South East European University
Faculty of Contemporary Sciences and Technologies
Doctoral Studies in Computer Sciences

Doctoral Dissertation Topic

Authorship Analysis in Albanian Texts

Mentor:

Prof. Dr. Arbana Kadriu

Author:

M. Sc. Arta Misini

Co-Mentor:

Assoc. Prof. Dr. Ercan Canhasi

Tetovo, 2024

Authorship Analysis in Albanian Texts

A thesis submitted to the Faculty of Contemporary Sciences and Technologies at South East European University (SEEU)

In partial fulfilment of the requirements for the degree of

Doctor of Science by

Arta Misini

Statement of authorship

I declare that this thesis, titled “Authorship Analysis in Albanian Texts” is my own work. In the parts where I have referred to the published work of others and quoted from the work of others, the source is always given.

M. Sc. Arta Misini

April, 2024

Publications

1. Arta Misini, Arbana Kadriu, and Ercan Canhasi. (2024). “Authorship classification techniques: Bridging textual domains and languages”. In: International Journal on Information Technologies and Security 16.1, pp. 27–38. doi: 10.59035/UKBE1226.

2. Arta Misini, Arbana Kadriu, and Ercan Canhasi. (2023). “A3C: Albanian Authorship Attribution Corpus”. In: Economic Recovery, Consolidation, and Sustainable Growth. Cham: Springer Nature Switzerland, pp. 755–763. doi: 10.1007/978-3-031-42511-0_49.

3. Arta Misini, Arbana Kadriu, and Ercan Canhasi. (2023). “Albanian Authorship Attribution Model”. In: 2023 12th Mediterranean Conference on Embedded Computing (MECO), pp. 1–5. doi: 10.1109/MECO58584.2023.10155046.

4. Arta Misini, Arbana Kadriu, and Ercan Canhasi. (2022). “A survey on authorship analysis tasks and techniques”. In: SEEU Review 17.2, pp. 153–167. doi: 10.2478/seeur-2022-0100.

Abstract

Authorship analysis is a natural language processing field that aims to identify distinct characteristics of authors in different textual contexts. Each natural language has specific attributes that indicate the uniqueness of each author's writing style. The increase of anonymous online content highlights the importance of analyzing authorship. It has applications in legal linguistics, criminal law, forensic investigations, and computer forensics.

This thesis makes a significant contribution to the analysis of Albanian authorship by focusing on three essential tasks: authorship identification, genre classification, and gender identification. The primary objective of this work is to address the lack of data in the Albanian language by compiling the first Albanian corpus for various tasks related to authorship. The corpus represents different aspects of the written language through different textual sources, such as literary writings, newsroom columns, and social media posts. We create an extensive set of features from different categories to capture the characteristics of the authors' writing style. Machine learning and deep learning techniques are used to discover patterns that indicate authorship, genre, and gender. Our approach uses two classification strategies to handle multi-class problems.

This study provides a comprehensive empirical analysis of authorship through a series of experiments. Experimental results are reported and further analyzed using learning curves, class prediction error, and feature analysis. We use feature selection and dimensionality reduction techniques to reduce the feature space and improve the model's effectiveness. With careful experimentation and analysis, this study enhances our understanding of authorship analysis in Albanian, providing valuable insights for further investigations in natural language processing and computational linguistics.

Keywords: authorship analysis · multi-class problem · cross-domain corpora · feature engineering · classification methods · author attribution · gender identification · genre classification · natural language processing · low-resource language.

Abstrakt

Analiza e autorësisë është një fushë e përpunimit të gjuhës natyrore që synon të identifikojë karakteristikat e dallueshme të autorëve në kontekste të ndryshme tekstuale. Secila gjuhë natyrore ka atribute specifike që tregojnë veçantinë e stilit të të shkruarit të secilit autor. Rritja e përmbajtjes anonime në internet nxjerr në pah rëndësinë e analizës së autorësisë. Ka aplikime në gjuhësinë juridike, të drejtën penale, hetimet mjeko-ligjore dhe forenzikën kompjuterike.

Kjo tezë jep një kontribut të rëndësishëm në analizën e autorëve në gjuhën shqipe duke u fokusuar në tre detyra thelbësore: identifikimi i autorësisë, klasifikimi i zhanrit dhe identifikimi i gjinisë. Objektivi parësor i kësaj pune është të trajtojë mungesën e të dhënave në gjuhën shqipe duke hartuar korpusin e parë në shqip për detyra të ndryshme të lidhura me autorësinë. Korpusi përfaqëson aspekte të ndryshme të gjuhës së shkruar përmes burimeve të ndryshme tekstuale, si shkrimet letrare, rubrikat e redaksisë dhe postimet në mediat sociale. Ne krijojmë një grup të gjerë të veçorive nga kategori të ndryshme për të kapur karakteristikat e stilit të të shkruarit të autorëve. Teknikat e të mësuarit të makinës dhe të mësuarit e thellë përdoren për të zbuluar modele që tregojnë autorësinë, zhanrin, dhe gjininë. Qasja jonë përdor dy strategji klasifikimi për të trajtuar problemet me shumë klasa.

Studimi jep një analizë gjithëpërfshirëse empirike të autorësisë përmes një sërë eksperimentesh. Rezultatet eksperimentale raportohen dhe analizohen më tej duke përdorur lakoret e të mësuarit, gabimet e parashikimit të klasës, dhe analizën e veçorive. Ne përdorim teknikat e përzgjedhjes së veçorive dhe reduktimit të dimensionalitetit për të reduktuar hapësirën e veçorive dhe për të përmirësuar efikasitetin e modelit. Me eksperimente dhe analiza të kujdesshme, studimi përmirëson të kuptuarit e analizës së autorësisë dhe ofron njohuri të vlefshme për shqyrtime të mëtejshme në përpunimin e gjuhës natyrore dhe gjuhësinë kompjuterike.

Fjalët kyçe: analiza e autorësisë · problemi me shumë klasa · korpuse ndër-domenesh · inxhinieri e veçorive · metodat e klasifikimit · atribuimi i autorit · identifikimi i gjinisë · klasifikimi i zhanrit · përpunimi i gjuhës natyrore · gjuhë me burime të pakëta.

Апстракт

Анализата на авторството е поле од обработка на природни јазик кое има за цел идентификување различни карактеристики на автори во различни текстуални контексти. Секој природен јазик има специфични атрибути кои укажуваат на уникатноста на стилот на пишување на секој автор. Зголемувањето на анонимните онлајн содржини ја нагласува важноста на анализата на авторството. Има апликации во правната лингвистика, кривичното право, форензичките истраги и компјутерската форензика. Оваа теза дава значаен придонес во анализата на албанското авторство преку фокусирање на три суштински задачи: идентификација на авторството, жанровска класификација и родова идентификација. Примарната цел на оваа работа е да се реши недостатокот на податоци на албанскиот јазик јазик со составување на првиот албански корпус за различни задачи поврзани со авторството. Корпусот претставува различни аспекти на пишаниот јазик преку различни текстуални извори, како што се литературни списи, колумни и објави на социјалните мрежи. Ние создаваме обемен збир на карактеристики од различни категории за да се доловат карактеристиките на пишувањето на авториски стил. Техниките за машинско учење и длабоко учење се користат за да се откријат обрасци што укажуваат на авторство, жанр и пол. Нашиот пристап користи две стратегии за класификација за справување со проблеми со повеќе класи.

Оваа студија обезбедува сеопфатна емпириска анализа на авторството преку серија експерименти. Експерименталните резултати се пријавени и дополнително се анализираат користејќи криви за учење, класа на грешка во предвидувањето и анализа на карактеристики. Ние користиме избор на карактеристики и намалување на димензија со техники за намалување на просторот за карактеристики и подобрување на ефективноста на моделот. Со внимателно експериментирање и анализа, оваа студија го подобрува нашето разбирање за анализата на авторството на албански, обезбедувајќи вредни сознанија за понатамошни истражувања во обработката на природниот јазик и компјутерска лингвистика.

Клучни зборови: анализа на авторство · проблем со повеќе класи · корпуси со вкрстени домени · инженерство на карактеристики · методи на класификација · авторска атрибуција · родова идентификација · жанровска класификација · обработка на природен јазик · јазик со ниски ресурси.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my mentor, Prof. Dr. Arbana Kadriu, and my co-mentor, Assoc. Prof. Dr. Ercan Canhasi, for their absolute support and guidance throughout my Ph.D. journey. Their extensive knowledge and tireless commitment have significantly influenced my research career.

My sincere appreciation also goes to my thesis committee members for their feedback, comments, and suggestions, which have greatly improved the quality of my work.

Lastly, I would like to thank my family, whose unconditional support, love, and encouragement have been a constant source of inspiration and motivation throughout my PhD.

M. Sc. Arta Misini

Tetovo, April 2024

Contents

List of Figures

List of Tables

List of Abbreviations

AA	Authorship Analysis
AAA	Albanian Authorship Analysis
AGnd	Authorship Gender identification
AGnr	Authorship Genre classification
AIdn	Authoship Identification
AP	Authorship Profiling
AV	Authorship Verification
CNN	Convolutional Neural Networks
DL	Deep Learning
DT	Decission Tree
FE	Feature Engineering
LIWC	Linguistic Inquiry and Word Count
LR	Linear Regression
ML	Machine Learning
NLP	Natural Language Processing
OCR	Optical Character Recognition
OvO	One verus One
OvR	One verus Rest
POS	Part of Speech
RF	Random Forest
RNN	Recurrent Neural Networks
SHAP	SHapley Additive exPlanations
SVC	Support Vector Classifier

SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
XGB	eXtreme Gradient Boosting

Chapter 1

Introduction

Authorship analysis has significant relevance in information retrieval and linguistic analysis. It is a field within NLP that examines authors' previous works to identify the authorship of a given text based on its language-related characteristics. The problem of determining the authorship of a written text has existed since the beginning of the written word [73, 95].

Authorship analysis is considered a multi-class, single-label text classification problem. It includes various methods and tasks and uses a feature set that shows a consistent pattern across texts by the same author. Different groups of individuals exhibit identifiable linguistic patterns in their speech and writing. The main focus of this thesis is the exploration of different tasks of AA in the Albanian language. We aim to investigate the elements and methods that demonstrate effectiveness in authorship-related tasks within the specific context of Albanian.

1.1 Aims of the research

The advancement of technology has impacted human interaction through digital platforms such as social media, blogs, and news portals. Researchers are working on developing models using machine learning and information technology [173, 68, 26, 82, 163, 61].

Similar studies of AA have been performed in English literature [94, 22, 140] and other languages [42, 93, 41, 52, 33, 97, 110, 171, 138, 164]. However, Albanian literature is a work in progress in the context of AA [118]. Due to its under-resourced nature, the lack of Albanian digital text content has been a limitation. This study aims to address the problem of AA in the

Albanian domain, trying to overcome the gap in the field. Also, this research introduces a novel corpus for AA in Albanian. The objectives of this thesis are outlined as follows:

- Compilation of a new corpus designed specifically for AA in the Albanian language.
- Identification and extraction of an extensive array of features relevant to AAA tasks.
- Identification of the optimal subset of stylometric features to capture the distinctive authors' writing patterns.
- Experimentation in different AA tasks using various features and traditional and DL-based methods.

This research explores AA in Albanian literature and provides the foundation for more sophisticated methods in this unexplored linguistic field.

1.2 Research questions and hypotheses

We formulate a set of research questions for our investigation and provide valuable insights:

RQ1: How could a novel authorship analysis corpus be compiled for a low-resource language like Albanian?

RQ2: Can the author, gender, and genre be identified from an Albanian text?

RQ3: Which linguistic features are crucial to determining the author's writing style?

RQ4: Are there linguistic features that indicate gender?

RQ5: Do authors naturally use distinct classes of language styles for different genres?

RQ6: What is the best feature set to identify the author of a text document?

RQ7: What are the methods for authorship analysis used in other languages?

RQ8: What are the best methods for authorship analysis in Albanian?

RQ9: How much data is sufficient to recognize authorial uniqueness?

We define three hypotheses for our research investigation:

H1: The novel Albanian authorship analysis corpus significantly impacts Albanian authorship-related tasks.

H2: There are highly relevant feature groups to be used in training effective and efficient machine learning models for each subtask of authorship analysis.

H3: Hand-crafted, task-adapted features-based traditional machine learning models will outperform automatic feature extraction-based deep learning models.

We evaluate the validity of hypotheses through a series of experiments conducted on Albanian textual data. Using different methods, we uncover patterns and distinctive features in Albanian texts that enhance our understanding of authorial characteristics.

1.3 Research methodology

AA relies on three fundamental components. The corpus of attributed texts from different sources serves as training set for the model. An ensemble of textual features that represent the distinct characteristics of an author's writing style is essential for AA. The last element is the selection of classification methods to predict the correct author based on the extracted features.

The proposed framework is a three-phase approach that includes text analysis, feature extraction, and author classification. The first step involves the analysis of the text in the corpus to identify distinct attributes and prepare them for further analysis. The second step includes the extraction of statistical information from the texts to represent the authors' writing habits. The third step uses the selected ML- and DL-based methods to predict the author of a given text using the extracted features [61, 50, 43, 62]. This research introduces a novel and, to our knowledge, the first Albanian AA corpus. This corpus is a collection of newsroom columns, social media posts, and literary works of different genres.

Authorship analysis is rooted in stylometry and focuses on determining an author's writing style through numerical analysis of linguistic elements [90, 130]. Corpus analysis reveals features that help define the original author of an anonymous document. Stylometric markers measure various aspects of a writer's style, including individual elements (n-grams), the characters and words they use (lexical) [116], how they organize sentences and paragraphs (structural) [67], patterns used to compose chunks of the text (syntactic) [173, 161, 162], particular keywords in the text (content-specific) [132], or specific elements for each author (idiosyncratic) [151]. These features contribute to the AA task, enabling the discrimination of authors based on their writing styles and patterns.

1.4 Importance of the thesis

The rapid evolution of technology has introduced new ways of communication through the Internet, social media, email, and other digital platforms [94, 151, 78, 133, 166]. These communication tools provide the advantage of anonymity, emphasizing the importance of this study. Since the text is the primary evidence, it is essential to develop new methodologies to identify authorship from digital data.

AA has a rich history and is used in authorship attribution, especially in the field of historical literature [159]. Researchers use authorship attribution to analyze anonymous or disputed material, such as works attributed to Shakespeare [96] or the Federalist Papers [60]. However, this field has a variety of applications, including legal linguistics, criminal law, forensic investigations, and computer forensics, among others [21, 2, 133, 119, 8, 121].

1.5 Contributions of the thesis

This study significantly contributes to the field of AA in several aspects. First, it provides a new resource for three tasks related to authorship in the Albanian language. These datasets are a reference point for investigating Albanian AA, providing valuable opportunities for future research. In addition to the new corpus, our work explores the process of feature engineering.

We engineer an extensive array of hand-crafted features that capture different textual patterns and stylistic aspects. This variety of features enables a more detailed analysis of the elements that characterize tasks related to authorship. Furthermore, our study conducts experiments across different configurations. The experimentation process explores the potential of the created corpus and features in evaluating AA methods. These experiments enable us to identify the most informative features and configurations. This study significantly contributes valuable resources and insights for future investigations in this field.

1.6 Structure of the thesis

This thesis is organized into several chapters. The following is a summary of the content and objectives of each chapter: Chapter 2 reviews related work in the AA field, examining authorship-related tasks, corpora in high- and low-resource languages, feature extraction, and authorship detection methodologies. Chapter 3 introduces the first corpus of the Albanian language for authorship analysis. We detail the data collection and creation of three distinct datasets. Chapter 4 presents the feature engineering process, describing the techniques used to create an extensive array of features for capturing different aspects of textual content and style. Chapter 5 discusses the classification methods used in our study. We explore strategies for handling multi-class classification and evaluate traditional ML- and DL-based techniques. Chapter 6 presents the experimental setup for authorship identification, providing an in-depth analysis of the findings. Chapter 7 documents our experiments conducted in the field of authorship genre classification, with the achieved results. Chapter 8 reports the experimental scenarios and our findings in the authorship gender identification task. Chapter 9 provides a more general discussion of our research findings. We validate the study's hypotheses and discuss the significance of our contributions to the field of AA. Chapter 10 concludes the thesis and outlines potential directions for future research in the field of AA.

1.7 Conclusion

Authorship analysis is a significant topic in the modern era, characterized by the rapid growth of anonymous sources of information. It is a pattern recognition problem that identifies text features to distinguish one author from another. Our primary objective is to extract diverse textual features from the data and detect the distinct writing patterns of different authors. The essence of AA tasks is the selection of specific features as the most relevant indicators of the writer's style. This thesis explores the complex relationship between different writing styles, textual patterns, and the unique features that make each author's writing distinctive.

Chapter 2

Background and related work

This chapter reviews the related work to authorship analysis and its tasks. Section 2.2 overviews diverse textual corpora used in different languages and the data pre-processing methods in authorship tasks. Section 2.3 examines several features utilized in various authorship-related tasks, including linguistic, domain-specific, psycho-linguistic, and stylometric attributes. The methods employed for authorship detection are provided in Section 2.4. Section 2.5 considers the methodologies used for authorship analysis and other related tasks in low-resource languages. Section 2.6 summarizes our findings and the insights gained from this review.

2.1 Authorship analysis and its related tasks

Determining the authorship of a written work through authorship analysis has been a research topic for many years. As a result, this analysis is conducted in three main subtasks: authorship profiling [48, 20, 85, 81], authorship verification [91, 125, 97, 126, 6], and authorship identification [79, 110, 134, 129, 61, 64, 50, 159, 38, 151, 118].

Authorship attribution, also known as authorship identification, identifies the author of a particular work by using its linguistic characteristics. However, in cases where accurate identification is not achievable, authorship profiling can help reduce the search space by identifying variables such as gender or age to focus the analysis on a more specific group. Identifying whether the claimed author of a particular document actually wrote it is known as authorship

verification. It can be performed by comparing the linguistic attributes of the documents in question to determine if they are stylistically similar.

Author identification, verification, and profiling are similarity detection problems because they measure the similarity between two written works. In these tasks, the writing styles of the analyzed texts are compared to make predictions about their authorship. The field of AA has developed and expanded recently, with researchers working on several real-world scenarios.

2.1.1 Authorship identification

Authorship identification is attributing the original author to a disputed text. This task is considered a text classification problem [62], in which documents' writing styles are analyzed to determine the potential author of a given piece of text. This task has a long history, with some of the earliest studies focusing on the statistical analysis of writing style in Shakespeare [96] and the Federalist Papers [60] in the 19th century. The methods used to determine the correct author depend on stylistic elements that define a specific writer's style. An authorship attribution framework consists of a dataset of texts written by different authors, a set of stylometric features, and a classification model to identify the potential author of an anonymous text. The dataset [110, 3, 112, 61] should represent the unique characteristics of the authors. The second component involves extracting features from the text to distinguish texts written by different authors [151]. The original author of the anonymous text [117] can be determined by training a classification model [61] on the extracted features. Author identification can be helpful in many contexts, including authenticating documents [27], detecting plagiarism, literary studies [45, 91, 18, 122, 162, 42], and forensic investigations [21, 119, 15].

2.1.2 Authorship profiling

Authorship profiling aims to identify an author's demographic characteristics based on their writing style. These attributes include age, gender, and educational level [160, 145]. This task is often used when the author of a text cannot be accurately identified through AIdn methods. AP tries to identify patterns in the language usage of different demographic groups to predict

the potential characteristics of an unknown author. Author profiling has a significant historical background dating back to the 19th century. Researchers have worked to develop new methods for extracting demographic information from written texts. Authorship profiling involves elements such as the dataset, feature set, and classification methods. The dataset is a collection of texts written by known authors [20, 21], with each text labeled with the demographic information of the author. Choosing a dataset that represents the author's characteristics is essential for training and evaluating the AP model. A set of stylometric features is extracted to differentiate between various demographic groups' writing styles [89]. These features are used to train a classification model to make predictions about the demographic characteristics of an unknown author. Identifying a writer's demographic characteristics can be valuable in real-world applications, including forensic investigations and identifying behavioral patterns to target specific user groups in social media marketing campaigns [85, 81].

2.1.3 Authorship verification

Authorship verification is a binary classification problem that determines whether an author wrote a claimed text. The goal is to identify patterns in an author's writing habits to make predictions about the attributed authorship of a text. The methodology used in AV includes the same elements as the previous two AA tasks. First, a dataset of texts written by known authors is required [91]. The texts are processed to identify relevant features to characterize the authors' writing style [91, 28]. The extracted features are used to train a classification model [125, 126, 124] to predict the probability that the given author wrote the disputed texts. The history of AV dates back to the 19th century, with early studies in stylometry focused on identifying the authors of literary works with contested authorship. This task is essential in various contexts, including identifying cases of plagiarism, ensuring the authenticity of written works [113], preventing and detecting illegal activity, and anonymous writing.

The following section examines various corpora in high- and low-source languages. The wide range of text types within corpora significantly defines the scope and depth of studies related to authorship.

2.2 Corpora for different authorship-related tasks

In AA, the selection of corpora is essential in analyzing different aspects of written texts and identifying authorship patterns. This section reviews different AA corpora in various languages, focusing on different types of texts, data pre-processing techniques, and challenges related to corpora in low-resource languages. A variety of corpora have been used for AA in many languages, including English [53, 175, 137, 168], Spanish [137, 168], French [168], and Chinese [178], among others. However, there is a lack of data for under-resourced languages, such as Urdu [110, 138], Bengali [61, 62, 79], Arabic [64, 3], Albanian [118], and Turkish [141]. Research on AA requires suitable corpora because the features and characteristics of the dataset affect the study's outcomes.

2.2.1 Corpora of different document types

This section explores the corpora of different text types and their relevance to different tasks related to authorship across languages. Document types encompass a wide range of textual sources, including literary works [178, 128, 64, 41, 130, 87, 118], news articles [110, 138, 129, 61, 43, 98, 123, 142], tweets [48, 3, 20, 9, 21, 7, 148, 114], emails [16, 172, 28], blog posts [79, 14, 153, 49, 38, 111, 145], and more [173, 134, 85, 97, 47, 119, 56, 167, 146, 147]. Each document type has distinct linguistic features, enabling researchers to explore different writing habits and stylometric aspects. Figure 2.1 visually represents the diverse document types in AA research.

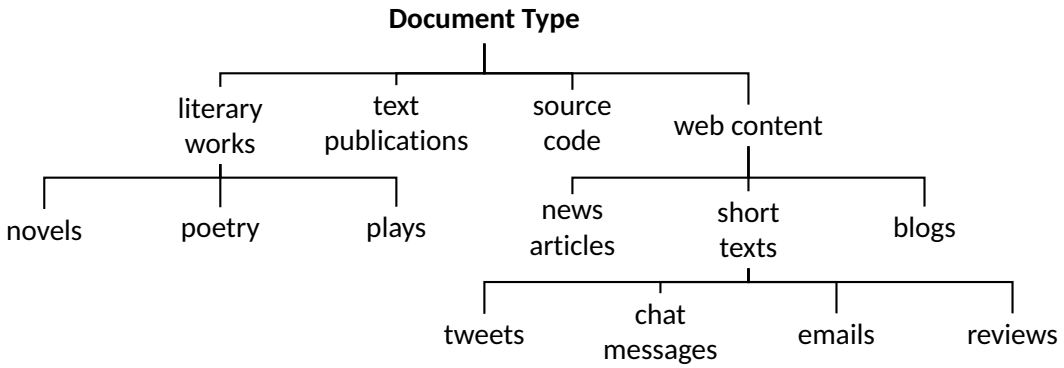


Figure 2.1: Different document types in AA corpora.

Different text types have been used in various authorship-related tasks, such as author gender identification [20, 85, 7, 148, 153], genre classification [178], author attribution [105, 106, 79, 110, 112, 173, 3, 64, 129, 61, 56, 118], author profiling [92, 167, 168, 145], and author verification [91, 97, 28]. Table 2.2 presents an overview of the document types used in AA research across various languages.

We categorize text types into three groups based on their characteristics: literary works use figurative style, non-literary texts demonstrate a formal language and online media texts use conversational language. Analyzing these datasets enables the exploration of authorship-related tasks in different contexts.

Researchers have utilized literary texts to identify the authorship of disputed writings [79, 128, 62, 112, 64, 69, 80, 118] and to classify an author's writings into specific genres [178]. Also, literary works have been used to explore author verification [91] and author profiling [152, 92, 167], providing deeper insights into the characteristics and background of writers [153].

News articles have also been used, particularly in recent years, as the availability of digital news articles has increased. Researchers have often investigated the identification of writers in online news platforms [105, 106, 110, 134, 173, 129, 61, 50, 142], or to profile the characteristics of writers [37].

Tweets, blog posts, and Facebook posts have also been used in the context of social media and online communication. Researchers have tried to identify the authors of anonymous posts [48, 3, 8, 49, 38], to study the writing styles of different authors to profile their demographic characteristics [48, 21, 20, 148, 7, 169, 168, 145] or to verify the authorship of the texts [137, 27].

In addition to these text types, researchers have also used a corpus of short written texts, such as emails [16, 172, 28, 37], reviews [167], discussions [97], or chat messages [85, 99], for different AA tasks. These texts present unique challenges due to their shortness and informality. In AA research, short written texts are often used for author identification [172, 167, 99], author profiling [85, 167, 37], or author verification [97, 28]. Table 2.1 presents an overview of the diverse corpora used in AA studies.

Table 2.2: Different document type corpora used for AA tasks in different languages.

Lang / Doc Type / Task	AGnd	AGnr	AIdn	AP	AV	literary works	text publications	news articles	tweets	discussions	messages	emails	reviews	blog posts
English	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
Spanish	✓		✓	✓	✓	✓			✓					✓
Russian	✓		✓	✓	✓	✓								
Japanese			✓						✓					
Chinese		✓				✓								
German	✓		✓	✓		✓			✓					
Dutch	✓	✓	✓	✓	✓	✓			✓				✓	✓
Greek	✓	✓	✓		✓				✓					✓
French	✓			✓	✓	✓			✓					
Italian	✓			✓					✓					
Portuguese	✓			✓					✓					
Bengali			✓			✓		✓						✓
Urdu			✓					✓						
Arabic	✓		✓			✓			✓					
Turkish	✓		✓					✓	✓		✓			
Swedish					✓					✓				
Thai			✓								✓			
Persian			✓			✓								✓
Mongolian			✓			✓								
Telugu		✓												
Albanian			✓			✓								

Table 2.1: Different document type corpora used for different AA tasks.

Document Type / AA Task	AGnd	AGnr	AIdn	AP	AV
literary works	✓	✓	✓	✓	✓
news articles	✓		✓		
text publications			✓		
blog posts	✓		✓	✓	✓
tweets	✓		✓	✓	✓
emails	✓		✓		✓
reviews	✓	✓	✓	✓	✓
messages	✓		✓		
discussions					✓

2.2.1.1 Data pre-processing for different document types

Different text types have been used for analysis in various authorship-related tasks. Researchers often utilize different data pre-processing techniques to prepare the data for analysis. Data pre-processing involves preparing the data for further analysis by cleaning and organizing it for the specific task. Numerous techniques are available for pre-processing data in different AA tasks. Table 2.3 presents different data preparation methods, considering the document types' specific characteristics used in various AA studies.

Table 2.3: Data pre-processing methods used for different document types in authorship-related tasks.

Technique	Method	Task	Reference	Document Type
Data Cleaning	lowercase conversion	AIdn	[37], [134]	emails, news stories, blog posts
	replacing chars	AIdn, AGnd	[3], [37]	tweets, emails, news stories
	removal of diacritic	AIdn	[37], [110]	emails, news stories

(continued on next page)

Table 2.3: Data pre-processing methods used for different document types in authorship-related tasks (continued).

Technique	Method	Task	Reference	Document Type
	removal of punctuations	AIdn	[3], [62]	tweets, blog posts
	removal of digits	AIdn	[47], [62], [110], [134]	movie reviews, message posts, news stories; blog posts
	removal of white spaces	AIdn, AGnd	[128], [85], [62], [134]	literary works, chat messages, blog posts
	removal of special chars	AIdn, AA	[3], [48], [62], [134]	tweets, blog posts
	removal of blank lines	AIdn	[62]	blog posts
	removal of HTML tags	AIdn	[62]	blog posts
	removal of URLs	AIdn	[3], [47], [62], [9]	tweets, movie reviews, message posts, blog posts
	removal of hashtags	AIdn	[3], [62], [9]	tweets, blog posts
	removal of emojis	AIdn	[3]	tweets
Text Normalization	tokenization	AIdn	[119], [123], [79], [49]	short written texts, literary works, news stories, blog posts
	stemming	AIdn, AGnd	[56], [14]	text publications, blog posts
	lemmatization	AIdn	[3], [119], [128]	tweets, short written texts, literary works
	stop words	AIdn, AGnd	[3], [128], [85], [20], [110], [14]	tweets, literary works, chat messages, news stories, blog posts
	function words	AIdn	[56]	text publications
	hapax legomenon	AIdn, AGnd	[85]	chat messages

(continued on next page)

Table 2.3: Data pre-processing methods used for different document types in authorship-related tasks (continued).

Technique	Method	Task	Reference	Document Type
Data Reduction	duplicate text	AIdn, AP, AGnd, AA	[3], [47], [37], [62], [9], [41], [48], [40]	tweets, movie reviews, message posts, emails, news stories, blog posts
	multi-topics	AIdn	[51]	collection of articles

In addition, techniques such as feature engineering [115, 5, 12] and feature selection [45, 112] have been used to extract and identify relevant features from the text data. In the next section, we will discuss feature extraction. Researchers have also experimented with function words or stopwords, revealing that including them in the analysis resulted in improved outcomes [105, 106, 122]. It is attributed to the fact that individual authors have little conscious control over these linguistic elements in their writing. Studies conducted by [105, 106, 10, 129] did not include any pre-processing steps in their analysis, as they used the TF-IDF text representation scheme. Therefore, the raw text data were used directly in the AA model without data cleaning.

2.2.2 Corpora in high-resource and low-resource languages

AA is conducted in various languages. The availability of resources and linguistic tools varies considerably between high- and low-resource languages.

2.2.2.1 High-resource languages

High-resource languages have extensive linguistic resources, annotated textual data, and specific tools readily available for research. The availability of corpora from various domains, different topics, and linguistic features makes it possible to understand the characteristics and unique patterns of the authors' writing style. Table 2.4 lists the high-resource language corpora used in AA research. It displays the specific AA tasks, dataset size, language, and document type.

Table 2.4: High-resource language corpora in authorship-related tasks.

