, subtracting the mean and dividing by the standard deviation. This step ensures that the features have a zero mean and unit variance. The *scikit-learn* documentation defines this scaling feature as a 'common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data […].'[8] The idea is that all features have roughly the same scope with a variance in the same order. Estimators might experience complications and perhaps be unable to learn from much larger features.

Both implemented models are trained via supervised learning. This means training the model with known data and supervised outcomes, such as the testing set. With the work so far, the prerequisites for this step have been laid out. So, ideally, the model produces an accurate output prediction based on the known output parameters while trying to learn this behaviour by the known input-output relationships, which can be used to classify unseen data samples; the algorithm tries to pinpoint the target values based on the training data (Cady, 2017; cf. Singh, 2023).

After both models are trained, the performances are evaluated with a classification report using the provided function by *scikit-learn*[9]. This summarises the results and includes accuracy scores and other metrics, including recall, precision, and F1-scores across the whole dataset and for each author. For this analysis, precision represents the number of correctly predicted authors amongst the predicted positives. The recall counts the actual authors who were correctly predicted. The F1-score is then a balanced, harmonic measure of precision and recall scores, while, lastly, the overall accuracy measures the proportion of predictions that were correct.

[8] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
[9] This function builds a text report visualizing the main classification metrics of the model as a table.

The first option is to train a logistic regression model using the features ('X') and the author labels ('Y'). Logistic regression is a classifier that assigns feature weights to decide which sample a class belongs to. It learns those weights for individual features, which are linearly combined to get the classification (Singh, 2023)

This works by splitting both inputs into training and testing datasets with an 80:20 ratio. The model then learns from the training data, which includes 80% of the total dataset. After that step, the model is tested on the previously unseen data in the testing set to evaluate its accuracy. Logistic regression works generally well with numerical data, specifically if it is scaled through measures like the earlier implemented StandardScaler (Cady, 2017). Further, logistic regression has additional advantages, e.g. regarding interpretability of the model and generally light computing costs (Cady, 2017; Shim & Lee, 2011).

The training data is used to predict values ('Y'), which are then compared to the actual labels (authors) of the testing set. Based on this comparison, the accuracy scores calculate the number of correct predictions made by the model. These scores can be summarised by the classification report, which can be generated with *Scikit-learn*.

Several options have been tested for running the logistic regression model to find the optimal configuration for the present data. One example of this is the already mentioned standard scaler. However, there are more possibilities, like changing specifics of the model training process, which have been undertaken and will be explained shortly:

- There are multiple possible solvers coming with the *scikit-learn* library. The documentation reveals possible solver choices based on the dataset[10]. For example, the 'saga' solver is well-suited for large datasets but not robust for unscaled datasets. Since the present data is scaled, that would not be an issue. However, it is computationally expensive and takes considerably longer than other solver options. Perhaps, with a larger dataset, this solver would have been the premier choice. Still, other solvers have been tested to circumvent this, ultimately settling on the 'lbfgs' solver. It is essentially an optimisation algorithm to find ideal model parameters. Specifically, this solver has significant computational advantages as it is memory efficient and thus it runs considerably faster while producing similar results.
- There is a L2 penalty assigned to make the model generalisable. This is standard with the 'lbfgs' solver to reduce overfitting by putting weights on specific values to prohibit

[10] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

the model from fitting training data too closely and to prevent a single feature from dominating predictions. It encourages small weights and penalises larger ones, while not removing any features (although it shrinks their potential impact). This penalty reduces model complexity and advocates noise reduction in the training data and finding patterns which are more likely to generalise well when faced with unseen data. The *scikit-learn* documentation describes the process as adding a constraint (penalty term) to the loss function, which is the sum of the squares of the model's coefficients, multiplied by the regularisation strength[11]. Doing that discourages single coefficients from being too large.

- Lastly, when splitting into training and testing sets, stratify=y was used. This metric influences the testing set for each author. Without implementing this, the train and test sets might have differing class distributions across the testing data. For example, even though 20% (the size of the testing split) would be 200 texts per author (with 1000 texts per author in the total dataset), this number might be distributed differently so that, say, one author might have 180 texts in the testing set while another has 220. This is circumvented by using stratification, which ensures the proportions in both sets are the same. This would also play a role with imbalanced datasets, where some authors might vanish entirely from the testing set when there are too few samples compared with other authors appearing more frequently in the original data. With stratification enabled, they would appear in the same ratio. Since the present dataset is balanced, each author appears with 200 testing texts and 800 training texts through the 80:20 split. In the testing phase, the balanced dataset led to a slightly enhanced accuracy when comparing the results.

The second implemented model is a Support Vector Machine (SVM). Support Vector Machines have been widely used as a popular model for authorship attribution problems (Juola, 2008; Neal et al., 2018; Overdorf and Greenstadt, 2016; Pan et al., 2023). They have produced good results and regularly outperform other models, such as decision trees (cf. Juola, 2008). This approach works similarly, with minor differences depending on the model choice. Like before, the data is split into 80% training and 20% testing data.

[11] https://scikit-learn.org/stable/auto_examples/linear_model/plot_ridge_coeffs.html

A Support Vector Machine ‘computes a hyperplane that best separates the datasample. […] This line is also called the decision boundary’ (Singh, 2023).

The decision boundary then separates different classes with their samples. The data gets separated accordingly depending on which kernel is used to train the SVM. Kernels can transform linearly inseparable data into linearly separable data (cf. Singh, 2023). The SVM learns stylistic patterns from the training set and then compares the predictions for the labels with the correct predictions from the testing set before it prints the performance score and the classification report.

Similar to logistic regression, the Support Vector Machine also employs some specification options tailored towards the given data, namely:

- Stratification (stratify=y) is also employed due to reasoning similar to logistic regression. This leads to an equal 80:20 split across the training and test sets, with the numbers per author being distributed in the same way.
- A linear kernel is used to train the Support Vector Machine to find decision boundaries between authors.
- The pipeline also employs the standard scaler to standardise all features before training.

For both models, there is an option to limit the feature selection to a number of the most influential features to fine-tune and account for features that might not influence the results by filtering them out and training the model only on the most impactful features. However, testing this did not improve results, so all 154 (157 features including VAD scores) have been used since the whole selection performed best.

4. Results

The result section focuses on discussing both model results of the training step and comparing the different model performances while taking the differences into account. It also focuses on presenting the derived accuracy metrics of model performance, namely the precision, recall, and F1-scores. To validate the performance and to make sure that the model is not overfitting, cross-validation is a valuable procedure to examine. Specifically, k-fold cross-validation is used to ensure that the models generalise well enough across different sets of data from the original dataset. This step splits the data into k parts, where each part's model performance is evaluated. With k = 5, these steps are repeated five times. For each iteration, one subset of the data is used for validation, while the other four subsets are used to train the model. This emerged as a good estimator for model validation, since it examines how consistent model performance is across different parts of the data (Berthold et al., 2020). Although since the process needs to be repeated five times, it is also a challenge as it increases the runtime through repeating the validation process. Usually, cross-validation is employed on an untrained model to determine the best model. This is done before the final model is trained on the whole dataset. Here, a 5-fold cross-validation was tested on all features, including the VAD scores, using logistic regression and a Support Vector Machine (Table 4).

Fold	1	2	3	4	5
LR	0.761	0.777	0.763	0.757	0.729
LR (w. VAD)	0.773	0.786	0.77	0.765	0.735
SVM	0.729	0.745	0.733	0.721	0.689
SVM (w. VAD)	0.732	0.751	0.736	0.73	0.694

Table 4. *Cross-Validation Accuracy Scores across five folds for Logistic Regression (LR) and Support Vector Machine (SVM), both with and without Valence, Arousal, and Dominance (VAD) features.*

After validating the potential results, both models seem to perform well across five test data subsets. The results for both models will be shown individually. First, the overall results will be examined, followed by the confusion matrices for both models, and the most important features that each model used to accurately predict the authors will be presented.

The confusion matrix is a popular tool to describe classification errors. Basically, it is a table representation showing the true and predicted classes for each author. Each table entry

then shows the entries from a given class that the model classified to that specific class. The correct classifications are shown in the diagonal entries (cf. Cady, 2017; Berthold et al., 2020). For the main evaluation, the logistic regression model is trained twice, once only with stylometric features, and once with valence, arousal, and dominance scores included. For the first evaluation, it produced an F1-score of 0.78, meaning a 78% accuracy rate across the whole dataset, with precision and recall scores also settling at 0.78. The scores are computed on 5.000 support documents (blog posts), meaning 200 per author. Table 5 presents the complete results for this training step, including precision, recall, and F1-scores for each author. All result tables were generated with the classification report function, the built-in *scikit-learn* function to evaluate the results of a model training.