Reference	Name	Document Type	Language	# of Authors	# of Samples	Task
[153]	Blog Dataset	blog posts	English	23 authors	4284 journalistic posts	AIdn, AGnd
[67], [145], [134]	BLOGS10, BLOGS50	blog posts	English	10, 50 authors	subset of 71,000 blogs	AIdn, AP, AGnd
[16], [172], [28], [37]	Enron	emails	English	158 users	200,399 messages	AGnd, AV, AIdn
[128]	Gungor 50	literary works (novels)	English	45 authors	25636 records	AIdn
[173], [47], [146], [134], [147]	IMDb62	movie reviews, message posts	English	62 users	62,000 movie reviews, 17,550 message posts	AIdn
[158], [147]	IMDB1M	posts and reviews	English	22,116 users	204,809 posts, 66,816 reviews	AIdn
[9]	ISOT	tweets	English	100 authors	varying	AIdn, AV
[62], [153], [152]	Literary Dataset	literary works (novels)	English	16 authors	48 different novels	AIdn, AV, AGnd
[123], [72], [69]	PAN-CLEF 2012	fragments of novels	English	25 authors	250 examples	AIdn
[56]	PLOS	text publications	English	203 authors	1506 publications	AIdn
[9]	SMF	tweets	English	10K authors	7.8M Tweets	AIdn
[123], [158], [98], [51]	Guardian corpus	collection of articles	English	13 authors	444 documents	AIdn

(continued on next page)

Table 2.4: High-resource language corpora in authorship-related tasks (continued).

Reference	Name	Document Type	Language	# of Authors	# of Samples	Task
[123], [37], [44], [98], [43], [67], [135]	RCV1	news articles	English	2,361 authors	109,433 texts	AIdn, AGnd
[129], [17], [14], [50]	Reuters C50 Corpus	news articles	English	50 authors	5000 documents	AIdn
[67], [156], [134], [173]	CCAT10, CCAT50	news articles	English	10, 50 authors	100 documents per author	AIdn
[53], [54], [177]	TREC corpus	news articles	English	7 authors	5600 documents	AIdn
[17], [111], [49]	CASIS	blog posts	English	1,000 authors	4,000 blog samples	AIdn
[112]	English essays corpus	essays	English	3 authors	27 essays	AIdn
[21]	Twitter dataset	tweets	English	40 users	120 to 200 tweets per user	AA
[48]	Poli Corpus-2020	tweets	Spanish	385 authors	241,864 tweets	AP, AIdn, AGnd
[119]	Rus Idio Style Corpus	short written texts	Russian	5 authors	55 documents	AIdn
[92], [143], [144]	Rus Personality	essays	Russian	1145 respondents	1850 documents	AP, AIdn, AGnd
[114]	Japanese Dataset	tweets	Japanese	10,000 users	2000 tweets per user	AIdn

(continued on next page)

Table 2.4: High-resource language corpora in authorship-related tasks (continued).

Reference	Name	Document Type	Language	# of Authors	# of Samples	Task
[178]	Chinese corpus	literary works (poems)	Chinese	20 poets	11,289 poems	AGnr
[87]	German Corpus	literary works (novels)	German	2 authors	3 novels per author	AIdn
[167]	CliPS	reviews and essays	Dutch	varying	749 documents	AP, AIdn, AGnd, AGnr
[20]	Greek corpus	tweets	Greek	463 users	45848 tweets	AGnd
[157]	Genre-based Corpus	varying	Greek	unknown	250 texts	AGnr

2.2.2.2 Low-resource languages

In contrast to high-resource languages, research in low-resource languages is challenging. The need for linguistic tools in under-resourced languages hinders the development of extensive and diverse corpora. Researchers must make further efforts to obtain textual data that is relevant and representative of the problem at hand.

Despite these challenges, researchers have made significant efforts to explore AA and its related tasks in low-resource languages. These corpora often include tweets [3, 7, 148], blog posts [79, 62, 14, 80, 38], articles [138, 110, 61, 15, 142], and literary works [79, 62, 64, 80, 41, 130, 118]. Table 2.5 presents the corpora used in studies for low-resource languages, providing the document type, language, dataset size, and the specific AA task.

Table 2.5: Low-resource language corpora in authorship-related tasks.

Reference	Name	Document Type	Language	# of Authors	# of Samples	Task
[80], [79], [62]	BAAD16	literary works	Bengali	16 authors	17,966 sample texts	AIdn
[38], [79], [80]	BAAD6	online posts and blog posts	Bengali	6 authors	2,100 sample texts	AIdn
[62]	BACC-18	blog posts	Bengali	18 authors	25,749 texts	AIdn
[61]	Articles of political writers	articles	Bengali	8 authors	1764 articles	AIdn
[110]	UNAAC-20	news articles	Urdu	94 authors	21,938 news articles	AIdn
[110], [15]	Urdu Corpus	articles	Urdu	15 authors	6,000 articles	AIdn
[138]	Urdu AIdn corpus	news articles	Urdu	90 authors	985 text samples	AIdn
[7]	SDTwittC	tweets	Arabic	200 authors	233926 replies (Male)	AGnd
[3]	their own	tweets	Arabic	45 authors	71,397 tweets	AIdn
[64]	AAAT dataset	literary works	Arabic	7 authors	7 books (71 paragraphs)	AIdn
[148]	Turkish dataset	tweets	Turkish	8071 users	>= 200 tweets per user	AGnd
[142]	Milliyet dataset	news articles	Turkish	9 authors	50 articles per author	AIdn
[142]	Kibris dataset	news articles	Turkish	7 authors	50 articles per author	AIdn
[142]	Radikal dataset	news articles	Turkish	7 authors	250 articles per author	AIdn

(continued on next page)

Table 2.5: Low-resource language corpora in authorship-related tasks (continued).

Reference	Name	Document Type	Language	# of Authors	# of Samples	Task
[85]	Online chat dataset	chat messages	Turkish	1616 user accounts	218,742 chat messages	AGnd
[97]	Intra-Domain, Domain Shift, Cross-Domain	discussion forums	Swedish	7726 authors, 8755 authors, 969 authors	16 topics, 17 topics, 2 topics	AV
[99]	Thai dataset	fanpage messages	Thai	6 authors	25 messages per author	AIdn
[139]	Thai AIdn corpus	online documents	Thai	200 authors	547 documents	AIdn
[130]	Persian Corpus	literary works	Persian	40 authors	5 documents per author	AIdn
[14]	PMBC	blogs	Persian	32 authors	4977 documents	AIdn
[41]	Mongolian dataset	literary works (novels)	Mongolian	13 authors	389 novels	AIdn
[86]	Telugu corpus	web pages	Telugu	unknown	50 web pages per 3-genre	AGnr
[118]	Albanian books	literary works	Albanian	4 authors	43 books	AIdn

2.3 Feature extraction

A critical component of AA analysis is feature engineering, which characterizes the distinctiveness of individual writing styles. People have consistent personal characteristics, leading to distinct writing habits. Writing style is an unconscious habit unique to each author, determined by

grammar usage, word choice, and punctuation preferences. A deeper examination in language-related charactreistics reveals that features can be classified into three categories: linguistic, domain-specific, and psycho-linguistic features, as illustrated in Figure 4.1. These feature categories provide valuable insights into the diverse aspects of authorship.

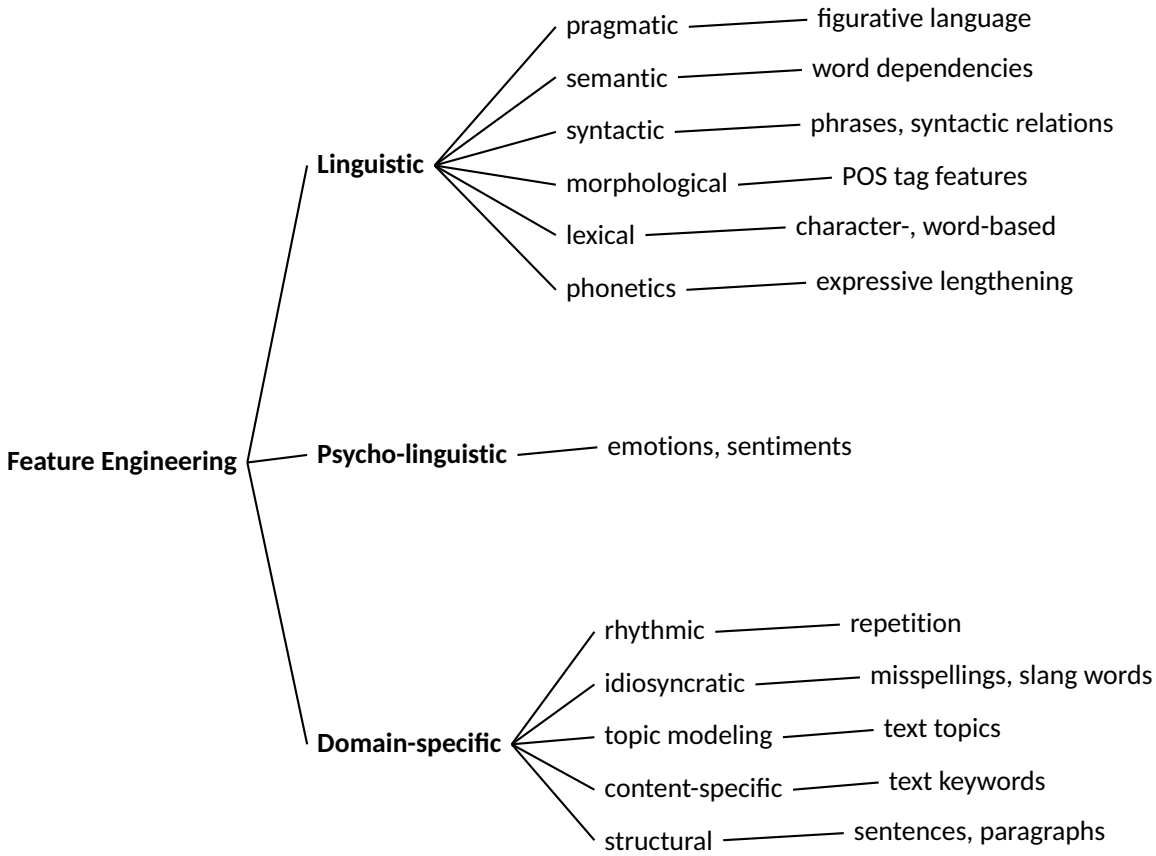


Figure 2.2: Feature categories used in different authorship-related tasks.

Researchers have made progress in tasks like AIdn, AP, AV, and AGnr by examining different sets of features. These tasks aim to reveal the attributes, stylistic patterns, and author-specific characteristics associated with authorship.

2.3.1 Linguistic features

One of the fundamental set of features used in AA is linguistic features, with a wide range of elements at different levels of text analysis. Researchers analyzed authors' phonetic patterns,

lexical and morphological features, syntactic structures, semantic attributes, and pragmatic elements to understand their unique writing style. Phonetic elements provide insights into dialectal variations, while lexical features examine word-level attributes like vocabulary richness and frequency. Morphological characteristics explore the author's structure and word forms. Syntactic structures focus on grammatical constructions and sentence structure. Semantic attributes analyze meaning and word dependencies, and pragmatic elements consider discourse structure and language use. Table 2.6 illustrates the extensive set of linguistic features considered when analyzing tasks related to authorship. These features encompass various language-related characteristics of the author's writing style, including grammar, vocabulary, and punctuation.

Table 2.6: Set of linguistic features used in AA tasks.

No.	Feature	Task	Reference
1.	**Phonetical**	AIdn, AP	[48]
	(1) expressive lengthening	AIdn, AP	[48]
	(2) repeating letters for emphasis	AIdn, AP	[48]
	(3) assonance (phonetical repetition)	AIdn	[64]
2.	**Lexical**	AIdn, AV, AGnd, AP	[7], [8], [122], [151], [6], [67], [139], [153], [24], [109], [56], [5], [155], [64], [11], [47], [48], [129], [116], [21], [130], [134], [176], [15], [14]
	Character-based	AIdn, AGnd, AV, AP	[8], [6], [109], [5], [155], [11], [47], [21], [130], [80], [7], [62], [44], [159], [110], [79], [153], [37], [173], [99]
	(1) total # of characters	AV, AIdn, AGnd	[6], [139], [110], [138], [153], [5], [155], [64], [37], [13], [11], [129], [132], [99], [130], [87], [16], [17]
	(2) total # of letters	AIdn, AV, AGnd	[8], [151], [139], [110], [149], [24], [5], [155], [37], [11], [173], [132], [99]
	(3) total # of vowels	AIdn	[8], [110], [138]
	(4) total # of consonants	AIdn	[110]

(continued on next page)

Table 2.6: Set of linguistic features used in AA tasks (continued).

No.	Feature	Task	Reference
	(5) total # of digits	AIdn, AGnd, AV	[8], [139], [110], [149], [138], [153], [24], [5], [155], [37], [13], [11], [173], [132], [99], [17]
	(6) total # of white spaces	AIdn, AV, AGnd	[8], [139], [110], [138], [42], [5], [37], [13], [11], [40], [132]
	(7) total # of special characters	AIdn, AV, AGnd	[8], [139], [110], [149], [24], [5], [155], [37], [13], [11], [40], [99], [134], [17]
	(8) total # of punctuations	AGnd, AIdn, AV, AP	[8], [151], [7], [110], [149], [112], [93], [61], [100], [153], [42], [24], [5], [155], [13], [173], [40], [129], [65], [99], [87], [134], [16], [17]
	(9) total # of elongations	AV, AIdn, AGnd	[6], [5], [13], [11]
	(10) character n-grams (most frequent)	AIdn, AV, AP, AGnr, AGnd	[26], [8], [122], [44], [149], [93], [3], [138], [100], [109], [56], [155], [64], [119], [47], [40], [48], [129], [142], [86], [131], [115], [4], [130], [114], [15], [10], [125], [51], [126]
	Word-based	AGnd, AIdn, AV, AP	[8], [139], [7], [62], [44], [110], [20], [79], [153], [5], [155], [37], [11], [47], [173], [99]
	(1) total # of words	AIdn, AGnd, AV, AP	[8], [62], [110], [112], [138], [153], [24], [5], [64], [37], [13], [11], [129], [21], [99], [130], [87], [16], [17]
	(2) total # of unique words	AIdn, AV, AGnd, AP	[8], [6], [139], [62], [61], [138], [100], [153], [5], [155], [13], [11], [116], [21], [99], [16]
	(3) total # of capitalized words	AIdn	[8], [24]
	(4) total # of hapax legomena	AV, AIdn, AGnd	[5], [155], [37], [13], [11], [173], [132], [17]
	(5) total # of hapax dislgomena	AV, AIdn, AGnd	[5], [155], [37], [11], [132], [17]
	(6) total # of elongated words	AV, AIdn, AGnd	[5], [13], [11]
	(7) average word length	AIdn, AV, AGnd, AP	[8], [151], [139], [112], [61], [138], [100], [153], [5], [64], [37], [13], [11], [173], [129], [132], [66], [21], [65], [99], [16]
	(8) average # of characters per word	AIdn, AGnd	[153], [173], [12], [87]

(continued on next page)

Table 2.6: Set of linguistic features used in AA tasks (continued).

No.	Feature	Task	Reference
	(9) word length frequency distribution	AV, AGnd, AIdn	[5], [37], [13], [11]
	(10) frequency of dictionary words	AIdn, AGnd, AP	[153], [24], [134]
	(11) frequency of long words	AIdn, AGnd, AV, AP	[8], [110], [153], [5], [37], [11], [21], [99]
	(12) frequency of short words	AIdn, AGnd, AV, AP	[8], [151], [153], [5], [37], [11], [21], [99]
	(13) LIWC features	AGnd	[37]
	(14) vocabulary richness	AIdn, AV, AGnd, AGnr	[140], [151], [138], [100], [153], [42], [155], [110], [37], [13], [173], [132], [115], [14], [17]
	(15) lexical density	AIdn	[110], [24]
	(16) readability measures	AIdn	[110]
	(17) the co-occurrence of word pairs	AIdn	[62]
	(18) word n-grams (most frequent)	AIdn, AP, AGnd, AV	[26], [122], [44], [110], [149], [93], [61], [3], [138], [123], [42], [109], [56], [155], [64], [47], [40], [48], [129], [116], [142], [66], [65], [131], [130], [114], [176], [15], [10], [14], [51], [126]
3.	**Morphological**	AIdn, AGnd, AV, AGnr, AP	[119], [144], [162], [157], [134], [48], [4]
	Morphems	AGnd, AP, AIdn	[144], [48], [4]
	(1) stems and affixes	AIdn, AP	[13], [48], [142]
	POS tag features	AV, AGnd, AIdn, AP	[5], [144], [162], [134], [112], [128], [48], [4]
	(1) total # of POS tags	AIdn, AV, AGnd, AP	[139], [3], [100], [56], [5], [40], [129], [161], [142], [65], [131], [4], [130], [14]
	(2) total # of function words	AGnd, AIdn, AV	[26], [151], [7], [100], [175], [5], [155], [64], [37], [13], [11], [4], [87], [176], [17]
	(3) total # of stop-words	AIdn, AGnd	[122], [149], [61], [153], [173], [40], [4], [16]
	(4) total # of foreign words	AV, AIdn	[5], [134]

(continued on next page)

Table 2.6: Set of linguistic features used in AA tasks (continued).

No.	Feature	Task	Reference
	(5) total # of gender-specific words	AGnd	[37]
	(6) frequency of abbreviations	AIdn, AGnd, AV	[153], [5], [37], [134]
	(7) frequencies of POS tags	AIdn	[56], [65]
	(8) POS n-grams (most frequent)	AIdn, AGnr, AP	[8], [151], [119], [47], [5], [115], [114], [14], [51]
	Gender features	AV, AIdn, AP, AGnd	[5], [11], [48], [162]
	(1) # of feminine words	AV, AIdn, AGnd	[5], [11], [162], [99]
	(2) # of masculine words	AV, AIdn, AGnd	[5], [11], [162], [99]
	(3) # of apologetic words	AGnd	[11]
	Number features	AV, AIdn, AP	[5], [48], [66]
	(1) # of singular words; plural words	AIdn, AGnd, AV	[153], [5], [162], [66]
	Grammatical person features	AV, AIdn	[5], [162]
	(1) 1st person	AIdn, AGnd, AV	[153], [5], [99]
	(2) 2nd person; 3rd person	AV, AIdn	[5], [162]
4.	**Syntactic**	AGnd, AIdn, AV, AP	[8], [151], [6], [67], [139], [7], [153], [24], [5], [155], [64], [37], [144], [11], [173], [129], [161], [162], [130], [134], [176], [14]
	(1) total # of syntactic phrases	AIdn, AGnr	[93], [64], [173], [129], [157], [176]
	(2) frequencies of syntactic writing rules	AIdn	[129], [142]
	(3) frequency of question sentences	AV, AGnd	[5], [37], [11]
	(4) frequency of exclamation sentences	AV, AGnd	[5], [37], [11]
	(5) syntactic relations of different types	AIdn, AGnd, AV, AP	[151], [153], [155], [144], [173], [162], [134]
	(6) dependency relations	AIdn, AGnd, AV, AP	[151], [153], [173], [40], [162], [176]

(continued on next page)

Table 2.6: Set of linguistic features used in AA tasks (continued).

No.	Feature	Task	Reference
5.	**Semantic**	AIdn, AP	[62], [64], [173], [48], [129], [130], [176], [14]
	(1) the # of active voice verbs	AIdn	[64], [176], [14]
	(2) the # of passive voice verbs	AIdn	[64], [176], [14]
6.	**Pragmatical**	AIdn, AP	[48]
	(1) the use of figurative language	AIdn, AP	[48]
	(2) discourse markers	AIdn, AGnd, AP	[93], [153], [48]
	(3) courtesy forms (greetings or condolences)	AGnd, AIdn, AP	[7], [48]

2.3.2 Domain-specific features

Domain-specific features capture authorship characteristics in a specific context. Structural features consider the organization and structure of the text. Content-specific features focus on the text's topic, such as specialized vocabulary, domain-specific terminology, or the presence of topic-related concepts. In AA, idiosyncratic features capture an author's distinctive stylistic choices. These features include slang words, misspellings, or grammatical errors. Rhythmic elements analyze patterns related to sentence rhythms. Social network-specific attributes consider social media elements, emoticons, and terminology related to social networks. Table 2.6 presents the domain-specific characteristics considered during the analysis.

Table 2.7: Set of domain-specific features used in AA tasks.

No.	Feature	Task	Reference
1.	**Structural**	AIdn, AV, AGnd, AP	[8], [6], [139], [5], [155], [64], [37], [102], [11], [40], [21], [99], [134], [176], [14]
	Sentence-based	AIdn	[110], [153], [173]
	(1) total # of sentences	AIdn, AV, AGnd	[8], [6], [139], [5], [155], [37], [11], [99], [176], [62], [112], [24]

(continued on next page)

Table 2.7: Set of domain-specific features used in AA tasks (continued).

No.	Feature	Task	Reference
	(2) total # of words per sentences	AIdn	[8], [176], [24], [87]
	(3) average sentence length	AIdn, AV, AGnd, AP	[151], [112], [61], [100], [153], [13], [173], [129], [66], [21], [65], [130]
	(4) average # of words per sentence	AIdn, AGnd, AV, AP	[110], [112], [153], [132], [21], [12], [130], [87], [139], [5], [37], [11], [99], [176]
	(5) average # of characters per sentence	AIdn, AV, AGnd	[110], [132], [12], [130]
	(6) sentences length frequency distribution	AV, AGnd	[5], [11]
	(7) frequency of sentences beginning with upper case	AIdn, AGnd, AP	[8], [37], [21], [24]
	(8) frequency of sentences beginning with lower case	AIdn, AGnd, AP	[8], [37], [21]
	Paragraph-based	AIdn	[110]
	(1) total # of paragraphs	AV, AIdn, AGnd	[5], [155], [37], [11], [62], [112]
	(2) total # of sentences per paragraph	AIdn	[110], [112], [155]
	(3) the # of words per paragraph	AIdn	[155]
	(4) the # of characters per paragraph	AIdn	[155]
	(5) the # of words per sentence	AIdn	[8], [176], [24], [87]
	(6) average paragraph length	AIdn	[151], [112]
	(7) average # of characters per paragraph	AIdn, AGnd, AV	[110], [12], [5], [37], [11], [99]
	(8) average # of words per paragraph	AIdn, AV, AGnd	[110], [112], [5], [37], [11], [99]
	(9) average # of sentences per paragraph	AIdn, AV, AGnd, AP	[110], [112], [5], [37], [11], [21], [99]
	Document-based	AIdn	[110]
	(1) total # of lines	AV, AIdn, AGnd	[5], [155], [37], [11], [99]

(continued on next page)

Table 2.7: Set of domain-specific features used in AA tasks (continued).

No.	Feature	Task	Reference
	(2) total # of characters	AIdn	[64]
	(3) total # of words	AV, AIdn	[6], [64]
	(4) total # of chunks (short phrases)	AV, AIdn, AGnd	[5], [64], [11]
	(5) total # of sentences per document	AIdn, AGnd	[110], [12]
	(6) total # of paragraphs per document	AIdn	[110]
	(7) average document length	AIdn	[173]
	(8) average word length per document	AIdn, AGnd	[110], [153], [12]
2.	**Content-specific**	AIdn, AV, AGnd	[151], [5], [11], [173], [99], [134]
	(1) total # of emotions	AIdn, AGnd, AV	[153], [5], [144], [12], [99], [134]
	(2) total # words in plural form	AGnd	[11]
	(3) frequency of content specific keywords	AIdn, AGnr, AGnd	[122], [173], [157], [134], [85]
	(4) total # sentiment-bearing words	AGnd, AIdn	[11], [134]
	(5) n-grams	AIdn	[173]
3.	**Topic modeling**	AIdn, AGnr, AV	[122], [15], [178], [126]
	(1) # of topical phrases	AIdn, AV, AGnd	[122], [5], [11]
4.	**Idiosyncratic**	AIdn, AP	[151], [48], [21], [134]
	(1) grammatical mistakes	AIdn	[134]
	(2) misspellings	AP, AIdn, AGnd	[21], [134], [85]
	(3) total # of slang words	AIdn, AP, AGnd	[21], [134], [85]
5.	**Rhythmic**	AV, AIdn, AGnr	[159], [91], [88], [19]
	(1) the use of repetition	AV, AGnr	[91], [88]
6.	**Social Network-specific**	AIdn, AP	[48]
	(1) emoticons	AIdn, AP, AGnd	[149], [24], [85]
	(2) abbreviations	AIdn, AGnd, AP	[149], [153], [21]

(continued on next page)

Table 2.7: Set of domain-specific features used in AA tasks (continued).

No.	Feature	Task	Reference
	(3) # of emojis per word	AIdn	[24]
	(4) frequency of emojis	AGnd, AIdn	[7], [24]
	(5) terminology related to social networks	AIdn, AP	[48]
	(6) hashtags, mentions, and hyperlinks usage	AIdn, AP	[24], [48]

2.3.3 Psycho-linguistic features

Psycho-linguistic features provide information about the author's emotional perspective, social and cognitive concepts, and other psychological aspects of their writing. Table 2.8 presents various psychological elements used in the AA analysis.

Table 2.8: Set of psycho-linguistic features used in AA tasks.

Feature	Task	Reference
PSYCHO-LINGUISTIC	AGnd, AIdn, AP	[7], [37], [48]
(1) # of positive emotions	AGnd, AIdn, AP	[37], [144], [48], [12], [99]
(2) # of negative emotions	AV, AGnd, AIdn, AP	[5], [37], [144], [11], [48], [12], [99]
(3) average # of positive emotions per sentence	AIdn, AGnd	[153]
(4) average # of negative emotions per sentence	AIdn, AGnd	[153]
(5) neutral sentiments and attitudes	AIdn, AP	[48]

2.3.4 Stylometric features

Stylometric features are a sub-class of linguistic features that measure specific stylistic elements within the text. These features include the length of the text, readability formulas, lexical diversity measures, word and sentence count, and number of syllables. Table 2.9 illustrates the diverse stylometric aspects considered in the analysis.

Table 2.9: Set of stylometric features used in AA tasks.

Feature	Task	Reference
STYLOMETRY		
(1) length of the text	AIdn, AP	[48]
(2) readability formulas	AIdn, AP	[110], [48]
(3) lexical diversity	AIdn, AGnd, AGnr, AP, AV	[48], [151], [138], [100], [153], [42], [155], [110], [37], [13], [173], [132], [115], [14], [17]
(4) number of sentences	AIdn, AP, AV, AGnd	[48], [8], [6], [139], [5], [155], [37], [11], [99], [176], [62], [112], [24]
(5) number of words	AIdn, AP, AGnd, AV	[48], [8], [62], [110], [112], [138], [153], [24], [5], [64], [37], [13], [11], [129], [21], [99], [130], [87], [16], [17]
(6) number of words in uppercase	AIdn, AP	[8], [24], [48]
(7) number of syllables	AIdn, AP, AGnr	[48], [109], [86]
(8) punctuation symbols	AIdn, AP, AGnd, AV	[48], [8], [151], [7], [110], [149], [112], [93], [61], [100], [153], [42], [24], [5], [155], [13], [173], [40], [129], [65], [99], [87], [134], [16], [17]

Word embeddings represent the word distributions that different authors use in the vector space and have proven to be powerful techniques for analyzing authorship. Word embedding models that have been effective in this task include Word2Vec [38, 62], GloVe [38, 62, 110], Skip-Gram model [38, 80], BERT [48], and fastText [38, 80, 62, 110]. Various tasks in authorship analysis have used these features to provide distinct perspectives on authorship characteristics.

2.4 Authorship-detection methodologies

This section focuses on the relationship between methods, datasets, and features and how their combination impacts the results in different authorship-related tasks. Researchers used computational models and learning algorithms to identify linguistic patterns and features in written texts. We investigated various methodologies that have demonstrated efficacy in understanding different aspects related to authorship. This study explores various authorship detection techniques, highlighting their strengths and limitations and providing insights on selecting the most effective method based on the task and available data. Figure 2.3 illustrates the various methods used in authorship-related tasks.

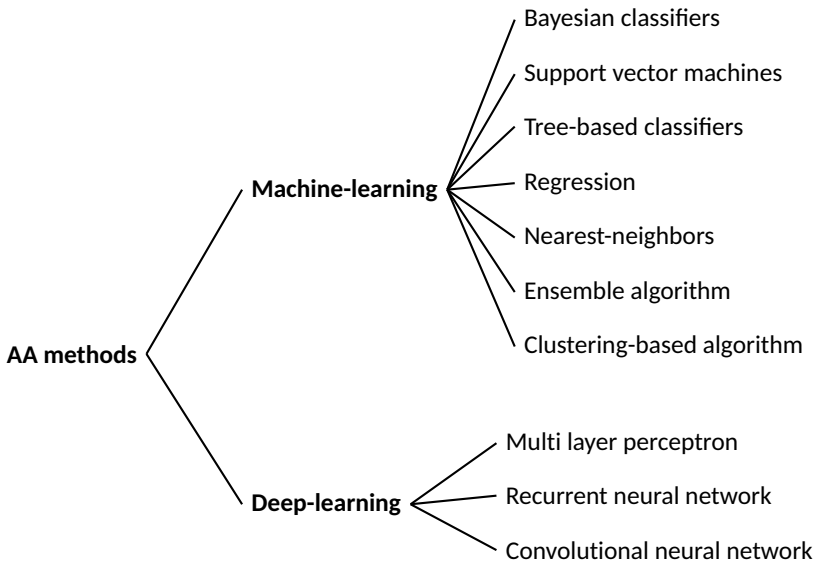


Figure 2.3: Methods used for AA-related tasks.

2.4.1 Machine learning-based methods

Machine learning provides powerful tools to learn patterns and features from data. Researchers have used a variety of ML-based algorithms within the authorship detection domain. This section explores various ML-based methods used for AA and its related tasks, including classification algorithms and ensemble-based models.

2.4.1.1 Classification-based algorithms

Classification algorithms are significant in analyzing authorship. They identify patterns within textual data to categorize texts into predefined classes. Through this exploration, we aim to investigate the characteristics, limitations, and applications of various classification algorithms in authorship tasks. Table 2.10 presents methodologies used in tasks related to authorship in different languages. It details the different types of algorithms, the data used for analysis, the features extracted, and the performance achieved.

Table 2.10: Machine learning-based methods in authorship-related tasks: classification.

Reference	AA task	Dataset	Number of classes	Language	Features	Algorithm	Performance
[10]	AIdn	tweets	100 authors	English	n-grams	SVM, kNN	98%, 97%
[14]	AIdn	news stories	50 authors	English	n-grams, POS-tags, lexical, syntactic, semantic, structural features	kNN	63.11%
[20]	AGnd	tweets	463 users	Greek	word features	NB, kNN	69%, 57%
[64]	AIdn	books	7 authors	Arabic	n-grams, lexical, syntactic, semantic, structural elements	DT	88.73%
[81]	AGnd	SMS messages	unknown	English	word features	NB, DT	66.59%, 79.1%
[85]	AGnd	chat messages	1616 users	Turkish	keywords, misspelled words, slang, emoticons, gender-specific terms	NB, SVM, DT, LR	65.2%, 83.5%, 79.5%, 83.2%

(continued on next page)

Table 2.10: Machine learning-based methods: classification (continued).

Reference	AA task	Dataset	Number of classes	Language	Features	Algorithm	Performance
[86]	AGnr	web pages	3 genres	Telugu	character n-grams	NB, SVM	97.3%, 98%
[110]	AIdn	news articles	94 authors	Urdu	lexical features, n-grams	NB, SVM, LR	74.5%, 55%, 86.9%
[112]	AIdn	essays	3 authors	English	word bigrams	SVM	100%
[115]	AGnr	product reviews	2 genres	English	n-grams, linguistic features	NB, SVM, LR, kNN	85%, 87.84%, 87.63%, 79.37%
[128]	AIdn	literary texts	3 authors	Bengali	word unigrams	NB,	97.60%
[128]	AIdn	books	45 authors	English	POS-based features	SVM, LR	77%, 74%
[129]	AIdn	news stories	50 authors	Persian	lexical attributes, n-grams, POS tags	NB	65.9%
[129]	AIdn	news stories	5 authors	English	lexical attributes, n-grams, POS tags	kNN	82.4%
[134]	AIdn	blog posts	10 authors	English	POS-related, lexical, syntactic, structural, content-specific, idiosyncractic, morphological features	SVM	97%
[138]	AIdn	online news	90 authors	Urdu	lexical features, n-grams	kNN	94.03%
[151]	AIdn	emails	2 authors	English	lexical, structural, content-specific features	SVM	87.43%

(continued on next page)

Table 2.10: Machine learning-based methods: classification (continued).

Reference	AA task	Dataset	Number of classes	Language	Features	Algorithm	Performance
[170]	AIdn	hotel and restaurant reviews	4,521 users	English	word n-grams	SVM	90.5%
[176]	AIdn	books	10 authors	English	lexical, syntactic, linguistic, morphological characteristics	SVM, kNN	97.52%, 79.63%
[176]	AIdn	news stories	50 authors	English	n-grams	SVM	78.67%

2.4.1.2 Ensemble machine learning algorithms

In AA, ensemble algorithms have become very effective tools that combine the insights from several individual models to improve prediction performance. These techniques come in different forms, such as bagging, boosting, and stacking, to address different aspects of classification challenges. We investigate the application of ensemble algorithms in the context of AA and provide insights into their effectiveness and performance. Table 2.11 presents the ensemble-based algorithms utilized in different tasks related to authorship.

Table 2.11: Machine learning-based methods in authorship-related tasks: ensemble algorithms.

Reference	AA task	Dataset	Number of classes	Language	Features	Algorithm	Performance
[144]	AGnd	essays	1145 authors	Russian	morphological, syntactic, emotion-based features	ExtraTrees	72%

(continued on next page)

Table 2.11: Machine learning-based methods: ensemble algorithms (continued).

Reference	AA task	Dataset	Number of classes	Language	Features	Algorithm	Performance
[129]	AIdn	news articles	5 authors	English	lexical features, n-grams	RF	97.6%
[110]	AIdn	news articles	15 authors	Urdu	multi-level text analysis	RF	97.9%
[151]	AIdn	emails	2 authors	English	lexical, structural, content-specific features	RF	90.10%
[85]	AGnd	chat messages	1616 users	Turkish	word-based features	RF	79.1%
[115]	AGnr	product reviews	2 genres	English	n-grams, linguistic features	RF	98%
[86]	AGnr	web pages	3 genres	Telugu	character n-grams	RF	98%
[91]	AV	literary texts	2 classes	Spanish	rhythmic features	RF	78.4%
[151]	AIdn	emails	2 authors	English	lexical, n-grams, structural, syntactic, content-specific attributes	Gradient Boosting	90.09%
[149]	AIdn	chat messages	441 users	English	n-grams, POS-related, social network-specific features	Bagging and Boosting	95.42%
[40]	AIdn	tweets	20 authors	English	n-grams, syntactic features	Stacked ensemble of LR classifiers	85%

(continued on next page)

Table 2.11: Machine learning-based methods: ensemble algorithms (continued).

Reference	AA task	Dataset	Number of classes	Language	Features	Algorithm	Performance
[37]	AGnd	emails	50 authors	English	lexical, syntactic, structural features, function words	AdaBoost	85.13%
[115]	AGnr	product reviews	2 genres	English	n-grams, linguistic features	AdaBoost	93.01%
[91]	AV	literary texts	2 classes	Spanish	rhythmic features	AdaBoost	88.3%

2.4.2 Deep learning-based methods

Recently, DL-based methods have become effective for authorship analysis. They aim to extract complex relationships and patterns from textual data. However, these approaches often require large amounts of data and significant computational resources for training and prediction. Table 2.12 provides an overview of DL-based methods used in AA-related tasks in different languages. It details the various algorithms, the data used for analysis, the features extracted, and the performance results.

Table 2.12: Deep learning-based methods in authorship-related tasks.

Reference	AA task	Dataset	Number of classes	Language	Algorithm	Performance
[66]	AIdn	news articles	5 authors	Bengali	MLP	90.2%
[38]	AIdn	online posts	6 authors	Bengali	CNN	92.9%
[50]	AIdn	news articles	unknown	English	GRU	96.65%
[69]	AIdn	fragments of novels	25 authors	English	CNN, LSTM	78.76%
[48]	AGnd	tweets	385 authors	Spanish	BERT	80.6%

(continued on next page)

Table 2.12: Deep learning-based methods in authorship-related tasks (continued).

Reference	AA task	Dataset	Number of classes	Language	Algorithm	Performance
[178]	AIdn	poetry	2 authors	Chinese	Transformers	94.6%
[134]	AIdn	news articles	10 authors	English	Transformers	94%

2.5 AA systems for low-resource languages

Authorship-related tasks present significant challenges when dealing with low-resource languages due to the limited availability of training data. This problem has been the subject of numerous studies. Misini et al. [104] have outlined various techniques applied in authorship-related tasks across diverse languages. Table 2.13 provides a detailed analysis of AA tasks across under-resourced languages, comparing datasets, features, methods, and performance and offering insights into the evolution of methodologies.

2.6 Conclusion

We conclude the chapter by reviewing previous research on AA and its tasks. Extensive research has been done on AIdn and AGnd tasks in many languages, demonstrating the applicability of various ML methods. On the other hand, AV and AGnr represent less explored areas, offering possible directions for future exploration.

Many experiments have been done with different groups of features in the task of authorship attribution, while psycho-linguistic features have played an important role in gender identification. Rhythmic and morphological features which represent literary elements and linguistic structures, have been used primarily in author verification and genre classification tasks.

Methods include traditional ML algorithms and more recent DL techniques. Modern approaches are used in author and gender identification tasks. Different methods are used in

Table 2.13: AA systems for low-resource languages (best results).

Reference	AA task	Document type	Language	Number of authors	Number of samples	Features	Method	Performance
[79]	AIdn	literary works	Bengali	16	17,966	linguistic features	LSTM, transformers	99.8%
[62]	AIdn	novels	Bengali	3	3,000	GloVe	CNN	98.67%
[38]	AIdn	online posts	Bengali	6	2,100	Skip-Gram by Word2Vec and fastText	RNN	92.9%
[66]	AIdn	articles	Bengali	5	1,973	average word and sentence-length	MLP	90.2%
[110]	AIdn	news articles	Urdu	94	21,938	word n-grams	NB	98.7%
[138]	AIdn	online news	Urdu	90	985	lexical features, n-grams	kNN, PkNN	94.03%
[15]	AIdn	articles	Urdu	15	6,000	n-grams	LDA instance-based	92.89%
[3]	AIdn	tweets	Arabic	45	71,397	n-grams, POS-related features	SVC model	99.24%
[64]	AIdn	books	Arabic	7	7	n-grams, lexical, syntactic, semantic, structural features	Logistic Model Tree	88.73%
[6]	AV	news articles	Arabic	2 labels	7686	lexical, syntactic, structural features	Bagging	98.8%

(continued on next page)

Table 2.13: AA systems for low-resource languages (best results) (continued).

Reference	AA task	Document type	Language	Number of authors	Number of samples	Features	Method	Performance
[129]	AIdn	short stories	Persian	5	unknown	word n-grams	Random Forest	99.6%
[14]	AIdn	blogs	Persian	32	4,977	word and POS n-grams, lexical, syntactic, semantic, structural features	Linear Discriminant Analysis	93.91%
[139]	AIdn	online documents	Thai	200	547	POS tags, lexical, syntactic, structural features	PkNN	91.02%
[85]	AGnd	chat messages	Turkish	1616	218,742	keywords, misspelled, slang and emoticon words, discriminating words	Linear SVM	83.5%
[97]	AV	discussion forums	Swedish	8,755	varying	character-to-word, word-to-sentence and sentence-to-document encoding	convolutional filter Bi-LSTMs and attention	97.7%
[86]	AGnr	web pages	Telugu	unknown	150	syllable extraction algorithm to extract char n-grams	SVM, RF	98%
[118]	AIdn	literary works	Albanian	4	43	syntactic structure, linguistic elements	Dmitri Khmelev algorithm	93.02%

languages with high resources, while specialized techniques are used in languages with low resources to overcome the challenges of data scarcity.

Different types of documents, including literary works, online discussions, and formal documents, were used for different analysis objectives. High-resource languages often use ready-made corpora, while low-resource languages focus on online texts such as tweets, news articles, and novels. These data provide valuable insights into various aspects of authorship.

Various features were used to capture authors' unique writing patterns, including lexical, structural, syntactic, and semantic elements. DL-based methods have been used with word embedding models such as Word2Vec, BERT, and fastText, which effectively characterize writing styles through authors' unique distributions of words.

In conclusion, AA exploration reflects diversity in tasks, methodologies, and linguistic elements. This chapter is a foundation for subsequent chapters, evaluating methodologies and contributing to a comprehensive understanding of the AA problem and related tasks. The dynamic nature of AA presents the potential for new knowledge, methods, and advancement of understanding in the field.

Chapter 3

Data collection and corpus creation

The chapter presents the book scanning and web scraping techniques and details the data collection and corpus creation process in Section 3.2. Section 3.3 considers the data cleaning and pre-processing techniques to prepare the data for further analysis. Section 3.4 provides an overview of the corpora specifications. In conclusion, Section 3.5 concludes the chapter and identifies potential directions for future research.

3.1 Introduction

Authorship analysis presents a significant challenge with considerable research within NLP. Despite the developments in analyzing authorship, the need for appropriate datasets for resource-limited languages hinders the evaluation of existing methods. To overcome this gap, we investigate AA in the Albanian language.

Albanian is a specific category in the Indo-European language family. It has a long history of literature and an expanding library of digital content, but there is limited research in this language. The literature review highlights the need for an explicitly designed corpus to investigate AA tasks in Albanian.

This chapter presents the methodology for compiling the novel AA dataset created for the low-resource Albanian language. The corpus creation process in the next section answers the following research question:

1. How could a novel authorship analysis corpus be compiled for a low-resource language like Albanian?

This chapter introduces the Albanian corpus (A3C3) explicitly designed for AA in Albanian. This dataset addresses the limited availability of data to enable Albanian AA research. The corpus is fundamental for analyzing authorship and identifying characteristic elements in Albanian written content.

The corpus includes newsroom columns, social media posts, and literary writings. This variety was selected to represent language use in various contexts, enabling the examination of AA techniques across different writing styles. Newsroom columns were selected for their extensive vocabulary from various topics. They are easily accessible to the public at no cost and with fewer restrictions on use. Newsroom columns have also been used in the classification of Albanian texts [76, 77], the identification of fake news [34], our work in authorship identification [105, 106], and other AA tasks [110, 138, 173, 129, 50, 15, 66]. Literary works enable the examination of AA techniques in more nuanced and figurative language. These texts [118] exhibit a unique linguistic style and provide a variety of genres and themes. On the other hand, the informal and spontaneous nature of social media posts provides a rich source of language use. The variety in language patterns can provide valuable insights into the author's style and preferences.

3.2 Data collection methodology

Authorship analysis requires a collection of textual works that have been attributed to different classes. Such corpora must be carefully chosen considering their size, the type of texts and classes they represent, and their alignment with the research questions [63, 101]. The dataset makes it possible to thoroughly analyze various AA strategies in different scenarios.

The data collection methodology uses a hybrid approach that combines web scraping and book scanning techniques. As illustrated in Figure 3.1, the process involves several steps to create the necessary datasets. The web scraping technique is used to extract textual content from news websites and social media, resulting in a list of newsroom columns and social media posts.

We used a different approach to data collection through book scanning. Various literary pieces were digitized, including novels, poems, and short stories. Collected newsroom columns, social media posts, and literary works were prepared for feature extraction and subsequent AA analysis. With mixed textual content, this collection is a valuable resource for analyzing Albanian authorial characteristics and writing styles.

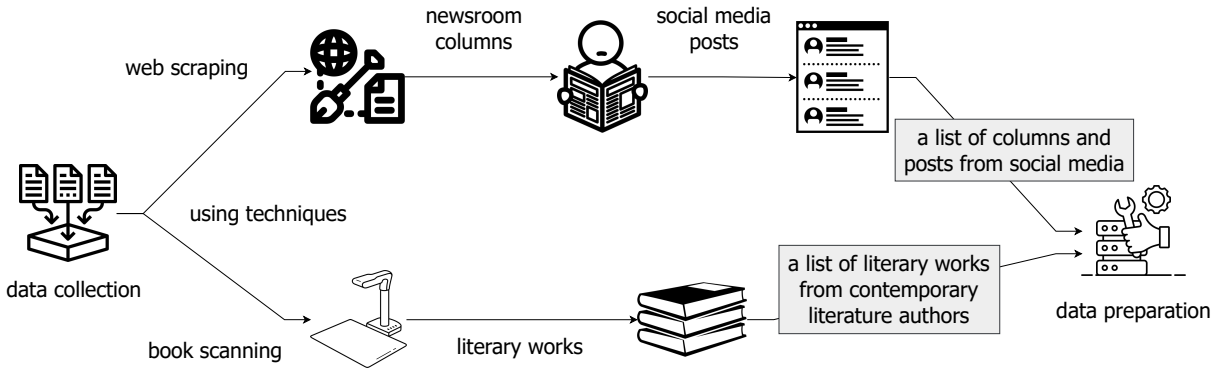


Figure 3.1: Process of building the A3C3 corpus.

3.2.1 Book scanning process

Our approach used book scanning to collect literary pieces from different contemporary authors. This process transforms physical books into digital format by capturing images of the pages and converting them to text using OCR technology. The process resulted in an extensive corpus of numerous Albanian literary works.

3.2.1.1 Literary works

Considering the need for a corpus, we focused on selecting books for scanning within the University of Prizren's library. This repository has an extensive collection of contemporary literary pieces in Albanian and was an ideal source for our objective.

The book scanning process began with carefully selecting books from the library's collection and identifying volumes by prominent writers of contemporary Albanian literature. We focused on authors with a significant writing presence within the library. The authors chosen from modern literature represent different literary styles and genres.

After selecting the writings from contemporary authors, the books were scanned using the Viisan scanner with the Cambook application. The Viisan scanner can transform significant volumes of printed books and manuscripts into digital form. Cambook software uses a high-resolution camera to capture the details of each page within a book.

We used a strategic method to handle the copyright issues during the book scanning. The approach was systematic, scanning every five pages of the book. This approach effectively limited the collected content following copyright rules while enabling the creation of a representative set of the author's writing style.

The sequential scanning took about a month, resulting in a complete and accurate scan of all the selected books. This approach ensured the quality and accuracy of the digitization process. After scanning, books were saved as digital files, with a separate image file for each page. The next step was to convert the image files to text format using UTF-8 encoding, which is compatible with Albanian text. We used OCR technology to convert the scanned images into editable text. The OCR process was performed using the CamBook application's integrated software designed to accurately recognize text in several languages, including Albanian.

Following the OCR procedure, we performed a manual evaluation to ensure the quality of the text files. This validation examined the scanned pages to identify any missing text, inaccuracies in the OCR output, or potential errors from the scanning process. Corrections, whenever necessary, were made manually, either re-scanning the specific pages or manually editing the scanned files. As a result, the corpus of literary texts is a valuable resource for studying AA in Albanian. Figure 3.2 illustrates the book scanning process.

3.2.2 Web scraping process

Web scraping enables the efficient extraction of substantial textual content from online sources. We created a custom web scraper to automate website information retrieval. The web scraping procedure used a semi-automated approach, combining automated data extraction with manual source identification to produce a thorough and effective process.

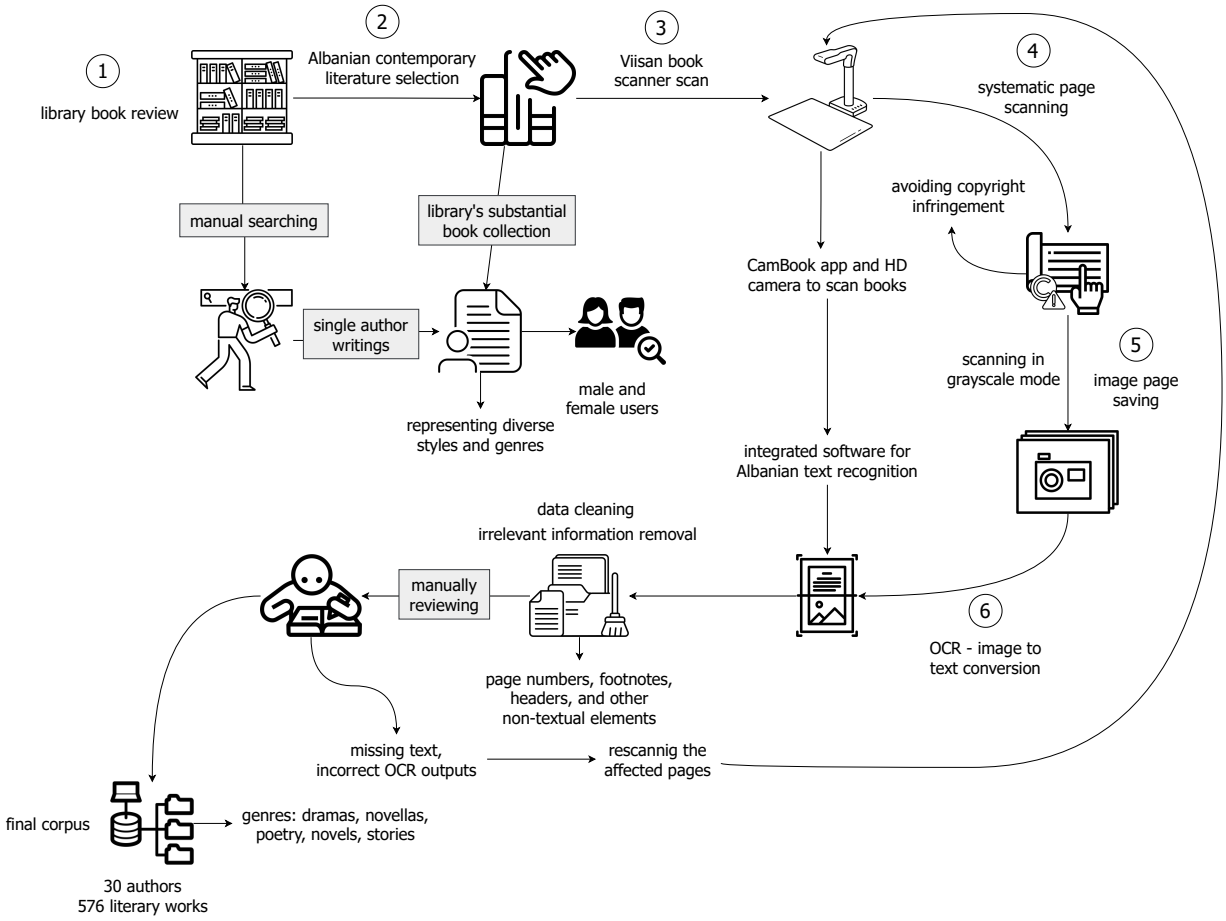


Figure 3.2: Process of building the A3C3 corpus (literary works).

3.2.2.1 Newsroom columns

The web scraping technique was used to collect the newsroom column data across diverse websites. First, we identified target web pages by searching online news portals that regularly publish Albanian newsroom columns. We selected multiple news websites to ensure a diverse set of columns. Among the primary sources selected for data collection were "javanews.al," "shenja.tv," "telegrafi.com," "koha.net," "syri.net," and "portalb.mk." The automated part of the web scraper used the "selenium" Python library to extract data from selected web pages. A list of column links is retrieved from particular web portals to ensure the collected data is unique by removing duplicates. After receiving the list of column links, the web scraper goes through each link and extracts the content of the corresponding columns. This process involved

sending HTTP requests to specific websites and extracting content from the columns by parsing the resulting HTML code. The data was saved in a CSV file with UTF-8 encoding compatible with the Albanian text.

We have extracted the metadata for the columns, including the title, author, and date. Then, data pre-processing was applied to the scraped data to eliminate irrelevant information and format it correctly under the corpus's specifications. It included removing URLs, HTML tags, and whitespaces from the text. To maintain the integrity of the corpus, columns hidden behind subscriptions or registrations, referred to as premium content, were excluded. It ensured that the resulting corpus was accessible to the public.

The web scraper facilitated the data collection process by reducing manual work and generating a significant volume of data in a short time. The result was a large and diverse corpus of Albanian columns designed explicitly for AA research. Figure 3.3 illustrates the web scraping procedure.

3.2.2.2 Social media posts

We followed a multi-step approach to collect a corpus of social media posts from Facebook pages. It included identifying numerous active user profiles of native Albanian speakers who regularly share content in the Albanian language.

To build a corpus for authorship investigations in Albanian, we identified a collection of Facebook pages attributed to public figures and everyday users. It provided profiles related to ongoing events and trending topics.

We used web scraping for data collection from Facebook pages through specialized software tools. The Python-based library *Facebook page scraper* was employed to automate this process. Considering Facebook's requirement for user login to access its content, we used a proxy service to simulate a user's presence during the data collection. The script was executed for each user profile, accessing Facebook posts through a proxy service to configure the connection. To ensure a balanced load on the server and prevent possible blocks, we incorporated an intentional delay of seconds between requests.

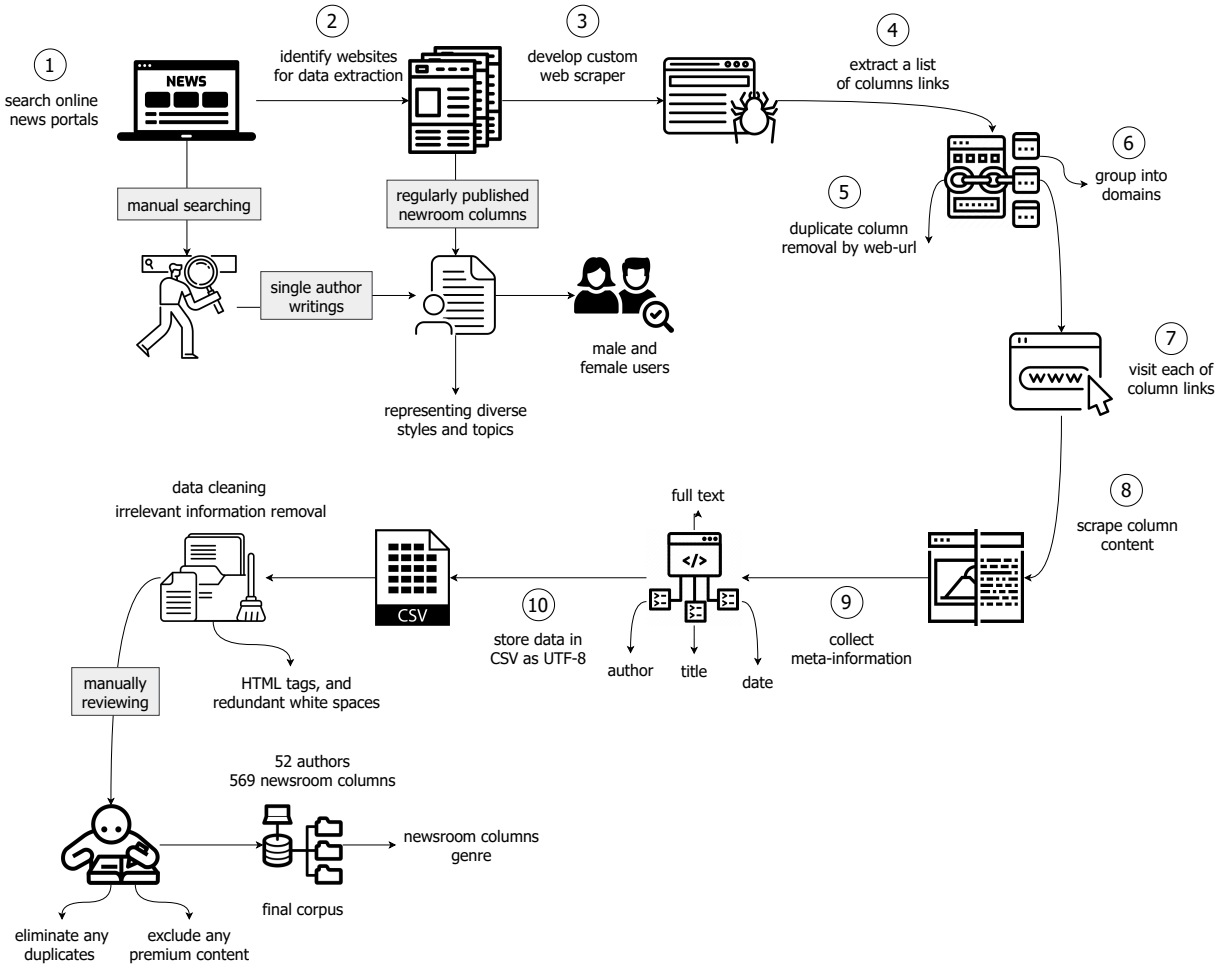


Figure 3.3: Process of building the A3C3 corpus (newsroom columns).

The scraping process began with a preliminary task: counting the posts associated with each user. Some user profiles were private, and we excluded them from the data collection process. Using the web scraper, we extracted different data elements from each post. These included the textual content of the post, the author's name, timestamp, and some associated metadata, including indicators such as likes, shares, and comments. The received data was stored in CSV files using UTF-8 encoding compatible with Albanian text. The collected data underwent cleaning and pre-processing, eliminating non-textual elements such as images, HTML tags, and extra white spaces. To maintain the quality of the corpus, we eliminated some irrelevant data, including non-Albanian content and posts shared by multiple users. Data collection and corpus creation resulted in a novel Albanian dataset designed for AA research. Figure 3.4

visually represents this process.

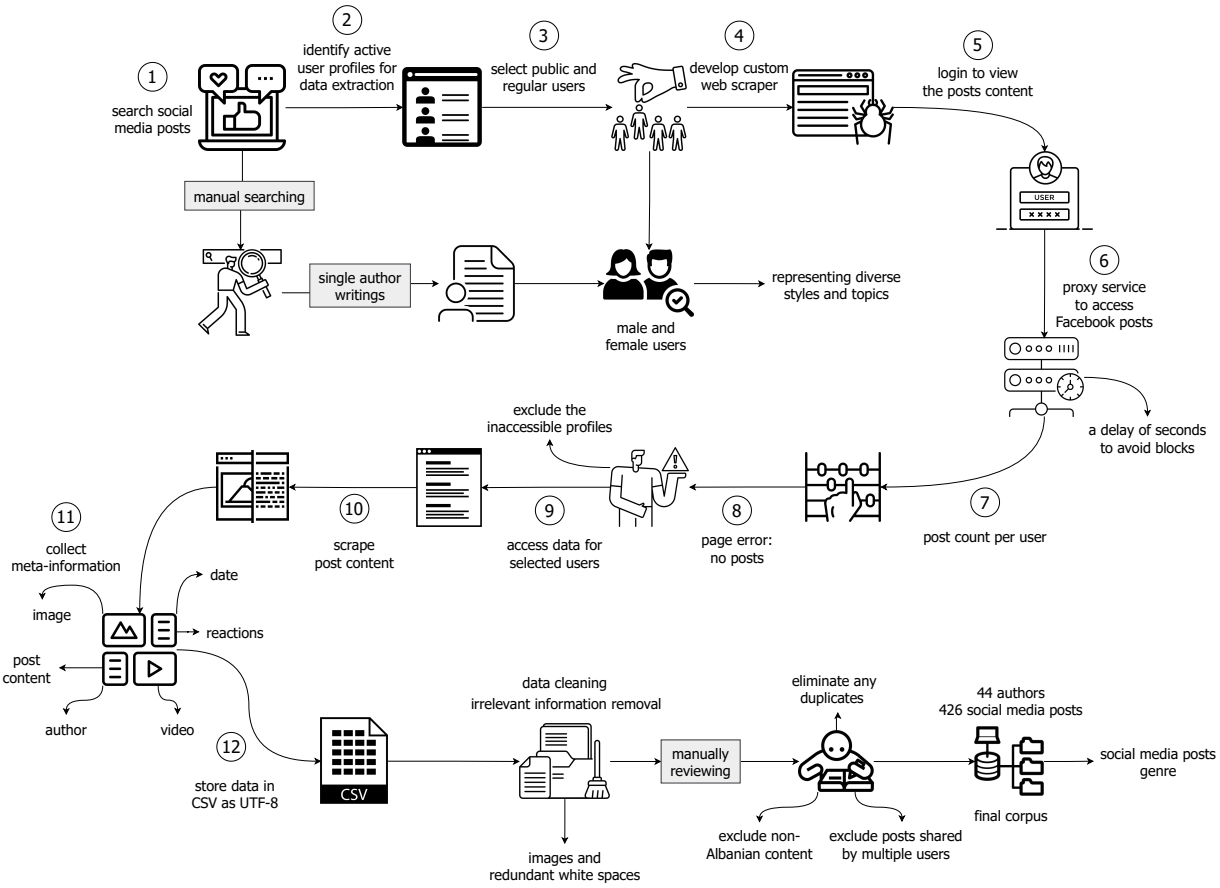


Figure 3.4: Process of building the A3C3 corpus (social media posts).

The icons used in the visual representations of the corpus creation for each dataset, feature engineering (Chapter 4), and methodology for AA tasks (Chapter 5) are sourced from the Noun Project [127].