Author	**Precision**	**Recall**	**F1-Score**	**Support**
baileysbeerblog	0.95	0.96	0.96	200
bluehousejournal	0.61	0.61	0.61	200
boomergirlsguide	0.7	0.77	0.73	200
cashonlyliving	0.82	0.8	0.81	200
christasrandomthoughts	0.74	0.76	0.75	200
crpgaddict	0.99	0.97	0.98	200
culturalsnow	0.79	0.77	0.78	200
cuponthebus	0.81	0.79	0.8	200
danabugseyeview	0.73	0.76	0.75	200
dorcassmucker	0.8	0.74	0.77	200
fatroland	0.75	0.73	0.74	200
frikosmusings	0.72	0.7	0.71	200
frugalistanbul	0.75	0.77	0.76	200
glassincarnate	0.86	0.81	0.83	200
interimarrangements	0.56	0.51	0.53	200
kimmy-cookingpleasure	0.99	0.99	0.99	200
lifeatgoldenpines	0.84	0.84	0.84	200
lindahoof	0.83	0.81	0.82	200
momsscribbles	0.67	0.7	0.69	200
mumssimplylivingblogat	0.67	0.69	0.68	200
myheartisalwayshome	0.82	0.84	0.83	200
newamusements	0.65	0.74	0.7	200
stitchesandseams	0.81	0.81	0.81	200
thriftathome	0.7	0.69	0.69	200
understandingsociety	0.98	0.96	0.97	200
Accuracy			0.78	5000
Macro Avg	0.78	0.78	0.78	5000
Weighted Avg	0.78	0.78	0.78	5000

Table 5. *Results for Logistic Regression*

Next, including the valence, arousal, and dominance scores led to a slight accuracy improvement of 1% (Table 6). After training the model on this data, the F1-score settled at 0.79 (with precision and recall scores producing 0.79, respectively).

Authors	**Precision**	**Recall**	**F1-Score**	**Support**
baileysbeerblog	0.95	0.96	0.96	200
bluehousejournal	0.59	0.63	0.61	200
boomergirlsguide	0.72	0.75	0.74	200
cashonlyliving	0.81	0.82	0.81	200
christasrandomthoughts	0.76	0.77	0.76	200
crpgaddict	0.99	0.97	0.98	200
culturalsnow	0.81	0.77	0.79	200
cuponthebus	0.82	0.81	0.82	200
danabugseyeview	0.73	0.78	0.75	200
dorcassmucker	0.79	0.74	0.77	200
fatroland	0.75	0.74	0.75	200
frikosmusings	0.74	0.71	0.73	200
frugalistanbul	0.76	0.78	0.77	200
glassincarnate	0.87	0.83	0.85	200
interimarrangements	0.59	0.54	0.56	200
kimmy-cookingpleasure	0.99	0.99	0.99	200
lifeatgoldenpines	0.82	0.84	0.83	200
lindahoof	0.84	0.82	0.83	200
momsscribbles	0.72	0.69	0.71	200
mumssimplylivingblogat	0.67	0.72	0.7	200
myheartisalwayshome	0.83	0.83	0.83	200
newamusements	0.65	0.71	0.68	200
stitchesandseams	0.84	0.8	0.82	200
thriftathome	0.7	0.72	0.71	200
understandingsociety	0.98	0.96	0.97	200
Accuracy			0.79	5000
Macro Avg	0.79	0.79	0.79	5000
Weighted Avg	0.79	0.79	0.79	5000

Table 6. *Results for Logistic Regression (including VAD Scores)*

When comparing the results, adding VAD scores led to slightly improved accuracy, precision, recall, and F1-score. This is consistent across most authors. Authors who were already classified well (e.g., *crpgaddict*) did not improve much. In contrast, other author labels were classified better by roughly one to three per cent when checking the individual F1-scores. Still, these results suggest that the emotional features provided useful values to differentiate blog authors.

We can examine the confusion matrix based on these model results to visualise these results for multi-class classification, including the VAD scores (Figure 9).

Each row represents the instances of the actual class (authors), while each column represents the instances of a prediction class, such as how often the model predicted a specific class. As mentioned, we can evaluate the number of correctly predicted cases for each class when looking at the diagonal elements. Then, the other elements represent the falsely predicted cases.

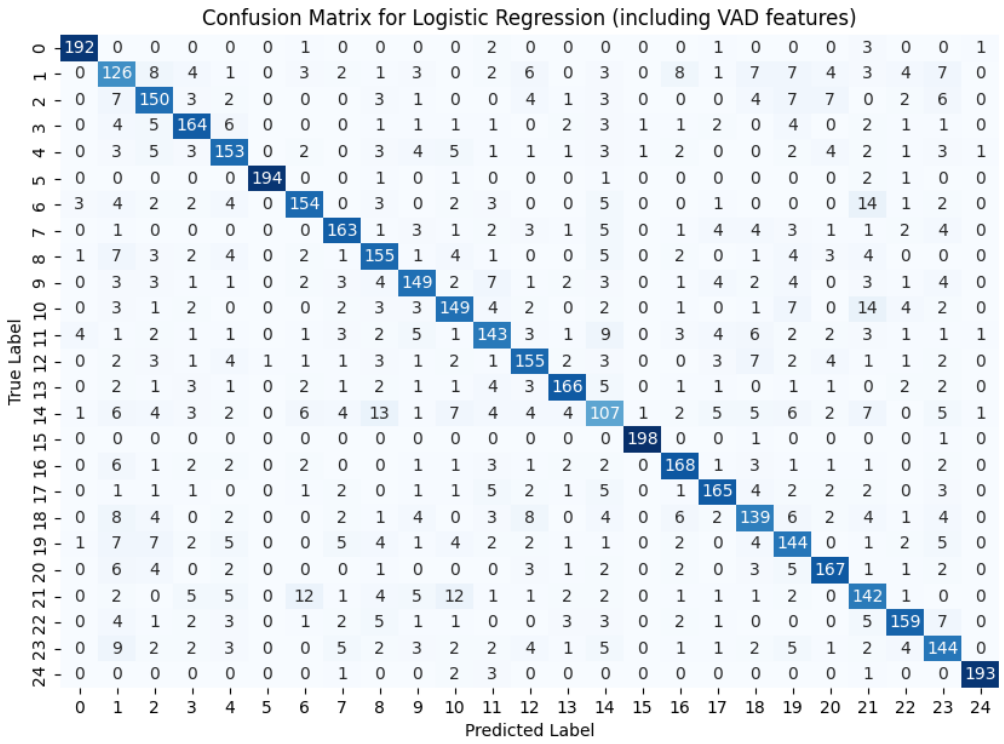

Figure 9. *Confusion Matrix for Logistic Regression (including VAD Scores).*

With the computed results in mind, we can examine the most important features that the model used to evaluate the labels. These importance scores can be evaluated after the model is trained since the trained model contains coefficients that are saved and represent the feature's importance in predicting the target label. Based on the coefficient score, the function then calculates the importance of the feature. Thus, features with higher coefficients are considered more important. To accurately get the column names, these names need to be passed together

with the feature matrix since these features were scaled, and thus, they were compiled in an array without the original column names. It is a necessary step to pass the scaled features, though, since those were used to train the model, and the computed coefficients correspond to these features. The most important features are displayed in Figure 10.

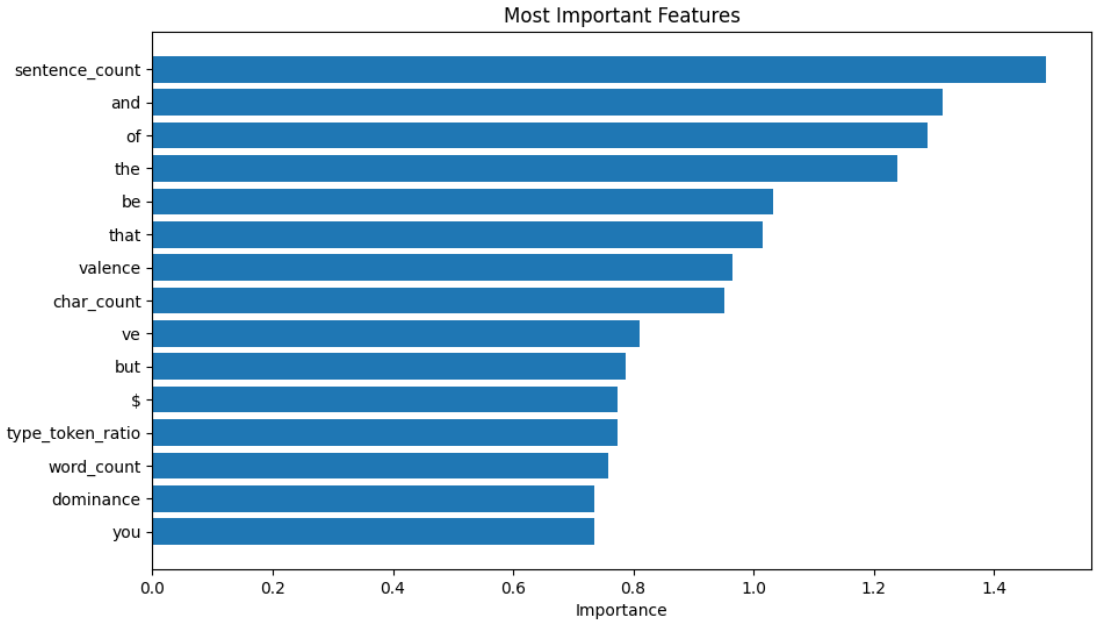

Figure 10. *Most important features for Logistic Regression (including VAD Scores).*

Judging from the most important features, we can see that sentence_count is considered the most important feature, with some other features coming close. Mainly, some of the computed word frequencies of common words such as 'and,' 'of,' and 'the.' However, other stylometric scores like the char_count and the word_count are also part of the top 15. As such, it seems plausible that many of the other features which the model was trained on and which are not displayed in the top 15 are other word frequencies for less important words.