3.3 Data cleaning and pre-processing techniques

The dataset we used for our experiments combines social media posts, newsroom columns, and literary works written in Albanian. The corpus underwent pre-processing to maintain data quality and consistency [30]. The first phase involved data cleaning to prepare the data for subsequent analysis, as described in Section 3.2. These operations include removing irrelevant

information, duplicate and null values, and correcting errors resulting from the OCR process. In addition, we processed the data to keep the complete information and the nuances of the author's writing style. By maintaining the original structure of lowercase letters, punctuation, and other special characters, it was possible to preserve the linguistic patterns of the authors.

3.4 Corpus description and organization

The A3C3 corpus is a collection of newsroom columns, social media posts, and literary writings. Subsequent chapters elaborate on the specifications for each dataset and explore experiments on authorship analysis and related tasks. The specifications describe three basic levels of language-related data from the corpus: corpus size, authorship-related task statistics, and POS-related data.

Prior studies have examined the POS tagging in Albanian [75, 74, 83]. However, this corpus extracts the POS-related information using an internally developed POS tagger. The tagger was trained on a wide range of texts written in Albanian using deep learning and achieved an impressive F1 score of 97%. The tagset in the corpus includes 58 distinct POS tags, such as abbreviations, adjectives, adverbs, articles, conjunctions, interjections, nouns, cardinal and decimal numbers, prefixes and suffixes, pronouns, prepositions, particles, verbs, and punctuation. These tags provide insights into the words' grammatical classification and contextual function.

The A3C3 corpus is an extensive collection of texts in Albanian from different domains for authorship analysis and research. The corpus presents a valuable resource for investigating the AA field and its tasks in the Albanian language. Appendix A provides representative text samples of the corpus.

3.5 Conclusion and future work

Albanian has received limited research as a linguistically underrepresented language in NLP. To address this problem, we presented the first reference corpus specifically designed for AA in Albanian. Our efforts focused on creating a diverse and representative corpus of literary

works, newsroom columns, and social media posts. Mixing these text types has resulted in a rich dataset that captures different writing styles, genres, and linguistic patterns. The diversity in our corpus makes it possible to better understand the relationship between different writing styles and linguistic characteristics. The corpus aims to advance research in a linguistically underrepresented context by revealing different aspects of authorship. It bridges the gap between language and technology, providing valuable insights into the Albanian literary and digital context.

Continuous progress and expansion of the corpus, including an extensive set of genres and language styles, will ensure its relevance. Creating language resources from our corpus will expand its impact within the research community. Future work promises diverse opportunities to advance AA and computational linguistics within the context of the Albanian language.

Chapter 4

Feature engineering

This chapter explores the feature engineering process for authorship analysis in Albanian. In Section 4.2, we detail the process of extracting a set of crafted features, including linguistic, domain-specific, psycho-linguistic, and stylometric attributes. Section 4.3 outlines the language models utilized for experiments. The feature selection techniques described in Section 4.4 reduce the dimensionality of the feature space to improve the model's effectiveness. Finally, in Section 4.5, we conclude the chapter and provide future directions in exploring AA.

4.1 Introduction

Authorship analysis task identifies the author of a written work based on stylistic features. Feature engineering is a critical process in this task, which involves extracting different text features to identify unique characteristics of writers. We carefully select a set of features to extract meaningful patterns about an author's writing style, linguistic elements, and psychological characteristics.

This study focused on three feature categories: linguistic, domain-specific, and psycho-linguistic features (Figure 4.1). Linguistic elements include a language's lexicon, morphology, and syntax. Domain-specific features consider the context of the text, such as gender-related or social network attributes. Psycho-linguistic features represent the psychological elements present in the text. This research aims to identify language-related patterns and distinctive

writing characteristics specific to each writer by analyzing various aspects of authorship in the Albanian language.

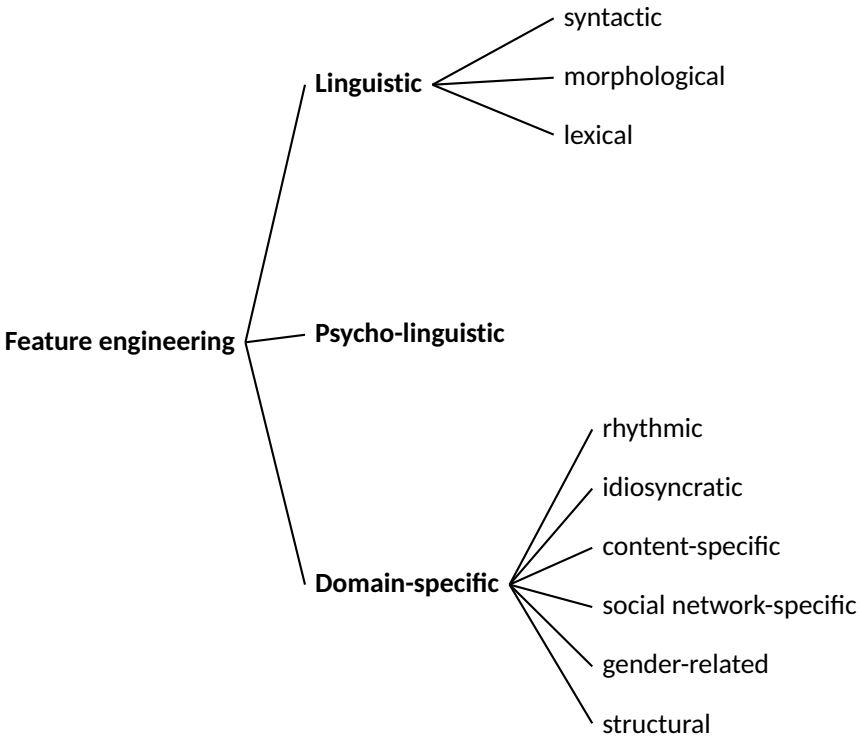


Figure 4.1: Feature categories used in different authorship-related tasks.

4.2 Feature extraction

The feature extraction process in this study involves several steps. Figure 4.2 illustrates the feature extraction pipeline. First, we calculate the most frequent words in the dataset, representing the dominant lexical elements. Then, using POS tagging, we analyze the author's syntactic preferences. Understanding grammatical writing rules and processing syntactical chunks provide significant information about the author's writing patterns and sentence structure. In addition, the feature extraction process includes searching for gender-related POS tags and words, content-specific word categories, and LIWC vocabulary words. Appendix B provides lists

of specific words in the feature set. The data is organized into folders and prepared for subsequent analysis. The feature extraction process enables authorship analysis, revealing different characteristics and patterns within textual data.

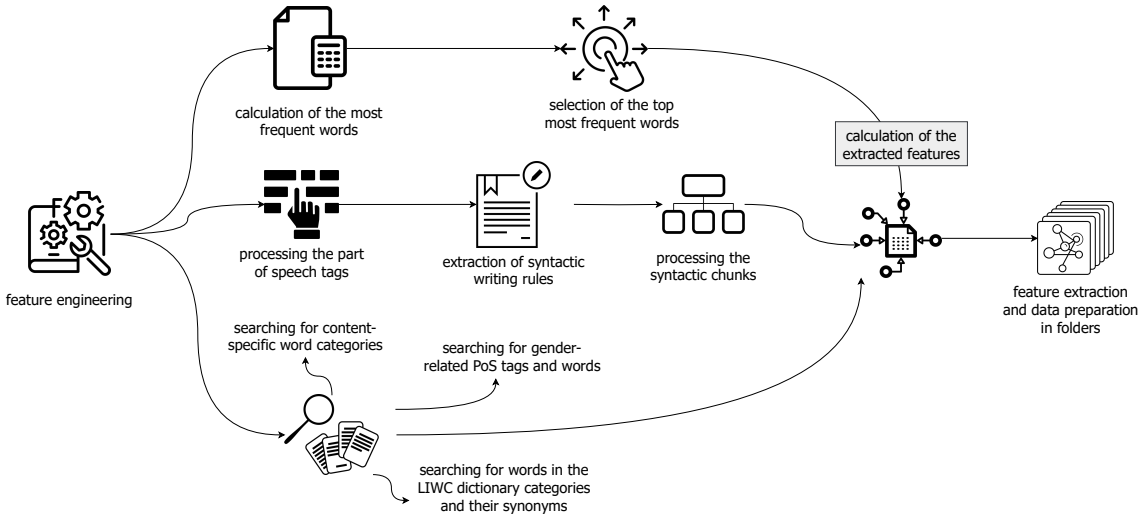


Figure 4.2: Process of feature engineering.

4.2.1 Linguistic features

Linguistic features represent the fundamental category of attributes, including lexical, morphological, and syntactic aspects of language. Its elements are detailed in Figure 4.3. Lexical features provide information about an author's vocabulary diversity and patterns when selecting words. On the other hand, morphological features examine the structure and formation of words, including inflections, prefixes, suffixes, and word stems. The syntactic analysis includes patterns of syntactic and grammatical structures.

4.2.2 Domain-specific features

Domain-specific features provide valuable information in identifying the author of a text within a specific context. Figure 4.4 represents the features within the domain-specific category. Gender-based features identify patterns in linguistic differences between male and female

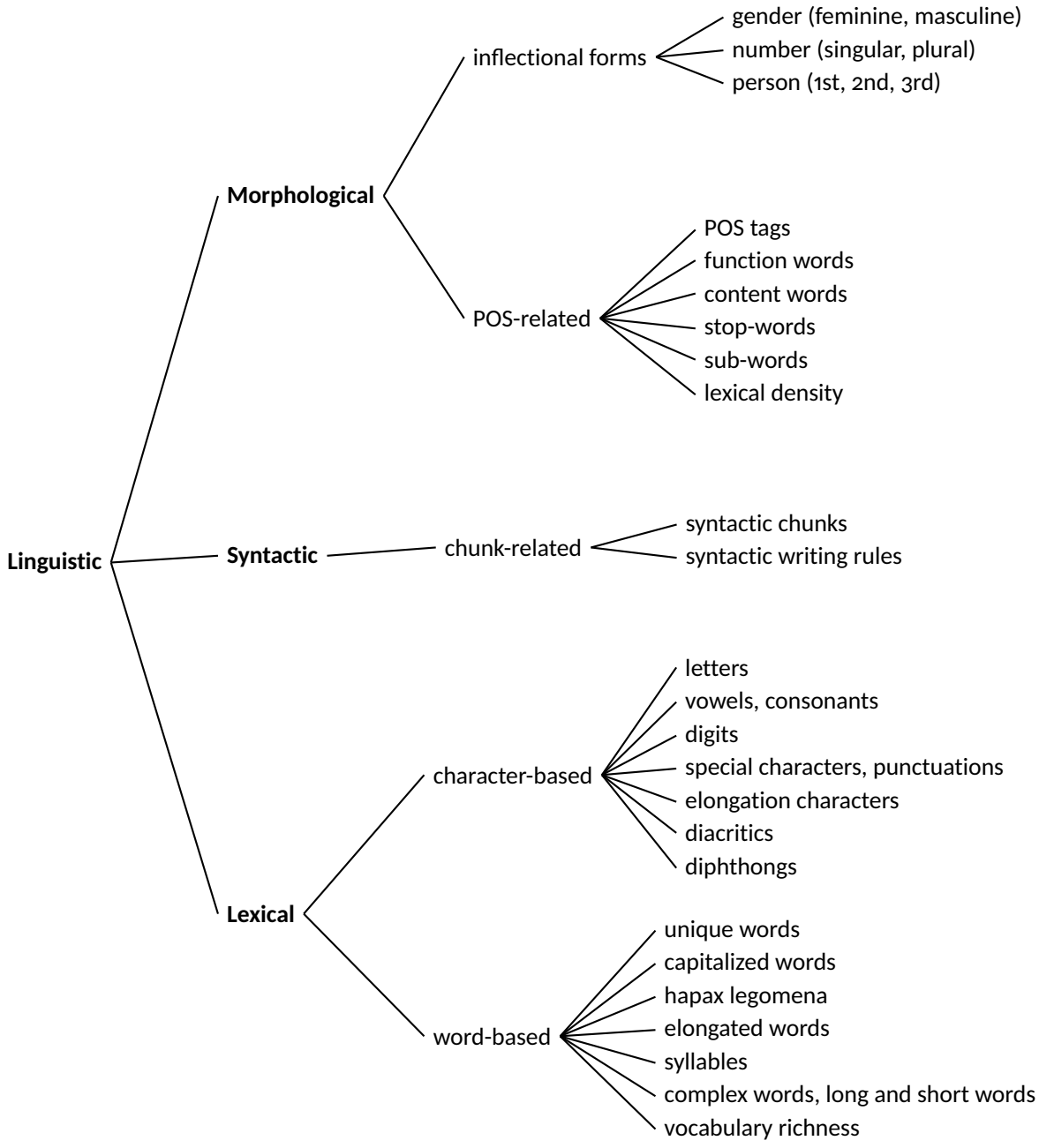


Figure 4.3: Linguistic features used in different authorship-related tasks.

authors. Social network-specific characteristics analyze writing patterns on social media platforms, including elements such as hashtags, mentions, emoticons, and URL links. On the other hand, idiosyncratic features display unique linguistic patterns or traits specific to individual authors, such as slang terms. Rhythmic features examine the emphasis and punctuation patterns in a text to identify the rhythm and repetition style of the writer. Content-specific features provide domain-specific information related to the analyzed text. These features include specific terminology or language related to a particular field. Structural features provide information about the organization of an author's writing and the general text structure.

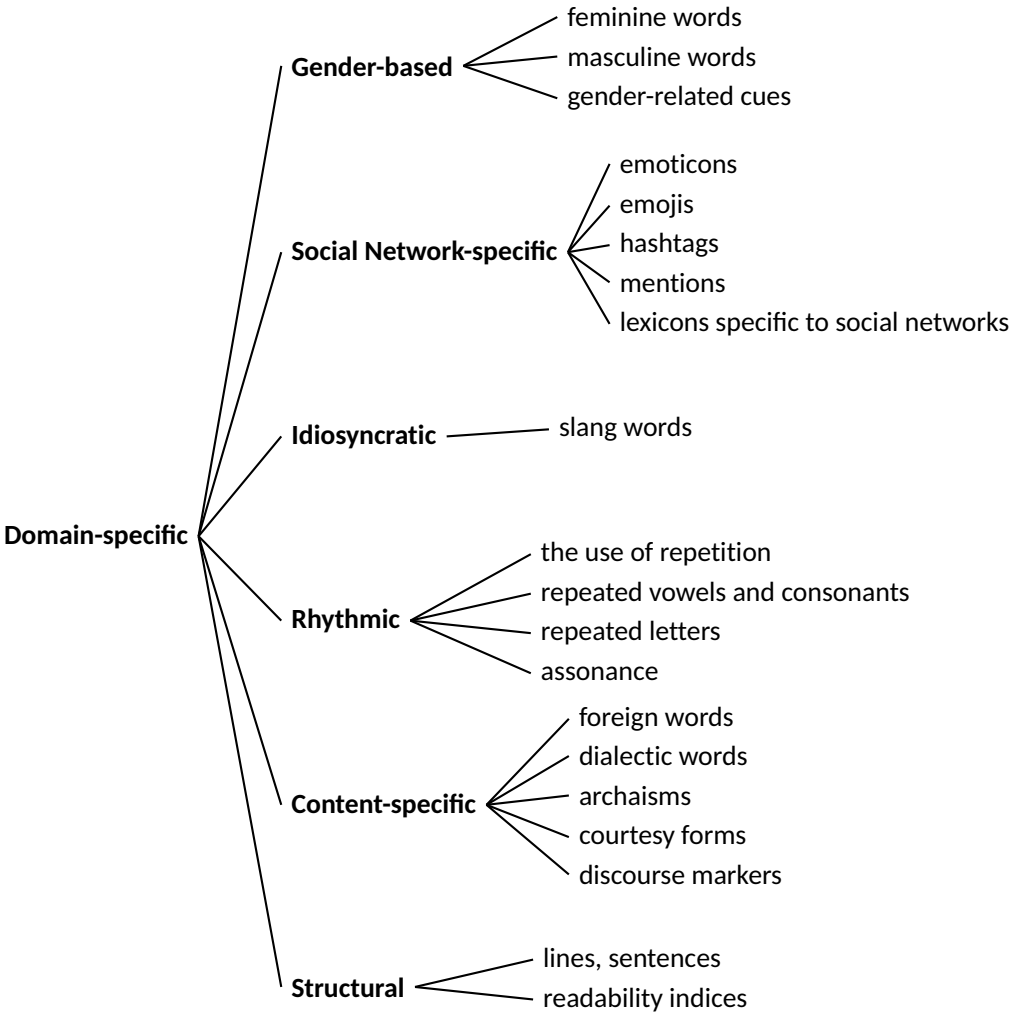


Figure 4.4: Domain-specific features used in different authorship-related tasks.

4.2.3 Psycho-linguistic features

Psycho-linguistic features explore the author's psychological state, emotional expressions, and cognitive and social processes. Figure 4.5 illustrates different psycho-linguistic features used in tasks related to authorship. LIWC features are used in the psycho-linguistic analysis of text based on psychological and linguistic dimensions. These characteristics include positive and negative emotions, cognitive processes, and social elements.

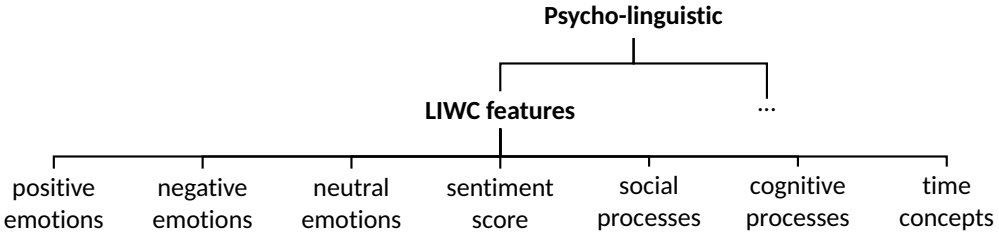


Figure 4.5: Psycho-linguistic features used in different authorship-related tasks.

4.2.4 Stylometric features

Stylometric attributes capture various aspects of an author's writing style, including writing habits, punctuation patterns, and syntactic elements. Figure 4.6 shows the stylometric features used in different authorship tasks.

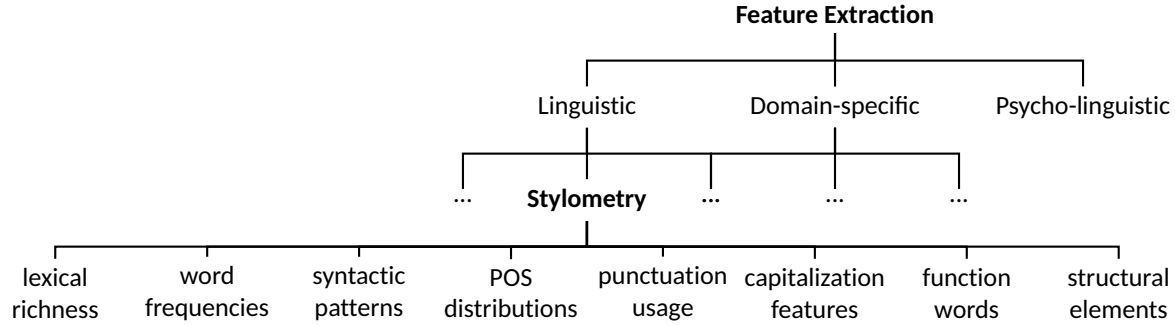


Figure 4.6: Stylometric features used in different authorship-related tasks.

Table 4.1 presents the complete set of extracted features carefully selected for their relevance to the AA tasks. These features were manually engineered and programmed with careful attention and precision.

Table 4.1: Proposed feature sets and description.

Nr.	Features	Total nr-F
1.	**LINGUISTIC FEATURES**	
1.1.	**Lexical features / Character-based**	
	total number of characters (1), lowercase letters (1), lowercase letters-digraphs (1), uppercase letters (1), uppercase letters-digraphs (1), vowels (1), consonants (1), digits (1), white spaces (1), special characters (1), punctuations (1), elongation characters (1), diacritics (1), pronouns short forms (1), diphthongs (1), non-alphabetic characters (1), non-punctuation characters (1), non space characters (1)	18
	percentage of lowercase letters (1), lowercase letters-digraphs (1), uppercase letters (1), uppercase letters-digraphs (1), vowels (1), consonants (1), digits (1), white spaces (1), special characters (1), punctuations (1), elongation characters (1), diacritics (1), pronouns short forms (1), diphthongs (1), non-alphabetic characters (1), non-punctuation characters (1), non space characters (1) out of total characters	17
	Herfindahl index of characters (1), lowercase letters (1), uppercase letters (1), vowels (1), consonants (1), digits (1), special characters (1), punctuations (1), diacritics (1)	9
	Gini index of characters (1), lowercase letters (1), uppercase letters (1), vowels (1), consonants (1), digits (1), special characters (1), punctuations (1), diacritics (1)	9
	Entropy of characters (1), lowercase letters (1), uppercase letters (1), vowels (1), consonants (1), digits (1), special characters (1), punctuations (1), diacritics (1)	9
	frequency of character lowercase unigrams (36), most common character lowercase bi- and trigrams (2*100), character uppercase unigrams (36), digits (0-9) (10), special characters (%,$,&) (23), punctuations (. , ? ! . . .) (9), diacritics (8), pronouns short forms (m', n', s') (3), diphthongs (ai, au, ei, eu, io, oi, ua, ue) (20)	345
1.2.	**Lexical features / Word-based**	
	total number of words (1), unique words (1), capitalized words (1), uppercase words (1), hapax legomena (1), hapax dislegomena (1), elongated words (1), syllables (1), complex words (1), long words (1), short words (1)	11
	percentage of unique words (1), capitalized words (1), uppercase words (1), hapax legomena (1), hapax dislegomena (1), elongated words (1), syllables (1), complex words (1), long words (1), short words (1) out of total words	8

(continued on next page)

Table 4.1: Proposed feature sets and description (continued).

Nr.	Features	Total nr-F
	ratio of capitalized words (1), uppercase words (1), hapax legomena (1), hapax dislegomena (1), elongated words (1), syllables (1), complex words (1), long words (1), short words (1) to number of unique words	9
	Herfindahl index of capitalized words (1), uppercase words (1), elongated words (1), complex words (1), long words (1), short words (1)	6
	Gini index of capitalized words (1), uppercase words (1), elongated words (1), complex words (1), long words (1), short words (1)	6
	Entropy of capitalized words (1), uppercase words (1), elongated words (1), complex words (1), long words (1), short words (1)	6
	most frequent word unigrams (50), bigrams (30), and trigrams (20), elongated words (20), complex words (20), long words (20), word length frequency distribution (m length words) (60)	220
	vocabulary richness - type-token ratio (1), Yule's K (1), Honore's R (1), Simpson's D (1), Sichel's S (1), Brunet's W (1), Entropy measure (1)	7
	average number of characters (1), vowels (1), consonants (1), elongation characters (1), syllables (1), diacritics (1), pronouns short forms (1) per word	7
1.3.	**Morphological features / POS-related**	
	total number of function words (1), content words (1), stopwords (1), conjunctions (2), abbreviations (1), suffixes and prefixes (1), adjectives (1), adverbs (1), articles (2), interjections (2), nouns (1), proper nouns (1), cardinal numbers (1), pronouns (1), prepositions (2), particles (2), modal verbs (2), imperative verbs (1), past participle verbs (1), auxiliary verbs (2), verbs (1)	28
	percentage of function words (1), content words (1), stopwords (1), conjuctions (2), abbreviations (1), sub-words (1), adjectives (1), adverbs (1), articles (2), interjections (2), nouns (1), proper nouns (1), cardinal numbers (1), pronouns (1), prepositions (2), particles (2), modal verbs (2), imperative verbs (1), past participle verbs (1), auxiliary verbs (2), verbs (1) out of total words	28

(continued on next page)

Table 4.1: Proposed feature sets and description (continued).

Nr.	Features	Total nr-F
	ratio of function words (1), content words (1), stopwords (1), conjuctions (2), abbreviations (1), sub-words (1), adjectives (1), adverbs (1), articles (2), interjections (2), nouns (1), proper nouns (1), cardinal numbers (1), pronouns (1), prepositions (2), particles (2), modal verbs (2), imperative verbs (1), past participle verbs (1), auxiliary verbs (2), verbs (1) to number of unique words	28
	Herfindahl index of function words (1), content words (1), stopwords (1), conjunctions (2), abbreviations (1), suffixes and prefixes (1), adjectives (1), adverbs (1), articles (2), interjections (2), nouns (1), proper nouns (1), cardinal numbers (1), pronouns (1), prepositions (2), particles (2), modal verbs (2), imperative verbs (1), past participle verbs (1), auxiliary verbs (2), verbs (1)	28
	Gini index of function words (1), content words (1), stopwords (1), conjunctions (2), abbreviations (1), suffixes and prefixes (1), adjectives (1), adverbs (1), articles (2), interjections (2), nouns (1), proper nouns (1), cardinal numbers (1), pronouns (1), prepositions (2), particles (2), modal verbs (2), imperative verbs (1), past participle verbs (1), auxiliary verbs (2), verbs (1)	28
	Entropy of function words (1), content words (1), stopwords (1), conjunctions (2), abbreviations (1), suffixes and prefixes (1), adjectives (1), adverbs (1), articles (2), interjections (2), nouns (1), proper nouns (1), cardinal numbers (1), pronouns (1), prepositions (2), particles (2), modal verbs (2), imperative verbs (1), past participle verbs (1), auxiliary verbs (2), verbs (1)	28
	frequency of POS unigrams (42), bigrams, and trigrams (2*20)	82
	average number of content words (1), function words (1), stopwords (1), conjunctions (2), abbreviations (1), suffixes and prefixes (1), unique pos types (1), adjectives (1), adverbs (1), articles (2), interjections (2), nouns (1), proper nouns (1), cardinal numbers (1), pronouns (1), prepositions (2), particles (2), modal verbs (2), imperative verbs (1), past participle verbs (1), auxiliary verbs (2), and verbs (1) per sentence	29

(continued on next page)

Table 4.1: Proposed feature sets and description (continued).

Nr.	Features	Total nr-F
	average number of content words (1), function words (1), stopwords (1), conjunctions (2), abbreviations (1), suffixes and prefixes (1), unique pos types (1), adjectives (1), adverbs (1), articles (2), interjections (2), nouns (1), proper nouns (1), cardinal numbers (1), pronouns (1), prepositions (2), particles (2), modal verbs (2), imperative verbs (1), past participle verbs (1), auxiliary verbs (2), and verbs (1) per line	29
	lexical density (1), ratio of unique words and unique pos types (1)	2
1.4.	**Morphological features / Inflectional forms**	
	total number of feminine gender pos types (1), masculine gender pos types (1), singular words (1), singular adjectives (1), singular nouns (1), singular all verbs (1), singular auxilary verbs (1), singular verbs (1), plural words (1), plural adjectives (1), plural nouns (1), plural all verbs (1), plural auxilary verbs (1), plural verbs (1), 1st person pos types (1), 2nd person pos types (1), 3rd person pos types (1)	17
	percentage of feminine gender pos types (1), masculine gender pos types (1), singular words (1), singular adjectives (1), singular nouns (1), singular all verbs (1), singular auxilary verbs (1), singular verbs (1), plural words (1), plural adjectives (1), plural nouns (1), plural all verbs (1), plural auxilary verbs (1), plural verbs (1), 1st person pos types (1), 2nd person pos types (1), 3rd person pos types (1) out of total words	17
	ratio of feminine gender pos types (1), masculine gender pos types (1), singular words (1), singular adjectives (1), singular nouns (1), singular all verbs (1), singular auxilary verbs (1), singular verbs (1), plural words (1), plural adjectives (1), plural nouns (1), plural all verbs (1), plural auxilary verbs (1), plural verbs (1), 1st person pos types (1), 2nd person pos types (1), 3rd person pos types (1) to number of unique words	17
	Herfindahl index of feminine gender pos types (1), masculine gender pos types (1), singular words (1), singular adjectives (1), singular nouns (1), singular all verbs (1), singular auxilary verbs (1), singular verbs (1), plural words (1), plural adjectives (1), plural nouns (1), plural all verbs (1), plural auxilary verbs (1), plural verbs (1), 1st person pos types (1), 2nd person pos types (1), 3rd person pos types (1)	17

(continued on next page)

Table 4.1: Proposed feature sets and description (continued).

Nr.	Features	Total nr-F
	Gini index of feminine gender pos types (1), masculine gender pos types (1), singular words (1), singular adjectives (1), singular nouns (1), singular all verbs (1), singular auxilary verbs (1), singular verbs (1), plural words (1), plural adjectives (1), plural nouns (1), plural all verbs (1), plural auxilary verbs (1), plural verbs (1), 1st person pos types (1), 2nd person pos types (1), 3rd person pos types (1)	17
	Entropy of feminine gender pos types (1), masculine gender pos types (1), singular words (1), singular adjectives (1), singular nouns (1), singular all verbs (1), singular auxilary verbs (1), singular verbs (1), plural words (1), plural adjectives (1), plural nouns (1), plural all verbs (1), plural auxilary verbs (1), plural verbs (1), 1st person pos types (1), 2nd person pos types (1), 3rd person pos types (1)	17
	average number of feminine gender pos types (1), masculine gender pos types (1), singular words (1), singular adjectives (1), singular nouns (1), singular all verbs (1), singular auxilary verbs (1), singular verbs (1), plural words (1), plural adjectives (1), plural nouns (1), plural all verbs (1), plural auxilary verbs (1), plural verbs (1), 1st person pos types (1), 2nd person pos types (1), and 3rd person pos types (1) per sentence	17
	average number of feminine gender pos types (1), masculine gender pos types (1), singular words (1), singular adjectives (1), singular nouns (1), singular all verbs (1), singular auxilary verbs (1), singular verbs (1), plural words (1), plural adjectives (1), plural nouns (1), plural all verbs (1), plural auxilary verbs (1), plural verbs (1), 1st person pos types (1), 2nd person pos types (1), and 3rd person pos types (1) per line	17
1.5.	**Syntactic features**	
	total number of syntactic features (1)	1
	Herfindahl index of syntactic features (1)	1
	Gini index of syntactic features (1)	1
	Entropy of syntactic features (1)	1
	frequency of nominal, verbal, prepositional, adjectival, and adverbal chunks (I-O-B tags) (9), syntactic writing rules (911)	920
2.	**DOMAIN-SPECIFIC FEATURES**	
2.1.	**Gender-based**	

(continued on next page)

Table 4.1: Proposed feature sets and description (continued).