What is interesting is that valence and dominance, two of the emotion features, apparently had a considerable impact on the model training. Both are among the most important features, ranking 7th and 14th, respectively. Only arousal is missing from this graph. Nevertheless, it seems that the emotional scores made a noticeable impact among the most important features.

These tables include training the model with all the aforementioned specifications, which stayed the same throughout the training phase. The same goes for the training process of the Support Vector Machine model, which will be evaluated next.

The Support Vector Machine was trained with the same two underlying datasets. The first run reached a F1-score of 0.75, with precision and recall scores also settled at 0.75 over 5.000 support documents (Table 7).

Author	**Precision**	**Recall**	**F1-Score**	**Support**
baileysbeerblog	0.93	0.95	0.94	200
bluehousejournal	0.55	0.65	0.6	200
boomergirlsguide	0.63	0.73	0.68	200
cashonlyliving	0.74	0.77	0.75	200
christasrandomthoughts	0.76	0.73	0.75	200
crpgaddict	0.98	0.96	0.97	200
culturalsnow	0.8	0.74	0.77	200
cuponthebus	0.73	0.79	0.76	200
danabugseyeview	0.72	0.72	0.72	200
dorcassmucker	0.68	0.74	0.71	200
fatroland	0.71	0.71	0.71	200
frikosmusings	0.68	0.67	0.67	200
frugalistanbul	0.73	0.71	0.72	200
glassincarnate	0.77	0.76	0.76	200
interimarrangements	0.52	0.51	0.52	200
kimmy-cookingpleasure	0.98	0.98	0.98	200
lifeatgoldenpines	0.81	0.8	0.8	200
lindahoof	0.8	0.72	0.76	200
momsscribbles	0.66	0.68	0.67	200
mumssimplylivingblogat	0.6	0.62	0.61	200
myheartisalwayshome	0.86	0.73	0.79	200
newamusements	0.63	0.67	0.65	200
stitchesandseams	0.82	0.77	0.8	200
thriftathome	0.72	0.61	0.66	200
understandingsociety	0.98	0.96	0.97	200
Accuracy			0.75	5000
Macro Avg	0.75	0.75	0.75	5000
Weighted Avg	0.75	0.75	0.75	5000

Table 7. *Results for the Support Vector Machine*

When VAD scores are introduced to the dataset, all three scores (F1, recall, and precision) increase by one per cent to 0.76 (Table 8).

Author	Precision	Recall	F1-Score	Support
baileysbeerblog	0.93	0.96	0.95	200
bluehousejournal	0.56	0.69	0.62	200
boomergirlsguide	0.63	0.76	0.69	200
cashonlyliving	0.74	0.8	0.77	200
christasrandomthoughts	0.82	0.74	0.78	200
crpgaddict	0.98	0.95	0.97	200
culturalsnow	0.79	0.72	0.76	200
cuponthebus	0.71	0.77	0.74	200
danabugseyeview	0.7	0.71	0.71	200
dorcassmucker	0.7	0.76	0.73	200
fatroland	0.73	0.73	0.73	200
frikosmusings	0.7	0.68	0.69	200
frugalistanbul	0.71	0.71	0.71	200
glassincarnate	0.79	0.72	0.76	200
interimarrangements	0.53	0.52	0.52	200
kimmy-cookingpleasure	0.99	0.99	0.99	200
lifeatgoldenpines	0.82	0.81	0.81	200
lindahoof	0.82	0.72	0.77	200
momsscribbles	0.68	0.69	0.68	200
mumssimplylivingblogat	0.64	0.68	0.66	200
myheartisalwayshome	0.89	0.77	0.82	200
newamusements	0.62	0.69	0.65	200
stitchesandseams	0.83	0.75	0.79	200
thriftathome	0.74	0.61	0.67	200
understandingsociety	0.97	0.96	0.97	200
Accuracy			0.76	5000
Macro Avg	0.76	0.76	0.76	5000
Weighted Avg	0.76	0.76	0.76	5000

Table 8. *Results for the Support Vector Machine (including VAD Scores)*

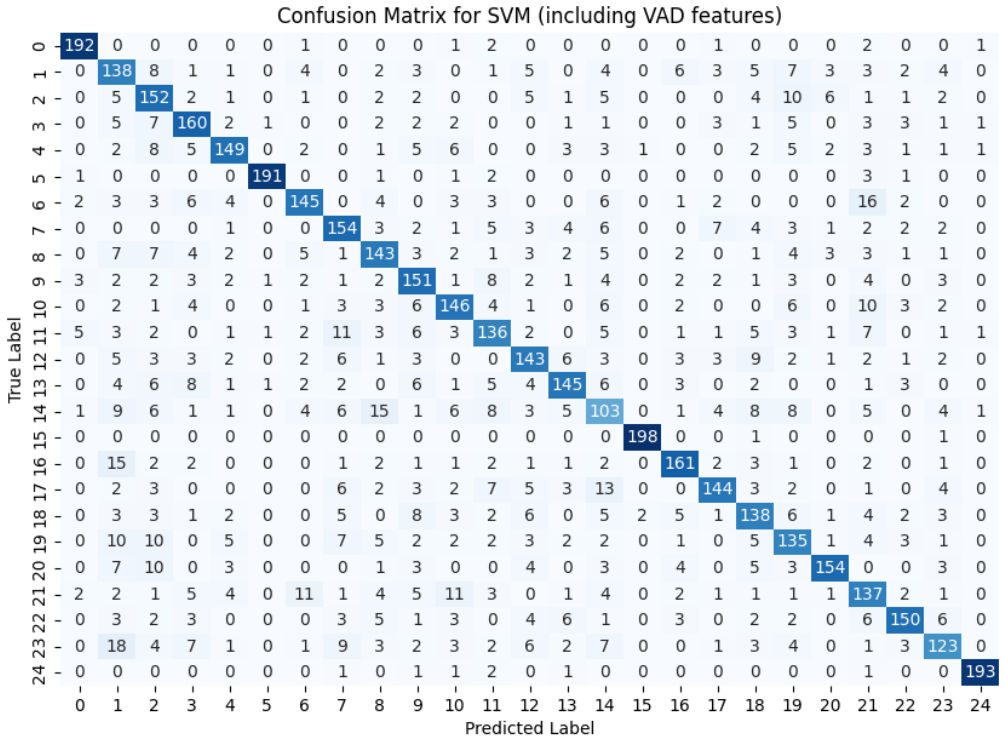

Figure 11. *Confusion Matrix for the Support Vector Machine (including VAD Scores).*

The results can again be visualised as a confusion matrix (Figure 11). Again, there is a slight improvement in accuracy when introducing VAD scores to the dataset with the Support Vector Machine. Most individual authors are classified better by one to two per cent, with outliers achieving up to five per cent (‘mumssimplylivingbotat’). The classification of some authors became worse, although overall, the model performs slightly better with the second set of features.

Considering the feature importance, unlike the logistic regression model, the Support Vector Machine usually does not have explicit feature importance scores. This is due to the support vectors created for decision boundaries instead of feature weights.

However, one advantage of using a linear kernel for the SVM is that it computes a weight vector, which can be interpreted similarly to the most important feature scores for the logistic regression model. Those scenes can again be visualised (Figure 12). This approach uses the weights of the trained SVM instead of decision boundaries.