Nr.	Features	Total nr-F
	total number of feminine adjectives (1), feminine nouns (1), masculine adjectives (1), masculine nouns (1), feminine pronouns (1), masculine pronouns (1), feminine suffixes (1), masculine suffixes (1), affective adjectives (1), exclamation words (1), expletive words (1), hedge words (1), intensive adverbs (1), judgmental adjectives (1), uncertainty verbs (1), independence related words (1), assertion words (1)	17
	percentage of feminine adjectives (1), feminine nouns (1), masculine adjectives (1), masculine nouns (1), feminine pronouns (1), masculine pronouns (1), affective adjectives (1), exclamation words (1), expletive words (1), hedge words (1), intensive adverbs (1), judgmental adjectives (1), uncertainty verbs (1), independence related words (1), assertion words (1) out of total words	15
	ratio of feminine adjectives (1), feminine nouns (1), masculine adjectives (1), masculine nouns (1), feminine pronouns (1), masculine pronouns (1), affective adjectives (1), exclamation words (1), expletive words (1), hedge words (1), intensive adverbs (1), judgmental adjectives (1), uncertainty verbs (1), independence related words (1), assertion words (1) to number of unique words	15
	Herfindahl index of feminine adjectives (1), feminine nouns (1), masculine adjectives (1), masculine nouns (1), feminine pronouns (1), masculine pronouns (1), affective adjectives (1), exclamation words (1), expletive words (1), hedge words (1), intensive adverbs (1), judgmental adjectives (1), uncertainty verbs (1), independence related words (1), assertion words (1)	15
	Gini index of feminine adjectives (1), feminine nouns (1), masculine adjectives (1), masculine nouns (1), feminine pronouns (1), masculine pronouns (1), affective adjectives (1), exclamation words (1), expletive words (1), hedge words (1), intensive adverbs (1), judgmental adjectives (1), uncertainty verbs (1), independence related words (1), assertion words (1)	15
	Entropy of feminine adjectives (1), feminine nouns (1), masculine adjectives (1), masculine nouns (1), feminine pronouns (1), masculine pronouns (1), affective adjectives (1), exclamation words (1), expletive words (1), hedge words (1), intensive adverbs (1), judgmental adjectives (1), uncertainty verbs (1), independence related words (1), assertion words (1)	15

(continued on next page)

Table 4.1: Proposed feature sets and description (continued).

Nr.	Features	Total nr-F
	average number of feminine adjectives (1), feminine nouns (1), masculine adjectives (1), masculine nouns (1), feminine pronouns (1), masculine pronouns (1), affective adjectives (1), exclamation words (1), expletive words (1), hedge words (1), intensive adverbs (1), judgmental adjectives (1), uncertainty verbs (1), independence related words (1), assertion words (1) per sentence	15
	average number of feminine adjectives (1), feminine nouns (1), masculine adjectives (1), masculine nouns (1), feminine pronouns (1), masculine pronouns (1), affective adjectives (1), exclamation words (1), expletive words (1), hedge words (1), intensive adverbs (1), judgmental adjectives (1), uncertainty verbs (1), independence related words (1), assertion words (1) per line	15
2.2.	**Social network-specific**	
	total number of emoticons (1), emojis (1), hashtags (1), mentions (1), hyperlinks (1), lexicons with terminology specific to social networks (1)	6
	percentage of emoticons (1), emojis (1), hashtags (1), mentions (1), hyperlinks (1), terms related to social networks (1) out of total words	6
	ratio of emoticons (1), emojis (1), hashtags (1), mentions (1), hyperlinks (1), terms related to social networks (1) to number of unique words	6
	Herfindahl index of emoticons (1), emojis (1), terms related to social networks (1)	3
	Gini index of emoticons (1), emojis (1), terms related to social networks (1)	3
	Entropy of emoticons (1), emojis (1), terms related to social networks (1)	3
	frequency of emoticons (1), emojis (9), lexicons with terminology specific to social networks (56)	66
	average number of emoticons (1), emojis (1), terms related to social networks (1), hashtags (1), mentions (1), hyperlinks (1) per sentence	6
	average number of emoticons (1), emojis (1), terms related to social networks (1), hashtags (1), mentions (1), hyperlinks (1) per line	6
2.3.	**Idiosyncratic**	
	total number of slang words (1)	1
	percentage of slang words (1) out of total words	1
	ratio of slang words (1) to number of unique words	1

(continued on next page)

Table 4.1: Proposed feature sets and description (continued).

Nr.	Features	Total nr-F
	Herfindahl index of slang words (1)	1
	Gini index of slang words (1)	1
	Entropy of slang words (1)	1
	average number of slang words per sentence (1)	1
	average number of slang words per line (1)	1
2.4.	**Rhythmic**	
	the use of repetition - anaphora (1), anadiplosis (1), diacope (1), epanalepsis (1), epiphora (1), epizeuxis (1), polysyndeton (1), symploce (1), aposiopesis (1), repeating interrogative sentences (1), repeating exclamation sentences (1), repeated vowels count (1), repeated consonants count (1), repeated letters count (1), assonance (phonetical repetition) (1)	15
2.5.	**Content-specific**	
	total number of apologetic words (1), dialectic words (1), archaisms (1), lexicons of international concepts (1), abstract concepts (1), courtesy forms (greetings or condolences) (1), indefinite articles (1), definite articles (1), discourse markers (1)	9
	percentage of apologetic words (1), dialectic words (1), archaisms (1), lexicons of international concepts (1), abstract concepts (1), courtesy forms (greetings or condolences) (1), indefinite articles (1), definite articles (1), discourse markers (1) out of total words	9
	ratio of apologetic words (1), dialectic words (1), archaisms (1), lexicons of international concepts (1), abstract concepts (1), courtesy forms (greetings or condolences) (1), indefinite articles (1), definite articles (1), discourse markers (1) to number of unique words	9
	Herfindahl index of apologetic words (1), dialectic words (1), archaisms (1), lexicons of international concepts (1), abstract concepts (1), courtesy forms (greetings or condolences) (1), indefinite articles (1), definite articles (1), discourse markers (1)	9
	Gini index of apologetic words (1), dialectic words (1), archaisms (1), lexicons of international concepts (1), abstract concepts (1), courtesy forms (greetings or condolences) (1), indefinite articles (1), definite articles (1), discourse markers (1)	9
	Entropy of apologetic words (1), dialectic words (1), archaisms (1), lexicons of international concepts (1), abstract concepts (1), courtesy forms (greetings or condolences) (1), indefinite articles (1), definite articles (1), discourse markers (1)	9

(continued on next page)

Table 4.1: Proposed feature sets and description (continued).

Nr.	Features	Total nr-F
	average number of apologetic words (1), dialectic words (1), archaisms (1), lexicons of international concepts (1), abstract concepts (1), courtesy forms (greetings or condolences) (1), indefinite articles (1), definite articles (1), discourse markers (1) per sentence	9
	average number of apologetic words (1), dialectic words (1), archaisms (1), lexicons of international concepts (1), abstract concepts (1), courtesy forms (greetings or condolences) (1), indefinite articles (1), definite articles (1), discourse markers (1) per line	9
2.6.	**Structural features**	
	total number of lines (1), blank lines (1), maximum length of words (1), minimum length of words (1), number of separators between paragraphs (1)	5
	sentence length frequency distribution (m length sentences) (397)	397
2.7	**Structural features / Sentence-based**	
	total number of sentences (1), sentences beginning with upper case (1), sentences beginning with lower case (1), short sentences (1), long sentences (1), maximum length of sentences (1), minimum length of sentences (1)	7
	average number of words (1), characters (1), white spaces (1), vowels (1), consonants (1), elongation characters (1), digits (1), special characters (1), punctuations (1), non-alphabetic characters (1), non-punctuation characters (1), non-space characters (1), unique words (1), capitalized words (1), uppercase words (1), hapax legomena (1), hapax dislegomena (1), elongated words (1), syllables (1), long words (1), and short words (1) per sentence	21
	average number of words (1), characters (1), white spaces (1), vowels (1), consonants (1), elongation characters (1), digits (1), special characters (1), punctuations (1), non-alphabetic characters (1), non-punctuation characters (1), non-space characters (1), unique words (1), capitalized words (1), uppercase words (1), hapax legomena (1), hapax dislegomena (1), elongated words (1), syllables (1), long words (1), and short words (1) per line	21
2.8	**Structural features / Document-based**	
	Gunning fog readability index (1), Flesch-Kincaid readability test (1)	2
3.	**PSYCHO-LINGUISTIC FEATURES**	
3.1.	**LIWC features**	

(continued on next page)

Table 4.1: Proposed feature sets and description (continued).

Nr.	Features	Total nr-F
	total number of words related to positive emotions (1), words related to negative emotions (1), words related to neutral emotions (1), social processes (1), cognitive processes (1), time concepts (1)	6
	percentage of words related to positive emotions (1), words related to negative emotions (1), words related to neutral emotions (1), social processes (1), cognitive processes (1), time concepts (1) out of total words	6
	ratio of words related to positive emotions (1), words related to negative emotions (1), words related to neutral emotions (1), social processes (1), cognitive processes (1), time concepts (1) to number of unique words	6
	Herfindahl index of words related to positive emotions (1), words related to negative emotions (1), words related to neutral emotions (1), social processes (1), cognitive processes (1), time concepts (1)	6
	Gini index of words related to positive emotions (1), words related to negative emotions (1), words related to neutral emotions (1), social processes (1), cognitive processes (1), time concepts (1)	6
	Entropy of words related to positive emotions (1), words related to negative emotions (1), words related to neutral emotions (1), social processes (1), cognitive processes (1), time concepts (1)	6
	average number of words related to positive emotions (1), words related to negative emotions (1), words related to neutral emotions (1), social processes (1), cognitive processes (1), time concepts (1) per sentence	6
	average number of words related to positive emotions (1), words related to negative emotions (1), words related to neutral emotions (1), social processes (1), cognitive processes (1), time concepts (1) per line	6
	sentiment score (1)	1
		Total: 2,881

We used a variety of imperative and reflective programming paradigms to compile the extensive array of extracted features. This approach enabled the execution of all user-defined functions embedded within the code (Code Listing B.1 in Appendix B).

Linguistic elements, domain-specific attributes, and language-based psychological patterns provide information on the distinct linguistic patterns and stylistic features that characterize an author's writing. The stylometric analysis is related to other linguistic dimensions, including different categories of features. The array of hand-crafted features in AA involves linguistic, domain-specific, and psycho-linguistic variables to analyze different aspects of authorship, such as writing style, domain knowledge, and psychological elements.

4.3 Language models

This section describes language models used to represent text data. Our experiments use three text representation models: TF-IDF, Word2Vec, and CountVetorizer.

The **TF-IDF (Term Frequency-Inverse Document Frequency)** [154] model is a widely employed method in feature extraction. The model calculates the frequency of tokens in a document and their inverse frequency in the corpus to assign weights to words. TF-IDF generates a feature matrix of the word's relevance score for each document.

Word2Vec [103] is another language model in NLP that defines the semantic relationships between words in a corpus. The model creates vector representations for each token to capture word relations in text data.

CountVectorizer [55] calculates the occurrence of each token in the corpus. This approach converts a collection of text documents into a matrix of word counts. The matrix helps analyze the usage of individual words and their significance in the text.

The text representation models are used in experiments to extract meaningful features and patterns. Each model represents a distinct aspect of the dataset to characterize the author's writing style.

4.4 Feature selection

Feature selection is a helpful technique for improving the effectiveness of the model's performance when using the feature set. It reduces feature dimensionality and computational

complexity by retaining the most relevant and informative attributes. The study investigates feature selection strategies to identify discriminative features for efficient analysis [31].

Feature importance from machine learning models [120]. Ensemble approaches use ranking mechanisms to provide information about the relevance of features—feature importance scores generated by the XGBoost model guide feature selection by identifying features with significant authorship information.

Recursive feature elimination (RFE) [36] is an iterative process that systematically removes the less relevant features from the dataset. By considering the impact of removing each feature on the model's performance, RFE identifies the feature subset that contributes to accurate AA.

SHapley Additive exPlanations (SHAP) [150, 136] is a machine learning method that uses game theory concepts to interpret the model decisions. It represents the contribution of each feature to the prediction and helps in understanding their influence on the model's output. The method is applied to feature importance analysis to enable the interpretability of ML models.

Principal Component Analysis (PCA) [1] is a dimensionality reduction technique that transforms original data into linear feature combinations known as principal components. The method makes it possible to visualize relationships and latent patterns in the data, like authorship clusters or trends.

Autoencoders [84, 57]. Our investigation uses autoencoders for feature learning and dimensionality reduction to explore latent representations of authors' writing styles. The learned representations identify complex relationships and non-explicitly engineered features. The model is classified as a neural network and will be further explored in the subsequent chapter.

We refine our models using various feature selection methodologies to minimize irrelevant attributes and reduce the computational complexity. This process reduces dimensionality by focusing on relevant data elements for efficient authorship detection across diverse linguistic settings.

4.5 Conclusion and future work

This chapter presented a wide range of manually crafted features for authorship analysis. We explored a variety of linguistic attributes, domain-specific characteristics, psycho-linguistic dimensions, and stylometric patterns that characterize an author's writing style in different linguistic scenarios.

Language models in feature engineering extend the process by utilizing numerical representation techniques and pre-trained models to capture complex relationships in the data. The feature selection process helps increase the model's efficiency and interpretability.

Subsequent chapters investigate the application of model-generated and manually engineered features with different classification techniques, highlighting their effectiveness in various authorship tasks and scenarios. Future exploration should expand the array of hand-crafted features to identify unique patterns for specialized topics and genres. Further research into pre-trained language models for specific authorship tasks could provide more contextual representations. Furthermore, exploring cross-linguistic and cross-domain feature extraction can provide new insights into authorship analysis.

Chapter 5

Classification methods

The chapter explores multi-class strategies to handle multiple classes in Section 5.2. Section 5.3 presents the traditional machine learning algorithms and state-of-the-art deep learning methods selected for the experimental framework. We conclude the chapter in Section 5.4 and provide the potential directions for further research.

5.1 Introduction

Machine learning and deep learning classification methods are critical to understand the relationship between language and authorship. These methods try to identify data patterns that define the writer's unique style to classify data points into predefined classes.

This chapter explores multi-class classification strategies in authorship analysis problems. In multi-class classification, texts correspond to multiple classes, genres, or authors. The following section presents their concepts, advantages, and considerations in classification tasks.

Machine learning techniques offer a variety of algorithms that can effectively capture patterns in an author's writing style. They learn to identify variations and characteristics in the data and apply this information to differentiate texts written by various authors. On the other hand, deep learning methods have transformed the AA field. They use neural networks to capture complex relationships in textual data. These techniques automatically extract features from raw data, facilitating the identification of linguistic elements in the text. ML- and DL-based techniques with prepared data and engineered features enable experimentation in different

classification tasks. Different configurations in experimental analysis provide valuable insights and interpretations from textual data.

5.2 Multi-class classification strategies

Multi-class classification handles situations with multiple classes. It is essential in various AA problems since the closed set of potential authors is extensive when attributing texts to their corresponding classes. Figure 5.1 illustrates machine learning algorithms classified according to multi-class strategies.

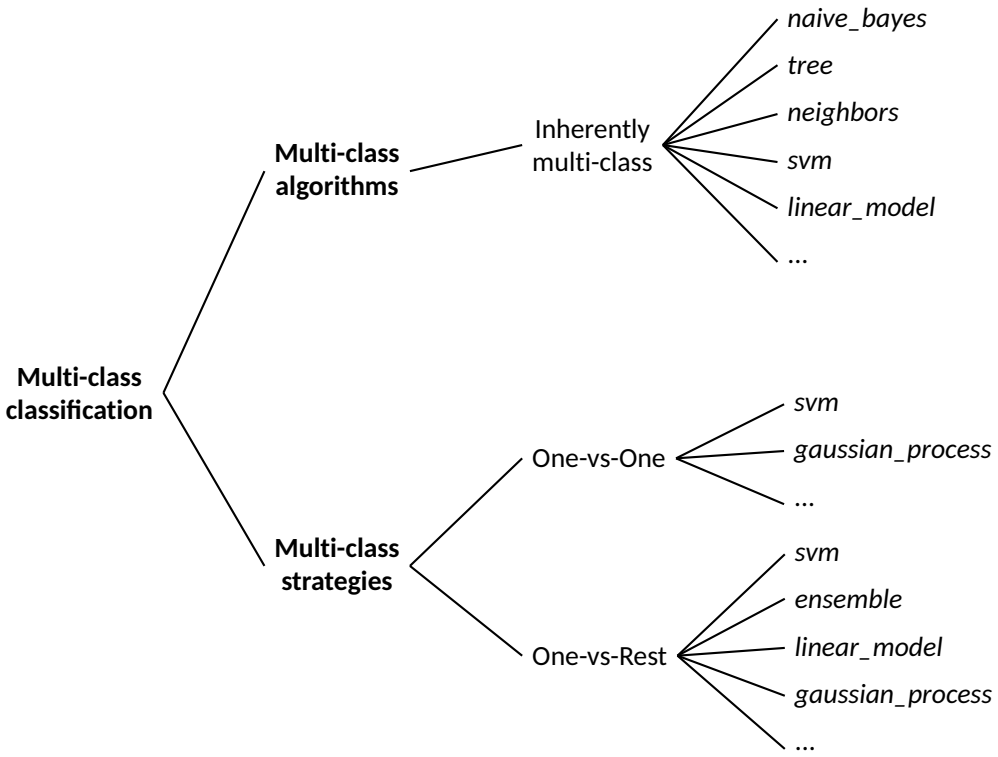


Figure 5.1: Multi-class classification strategies (according to `scikit-learn` library documentation).

One-versus-One (OvO) and One-versus-Rest (OvR) are popular strategies for multi-class classification in AA. These strategies provide techniques to handle multi-class complexity in classification algorithms [108].

5.2.1 One-versus-One strategy

The OvO strategy simplifies multi-class tasks into binary problems, where a binary classifier is trained for each pair of classes. There are N*(N-1)/2 binary classifiers for N classes, and the majority voting method determines the final predicted result. As a result, the number of required classifiers grows exponentially with the number of classes. It requires careful consideration of computational resources. The OvO strategy has several advantages [108, 32, 46]:

1. **Binary classification simplicity**: OvO simplifies multi-class problems into binary classification scenarios, enabling faster training and prediction processes using traditional binary classification algorithms.

2. **Class imbalance handling**: OvO helps handle imbalanced class distributions. Binary classifiers focus on specific class pairs, limiting the effect of imbalanced data and improving classification performance.

3. **Handling non-linear boundaries**: OvO enables the creation of complex decision boundaries between classes, which is especially useful for writing styles with complex patterns that are not easily separable.

OvO offers numerous advantages, but it is crucial to understand potential challenges:

1. **Computational complexity**: Binary classifiers grow exponentially with the number of classes, which increases memory requirements and computational complexity when handling multiple classes.

2. **Ambiguity and confusion**: Multiple binary classifiers making different predictions for the same input can lead to ambiguous situations in OvO.

3. **Risk of misclassification**: The process of providing the final prediction combines the binary classifier's results. Due to "ties" in the vote counts, the risk of misclassification increases in classes with similar characteristics.

The OvO strategy in AA simplifies multi-class problems into binary ones and addresses class imbalances. Still, it requires careful handling of computational requirements and complexity.

5.2.2 One-versus-Rest strategy

The OvR strategy uses a binary classifier to categorize samples into predefined classes, distinguishing each class from others. The class with the highest confidence score is used to determine the final prediction. Authorship analysis uses the OvR strategy to classify authors based on their writing style, using a binary classifier for each author to differentiate their style from the rest. The OvR strategy has several advantages when applied to AA [108, 32, 46]:

1. **Simplicity**: OvR is straightforward. It considers each author as an individual binary classification task, ensuring simplicity in the training and prediction procedures.

2. **Computational efficiency**: Training several individual binary classifiers requires fewer resources compared to the OvO approach. This makes it a helpful option when there are limited resources available and when dealing with a large number of classes.

3. **Distinct author profiles**: OvR can identify unique author characteristics by differentiating each author's writing style from the rest.

In addition, we should take into consideration the difficulties associated with OvR:

1. **Imbalanced classes**: Class imbalance is a significant concern in the OvR strategy when particular authors have a limited number of instances. This problem could potentially influence the performance of classifiers in minority classes.

2. **Class overlap**: Misclassifications can occur when stylistic similarities cause decision regions to overlap in the feature space.

3. **Class ambiguity**: OvR may need help differentiating the writing styles of similar authors, potentially causing confusion between classes.

OvR is a straightforward and efficient method for multi-class classification. However, it still has to deal with issues like class imbalance, ambiguity, and confusion. Addressing these problems can improve its effectiveness in identifying unique characteristics in writing styles.

The selection of OvO and OvR strategies depends on the dataset size, the available computational capabilities, and the nature of the classification problem. These techniques are essential in AA for correctly assigning texts to their classes in a multi-class scenario.

5.3 Classification models

Authorship analysis aims to accurately classify texts into their respective classes based on linguistic attributes. Classification methods help understand and identify patterns of language use and writing habits in text data. This section discusses machine learning and deep learning methods in authorship analysis tasks to extract meaningful patterns from textual data [29].

5.3.1 Machine learning-based models

This section considers several ML-based algorithms used in our experiments. These include ensemble methods like XGBoost and Random Forest, Logistic Regression and Ridge Classifier as linear models, and Linear Support Vector Classifier. ML-based algorithms were selected based on an extensive evaluation of various methods, comparing their performance across different configurations.

XGBoost [35] is an ensemble algorithm that uses gradient boosting to handle complex data relationships and analyze authorship in scenarios with multiple classes. **Logistic Regression** [23] is a linear model that provides information about the probability distributions across different classes. The algorithm is adaptable for multi-class analysis. The **Ridge Classifier** [59] improves generalization by introducing regularization to the linear model. It is particularly useful in multi-class classification settings with OvO and OvR strategies to handle multiple classes. **Linear SVC** [39] effectively handles high-dimensional data and class separation, improving its performance in multi-class classification problems. **Random Forest** [58] is an ensemble of decision trees that captures complex relationships and overcomes the limitations of a single model, resulting in more reliable results.

Our investigation maintains default model parameters to reflect their inherent strengths for consistency and comparison. The models are applied across two multi-class classification strategies, OvO and OvR, for effective multi-class authorship analysis.

5.3.2 Deep learning-based models

This section extends the investigation by using deep learning techniques that have significantly transformed the field of NLP. Our study utilized fasttext, autoencoders, and transformer models to understand new aspects of authorship analysis and its related tasks.

Fasttext [25, 70, 71] was utilized for AA tasks due to its ability to capture hidden patterns in text data. It effectively handles out-of-vocabulary words and semantic relationships in authors' writings, which makes it suitable for different linguistic analyses.

The study uses **autoencoders** [84, 57] to capture complex relationships and hidden features. The model inputs the feature set into the encoder part, resulting in a compressed representation in the bottleneck layer. The trained autoencoder effectively learns abstract features, reduces dimensionality, and transfers knowledge to other authorship-related tasks.

Our research investigates the use of **transformers** [165] in analyzing authorship. These models capture long-range dependencies and contextual relationships within data points. They also use pre-trained language models such as BERT to extract characteristics that contribute to analyzing authorship.

Figure 5.2 and Figure 5.3 provide an overview of the architectural configurations of autoencoders and transformers.

Deep learning models were selected through experimentation and evaluation of their performance in the AA domain. Using fasttext, autoencoders, and transformers, we present new insights and findings in understanding authorial patterns across literary works, newsroom columns, and social media posts.

Figure 5.4 shows the methodology used for authorship-related tasks. The figure presents a visual overview of the critical steps of our approach to achieve our objectives in exploring authorship analysis and its tasks across various settings.

5.4 Conclusion and future work

This chapter considered various ML- and DL-based methods to analyze authorship in Albanian. We identified the best-performing algorithms in this context through extensive experimentation and provided a reliable foundation for further research. The study explored multi-class classification strategies, OvO and OvR, to handle scenarios with multiple classes.

This investigation provides potential directions for future research. Transfer learning techniques can address limited data issues in various situations, enabling models trained in other languages or domains to adapt to Albanian AA tasks. We can further fine-tune model parameters, explore advanced architectures, and integrate domain knowledge.

This work conducts experiments on authorship-related tasks using classification methods and integrates them with hand-crafted features and language models to create a robust approach to authorship analysis.

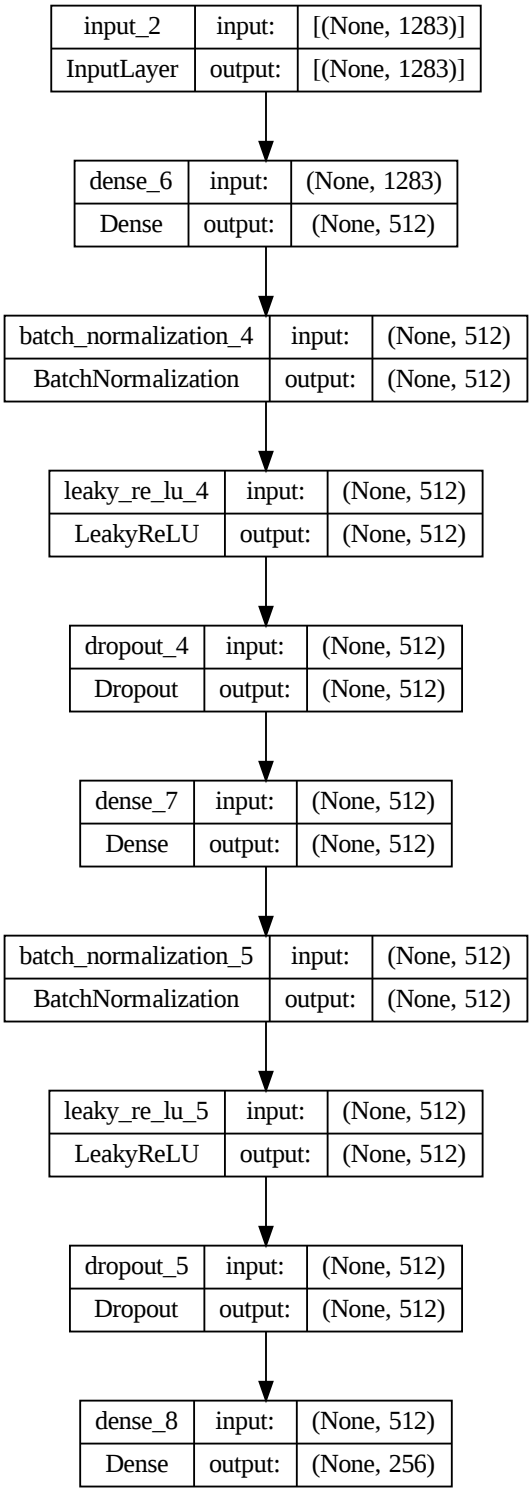


Figure 5.2: The architecture of autoencoder model.

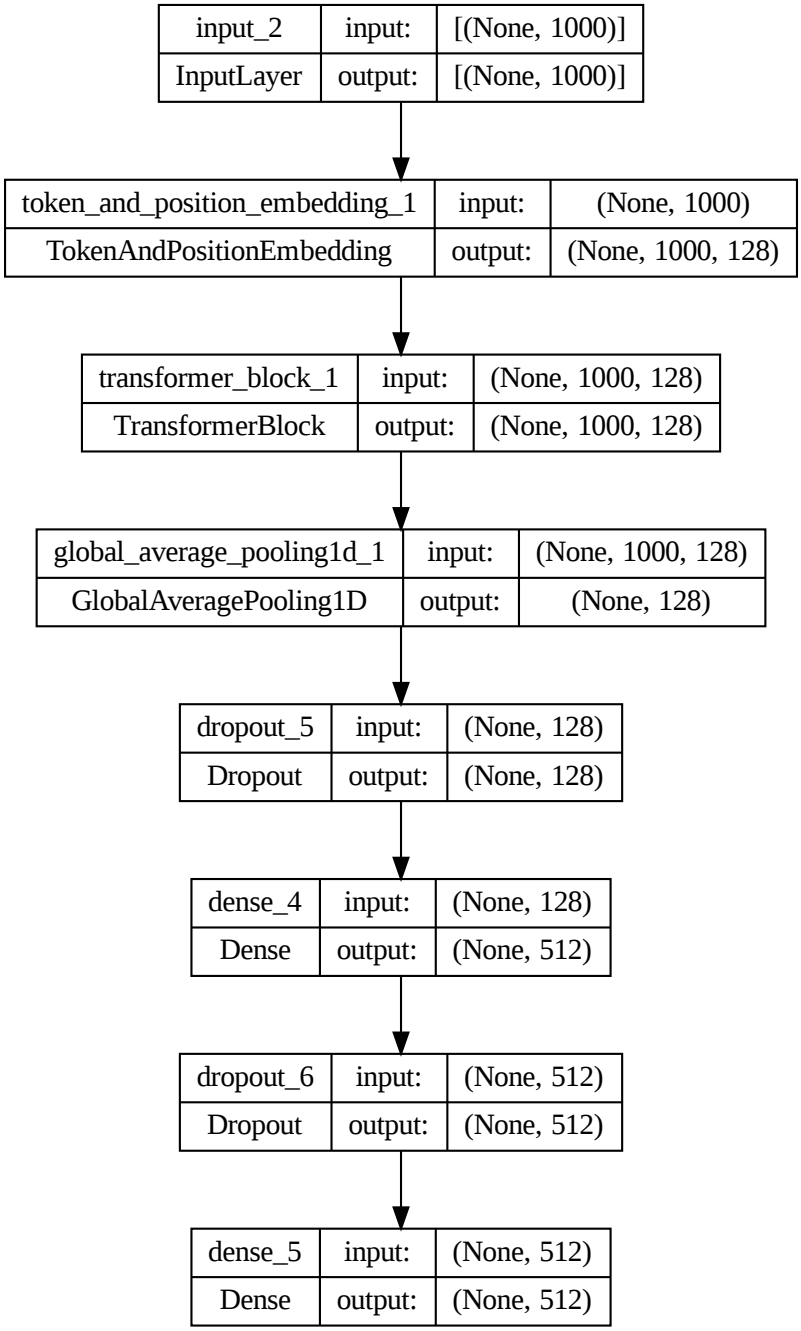


Figure 5.3: The architecture of transformer-based model.

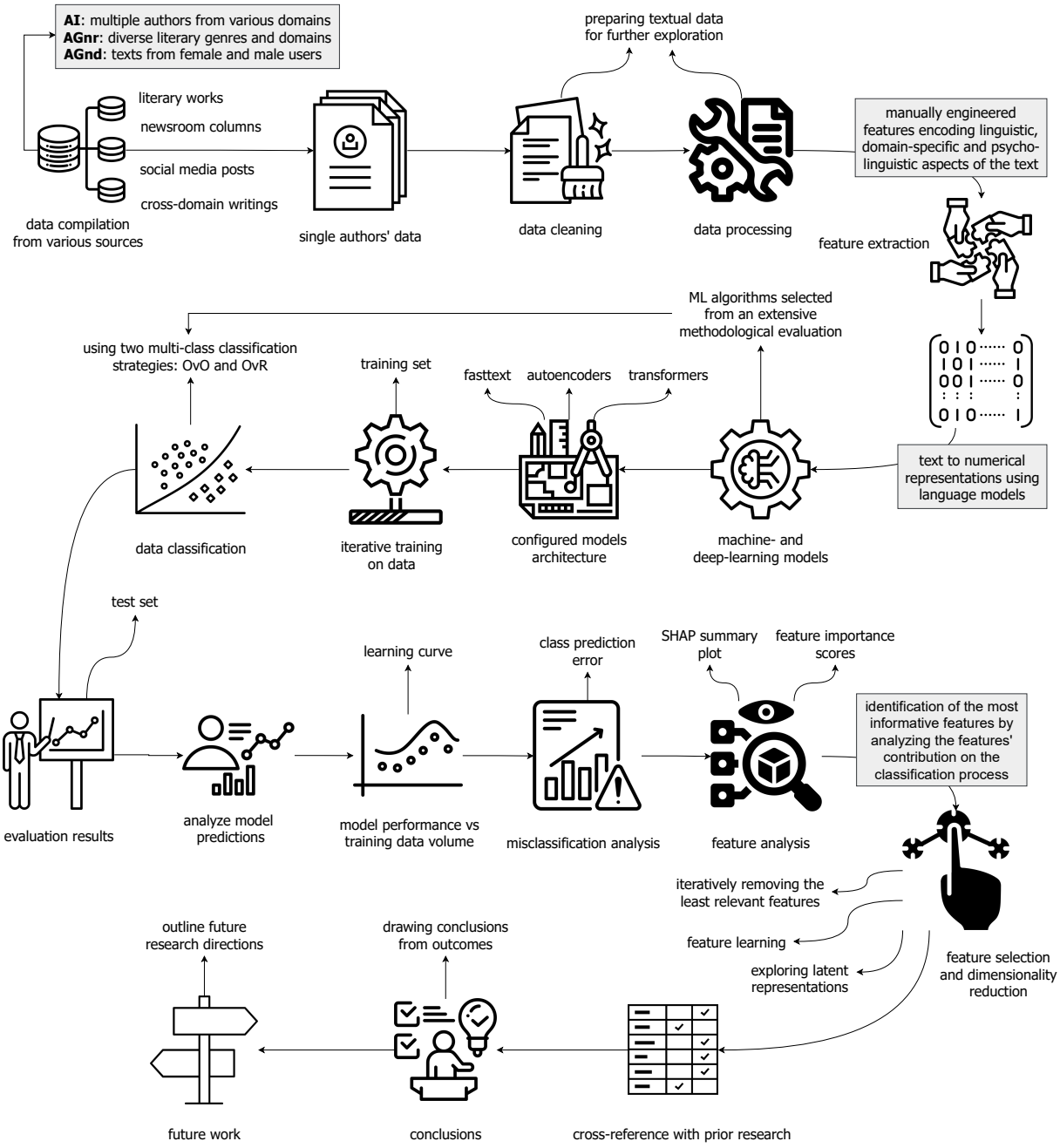


Figure 5.4: The methodology employed for authorship-related tasks.

Chapter 6

Authorship identification task in Albanian

This chapter aims to relate the theory and practical experimentation of authorship identification [106, 105, 107]. Section 6.2 discusses the foundational elements of our experiments, including the corpus, hand-crafted features, classification methods [120], and evaluation metrics for model evaluation. Section 6.3 presents the experimental results of our investigations, considering class prediction errors, feature analysis, and feature selection [174]. In Section 6.4, we compare with earlier studies to underscore the significance and originality of our work. The study's findings provide valuable insights and potential directions for future research in Section 6.5.

6.1 Introduction

Authorship identification is crucial in computational linguistics and text analysis. This task aims to identify unique authorship characteristics and patterns in written works to conclude their authorship. In previous chapters, we discussed the theory behind the task, which guides the experimental investigation and analysis of attributing authorship.

Authorship identification is a branch of research that combines linguistics, computational analysis, literature, and forensics to investigate the characteristics of human writing through text. Recent progress in machine learning, natural language processing, and computational linguistics has significantly transformed the methods for authorship identification.

Substantial prior research using various approaches has focused on authorship identification. Examining this volume of work helps understand field evolution and identify areas for future contributions.

Our research investigates manual feature engineering, machine learning, and deep learning in Albanian and addresses open issues to enhance authorship identification. The objective of this study is to assess the robustness and generalizability of classification models by determining authorship among a closed set of authors from different linguistic contexts, genres, and time periods.

Our experiments aim to evaluate the effectiveness of different algorithms and methodologies in authorship identification and to understand the elements that define authors' writing styles. This chapter revisits the research questions and hypotheses from Chapter 1.

1. Can the author be identified from an Albanian text?
2. Which linguistic features are crucial to determine the author's writing style?
3. What is the best feature set identifying the author of a text document?
4. What are the best methods for authorship analysis in Albanian?
5. How much data is sufficient to recognize authorial uniqueness?

This research contributes by compiling a novel corpus for classifying authorship, using diverse textual works for analysis, and demonstrating its potential across specific contexts. A significant contribution is the manual engineering of hand-crafted features that, through experiments, uncover authorial patterns in Albanian texts and provide valuable insights into language-related characteristics. Our research methodology involves cross-validation with previous studies to demonstrate the relevance and originality of our contributions to authorship analysis.

6.2 Experimental setup

This section outlines the experimental setup, corpus, classification methods, and selected features for defining authors' writing characteristics and evaluating the methods' effectiveness.

6.2.1 Materials

This subsection provides a detailed overview of dataset characteristics, hand-crafted features, and various models used for textual data representation.

6.2.1.1 Data preparation

Chapter 3 provides a detailed overview of the A3C3 corpus. This section outlines the corpus's statistics about various NLP attributes. The A3C3 corpus is a diverse collection of newsroom columns, social media posts, and literary writings, as detailed in Table 6.1.

Table 6.1: The corpus size and authorship statistics of the A3C3 corpus.

Level / Features	**Literary**	**Columns**	**Posts**	**All writings**
Corpus size				
total number of samples	576	569	426	1,107
average number of words	7,295.61	8,608.14	8,168.46	10,562.32
average number of unique words	308.05	252.41	307.49	265.89
average word length (in characters)	5.79	5.92	6.18	5.94
average sentence length (in words)	16.20	22.93	20.32	19.63
average unique words per sentence	0.68	0.67	0.77	0.49
average number of sentences	45.02	37.55	40.20	53.81
average syllables per word	1.91	2.05	2.14	2.02
max length word	69	36	98	98
Authorship-related task statistics				
total number of authors	30	52	44	122
average number of samples per author	19.20	10.94	9.68	9.07
average number of tokens per author	243.06	165.54	185.65	86.59
average number of unique words per author	10.27	4.85	6.99	2.18
POS-related information				
lexical density (proportion of content words)	49.94	49.16	50.40	49.82
average number of function words (TBN)	47.24	47.22	45.49	46.72
average number of verbs (TBN)	12.16	10.46	9.55	10.80
average number of nouns (TBN)	19.04	18.96	20.24	19.36
average number of adjectives (TBN)	5.63	6.48	6.66	6.23
average number of adverbs (TBN)	4.67	4.14	3.14	4.06
average number of pronouns (TBN)	4.05	4.76	7.06	5.13

TBN = Token-Based Normalization.

The A3C3 corpus is a valuable resource for research and analysis in Albanian. It contains 11,692,491 words and 294,284 unique tokens, representing a wide range of linguistic domains.

6.2.1.2 Combination of author-related hand-crafted features

An extensive set of language-related features have been manually engineered. This feature set mainly includes linguistic attributes because they can capture language usage patterns that the written texts manifest. The extensive set of linguistic features constitutes a foundational element in our approach.

6.2.1.3 Language models

This section describes text representation models, which convert raw text into numerical matrices for experimental analysis. These models include CountVectorizer, TF-IDF, and Word2Vec.

6.2.2 Classification methods and parameter settings

The experiments employ machine learning algorithms, including ensemble techniques, linear models, and Support Vector Machines. They use OvO and OvR multi-class classification strategies to address authorship identification with multiple classes. The study maintains model parameters in default configurations for comparative analysis. The research also includes state-of-the-art deep learning techniques like fastText, autoencoders, and transformers to understand new aspects of authorship patterns.

6.2.3 Evaluation metrics

The study used cross-validation to evaluate the effectiveness of the selected methods. The models were trained using a 5-fold technique, and their performance was evaluated using the F1 score. We applied the weighted averaging scheme of the F1 score was applied, due to class imbalance, with weights defined proportionally to instance counts for each class.

6.3 Author identification results

This section reports the results of our experiments, which were designed to evaluate the efficiency of different ML-based methods in identifying authorship patterns in Albanian texts. We examine the learning curves and errors in class predictions for each classification algorithm to evaluate their reliability, ability to generalize, and areas for improvement.

6.3.1 Experimental results

We provide an overview of the classification results from our experiments on the recently introduced A3C3 corpus. This dataset investigates data variations in attributing authorship to different text types, including literary works, newsroom columns, and social media posts. We also used a combined configuration to evaluate the reliability of our methods by incorporating all text data.

The study uses two levels of text analysis, including chapter- and paragraph-level, to evaluate the performance of classification methods with different text sizes. We employed eight distinct feature configurations to represent the characteristics of text data. This approach contributes to a comprehensive understanding of authorship identification by presenting our experimental findings in various aspects, methods, and contexts. The research aims to optimize authorship attribution in Albanian texts by defining a minimal feature set that reduces computational complexity and maintains accuracy.

Literary works. The first experiments were performed on literary works employing various classification methods and feature settings, as presented in Table 6.2. The TF-IDF model is the most effective for feature extraction and text representation in model-generated features. The model achieved a perfect F1 score of 1.00 using the OvR strategy, LR, and LinSVC algorithms at the chapter level. At the paragraph level, we notice a dropout of 0.05 using the LR algorithm with the TF-IDF model. A substantial drop in classification performance of 0.08 is observed when using the Word2Vec language model with the Ridge algorithm. This reduction in results can be attributed to the shorter length of literary texts at the paragraph level, which

increases the variability and negatively impacting the algorithms' classification performance. The CountVectorizer model performs similarly, while Word2Vec's performance is lower due to the lack of data, limiting generalization to new instances in smaller datasets. The Ridge algorithm with the OvR multi-class strategy obtains an F1 score of 0.997 at the chapter level with hand-crafted features. This trend demonstrates the model's efficacy in classifying authorship. The distinct genres of these literary writings ensure a clear distinction between them, which contributes to the higher and more consistent accuracy scores in the classification of literary works.

Table 6.2: Comparing accuracy: Performance of ML-based algorithms on literary works classification. Bold indicates the top-performing results in each experimental setup.

	One-versus-One					**One-versus-Rest**				
Level of analysis	**XGB**	**LR-CV**	**RC-CV**	**LinSVC**	**RF**	**XGB**	**LR-CV**	**RC-CV**	**LinSVC**	**RF**
Chapter-level										
TF-IDF	0.959	**0.999**	**0.999**	**0.999**	0.972	0.991	**1.000**	0.999	**1.000**	0.967
CountVec	0.964	0.988	0.996	0.993	0.893	0.995	0.999	0.999	0.999	0.985
Word2Vec	0.957	0.974	0.997	0.995	0.976	0.986	0.978	0.953	0.996	0.976
FE	0.968	0.979	**0.995**	0.967	0.953	0.992	0.979	**0.997**	0.984	0.980
Paragraph-level										
TF-IDF	0.914	0.996	**0.997**	0.995	0.886	0.976	0.941	**0.999**	**0.999**	0.963
CountVec	0.922	0.970	0.984	0.979	0.914	0.981	0.995	0.995	0.997	0.963
Word2Vec	0.925	0.955	0.984	0.980	0.915	0.952	0.962	0.870	0.982	0.919
FE	0.936	0.961	**0.977**	0.902	0.891	0.978	0.979	**0.989**	0.953	0.935

TF-IDF = Term Frequency-Inverse Document Frequency; **CountVec** = Count Vectorizer; **Word2Vec** = Word to Vector; **FE** = Feature Engineering (hand-crafted features); **XGB** = eXtreme Gradient Boosting; **LR-CV** = Logistic Regression Cross Validation; **RC-CV** = Ridge Classifier Cross Validation; **LinSVC** = Linear Support Vector Classification; **RF** = Random Forest.

Newsroom columns. Table 6.3 presents similar experimental settings applied to the newsroom columns subset of the original corpus. The table demonstrates results that are consistent with previous experiments. Ridge classifier accuracy decreases more at the paragraph level in newsroom columns due to shorter and fragmented paragraphs. In such a situation, the model struggles to capture relevant information.

Social media posts. Table 6.4 presents the experimental results in the dataset of social media posts. However, the outcomes are consistent with the findings observed in previous tables.

Table 6.3: Comparing accuracy: Performance of ML-based algorithms on newsroom columns classification. Bold indicates the top-performing results in each experimental setup.

	One-versus-One					One-versus-Rest				
Level of analysis	**XGB**	**LR-CV**	**RC-CV**	**LinSVC**	**RF**	**XGB**	**LR-CV**	**RC-CV**	**LinSVC**	**RF**
Chapter-level										
TF-IDF	0.891	0.993	**0.998**	**0.998**	0.947	0.972	0.958	**0.998**	0.997	0.936
CountVec	0.896	0.946	0.987	0.979	0.946	0.976	0.991	0.996	0.997	0.954
Word2Vec	0.871	0.956	0.987	0.977	0.921	0.937	0.975	0.896	0.981	0.936
FE	0.897	0.939	**0.976**	0.857	0.900	0.968	0.987	**0.992**	0.882	0.942
Paragraph-level										
TF-IDF	0.825	0.980	**0.987**	0.977	0.869	0.919	0.924	**0.994**	0.993	0.876
CountVec	0.819	0.932	0.937	0.929	0.864	0.927	0.961	0.983	0.989	0.861
Word2Vec	0.779	0.902	0.926	0.898	0.772	0.828	0.869	0.892	0.909	0.766
FE	0.837	0.878	**0.919**	0.792	0.783	0.918	0.945	**0.963**	0.865	0.838

TF-IDF = Term Frequency-Inverse Document Frequency; **CountVec** = Count Vectorizer; **Word2Vec** = Word to Vector; **FE** = Feature Engineering (hand-crafted features); **XGB** = eXtreme Gradient Boosting; **LR-CV** = Logistic Regression Cross Validation; **RC-CV** = Ridge Classifier Cross Validation; **LinSVC** = Linear Support Vector Classification; **RF** = Random Forest.

The brevity of social media posts resulted in a minor reduction in performance compared to other data sources.

Cross-domain. The final stage of our experiments is performed in a combined setting with all texts. Due to the large number of classes, the binary classifiers required by the OvO strategy increased exponentially. Consequently, we use the OvR strategy in the cross-domain scenario. Table 6.5 records the results of this evaluation, which are consistent with findings observed in the previous tables.

In the following series of experiments, we used three DL-based models, fastText, autoencoders, and transformers, to identify authorship. The results achieved in the A3C3 dataset are presented in Table 6.6. We observe lower F1 scores in DL-based models than the results with model- and manually-engineered features in various scenarios. It suggests further improvement in the architecture configurations and hyperparameter optimization to surpass traditional ML techniques. The transformer-based model shows decreased scores in newsroom columns and social media posts compared to previous results. Its effectiveness is evident in tasks with fewer classes due to data availability per class. The autoencoder-based model experiences a

Table 6.4: Comparing accuracy: Performance of ML-based algorithms on social media posts classification. Bold indicates the top-performing results in each experimental setup.

	One-versus-One					One-versus-Rest				
Level of analysis	**XGB**	**LR-CV**	**RC-CV**	**LinSVC**	**RF**	**XGB**	**LR-CV**	**RC-CV**	**LinSVC**	**RF**
Chapter-level										
TF-IDF	0.891	0.991	**0.996**	0.992	0.907	0.970	0.947	**0.999**	**0.999**	0.974
CountVec	0.903	0.946	0.977	0.970	0.906	0.976	0.979	0.998	0.996	0.968
Word2Vec	0.830	0.917	0.947	0.935	0.869	0.889	0.899	0.887	0.947	0.880
FE	0.860	0.879	**0.947**	0.770	0.851	0.949	0.972	**0.985**	0.788	0.902
Paragraph-level										
TF-IDF	0.814	0.963	**0.978**	0.967	0.814	0.903	0.903	0.991	**0.992**	0.915
CountVec	0.824	0.894	0.919	0.918	0.816	0.917	0.949	0.979	0.979	0.909
Word2Vec	0.735	0.816	0.863	0.844	0.744	0.778	0.788	0.830	0.856	0.730
FE	0.802	0.818	**0.853**	0.717	0.734	0.876	0.912	**0.945**	0.770	0.776

TF-IDF = Term Frequency-Inverse Document Frequency; **CountVec** = Count Vectorizer; **Word2Vec** = Word to Vector; **FE** = Feature Engineering (hand-crafted features); **XGB** = eXtreme Gradient Boosting; **LR-CV** = Logistic Regression Cross Validation; **RC-CV** = Ridge Classifier Cross Validation; **LinSVC** = Linear Support Vector Classification; **RF** = Random Forest.

Table 6.5: Comparing accuracy: Performance of ML-based algorithms on cross-domain classification. Bold indicates the top-performing results in each experimental setup.

Level of analysis	**Set of features**	**XGB**	**LR-CV**	**RC-CV**	**LinSVC**	**RF**
Chapter-level	TF-IDF	0.984	**0.999**	**0.999**	**0.999**	0.935
	CountVec	0.956	**0.999**	**0.999**	**0.999**	0.889
	Word2Vec	0.926	0.993	0.950	0.993	0.957
	FE	0.944	0.968	**0.994**	0.930	0.942
Paragraph-level	TF-IDF	0.890	**0.996**	0.970	0.994	0.754
	CountVec	0.908	0.993	0.954	**0.996**	0.760
	Word2Vec	0.831	0.970	0.846	0.962	0.849
	FE	0.879	0.873	**0.969**	0.916	0.816

TF-IDF = Term Frequency-Inverse Document Frequency; **CountVec** = Count Vectorizer; **Word2Vec** = Word to Vector; **FE** = Feature Engineering (hand-crafted features); **XGB** = eXtreme Gradient Boosting; **LR-CV** = Logistic Regression Cross Validation; **RC-CV** = Ridge Classifier Cross Validation; **LinSVC** = Linear Support Vector Classification; **RF** = Random Forest.

reduction of 0.155 in the cross-domain scenario in both levels of text analysis. On the other hand, the fastText model demonstrates its ability to capture meaningful information and generalize effectively in different settings. The model yields comparable results to those reported

in ML-based experiments.

Table 6.6: Comparing accuracy: Performance of DL-based algorithms on cross-domain classification. Bold indicates the top-performing results in each experimental setup.

Models	**fasttext**		**autoencoders**		**transformers**	
Level of analysis	**Chapter**	**Paragraph**	**Chapter**	**Paragraph**	**Chapter**	**Paragraph**
Literary works	0.994	0.975	0.996	0.970	0.993	0.953
Newsroom columns	0.972	0.940	0.982	0.903	0.868	0.754
Social media posts	0.946	0.901	0.973	0.899	0.874	0.786
All writings	0.980	0.951	0.956	0.801	0.881	0.863

The Ridge algorithm was selected for further investigation due to its consistent performance in different configurations. In the chapter-level analysis, we examined the algorithm's effectiveness using the OvR multi-class classification strategy in four different scenarios. The analysis excludes words that can indicate authorship, such as n-grams and capitalized words, from the engineered features and language models. These words correspond to artistic characters or prominent figures from online sources. This experimentation considers the effect of these words on classification, with results presented in Table 6.7.

Table 6.7: Comparing accuracy: Performance of the Ridge classifier on cross-domain classification.

Form of analysis	**no-ngrams**	**no-uppercase**		
Data	**FE**	**TF-IDF**	**CountVec**	**Word2Vec**
Literary works	0.989	0.994	0.993	0.967
Newsroom columns	0.979	0.982	0.983	0.953
Social media posts	0.936	0.968	0.943	0.815
All writings	0.980	0.970	0.936	0.922

The study found no significant decrease in results after removing n-grams from manually-crafted attributes. We achieved the most reliable performance in the literary works dataset. The impact of reducing the usage of capitalized words was evident in different scenarios. The Word2Vec model demonstrated a different trend in literary works and newsroom column datasets. The improvement in the model's performance was evident after excluding the capitalized words from the text. With the exclusion of these words, the model focused on linguistic patterns and

contextual relations, enhancing its ability to capture more significant details. The experimental findings reveal valuable insights into the performance of different methods across various configurations.

6.3.2 Learning curves

This subsection uses learning curves to illustrate the relationship between training and cross-validated test scores. The Ridge algorithm with the OvR strategy is used on different training sample sizes. As illustrated in Figure 6.1, the training result consistently outperforms the validation score, reducing the difference over learning. This improvement can be attributed to increased training data and diverse examples, enabling the model to capture more relevant patterns in data points.

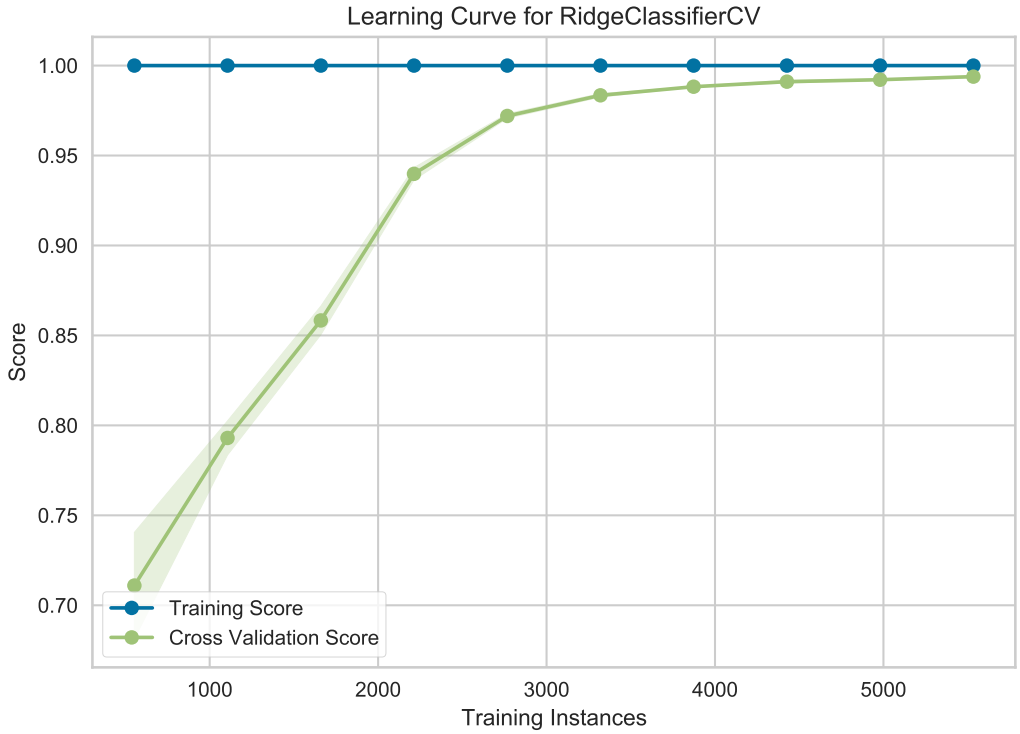


Figure 6.1: Learning curve for author-based all writings chapter-level classification using Ridge classifier and feature engineering.

The figure represents the learning curve for cross-domain classification and indicates a consistent improvement in the learning capability of the model in other datasets.

6.3.3 Class prediction error analysis

This subsection investigates the performance of ML-based models in accurately identifying authorship through the analysis of class prediction errors. This study introduces a new confusion matrix to represent classification model performance, which shows correct and incorrect predictions for each author. Considering the large number of classes, traditional confusion matrices would be complex and unmanageable, so continuous error plots are used to communicate information.

Literary works. Our model shows reliable performance, demonstrating its effectiveness in handling diverse data instances. It is supported by the low error rate in the literary works dataset presented in Figure 6.2. The analysis highlights the generalization ability of the model and the high accuracy of the features and techniques. Continuous error analysis in literary works demonstrates some struggles to accurately attribute authorship within the same genres due to similar language-related patterns and narrative styles.

Newsroom columns. In newsroom columns, errors in classification are associated with a specific author, as presented in Figure 6.3. It shows that the model has difficulty appropriately identifying the author's writing style. It can be attributed to similar linguistic elements and particular topics. The model accurately differentiates between columns of other authors, considering potential misclassifications due to shared linguistic attributes, specific topics, and limited training data for particular authors.

In literary works and newsroom columns, misclassifications were reported within the same gender category without confusion between male and female authors.

Social media posts. The distribution of errors in Figure 6.4 suggests that authors on social media focus on similar topics or subject matter, consequently confusing in classifying authorship.

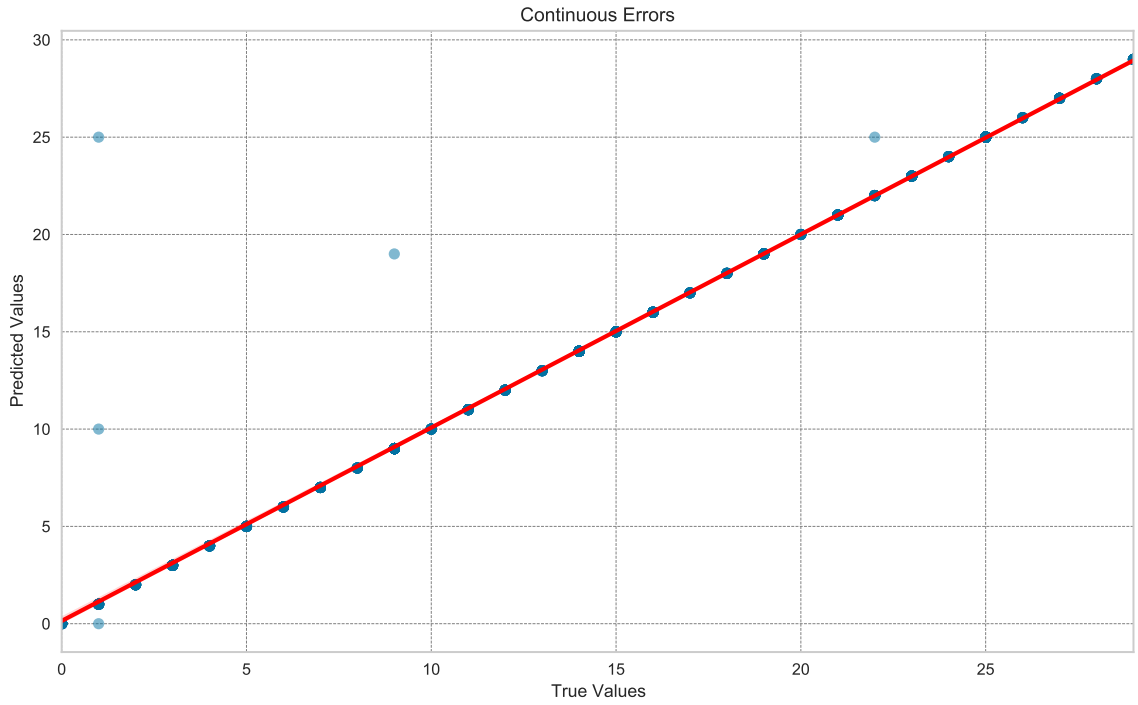


Figure 6.2: Continuous error for author-based literary chapter-level classification using Ridge classifier and feature engineering.

Cross-domain. The incorrect classifications in the cross-domain scenario shown in Figure 6.5 mainly occur between authors within the same dataset. This trend demonstrates authors' unique creative and figurative language in their literary writings, distinct from the formal language used in newsroom columns and informal communication used in social media posts.

Literary language, distinct from other content types, can sometimes be confused with journalistic expressions in specific genres like analysis and criticism. Political opinions may result in misclassifications across columns and posts. However, formal language standards and similar thematic considerations can result in misclassifications in the cross-domain context.

6.3.4 Feature analysis

This study examines the importance of features in authorship classification using SHAP values and waterfall plots for each dataset. The SHAP waterfall plots visualize feature significance,

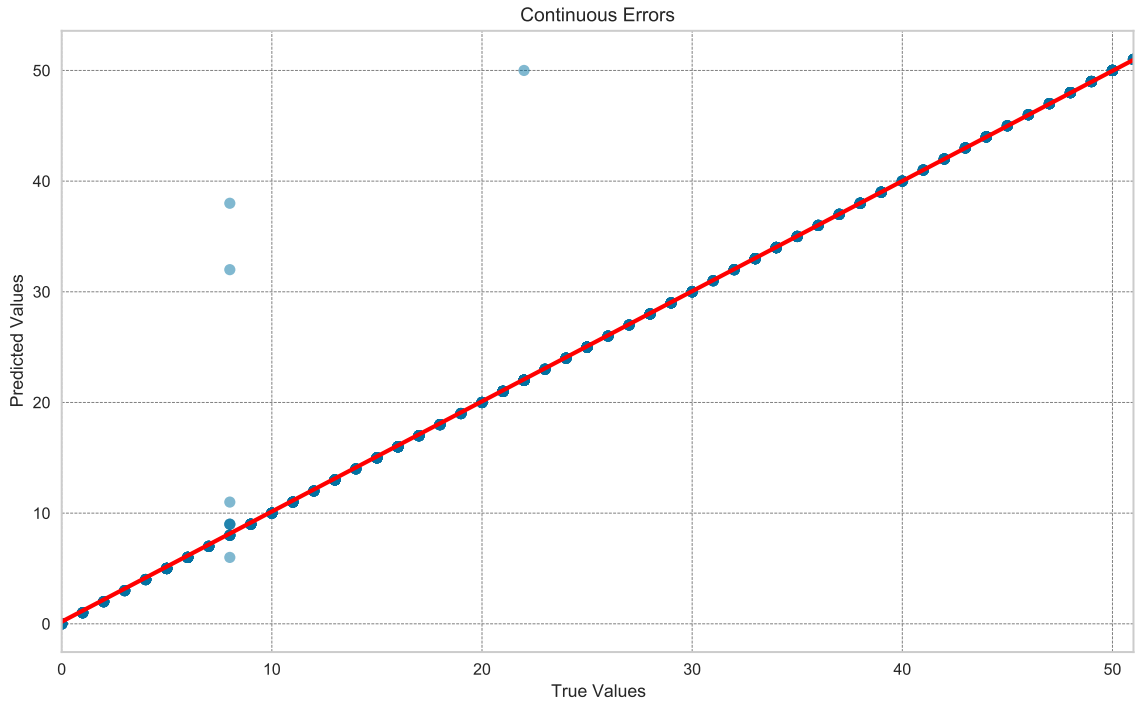


Figure 6.3: Continuous error for author-based columns chapter-level classification using Ridge classifier and feature engineering.

with each bar representing the features' impact on the model's output for a specific input. The plot shows that certain features significantly influence the model's predictions.