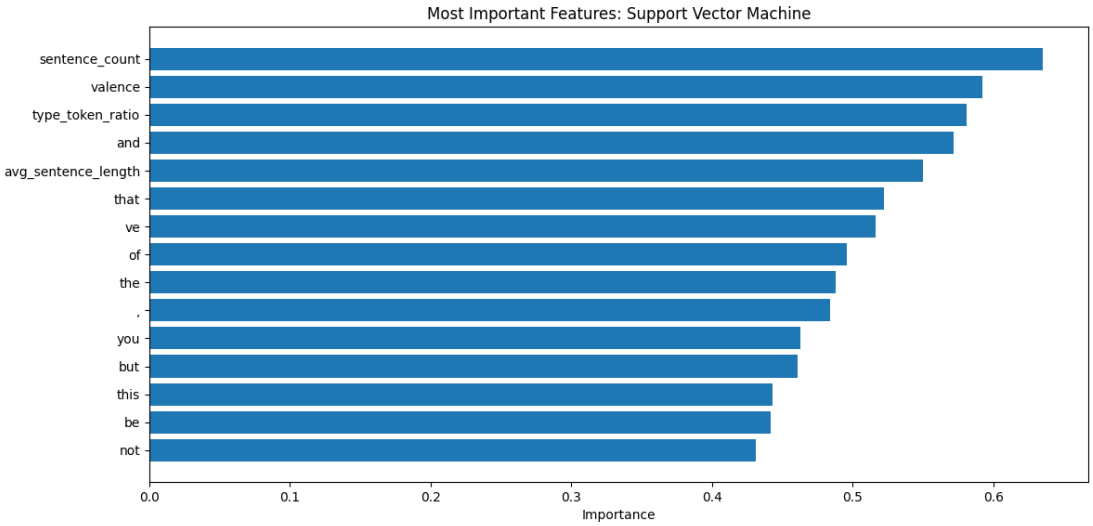

Figure 12. *Most important features for the Support Vector Machine (including VAD Scores).*

Again, the table shows that sentence count (sentence_count) is the most important feature for distinguishing between the authors. Type-token-ratio (type_token_ratio) and the average sentence length (avg_sentence_length), as averaged stylometric features, are also highly rated, together with some word frequencies for common stopwords. However, valence is again an important metric for the model, scoring second place among the features, thus making it even more relevant than it was for the first model. While arousal and dominance features are missing from those important features, valence seems to make a noticeable impact again, as it did when looking at the most important features of the logistic regression model. Apart from the VAD scores, stylometric features also score high across both model implementations.

Across both models, the authors who showed improvements might have a more emotional writing style, even though the improvements are only minor. This could also indicate the other extreme: the blogs do not show a distinct enough emotional style across their texts for classification to improve solely based on these metrics. While emotional features helped to improve classification for both models, the effects were relatively small.

So, integrating the additional VAD scores into the dataset increases performance for both models, albeit small, but consistently.

5. Discussion

This section will focus on discussing and reflecting on the main results and the implementation of the original idea. Further, the research questions are addressed in the context of the results, the implemented programming pipeline and the underlying dataset.

Coming from the enormous size of the original dataset, it appears to have been a reasonable approach to reduce the size to 25 authors, with a total of 25.000 texts in the dataset used for this thesis. This is still a significant number in terms of possible candidate authors. By capping the number of texts per author to 1.000 each, it was guaranteed that a balanced dataset was used to have an equal ratio of texts (blog posts) per author for model training and testing. Further, the words per author exceeded 100.000 across the whole dataset, even after preprocessing and reducing the dataset size. For authorship attribution, this seems to be a generous number and a large enough sample size from which to draw conclusions. Other popular datasets used by researchers often fall far below this threshold number. For example, the collection of publicly available datasets for stylometry, which Neal et al. (2018) summarised, showed that most authors were represented by a few thousand words in those datasets. The exact number of words which are necessary to derive authorship from seems to be disputed, with some researchers suggesting that 6.500 words are enough to get adequate results from (Rao and Rohatgi, 2000) while other authors talk about more significant numbers which are needed (Eder, 2015). Other research datasets are more extensive, but as already discussed earlier, the authorship attribution studies focusing on multiple authors were also mostly limiting their research to a small selection of authors, whereas few studies exceeded the number to more than a few dozen, with the accuracy dropping significantly when including more authors. So, the dataset for this thesis seems to strike an adequate middle ground from which to draw appropriate conclusions, while also being on the larger side compared with other datasets used for authorship attribution.

Having tested the complete dataset at first, reducing the dataset to an appropriate size also significantly improved training times, specifically for the Support Vector Machine. Leaving out the many authors in the original corpus who produced fewer than 1.000 texts was the right choice since the dataset would have become too unbalanced otherwise. After balancing the dataset, the computational costs improved, and the training and computing time were considerably faster. Even though class imbalance can be somewhat circumvented by using

specific models[12] that can, to some extent, deal with unbalanced data, these specific models were not needed then, especially since it would have been another question if these models had led to a higher accuracy after all.

By comparing 25 authors, an accurate overview was still possible. In comparison, with the full dataset of over 100 authors, this would have been difficult to achieve, and introducing more authors might have convoluted the interpretation of the results. This would have also led to a lower accuracy score, even without the computational limitations, judging by the earlier-mentioned research, which compared authorship attribution studies with more authors, finding that it gets more complicated as more possible authors are introduced (Abazari et al., 2023). With this approach, it was still possible not to make the study too expensive in terms of computational resources. On the contrary, the scope was also somewhat limited by that factor and the circumstances of the machine on which the programming part of this thesis was constructed. As such, the corpus size was acceptable, and the amount of data was adequate for the scope of this thesis, even being on the larger end compared with the other mentioned datasets for authorship attribution (Neal et al., 2018). Each author had at least 150.000-word tokens before the preprocessing phase – and in many cases, many more – in their text corpus to extract robust stylometric features.

The preprocessing pipeline cleaned the text of the most unnecessary text fragments while still preserving those which are essential for stylometry. Also, it is noteworthy that the inclusion of stopwords was an impactful choice, with many stopwords appearing in the lists of the most important features across both models (Figure 10; Figure 12). While many other studies implement a stricter stance of preprocessing, focusing on stylometry gives certain limitations in that regard, explicitly concerning text capitalisation and punctuation, which are needed for stylometry but perhaps not for other forms of text analysis in natural language processing applications. So, many stylometric features could not have been extracted correctly without these careful preprocessing steps. Still, after analysing the results, some noise fragments were found in the data. For example, when looking at the models' most prominent features for training, one of those was the most common word ‘ve,’ which is not a word in itself but rather the shortened, contracted form of ‘have’ in the perfect tense. Typically, this should have been cleaned. This specific example, though, seemed to have been helpful for the model

[12] For example, a specific naïve Bayes classifier was implemented during the testing phase, which could handle class imbalances specifically well and produced decent results. Still, the imbalanced data remained a concern since the differences between authors were significant based on their text output.

training. It might even make sense to include this contracted form for authorship attribution. Some authors might use it while others do not (e.g. some might write ‘could have gone’ and others ‘could’ve gone’), so they might be differentiated by the model, realising that authors use these forms differently. This seems to align with the form of writing used in blogs, which can specifically be considered as it is often informal, and thus, the writing style of many authors might be more prone to using shortened word forms. In short, the context of the platform authors write on and the topics they write about influence the writing style (Overdorf and Greenstadt, 2016). So, while the preprocessing pipeline worked, there is still room for improvement and refinement depending on what preprocessing steps specifically to implement and what to exclude for the desired research questions.

Regarding the word frequency scores in this thesis, all frequencies were extracted per text in the corpus. After preprocessing the dataset, all authors had the same number of texts in the dataset, although the lengths of the individual texts vary. This variation, however, was thought of as a potential stylistic factor by itself, as text lengths may reveal an author’s writing tendencies and preferences, and the models learn from the individual writing instances. This is why the raw word frequencies were retained in the feature set without normalising them, as they are computed on a text basis, and not on the whole text corpus per author altogether. This means the models still have access to a thousand writing instances for each author. So, the design of the dataset also played a crucial role in this argument, as a fixed number of texts per author ensured that all authors were represented equally, with a minimum word token count, even though the corpus size differed, although much less than in the original unprocessed dataset (Table 1). So, the dataset is balanced in terms of the number of texts per author. Still, the raw frequency scores can correlate with text length. To address that, the raw frequency scores and their potential bias were combined by including several normalised and averaged features, such as the type-token ratio or the average sentence length. Through this, stylometric features and their variations were captured independently of text length, and the feature set combines raw and normalised features. Noteworthy is that the feature importance results (Figure 10; Figure 12) show that the normalised features rank close behind the text-length-based features. This suggests that the models do not rely on text-length dependent features alone, and instead, they utilise the combination of all features to form predictions accurately.

While there are more stylometric markers that can be extracted from texts, this thesis aimed to inspect the influence of emotional markers in texts and whether they improve authorship attribution results, so focusing on just some of the possible stylometric features was

sufficient since they already produced good results. In addition, the chosen stylometric features align with some of the most used ones for decades, returning to the earlier works on stylometry and the computational possibilities that can extract those features (Holmes, 1994; 1998). Following up on that thought, it might have been counterproductive even to extract more stylometric measures since the emotional language markers just produced three additional features for the model training. Thus, with other markers, the chances might have been higher that these features would have overshadowed the emotion features. Even when reducing the feature sizes with specific scalers (such as the used standard scaler), which focus on the top features for model training, chances are that by using these functions, the emotional language features might have just been excluded entirely. As we have seen, both models excluded arousal from the most important features. So, this feature would probably not have been chosen for model training in the first place when reducing the features further.

Regarding the results, both models performed reasonably well, producing accuracy scores of 79% and 76%, respectively, with included VAD scores. Both models proved to function well enough with the used features and could attribute authorship well, especially compared to other tested models. Further, logistic regression and the SVM both provide consistent results with all three metrics (F1-score, precision, and recall), showing improvements. Other models, like Naïve Bayes and a Random Forest implementation, were tried for the initial testing phase. All other models had worse classification results and high computational costs, even when training them on the reduced dataset.

While both models produced acceptable results, logistic regression scored slightly better across both variations. This aligns with other studies comparing SVM and logistic regression (as well as, for example, the Naïve Bayes approach during the testing phase, which performed worse than the two used models) (Zhang and Oles, 2001). For example, Overdorf and Greenstadt (2016) note, "Generally speaking, for linearly separable problems, regularized linear classifiers will perform similarly. Our domain results and those in previous studies, show that stylometry is a linearly separable problem on which linear classifiers perform well." Further, they also note that logistic regression has an edge in performance and "[…] logistic regression provides estimates of the posterior probabilities for the classes that are not entirely based on the discriminant function, as the estimates provided by an SVM are. Our methods which make use of ensembles depend on these probabilities and there is reason to suspect that they might be (slightly) more accurate in the logistic regression case" (Overdorf and Greenstadt 2016, 161).

When looking at the results specifically, the integration of valence, arousal, and dominance led to a slight increase of 1% across both models. Across all author performance scores, a slight but noteworthy improvement is visible, for most authors around 1% to 3%. These results might suggest that these authors have a more unique emotional style. When looking at individual rating jumps (the SVM classification had the highest individual accuracy jump of 5%), these can be explained by the fact that authors with less distinctive writing styles profit the most when including VAD scores, for example, if their texts are better classifiable by their emotional writing style. This is further underlined when looking at the other most improved classes after including VAD scores. These classes initially had modest performances on stylometric features alone. In contrast, a few authors' results did not improve or were slightly negative. For these authors, either their writing style could not benefit from emotion scores directly, or the other authors had a more emotion-driven style, which led to the model being more accurate based on their specific VAD scores across their texts. Generally, classes with roughly the same rating distribution might already be well classified because of their distinct stylometric writing patterns, making additional emotional language markers redundant. So, emotional language is more important for some authors than others when attributing authorship.

One thing that could be considered in that case is that the topics the authors write about could give additional insights at that specific point. Since the context influences the writing style (Overdorf and Greenstadt, 2016), it could be the case that some blog authors write more enthusiastically about their topic of choice, for example, when their blog is about things they enjoy. Thus, their writing style might be more emotion-driven. This might be a potential challenge with the given dataset, which bridges multiple topics due to the diverse nature of the collected blogs. As mentioned, authorship attribution becomes a challenge when multiple topics or genres are mixed (Stamatatos, 2009). This thesis had to tackle this, for example, with the chosen stylometric features, which try to avoid topic-specific information so that they are more reliable when faced with texts about multiple topics (Stamatatos, 2009). Cross-domain authorship attribution is also a challenge for accurate authorship attribution, considering that different domains have different requirements for text (e.g., short texts on social media platforms versus long blog posts) (Overdorf and Greenstadt, 2016). However, this problem of cross-domain authorship attribution is circumvented by the original dataset comprising only blog posts from the same blogging platform, albeit scattered across different blogs.

Still, theoretically, the integration of emotion scores improved both model performances, even though the improvements are minor. Further, it is difficult to tell if the

improvement is precisely because of emotion scores and not rather because the models have more features to classify authors more accurately. Earlier testing has shown that until a certain threshold, more features led to better results. However, one indicator that the VAD scores are helpful and not simply provide more features is that some of them have been listed as influential features for both models, even ranking high up in the list, with valence scoring high for both models, even being on the second place for the Support Vector Machine (Figure 10; Figure 12). So, even when assuming that these are just more features for the models, they seem to have been impactful. The results also provide a decent improvement from another angle, considering that this authorship attribution task considers many different classes with the 25 authors in the dataset. That makes the results non-trivial compared to other research, showing that the tasks get more challenging as more authors are considered, and not many authorship attribution studies include that many possible candidate authors in their research (Abazari et al., 2023).

If the goal is to build a model for authorship attribution, aiming to achieve the highest accuracy scores possible, then the chosen method might not be the most promising. In that case, it seems reasonable to argue that a model based on a few stylometric features alone might not be the best approach. Even when incorporating emotion scores, state-of-the-art approaches outperform the model and deliver higher accuracy scores (Abbasi & Chen, 2008). Still, it is difficult to compare studies ranging across different methodologies and different datasets.

It might be a worthwhile idea to expand the original research with other features focusing on emotion and authorship attribution based on the promising results of this initial thesis to see which other features and analysis options might emerge providing better results, let alone for the reasoning that the current research evaluating these two topics together is currently limited. Other research could overhaul the current pipeline and try out other specifications, from incorporating more stylometric features to a different take on attributing VAD scores to the texts, fine-tune the machine learning models further, or, perhaps, try out other models that might function more effectively, such as deep learning methods. Other emotional theories, such as those described in the Emotion Analysis chapter of the literature review, could be considered apart from VAD scores. The existing pipeline could also be tested on different text genres to see how the results are influenced if the dataset contains not blogs but historical literary texts or academic writing. This might also yield interesting results regarding the interpretation of the use of emotional language by authors and how this might differ when the genres are different. Martins et al. (2018) showed that these emotion scores also differ when authors come from different backgrounds, as their study showed divergent uses of emotional

language between public-life personas and politicians. This could be interconnected with other measures, such as the research by Koppel et al. (2002) focusing on deducing demographic information about authors from their texts. So, the current approach offers many possible avenues for future research to explore, while laying the groundwork with the undertaken approach.

6. Limitations

Several potential limitations exist regarding the project in this thesis. These will be briefly discussed here. However, the mentioned points are not exhaustive, and other possible limitations remain to be considered.

First, there are some basic limitations to the underlying dataset. While it provides enough data, with many authors and each of these authors having a plethora of available texts, it is a dataset containing texts from online blogs. As mentioned, these are often personal and informal in nature, while covering several areas of interest or potential recreational activities undertaken by the specific author. Thus, on average, they might contain a mostly informal and personal writing style, which might differ from other writing styles in literature, academic writing, or political speeches. Thus, the results may vary when inspecting texts from these areas, specifically when considering that emotional scores also might differ depending on what areas authors write about. That might lead to worse results when just looking at emotional features in text genres deemed more formal, or that follow a fixed structure like texts written for academic purposes. However, this is just a reasonable assumption which could be tested in future research.

Secondly, stylometric features represent an asset and a limitation. Since so many vastly different stylometric features can be examined (Rudman, 1998), there are always other potential features to look at and analyse. Also, the blogs were written in a timeframe of roughly two decades. This is not necessarily accounted for by simply extracting stylometric measures when considering that an author's writing style might evolve in several years and change, especially when looking at such a long timeframe. Still, suppose more stylometric features had been introduced to the pipeline. In that case, they might have convoluted the dataset to a point where the emotion features might no longer have impacted the accuracy of the results, since the number of other features would have even further devalued them. This also remains the case for the word frequencies. Currently, the top 100 most frequent words are analysed. Possibly, this number could also be increased while adding more features to the DataFrame. So, while adding more stylometric features to the dataset is possible, it is not certain if these would improve the results. At a certain point, that undertaking would have exceeded the scope of the thesis, with the research question in mind whether emotional features can improve stylometric authorship attribution. Also, even though the accuracy of the results could be improved in this thesis, there might still be other ways of approaching the problem from a purely results-driven perspective, which might provide more accurate and better results.