Literary works. The analysis of literary works reveals that linguistic features like consonants, syntactic phrases, and punctuations significantly contribute to authorship identification, as illustrated in the SHAP waterfall plot in Figure 6.6.

The authors displayed unique patterns in consonant usage, highlighting the unique phonetic and stylistic variations that differentiate them in the literary domain.

Syntactic phrases, including nominal, verbal, and prepositional phrases, are crucial for structuring sentences in natural language. They provide helpful information about how authors create their works. Different authors demonstrated various levels of complexity and variance in their syntactic structures, highlighting different stylistic preferences.

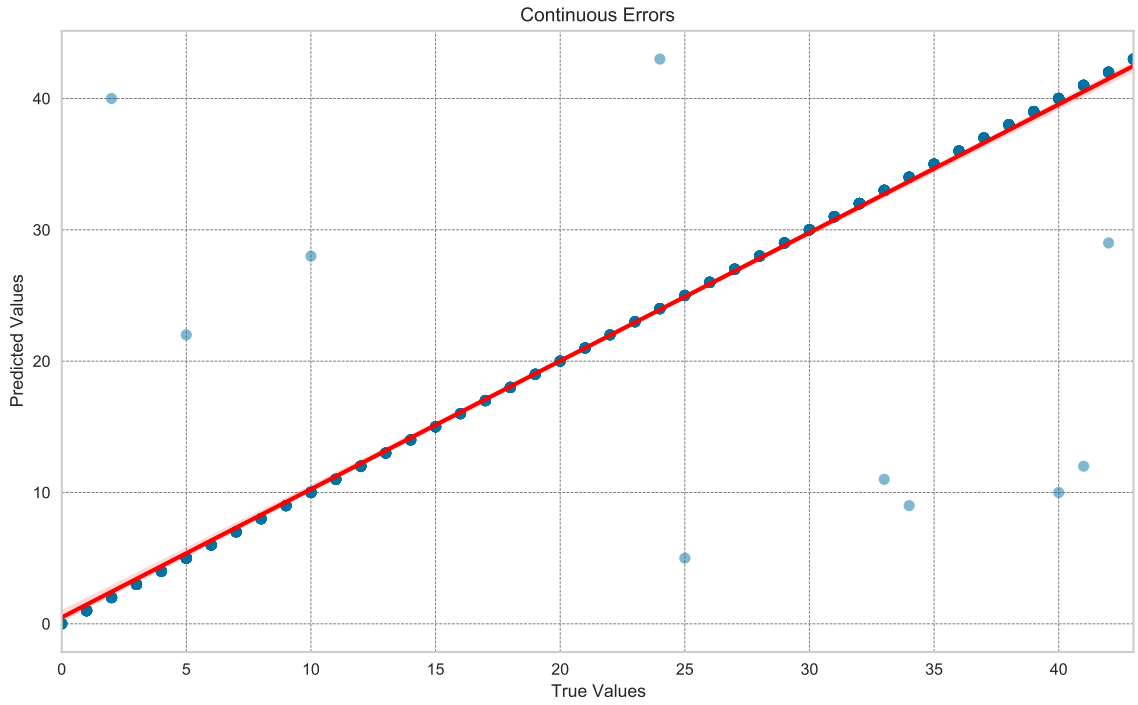


Figure 6.4: Continuous error for author-based posts chapter-level classification using Ridge classifier and feature engineering.

Exclamation marks can represent expressive tendencies, and commas may indicate sentence construction and punctuation style preferences. These features reflect the author's artistic and creative aspects.

According to the SHAP analysis, specific characteristics negatively impact the model's predictions; this indicates a recurring pattern across authors. A variable that demonstrates a negative impact is the 'punctuations Herfindahl index' variable, which suggests the presence of a shared writing standard or stylistic preference.

In general, the analysis reveals that linguistic features, such as phonetic attributes, syntactic elements, and punctuation habits, significantly influence the determination of authorial patterns in literary texts. These features provide a deeper understanding of authors' writing habits and styles, highlighting the importance of linguistic characteristics in literary texts and authorship identification.

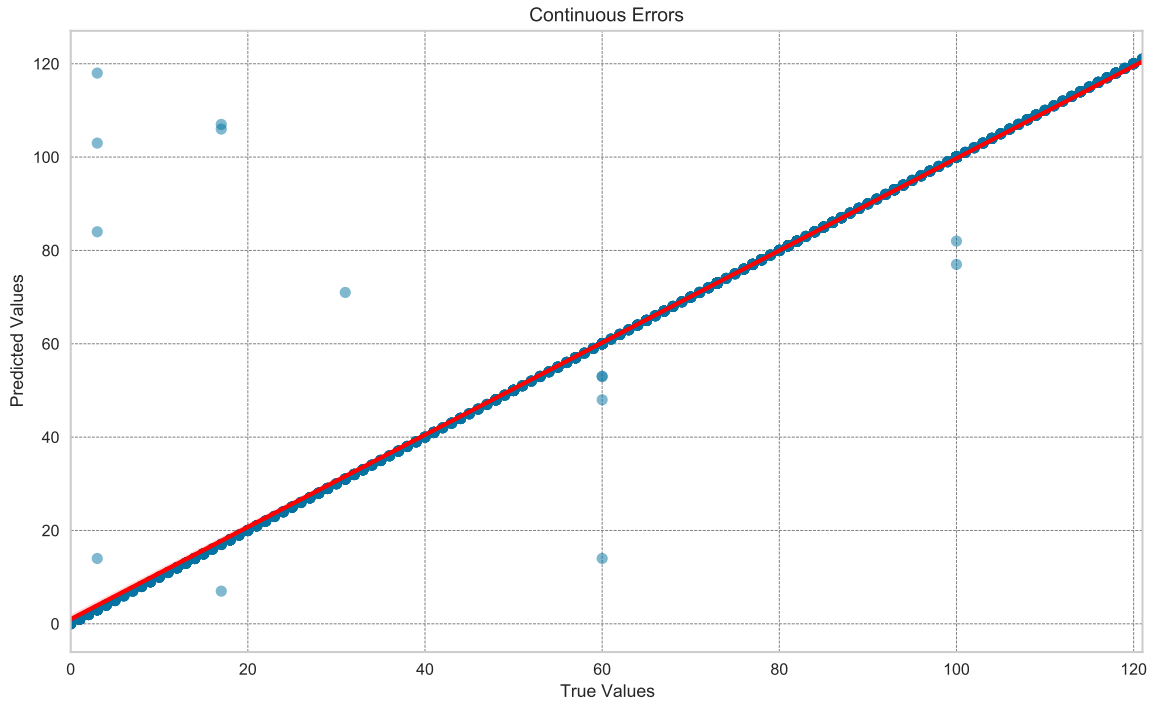


Figure 6.5: Continuous error for author-based all writings chapter-level classification using Ridge classifier and feature engineering.

Newsroom columns. The "exclamation" feature and other lexical and morphological features, indicated an impact on the model's predictions in newsroom columns. It is shown in the SHAP waterfall plot in Figure 6.7.

Uppercase digraphs, stopword entropy, and complex words significantly influence the model's performance in classifying newsroom columns. The exclamation variable is essential in representing the emotional elements of the text. It suggests that columnists use this punctuation to change the mood and tone of their writing.

Stopword entropy shows the variation in using stopwords for different authors, highlighting nuances in writing style and content. The 'percentage of complex words' feature characterizes the proportion of complex words in the text and suggests variations in linguistic complexity in newsroom columns.

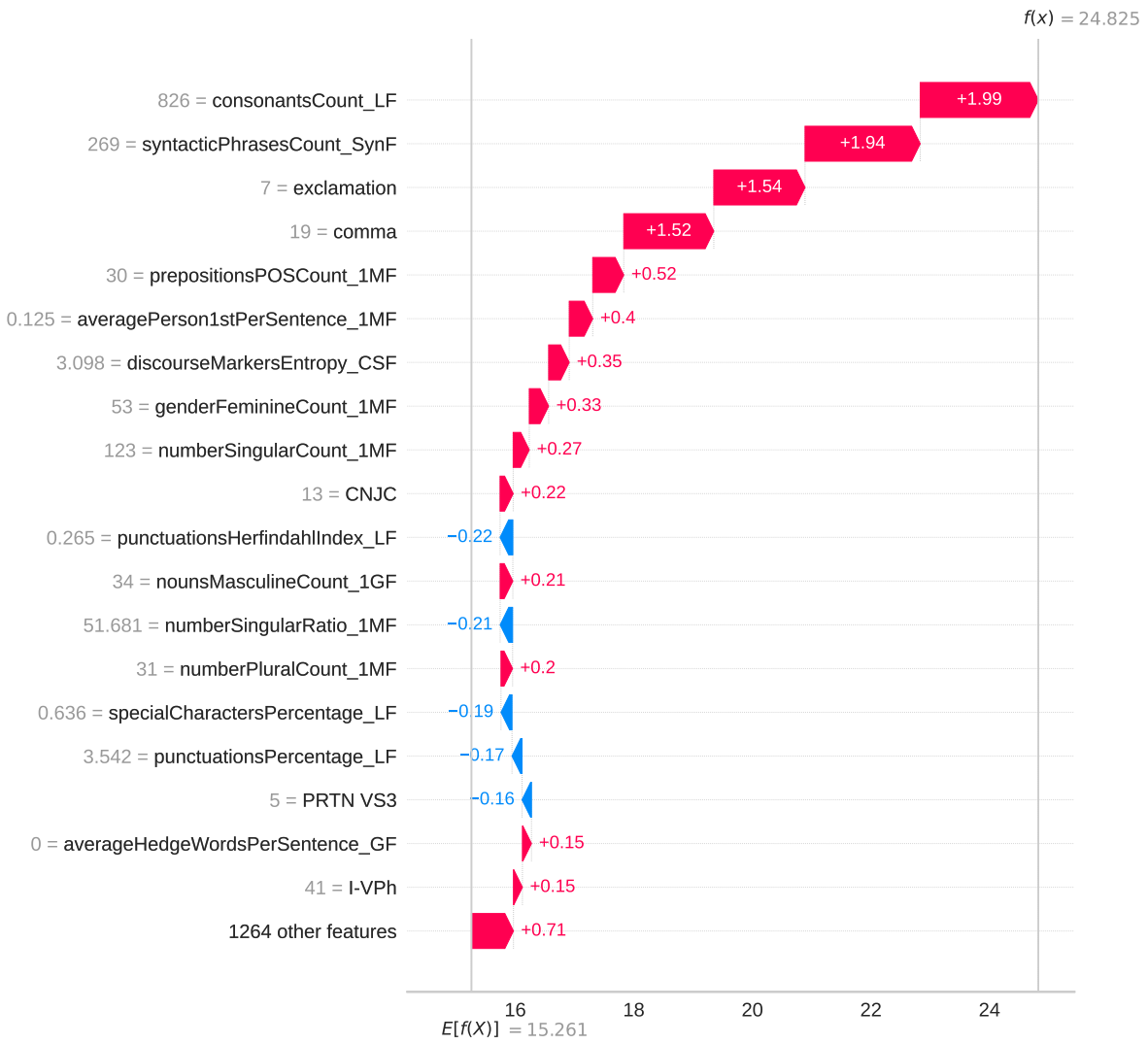


Figure 6.6: SHAP waterfall plot for author-based literary chapter-level classification using XGBoost classifier and feature engineering.

The high average of hapax legomena per sentence negatively impacts the model's prediction. This feature reflects the diversity of vocabulary used in the text, making it difficult for the model to capture a consistent pattern specific to an author and attribute the text to that individual writer.

The newsroom column corpus demonstrates unique linguistic and stylistic elements, each influencing the model's predictions. These attributes contribute essential information for successful authorship attribution in the newsroom column dataset.

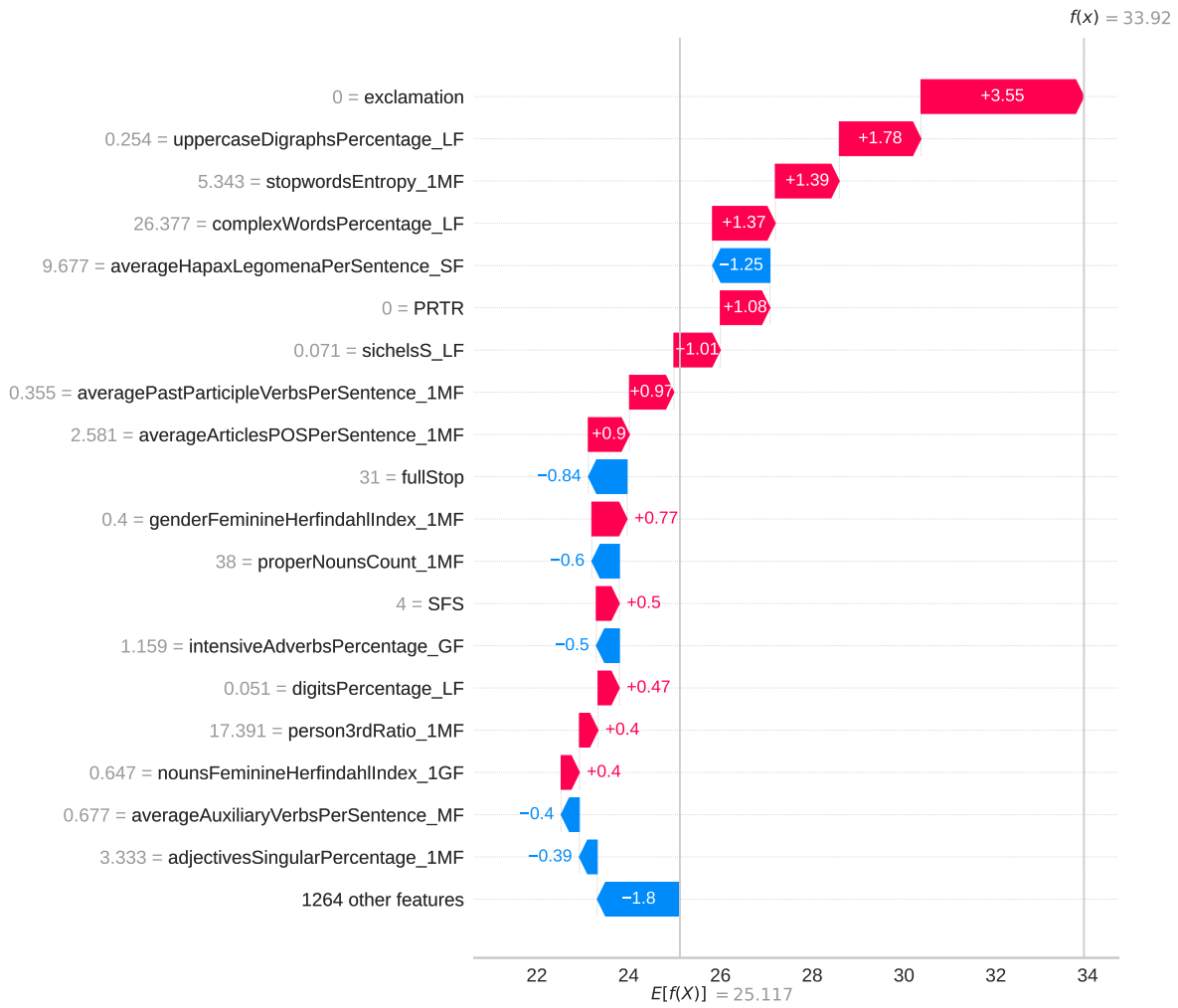


Figure 6.7: SHAP waterfall plot for author-based columns chapter-level classification using XGBoost classifier and feature engineering.

Social media posts. Morphological, lexical, and structural features are essential in the model's predictive decision in social media texts. Figure 6.8 shows the SHAP waterfall plot adapted explicitly for social media content.

The 'third-person ratio' measures the proportion of third-person POS tags. This feature has a significant impact on the model's performance. A higher score indicates a distinct characteristic of authors in social media texts regarding their unique narrative perspectives.

The distribution of negative particles in social media texts can influence the tone and sentiment of the text, indicating a pessimistic or critical attitude.

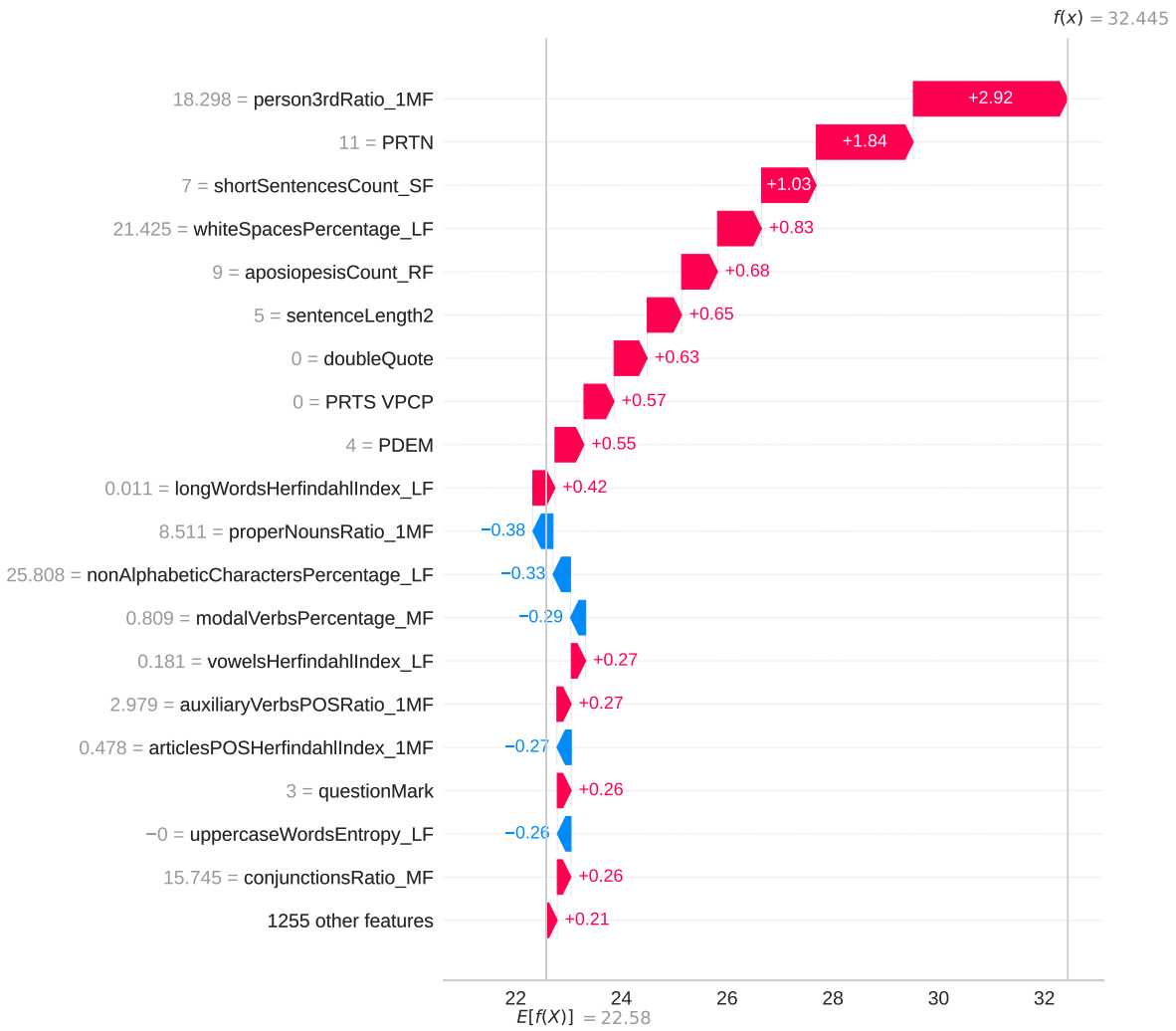


Figure 6.8: SHAP waterfall plot for author-based posts chapter-level classification using XGBoost classifier and feature engineering.

Short sentences significantly affect readability and attention in social media content, suggesting a dynamic or concise communication style.

The percentage of white spaces in the text can indicate organization, format, or layout preferences, characterizing the overall structure and content presentation to the public. The use of proper nouns varies substantially in social media posts. Authors use a diverse array of proper nouns corresponding to different topics, making it more difficult for the model to attribute authorship adequately.

The study highlights the significance of linguistic and structural features in understanding

social media writing and authorship styles. It emphasizes the role of conciseness, sentiment, and attention in attributing written work to distinct authors, providing insights into the unique elements that define their writing style.

Cross-domain. The cross-domain analysis reveals that linguistic and structural features significantly contribute to the model's predictions. Figure 6.9 presents the SHAP waterfall graph, illustrating feature significance analysis for the mixed scenario.

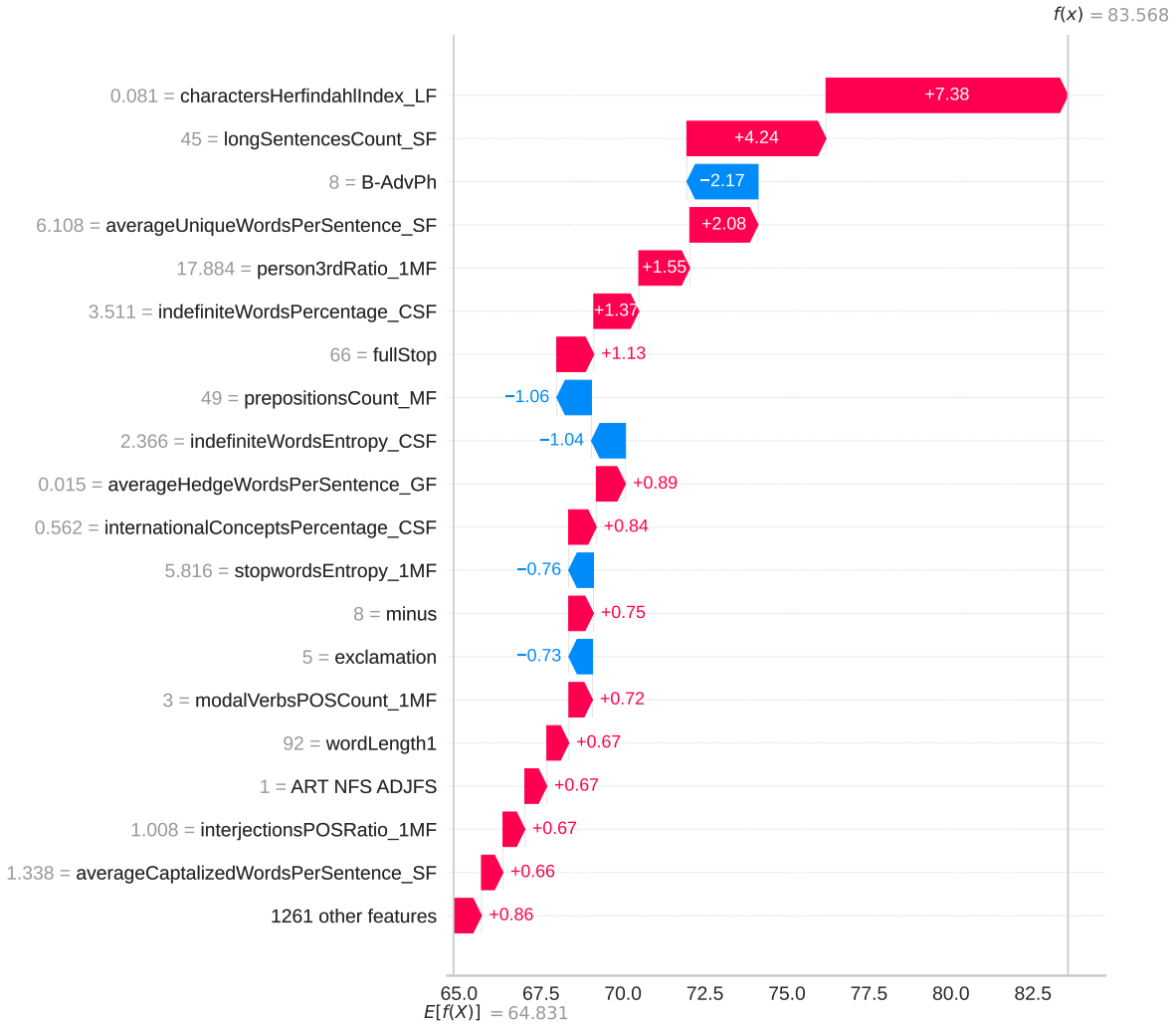


Figure 6.9: SHAP waterfall plot for author-based all writings chapter-level classification using XGBoost classifier and feature engineering.

The 'characters Herfindahl index' measures the distribution and concentration of specific characters, indicating dominant usage within authors' writings. It can suggest unique writing

styles and influence the model's performance in classifying authorship.

The 'long sentence count' affects the text's structure and complexity, as authors may have distinct habits in crafting long and complex sentences, demonstrating the importance of sentence structure in different contexts.

The analysis reveals that certain features are crucial in authorship attribution for different text types, as they represent fundamental aspects of an author's style in different scenarios. This information helps with successful attribution in the cross-domain corpus and extends our understanding of authorship classification.

The visualization of feature importance to the model decision-making process through SHAP summary plots for each dataset is provided in Appendix C.1.3. Feature importance scores obtained using the XGBoost ranking system are presented in Appendix B.5.

However, certain features have a minimal effect on the model's predictive results. These features characterize common linguistic characteristics and provide less informative elements. The following section explores some feature selection techniques to minimize the presence of these features and reduce the model's complexity without affecting its performance.

6.3.5 Feature selection

To improve the model's effectiveness, we performed feature selection evaluations on four dataset scenarios: literary works, columns, social media posts, and a cross-domain dataset. The objective was to determine the most relevant subsets of features to optimize model performance. We used the Ridge classifier with hand-crafted features and applied RFE to the cross-domain scenario.

RFE is an iterative feature elimination process that repeatedly examines the model's performance. The x-axis corresponds to the number of features, and the y-axis represents the cross-validation score of the model. The RFE approach highlights performance evolution with each gradual increase in the number of features.

A significant element of these curves is the determination of two key points. First, we define the exact number of features with which the Ridge model achieves maximum performance.

Second, we set the threshold number of features that indicate a tolerance drop in performance of 0.02 from the peak, marked with a dashed line on the curves. This threshold demonstrates the compromise between feature inclusion and model performance reliability.

The figures below (Figure 6.10, Figure 6.11, Figure 6.12, and Figure 6.13) illustrate the relationship between features and the model's performance in literary, column, social media, and cross-domain contexts, respectively. These figures provide valuable information on how feature selection affects the Ridge model's performance on each dataset. We can identify the minimal number of features needed for effective model performance and determine the impact of feature space on model reliability.

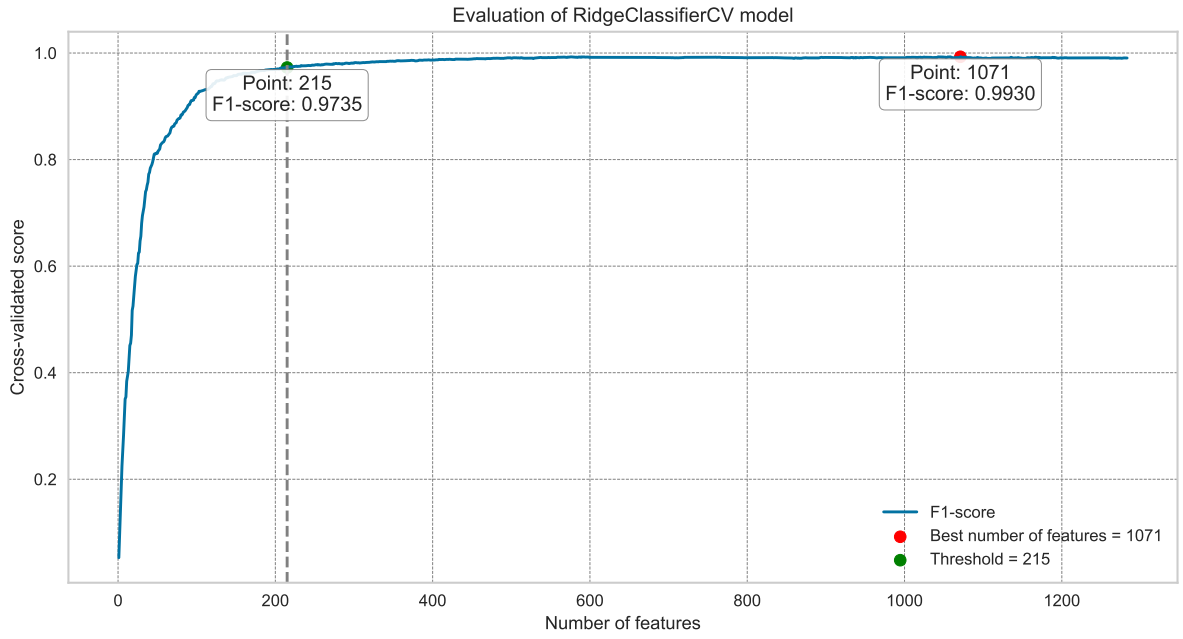


Figure 6.10: RFE for author-based literary chapter-level classification using Ridge classifier and feature engineering.

We also provide a table with the results of the RFE approach to feature selection across different settings. The information presented in Table 6.8 indicates the exact number of features with the best F1 score for each dataset. The threshold set on the model F1 score with a drop of 0.02 defines the number of features for optimal model outcomes.