Regarding the valence, arousal, and dominance scores, the aggregation method used in this thesis might also have its limitations, since a better representation might have led to better results. Ultimately, those scores represent three features per text across nearly 160 other features. Also, not all words in the dataset could be linked to accurate VAD scores, so that some emotional expressions might have gone missing during that phase. Also, other emotional markers in the text could be considered, such as an author's use of question- and exclamation marks, emoticons, or specific words written in all caps. This would inject other emotional features into the pipeline. This is a possibility that further research could explore as well. It remains to be seen whether these measures would impact the results. One notable restriction regarding the computational process of evaluating VAD scores is how words were handled, which could not be attributed to the computed dictionary because they were either missing in the dictionary, or because the preprocessing pipeline could not attribute these words correctly to their lemmatised versions found in the scores by Warriner et al. (2013). These were accounted for by attributing default values of (0.0, 0.0, 0.0) to them. This was initially done as a pragmatic choice to prevent missing values in the dataset and to maintain data consistency. However, the meaning of VAD scores might have been negatively influenced by that decision by introducing a potential bias, as these scores are not neutral and instead could imply low valence, low arousal, and low dominance. As such, these could slightly distort the aggregated emotional scores of a text. Still, this implementation ensures a uniform treatment of missing data across all texts, even though imperfect.

Lastly, the model constraints should be considered a potential limitation. On the one hand, the results could be increased through further feature engineering of the existing models, and on the other hand, other models might lead to better results altogether. Also, for example, deep learning or transformer-based models like BERT have not been tested in this thesis's scope and could also potentially perform well.

Overall, the results of the current pipeline are satisfactory. However, future research on this topic could still improve the results by implementing some of the limiting factors mentioned.

7. Conclusion

This thesis aimed to build and compare an authorship attribution model incorporating text features based on emotional language and to analyse whether these features improve the accuracy of correctly attributing unknown texts to their respective authors.

With this goal in mind, the central research question was formulated: Can the implementation of emotional language features improve the accuracy of authorship attribution methods?

To accomplish this goal, two pipelines for authorship attribution were implemented. The first was based on stylometric features, which were extracted from preprocessed blog posts drawn from a diverse dataset containing thousands of texts written by 25 authors. The second pipeline then built on this approach by also incorporating valence, arousal, and dominance scores as emotional features, which were derived from the dataset while comparing the words in the dataset to a word-lemma-based dictionary of almost 14.000 English words with their computed VAD scores developed by Warriner et al. (2013).

One decision was to balance the dataset with the original blog corpus being vastly imbalanced, with some authors having produced much more than others. During the initial testing, this led to a negative effect on the results. Furthermore, an imbalanced dataset increases the computational costs, meaning that the training of the models took significantly longer, even though models such as logistic regression theoretically have efficient training times. After balancing the dataset to a thousand texts per author, those problems vanished, and the results became more stable across both models, so that this seems to have been the correct decision.

Then, based on all extracted features for authorship attribution, two machine learning models, logistic regression and a Support Vector Machine, were trained. Both models produced good results, with logistic regression scoring slightly higher (79%) than the SVM (76%) with incorporated VAD scores. The baseline results only using stylometric features produced a slightly worse result for both implementations, so the incorporated emotion scores led to a slight performance increase in general and across most individual authors, even though the improvements are rather small.

Since the research goal was to incorporate emotion scores and stylometry, other authorship attribution methods could be neglected, even though it should be noted that there are implementations utilising other features which can yield higher accuracies, if accuracy is the only goal of the authorship attribution study.

Regarding the answer to the original research question, it seems reasonable to conclude that integrating VAD scores as emotional text features improves the model training results. However, as said, the performance does not increase significantly.

When talking about the most significant emotional features which potentially led to a performance increase, valence seemed to be most important for both machine learning models, even scoring as the second most important feature in the SVM training. Dominance was still included as one of the most important features for logistic regression as well. Other than that, stylometric features were important features across both models, with specific word frequencies also being highly important for both models.

Still, these initial results highlight the need for further research into the relationship between emotional language and authorship attribution techniques. This thesis contributes to the field by offering a functioning, modular pipeline to attribute authorship, combining traditional stylometric features with emotion-based metrics for attributing authorship. It documents the approach and makes it reproducible. By integrating these two approaches, it offers a novel perspective on the field, especially considering the sparse existing research on this particular niche of authorship attribution. Additionally, this study also provides a publicly available dataset of blog posts spanning roughly two decades and incorporating thousands of texts by many authors. As such, the dataset offers a vast resource for researchers and future research on this topic.

The findings suggest that emotional language can be a potentially important stylistic marker in authorship attribution, opening a new layer of text understanding and bridging the gap between purely stylistic or lexical features to a more nuanced understanding of an author's writing style, incorporating emotion. Building on current studies and the accessibility of emotional lexica and dictionaries, it is an accomplishable goal to integrate those research-backed scores into a programming pipeline for efficient authorship attribution. Following these directions offers rich potential to consider other approaches, incorporating emotional language with more potential emotional categories, or when considering other emotional and sentiment theories for future research.

References

Abazari, Farzaneh, Enrico Branca, Norah Ridley, Natalia Stakhanova and Mila Dalla Preda. "Dataset Characteristics for Reliable Code Authorship Attribution." *IEEE Transactions on Dependable and Secure Computing* 20 (2023): 506–521. https://doi.org/10.1109/TDSC.2021.3138700.

Abbasi, Ahmed, and Hsinchun Chen. "Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace." *ACM Transactions on Information Systems* 26, no. 2 (2008): 1–29. https://doi.org/10.1145/1344411.1344413.

Abbasi, Ahmed, Abdul Rehman Javed, Farkhund Iqbal, Zunera Jalil, Thippa Reddy Gadekallu, and Natalia Kryvinska. "Authorship Identification Using Ensemble Learning." *Scientific Reports* 12, no. 1 (2022): 9537. https://doi.org/10.1038/s41598-022-13690-4.

Al-Sarem, Mohammed, and Abdel-Hamid Emara. "The Effect of Training Set Size in Authorship Attribution: Application on Short Arabic Texts." *International Journal of Electrical and Computer Engineering (IJECE)* 9, no. 1 (2019): 652–659. https://doi.org/10.11591/ijece.v9i1.pp652-659.

Azarbonyad, Hosein, Mostafa Dehghani, Maarten Marx, and Jaap Kamps. "Time-Aware Authorship Attribution for Short Text Streams." In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2015): 727–30. https://doi.org/10.1145/2766462.2767799.

Balakrishnan, Vimala, and Lloyd-Yemoh Ethel. "Stemming and Lemmatization: A Comparison of Retrieval Performances." *Lecture Notes on Software Engineering* 2, no. 3 (2014): 262–267. https://doi.org/10.7763/LNSE.2014.V2.134.

Behrens, John T. "Principles and Procedures of Exploratory Data Analysis." *Psychological Methods* 2, no. 2 (1997): 131–160. http://dx.doi.org/10.1037/1082-989X.2.2.131.

Berthold, Michael R., Christian Borgelt, Frank Höppner, Frank Klawonn, and Rosaria Silipo. *Guide to Intelligent Data Science: How to Intelligently Make Use of Real Data*. Springer Cham, 2020. https://doi.org/10.1007/978-3-030-45574-3.

Bird, Steven, Edward Loper and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc, 2009.

Boyd, Ryan L. “Mental Profile Mapping: A Psychological Single-Candidate Authorship Attribution Method.” *PLOS ONE* 13, no. 7 (2018): e0200588. https://doi.org/10.1371/journal.pone.0200588.

Bradley, Margaret M, and Peter J. Lang. “Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings.” Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, 1999.

Cady, Field. *The Data Science Handbook*. John Wiley & Sons; Inc, 2017. https://doi.org/10.1002/9781119092919.

Cambria, Erik, Amir Hussain, Catherine Havasi, Chris Eckl. “SenticSpace: Visualizing Opinions and Sentiments in a Multi-dimensional Vector Space.” *Knowledge-based and Intelligent Information and Engineering Systems. KES 2010. Lecture Notes in Computer Science* 6279 (2010). https://doi.org/10.1007/978-3-642-15384-6_41.

Chaski, Carola E. “Empirical evaluations of language-based author identification.” *Forensic Linguistics* 8, no. 1 (2001): 1–65. https://doi.org/10.1558/sll.2001.8.1.1 .

Eder, Maciej. “Does Size Matter? Authorship Attribution, Small Samples, Big Problem.” *Digital Scholarship in the Humanities* 30, no. 2 (2015): 167–82. https://doi.org/10.1093/llc/fqt066.

Ekman, Paul. “An Argument for Basic Emotions.” *Cognition and Emotion* 6, no. 3–4 (1992): 169–200. http://dx.doi.org/10.1080/02699939208411068.

Ekman, Paul, and Daniel Cordaro. “What Is Meant by Calling Emotions Basic.” *Emotion Review* 3, no. 4 (2011): 364–70. https://doi.org/10.1177/1754073911410740.

Esuli, Andrea, and Fabrizio Sebastiani. “SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining.” *In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*. European Language Resources Association (ELRA), 2006.

Frantzeskou, Georgia, Efstathios Stamatatos, Stefanos Gritzalis, and Sokratis Katsikas. “Effective identification of source code authors using byte-level information.” *Proceedings of the 28th International Conference on Software Engineering* (2006): 893–896. https://doi.org/10.1145/1134285.1134445.

Gaston, Joshua, Mina Narayanan, Gerry Dozier, D. Lisa Cothran, Clarissa Arms-Chavez, Marcia Rossi, Michael C. King, and Jinsheng Xu. “Authorship Attribution via Evolutionary Hybridization of Sentiment Analysis, LIWC, and Topic Modeling Features.” In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 933–40. IEEE, 2018. https://doi.org/10.1109/SSCI.2018.8628647.

Geetha A.V., Mala T., Priyanka D., and Uma E. “Multimodal Emotion Recognition with Deep Learning: Advancements, Challenges, and Future Directions.” *Information Fusion* 105 (2024): 102218. https://doi.org/10.1016/j.inffus.2023.102218.

“Glossary of psychological terms,” A.P. Association, accessed April 17, 2025, https://dictionary.apa.org/emotion.

Graham, Neil, Graeme Hirst, and Bhaskara Marthi. “Segmenting documents by stylistic character.” *Journal of Natural Language Engineering* 11, no. 4 (2005): 397–415. https://doi.org/10.1017/S1351324905003694 .

Hammoud, Khodor, Salima Benbernou, and Mourad Ouziri. “A Sentiment-Based Author Verification Model Against Social Media Fraud.” *Atlantis Studies in Uncertainty Modelling* 3 (2021): 219–226. https://doi.org/10.2991/asum.k.210827.030.

Hassler, Marcus, and Günther Fliedl. “Text Preparation through Extended Tokenization.” In *Data Mining VII: Data, Text and Web Mining and Their Business Applications* 37 (2006): 13–21. https://doi.org/10.2495/DATA060021.

Holmes, David I. “Authorship Attribution.” Computers and the Humanities 28, no. 2 (1994): 87-106, accessed April 9, 2025, https://www.jstor.org/stable/30200315.

Holmes, David I. “The Evolution of Stylometry in Humanities Scholarship.” *Literary and Linguistic Computing* 13, no. 3 (1998): 111–17. https://doi.org/10.1093/llc/13.3.111.

Hunter, John D. “Matplotlib: A 2D Graphics Environment.” *Computing in Science & Engineering* 9, no. 3 (2007): 90–95. https://doi.org/10.1109/MCSE.2007.55.

Hurtado, Jose, Napat Taweewitchakreeya, and Xingquan Zhu. “Who Wrote This Paper? Learning for Authorship de-Identification Using Stylometric Features.” In *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI)* (2014): 859–62. https://doi.org/10.1109/IRI.2014.7051981.

Ivanov, Lubomir, and Felix Perez. “Authorship Attribution of English Poetry Using Sentiment Analysis.” *The International FLAIRS Conference Proceedings* 37 (2024). https://doi.org/10.32473/flairs.37.1.135272.

Juola, Patrick. “Ad-hoc authorship attribution competition.” In *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing* (2004): 175–176, accessed April 8, 2025, https://dh-abstracts.library.virginia.edu/works/365.

Juola, Patrick. “Authorship Attribution.” *Foundations and Trends in Information Retrieval* 1 (2008): 233–334. https://doi.org/10.1561/1500000005.

Juola, Patrick, George K. Mikros, and Sean Vinsick. “A comparative assessment of the difficulty of authorship attribution in Greek and in English.” *Journal of the Association for Information Science and Technology* 1, no. 70 (2018): 61-70. https://doi.org/10.1002/asi.24073.

Kalgutkar, Vaibhavi, Ratinder Kaur, Hugo Gonzalez, Natalia Stakhanova, and Alina Matyukhina. “Code Authorship Attribution: Methods and Challenges.” *ACM Computing Surveys* 52, no. 1 (2020): 1–36. https://doi.org/10.1145/3292577.

Koppel, Moshe. “Automatically categorizing written texts by author gender.” *Literary and Linguistic Computing* 17, no. 4 (2002): 401–412. https://doi.org/10.1093/llc/17.4.401.

Koppel, Moshe, Jonathan Schler, Elisheva Bonchek-Dokow. “Measuring Differentiability: Unmasking Pseudonymous Authors.” *Journal of Machine Learning Research* 8, no. 45 (2007): 1261-1276.

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. “Computational Methods in Authorship Attribution.” *Journal of the American Society for Information Science and Technology* 60, no. 1 (2009): 9-26. http://dx.doi.org/10.1002/asi.20961.

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. “Authorship attribution in the wild.” *Language Resources and Evaluation* 1, no. 45 (2010): 83–94. https://doi.org/10.1007/s10579-009-9111-2.

Kim Luyckx and Walter Daelemans. “Authorship attribution and verification with many authors and limited data.” In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING '08)* 1 (2008): 513–520. https://doi.org/10.3115/1599081.1599146.

Lange, Jens and Janis H. Zickfeld. “Emotions as overlapping causal networks of emotion components: implications and methodological approaches.” *Emotion Review* 13, no. 2 (2021): 157–167. https://doi.org/10.1177/1754073920988787.

Madigan, David, Alexander Genkin, David Lewis, Shlomo Engelson Argamon, Dmitriy Fradkin and Li Ye. “Author Identification on the Large Scale.” In *Proceedings 2005 Conference of the Classification Society of North America* (2005).

Mahmood, Asad, Zubair Shafiq, and Padmini Srinivasan. “A Girl Has A Name: Detecting Authorship Obfuscation.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics (2020): 2235–2245. https://doi.org/10.18653/v1/2020.acl-main.203.

Martins, Ricardo, José Almeida, Pedro Henriques, and Paulo Novais. “Increasing Authorship Identification Through Emotional Analysis.” *WorldCIST’18: Trends and Advances in Information Systems and Technologies*, In *Advances in Intelligent Systems and Computing* 745 (2018): 763-772. https://doi.org/10.1007/978-3-319-77703-0_76.

Martins, Ricardo, Pedro Henriques, and Paulo Novais. “Determining Emotional Profile Based on Microblogging Analysis.” In *Progress in Artificial Intelligence*. Lecture Notes in Computer Science 11805: 159–71. Springer Cham, 2019. https://doi.org/10.1007/978-3-030-30244-3_14.

Martins, Ricardo, Jose Joao Almeida, Pedro Henriques, and Paulo Novais. “A Sentiment Analysis Approach to Increase Authorship Identification.” *Special Sections: WorldCist18/Recent advances in Data Science and Systems* 38, no. 5 (2021). https://doi.org/10.1111/exsy.12469.

Marton, Yuval, Ning Wu, and Lisa Hellerstein. “On compression-based text classification.” In *Proceedings of the European Conference on Information Retrieval* (2005): 300–314. https://doi.org/10.1007/978-3-540-31865-1_22.

Mascol, Conrad. “Curves of Pauline and pseudo-pauline style i.” *Unitarian Review* 30 (1888): 539–546.

Mascol, Conrad. “Curves of Pauline and pseudo-pauline style ii.” *Unitarian Review* 30 (1888): 452–460.

McKinney, Wes. “Data Structures for Statistical Computing in Python.” In *Proceedings of the 9th Phyton in Science Conference* (2010): 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

Mendenhall, Thomas. “The characteristic curves of composition.” *Science* 214 (1887): 237–249.

Mehrabian, Albert, and James A. Russell. *An Approach to Environmental Psychology*. MIT Press, 1974.

Mehrabian, Albert. “Basic dimensions for a general psychological theory.” Oelgeschlager, Gunn & Hain, 1980: 39–53.

Meyer zu Eissen, Sven, Benno Stein, and Marion Kulig. “Plagiarism detection without reference collections.” *Advances in data analysis* (2007): 359–366. https://doi.org/10.1007/978-3-540-70981-7_40.

Mohammad, Saif M., and Peter D. Turney. “Crowdsourcing a Word-Emotion Association Lexicon.” Computational Intelligence 29, no. 3 (2013): 436–465. https://doi.org/10.48550/arXiv.1308.6297.

Moors, Agnes. “Appraisal Theory of Emotion.” In *Encyclopedia of Personality and Individual Differences*: 1–9. Springer, Cham, 2017. https://doi.org/10.1007/978-3-319-28099-8_493-1.

Mosteller, Frederick, and David Wallace. “Inference and disputed authorship.” *The Federalist*. Addison-Wesley, 1964.

Munezero, Myriam, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. “Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text.” *IEEE Transactions on Affective Computing* 5, no. 2 (2014): 101–11. https://doi.org/10.1109/TAFFC.2014.2317187.

Nandwani, Pansy, and Rupali Verma. “A Review on Sentiment Analysis and Emotion Detection from Text.” *Social Network Analysis and Mining* 11, no. 81 (2021). https://doi.org/10.1007/s13278-021-00776-6.

Narayanan, Mina, Joshua Gaston, Gerry Dozier, Lisa Cothran, Clarissa Arms-Chavez, Marcia Rossi, Michael C. King, and Kelvin Bryant. “Adversarial Authorship, Sentiment Analysis, and the AuthorWeb Zoo.” In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* (2018): 928–32. https://doi.org/10.1109/SSCI.2018.8628806.

Nazir, Shahzad, Muhammad Asif, Mariam Rehman, and Shahbaz Ahmad. “Machine Learning Based Framework for Fine-Grained Word Segmentation and Enhanced Text Normalization for Low Resourced Language.” *PeerJ Computer Science* 10 (2024). https://doi.org/10.7717/peerj-cs.1704.

Neal, Tempestt, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. “Surveying Stylometry Techniques and Applications.” *ACM Computing Surveys* 50, no. 6 (2018): 1–36. https://doi.org/10.1145/3132039.

Osgood, Charles E. “Dimensionality of the Semantic Space for Communication of facial applications.” *Scandinavian Journal of Psychology* 7, no. 1 (1966): 1–30. https://doi.org/10.1111/j.1467-9450.1966.tb01334.x.

Overdorf, Rebekah, and Rachel Greenstadt. “Blogs, Twitter Feeds, and Reddit Comments: Cross-Domain Authorship Attribution.” *Proceedings on Privacy Enhancing Technologies*, no. 3 (2016): 155–171. https://doi.org/10.1515/popets-2016-0021.

Pan, Bei, Kaoru Hirota, Zhiyang Jia, and Yaping Dai. “A Review of Multimodal Emotion Recognition from Datasets, Preprocessing, Features, and Fusion Methods.” *Neurocomputing* 561 (2023): 126866. https://doi.org/10.1016/j.neucom.2023.126866.

Petersohn, Devin, Stephen Macke, Doris Xin, William Ma, Doris Lee, Xiangxi Mo, Joseph E. Gonzalez, Joseph M. Hellerstein, Anthony D. Joseph, and Aditya Parameswaran. “Towards Scalable Dataframe Systems.” *Proceedings of the VLDB Endowment* 13, no. 12 (2020): 2033–2046. https://doi.org/10.14778/3407790.3407807.

Puig, Xavier, Martí Font, and Josep Ginebra. “A Unified Approach to Authorship Attribution and Verification.” *The American Statistician* 70, no. 3 (2016): 232–242. https://doi.org/10.1080/00031305.2016.1148630.

Rao, Jusyula R., and Pankaj Rohatgi. “Can pseudonymity really guarantee privacy?” In *Proceedings of the 9th Conference on USENIX Security Symposium* (2000).

Roseman, Ira J. “Appraisal in the Emotion System: Coherence in Strategies for Coping.” *Emotion Review* 5, no. 2 (2013): 141–49. https://doi.org/10.1177/1754073912469591.

Rubin, David C., and Jennifer M. Talarico. “A Comparison of Dimensional Models of Emotion: Evidence from Emotions, Prototypical Events, Autobiographical Memories, and Words.” *Memory* 17, no. 8 (2009): 802–8. https://doi.org/10.1080/09658210903130764.

Rude, Stephanie, Eva-Maria Gortner, and James Pennebaker. “Language use of depressed and depression-vulnerable college students.” *Cognition and Emotion* 18, no. 8 (2004): 1121–1133. https://doi.org/10.1080/02699930441000030.

Rudman, Joseph. “The state of authorship attribution studies: Some problems and solutions.” *Computers and the Humanities* 31 (1998): 351–365.

Sailunaz, Kashifa, and Reda Alhajj. “Emotion and Sentiment Analysis from Twitter Text.” *Journal of Computational Science* 36 (2019). https://doi.org/10.1016/j.jocs.2019.05.009.

Scherer, Klaus R. „Appraisal Theory.“ In *Handbook of Cognition and Emotion* (1999): 637–63. https://doi.org/10.1002/0470013494.ch30.

Scherer, Klaus R. “Emotions, Psychological Structure Of.” *International Encyyclopedia of the Social & Behavioral Sciences* (2001): 4472–4477. https://doi.org/10.1016/B0-08-043076-7/01711-3.

Shim, Ju-Hyun, and Lee Young-K. “Generalized Partially Linear Additive Models for Credit Scoring.” *Korean Journal of Applied Statistics* 24, no. 4 (2011): 587–595. https://doi.org/10.5351/KJAS.2011.24.4.587.

Sidorov, Grigori, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, Liliana Chanona-Hernández. “Syntactic Dependency-Based N-grams: More Evidence of Usefulness in Classification.” *In Computational Linguistics and Intelligent Text Processing. CICLing 2013. Lecture Notes in Computer Science* 7816 (2013): 13-24. https://doi.org/10.1007/978-3-642-37247-6_2.

Singh, Jyotika. *Natural Language Processing in the Real World: Text Processing, Analytics, and Classification*. Chapman and Hall/CRC, 2023. https://doi.org/10.1201/9781003264774.

Stamatatos, Efstathios. “Authorship Attribution Based on Feature Set Subspacing Ensembles.” *International Journal on Artificial Intelligence Tools* 15, no. 05 (2006): 823–38. https://doi.org/10.1142/S0218213006002965.

Stamatatos, Efstathios. “A survey of modern authorship attribution methods.” *Journal of the American Society for Information Science and Technology* 60, no. 3 (2009): 538-556. http://dx.doi.org/10.1002/asi.21001.

Stamatatos, Efstathios. “Authorship Verification: A Review of Recent Advances.” *Research in Computing Science* 123, no. 1 (2016): 9–25. https://doi.org/10.13053/rcs-123-1-1.

Statamatos, Efstathios. “Authorship Attribution Using Text Distortion.” In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* 1 (2017): 1138–1149. https://doi.org/10.18653/v1/e17-1107.

Strapparava, Carlo, and Alessandro Valitutti. “WordNet-Affect: An Affective Extension of WordNet.” In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04). European Language Resources Association (ELRA), 2004.

Somers, Harold, and Fiona Tweedie. “Authorship Attribution and Pastiche.” *Computers and the Humanities* 37 (2003): 407–429. http://dx.doi.org/10.1023/A:1025786724466.

Song, Wei, Chen Zhao, and Lizhen Liu. “Multi-Task Learning for Authorship Attribution via Topic Approximation and Competitive Attention.” *IEEE Access* 7 (2019): 177114–21. https://doi.org/10.1109/ACCESS.2019.2957152.

Stein, Benno, Sven Meyer zu Eissen. “Intrinsic Plagiarism Analysis with Meta Learning.” In *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection* (2007).

Stolerman, Ariel, Rebekah Overdorf, Sadia Afroz, and Rachel Greenstadt. “Breaking the Closed-World Assumption in Stylometric Authorship Attribution.” In *Advances in Digital Forensics X* (2014): 185-205. https://doi.org/10.1007/978-3-662-44952-3_13.

Taboada, Maite. “Sentiment Analysis: An Overview from Linguistics.” *Annual Review of Linguistics* 2, no. 1 (2016): 325–47. https://doi.org/10.1146/annurev-linguistics-011415-040518.

The pandas development team. 2024. pandas-dev/pandas: Pandas (v2.2.3). Zenodo. https://doi.org/10.5281/zenodo.13819579.

Warriner, Amy Beth, Victor Kuperman, and Marc Brysbaert. “Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas.” *Behavior Research Methods* 45, no. 4 (2013): 1191–1207. https://doi.org/10.3758/s13428-012-0314-x.

Waskom, Michael. “Seaborn: Statistical Data Visualization.” *Journal of Open Source Software* 6, no. 60 (2021): 3021. https://doi.org/10.21105/joss.03021.

Yang, Min and Kam-Pui Chow. “Authorship attribution for forensic investigation with thousands of authors.” *ICT Systems Security and Privacy Protection* (2014): 339-350. https://doi.org/10.1007/978-3-642-55415-5_28.

Zhang, Tong and Frank J. Oles. “Text Categorization Based on Regularized Linear Classification Methods.” *Information Retrieval Journal* 4, no. 1 (2001): 5–31. http://dx.doi.org/10.1023/A:1011441423217.

Zhang, Shaomin. *Authorship Analysis in Chinese Social Media Texts*. Cambridge University Press, 2024. https://doi.org/10.1017/9781009324298.

Zhou, Deyu, Zhikai Zhang, Min-Ling Zhang, and Yulan He. “Weakly Supervised POS Tagging without Disambiguation.” *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 17, no. 4 (2018): 1–19. https://doi.org/10.1145/3214707.

Appendix

A.1 Code and Data availability

The complete code and dataset used for this thesis are available via a public repository hosted on the University of Vienna's GitLab instance:
https://gitlab.phaidra.org/berensm23/authorship_attribution_vad_scores_thesis

A.2 List of Figures

A.3 List of Tables