The study aimed to identify informative feature subsets (Appendix B.5) in different text data using RFE. The analyses demonstrate the impact of feature selection on the Ridge algorithm's

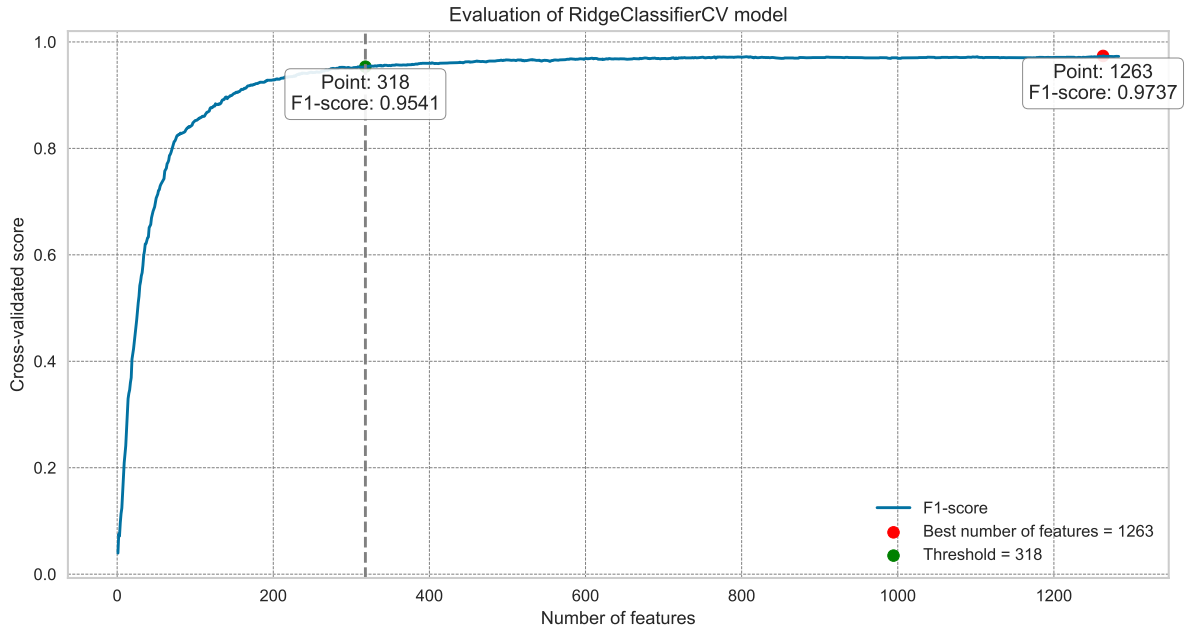


Figure 6.11: RFE for author-based columns chapter-level classification using Ridge classifier and feature engineering.

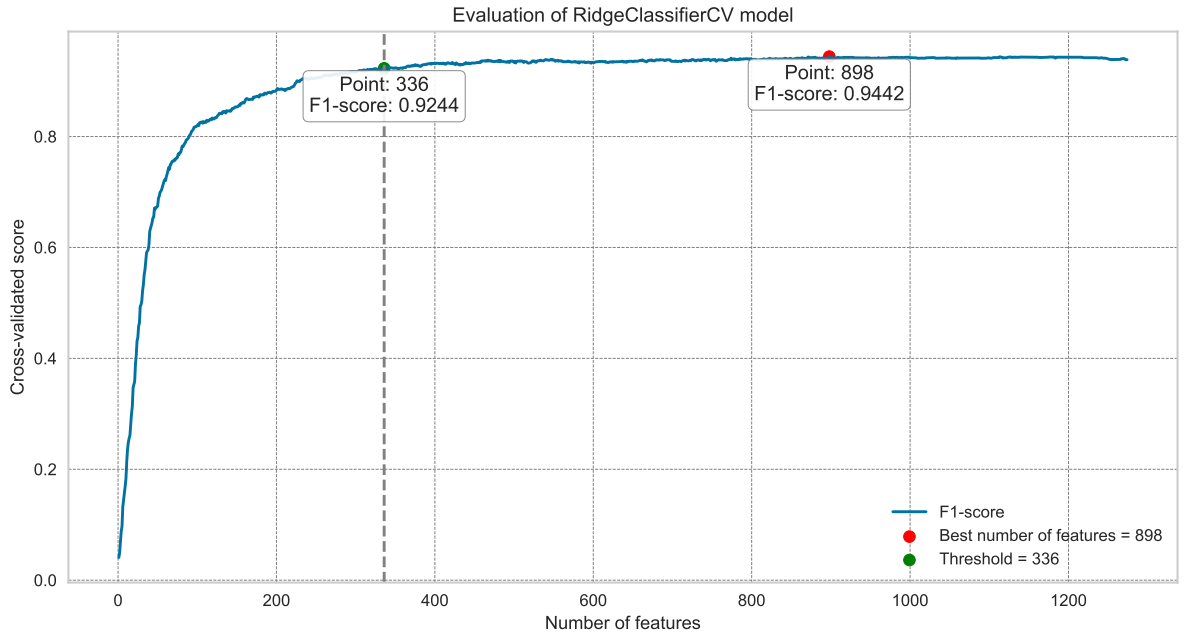


Figure 6.12: RFE for author-based posts chapter-level classification using Ridge classifier and feature engineering.

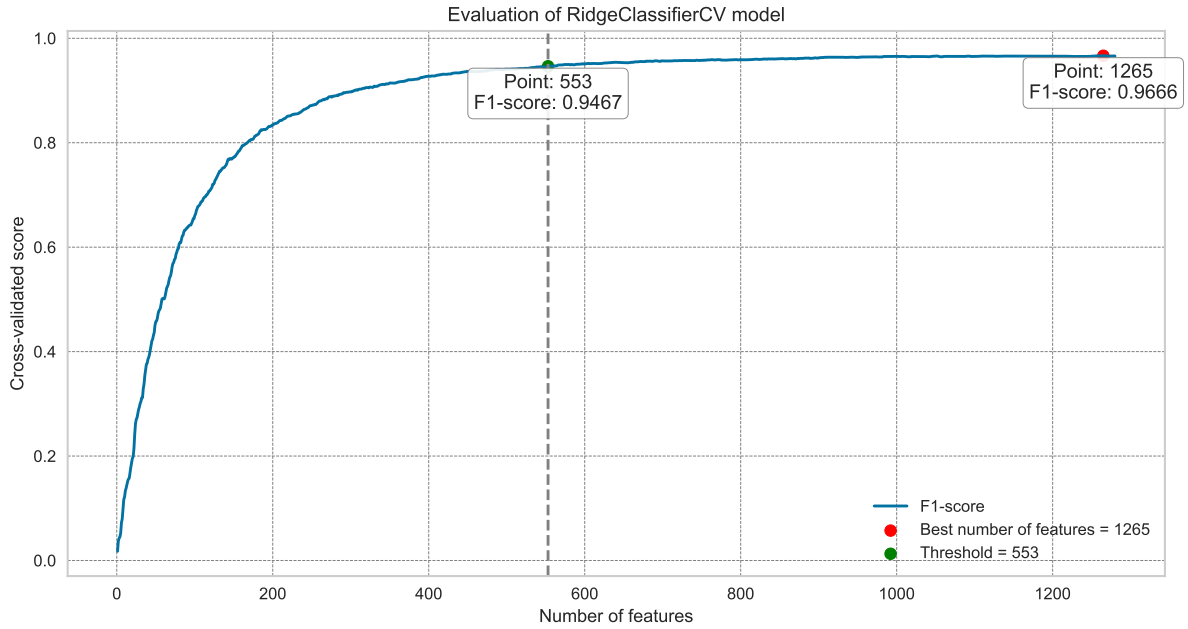


Figure 6.13: RFE for author-based all writings chapter-level classification using Ridge classifier and feature engineering.

Table 6.8: Feature selection outcomes using RFE in a cross-domain configuration.

Corpora	Number of features	F1-score	Number of features (threshold)	F1-score (threshold)
Literary works	1071	0.993	215	0.974
Newsroom columns	1263	0.974	318	0.954
Social media posts	898	0.944	336	0.924
All writings	1265	0.967	553	0.947

performance, providing insights into optimal feature sets for each dataset. This investigation validates the Ridge model's reliability and performance in authorship attribution, enhancing its effectiveness in this task.

6.3.6 Dimensionality reduction

This section presents the process of dimensionality reduction using PCA and autoencoder models. Similar to the feature selection section, we present the results of these techniques in the tables below.

We first performed experiments in the PCA model (Table 6.9) to define the number of components needed to maintain 1.0 of the data's total variance. Next, we determined the number of components with the highest F1 score and set a threshold with a performance drop of 0.02 from the peak. Appendix C.1.1 provides the performance curves illustrating the model performance scores for the number of components in each dataset.

Table 6.9: Dimensionality reduction outcomes using PCA in a cross-domain configuration.

Corpora	Number of components - PCA cumulative variance=1.0	Number of components (best)	F1-score (best)	Number of components (threshold)	F1-score (threshold)
Literary works	517	502	0.985	281	0.969
Newsroom columns	537	532	0.943	443	0.924
Social media posts	532	516	0.919	293	0.900
All writings	516	507	0.853	477	0.837

The autoencoder model uses the same approach for dimensionality reduction (Table 6.10). We include the complete feature set in this process and constantly increase the number of features by 50. The performance curves for the autoencoder model, which show the model's effectiveness with different feature subsets for each dataset, are provided in Appendix C.1.2.

Table 6.10: Dimensionality reduction outcomes using autoencoders in a cross-domain configuration.

Corpora	Best point	Best F1-score	Threshold point	Threshold F1-score	Epochs, early stopping (avg)
Literary works	650	0.996	100	0.987	24
Newsroom columns	1100	0.988	200	0.972	21
Social media posts	1000	0.981	150	0.965	28
All writings	1200	0.990	450	0.971	20

As the tables show, the autoencoder model experiences superior performance with a limited set of features and minimal training epochs enabled by the early stopping mechanism.

The PCA model enables us to understand the structure and relationship between the data by reducing the feature space.

Figure 6.14 shows the PCA model visualization for literary texts. It represents the distribution of data clusters in a two-dimensional space determined by the first two principal components (PC1 and PC2). The visualization presents the structure of the data and simplifies the identification of possible literary clusters or particular writing patterns.

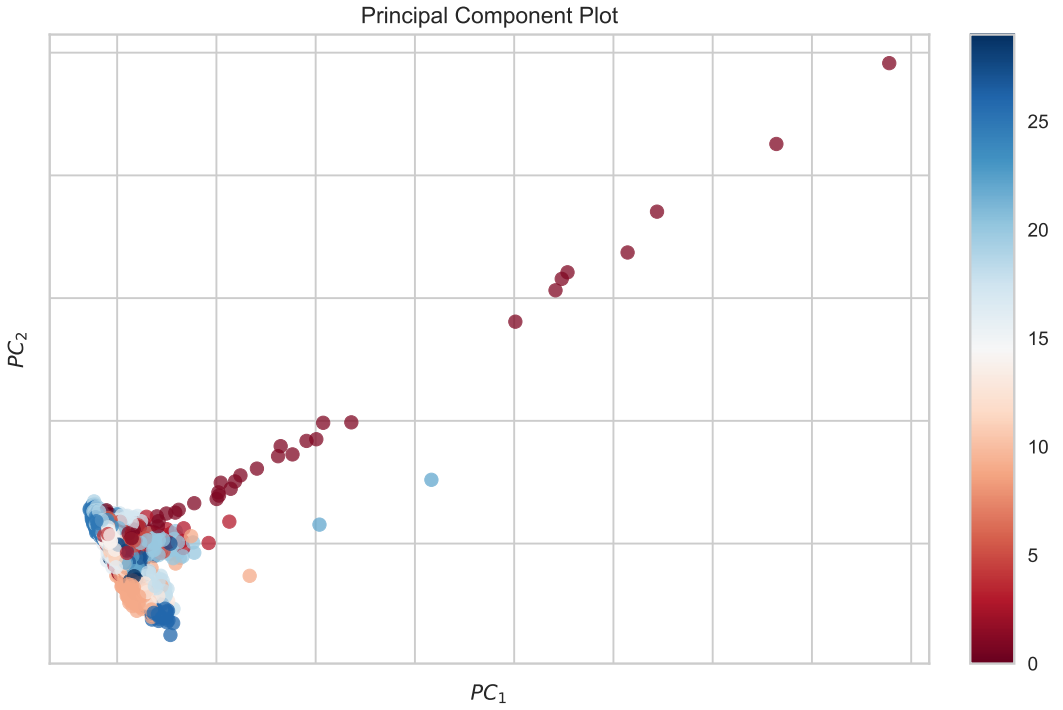


Figure 6.14: Illustration of literary works data via PCA dimensionality reduction.

The analysis exhibits the presence of overlapped clusters in newsroom columns and social media posts. These findings, illustrated in Figure 6.15 and Figure 6.16, indicate similar topic dominance distinct from the literary scenario. The data points from newsroom columns and social media posts show recurring discussion patterns in both domains.

The cross-domain scenario in Figure 6.17 shows a significant overlap among data points, indicating a shared variation in authorship patterns across different contexts.

This subsection explored the application of PCA and the autoencoder to the dimensionality reduction process while maintaining relevant information. The selection of the dimensionality reduction technique influenced the model performance and provided significant implications for further research.

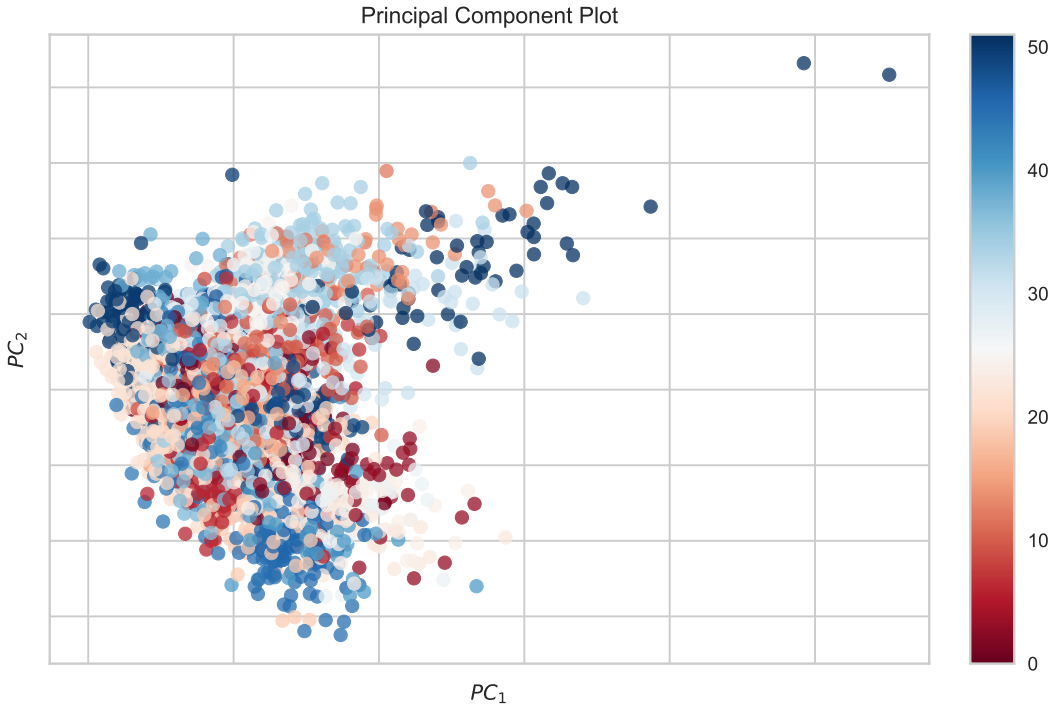


Figure 6.15: Illustration of newsroom columns data via PCA dimensionality reduction.

6.4 Comparison with previous works

Cross-validation with previous studies is critical to our research methodology, comparing our experimental findings to those reported in previous research (Table 6.11). This investigation discusses the reliability, originality, and improvement of our results within the field of authorship attribution in Albanian. This analysis indicates the ongoing nature of the research and the significance of our contributions.

Table 6.11: Comparative analysis with prior studies in cross-domain classification settings (top-performing results).

Reference	Document type	Language	Number of authors	Number of samples	Features	Method	Performance
[79]	literary works	Bengali*	16	17,966	sub-word features	Transfer learning	0.998 F1 score

(continued on next page)

Table 6.11: Comparison with previous works (continued).

Reference	Document type	Language	Number of authors	Number of samples	Features	Method	Performance
[79]	online posts	Bengali*	6	2,100	sub-word features	Transfer learning	0.9536 F1 score
[69]	fragments of novels	English	25	250	syntactic features	CNN, LSTM	0.9888 accuracy
[118]	books	Albanian*	4	43	syntactic structure, linguistic elements	Dmitri Khmelev algorithm	0.9302 accuracy
[139]	online documents	Thai*	200	547	lexical, structural, syntactic features, PoS tags	StyloThai-LSS	0.9102 accuracy
[64]	books	Arabic*	7	7	lexical, structural, semantic, syntactic features, n-grams, function words, assonance	Logistic Model Tree	0.8873 accuracy
[15]	articles	Urdu*	15	6,000	lexical features, n-grams, topic modeling	LDA, SQRT-cosine similarity (proposed method)	0.93 F1-measure
Our method	literary works	Albanian	30	576	full set of features	OvR RC-CV	0.997 F1 score

(continued on next page)

Table 6.11: Comparison with previous works (continued).

Reference	Document type	Language	Number of authors	Number of samples	Features	Method	Performance
Our method	newsroom columns	Albanian	52	569	full set of features	OvR RC-CV	0.992 F1 score
Our method	social media posts	Albanian	44	426	full set of features	OvR RC-CV	0.985 F1 score
Our method	all writings (mixed)	Albanian	122	1,107	full set of features	OvR RC-CV	0.994 F1 score

6.5 Conclusion and future work

The investigation of the authorship identification task in various written works in Albanian provided a valuable perspective on the field's potential and limitations. This study analyzed literary works, newsroom columns, and social media posts with an extensive examination of language-related characteristics and ML-based methodologies. The findings from our analysis provide valuable information that demonstrates the careful design of our experiments to address the formulated research questions:

1. *Can the author be identified from an Albanian text?*

   Our research includes an extensive investigation using various classification algorithms and a specifically created set of language-related elements. The complete set of attributes consists of numerous functions that define the authors' unique writing style and characteristics. The results indicate the effectiveness of selected methods and the discriminatory ability of linguistic features. Our findings highlight the potential of authorship identification in Albanian texts.

2. *Which linguistic features are crucial to determine the author's writing style?*

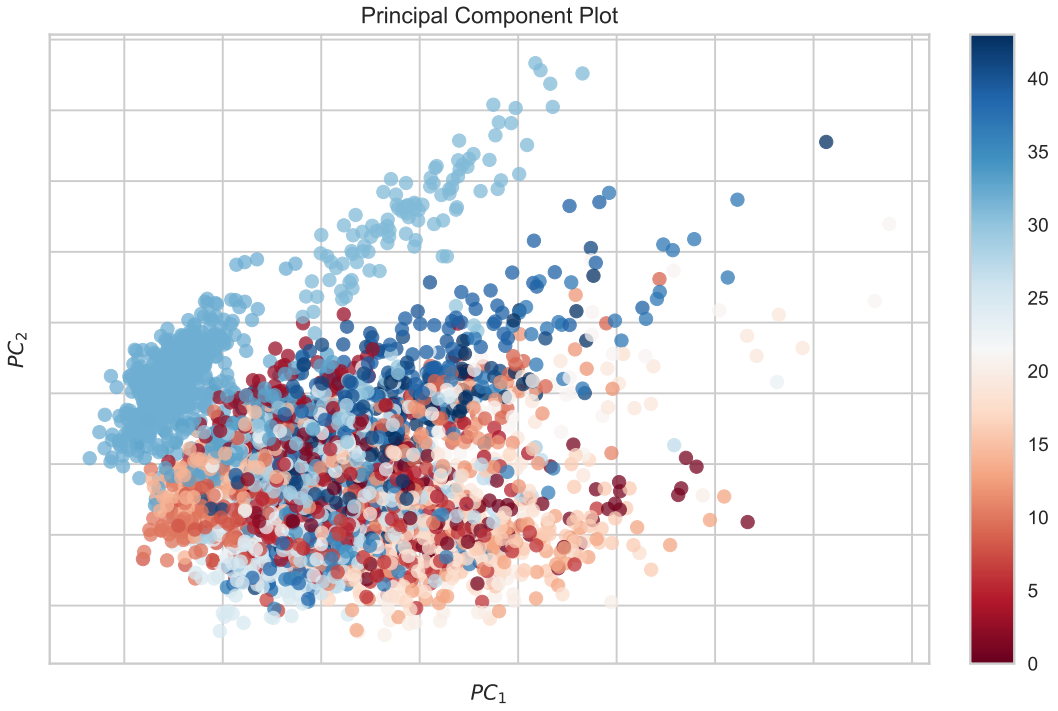


Figure 6.16: Illustration of social media posts data via PCA dimensionality reduction.

The study analyzes stylometric features in Albanian texts to identify the unique patterns that characterize an author's writing habits. It reveals that sound distribution, structural traits, punctuation preferences, POS tags, capitalization, and syntax are critical elements in defining authorial styles. These features contribute to a comprehensive understanding of an author's writing style in Albanian texts.

3. *What is the best feature set identifying the author of a text document?*

   The research reveals that there is no universally optimal feature set for attributing authorship, as its effectiveness depends on the domain and characteristics of the text documents. The results indicate the significance of creating features in line with the requirements and specifics of the task.

4. *What are the best methods for authorship analysis (authorship identification) in Albanian?*

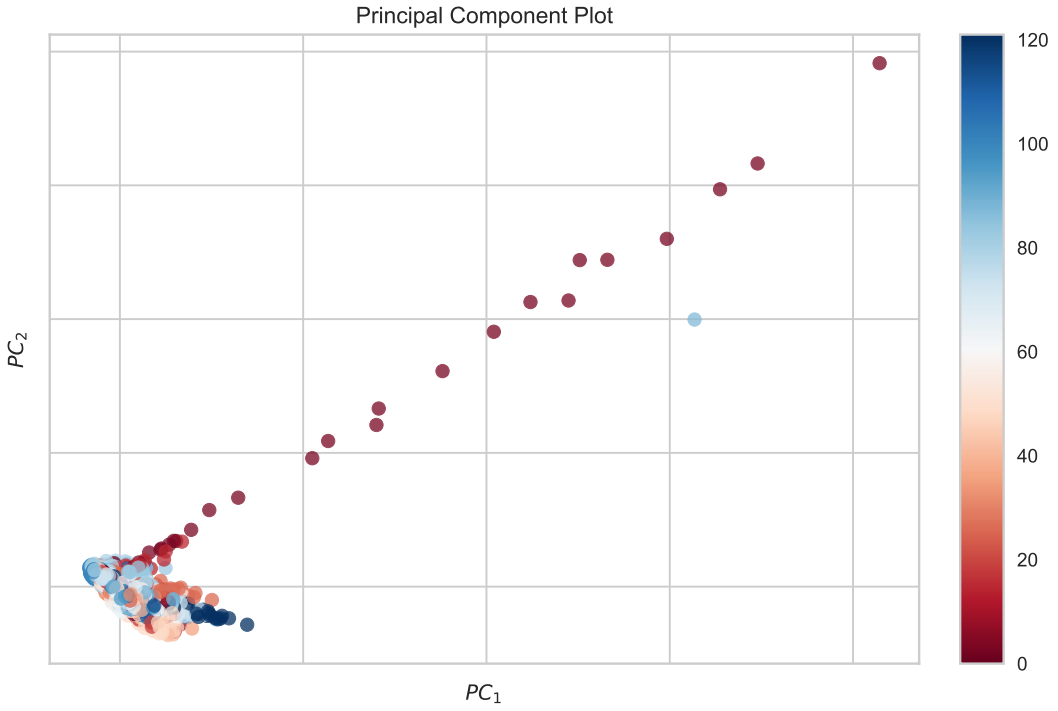


Figure 6.17: Illustration of all writings data via PCA dimensionality reduction.

Our investigation considers the task of authorship attribution in written texts to identify reliable methods in the Albanian language. A significant observation is the efficiency of the Ridge linear model and its ability to handle datasets with an extensive set of features. Other promising methods involve autoencoders for data compression and feature learning and FastText to capture contextual relationships in the data. These methods have the potential to analyze Albanian authorship and understand the nuances of written language.

5. *How much data is sufficient to recognize authorial uniqueness?*

   The research explores the learning curve and its contribution to understanding the relationship between the varying size of training data and the model's capability to capture authorship patterns. A valuable observation is the effect of the gap between training and validation scores and the need for more data for model generalization. The research highlights a critical point: no fixed universal data threshold guarantees recognition of

authorial uniqueness. The sufficiency of a dataset depends on the task complexity and language-related characteristics.

This research enables authorship classification and demonstrates the importance of specific features in identifying writing patterns in Albanian texts. The analysis of feature selection and dimensionality reduction processes revealed their ability to improve model effectiveness while maintaining significant information. Our study indicates the adaptability of methodologies that can be extended to other languages and contribute to cross-linguistic analyses in the field.

This investigation provides the foundation for further progress in Albanian authorship identification. It suggests expanding the dataset's representativeness by including written texts from various authors and refining the ensemble of features to improve discriminatory power. The goal is to explore authorship identification in Albanian further and provide significant information on natural language processing and computational linguistics.

Chapter 7

Genre classification task in Albanian

This chapter explores genre classification in Albanian texts and combines theory with empirical analysis. Section 7.2 outlines the experimental elements, including the corpus, selected features, and methodologies [120]. The section also presents the metrics used for the models' performance evaluation. Section 7.3 reports the results from our experiments and focuses on feature analysis and selection to reduce dimensionality [174]. In Section 7.4, we compare the study with previous works, highlighting its relevance and originality within the research field. The chapter concludes with a discussion of our findings and future research objectives in Section 7.5.

7.1 Introduction

The genre classification task categorizes texts into specific genres, such as dramas, novels, poems, and more, based on their writing style and theme-related characteristics. This task is significant for NLP and text analysis. Numerous studies in NLP, machine learning, and literary analysis have focused on genre classification. Various techniques, including traditional ML algorithms and modern DL methods, have been investigated to perform this task.

We explore this task with extensive experimentation in the context of the Albanian language. Our experiments contribute novel insights and progress in this field, particularly in the Albanian language. This research extensively investigates genre classification, experimenting with various crafted features, methods, and model architectures.

This study aims to evaluate the effectiveness of cutting-edge genre classification methods and examine feature engineering and selection to identify informative characteristics for different genres. To achieve these goals, we provide a series of experiments designed to explore different aspects of genre classification and contribute important information to the field.

This chapter addresses the research questions and hypotheses presented in Chapter 1. We aim to provide a reliable foundation for presenting and interpreting our findings.

1. Can the genre be identified from an Albanian text?
2. Do authors naturally use distinct classes of language styles for different genres?
3. What are the best methods for authorship analysis (genre classification) in Albanian?

We significantly contribute to the AA context by compiling the first-of-its-kind corpus designed explicitly for genre classification in Albanian texts. Our study's experiments and analyses are built on this corpus of varied and representative written works. Cross-domain analysis demonstrates the robustness of our model and its reliability across various domains.

The experimentation framework significantly contributes to Albanian text genre classification by providing new perspectives into feature analysis, modern techniques, and genre-related writing patterns. The fundamental aspect of our contribution is the manual engineering of hand-crafted features that represent characteristics of various genres in the text. Using new visualization tools, we identified relevant features that contribute to the decision-making process of classification methods. Experiments reveal language-related characteristics crucial for genre classification and identify linguistic variations in Albanian writings. The research methodology extends to the cross-validation of our findings with prior works to validate their significance within the authorship analysis field.

7.2 Experimental setup

This section outlines the elements of our exploration, including the recently designed corpus, selected features, and the techniques used to identify unique patterns in literary genres. We

also present evaluation metrics to evaluate the performance of selected methods in genre categorization.

7.2.1 Materials

This section provides a detailed review of the dataset's specifications, manually engineered features, and various language models essential for text representation.

7.2.1.1 Data preparation

The A3C3 corpus, created explicitly for authorship analysis tasks, is introduced in Chapter 3. This subsection summarizes statistics about the various NLP attributes. The corpus is a mixed collection of newsroom columns, social media posts, and literary texts. Its details, including corpus size, genre-related statistics, and POS data, are presented in Table 7.1.

The A3C3 corpus is a comprehensive dataset with 20,185,632 words and 423,285 unique tokens, covering various aspects of language use habits. It is a valuable resource for studying genre classification when considering textual content in the Albanian language.

7.2.1.2 Combination of genre-related hand-crafted features

The genre classification task relies on an extensive ensemble of features, primarily linguistic attributes. These features encompass various characteristics, such as lexical, morphological, and syntactic structures. Numerous attributes are designed for specific literary genres and linguistic contexts. Analyzing these features is crucial for authorship-related research, providing a deep understanding of text genres.

7.2.1.3 Language models

This section outlines the text representation models essential for transforming raw text into numerical data for experimental analysis. These language models include CountVectorizer, TF-IDF, and Word2Vec.

Table 7.1: The corpus size and genre-related statistics of the A3C3 corpus.

Level / Features	Literary	Writings
Corpus size		
total number of samples	392	1,936
average number of words	5,727.78	9,354.34
average number of unique words	290.09	185.08
average word length (in characters)	5.63	5.94
average sentence length (in words)	15.58	19.79
average unique words per sentence	0.79	0.39
average number of sentences	367.62	472.61
average syllables per word	1.84	2.03
max length word	69	98
Genre-related task statistics		
total number of genres	5	3
average number of samples per genre	78.40	645.33
average number of tokens per genre	1,145.56	3,118.11
average number of unique words per genre	58.02	61.69
POS-related information		
lexical density (proportion of content words)	49.84	49.74
average number of function words (token-based normalization)	48.03	46.82
average number of verbs (token-based normalization)	13.41	10.86
average number of nouns (token-based normalization)	18.85	19.32
average number of adjectives (token-based normalization)	5.09	6.20
average number of adverbs (token-based normalization)	5.03	4.05
average number of pronouns (token-based normalization)	2.89	8.24

7.2.2 Classification methods and parameter settings

The research uses a variety of ML- and DL-based techniques, which were methodologically selected after a careful performance evaluation to ensure their effectiveness and relevance within our experimental framework. Various algorithms from different categories are used in experiments, including XGBoost and Random Forest ensemble techniques, linear models like Logistic Regression and Ridge Classifier, and Linear Support Vector Classifier. We use two multi-class classification strategies, OvO and OvR, to deal with multiple classes. The model parameters

are kept in default settings for comparative examination. The research also employs state-of-the-art DL techniques like fastText, autoencoders, and transformers to expand the scope of analysis. These methods automatically extract features from text data, capturing complex relationships and genre-related patterns in writing styles.

7.2.3 Evaluation metrics

The study used a cross-validation scheme to evaluate the performance of our models. The models were trained using a 5-fold cross-validation technique and evaluated using the F1 score, which takes the average of precision and recall. Weighted averaging was applied to the F1 score to handle imbalanced data distributions.

7.3 Genre identification results

This section documents the experiments performed to evaluate the efficiency and quality of the A3C3 dataset. We experiment with different ML-based algorithms to identify genre-related variations in Albanian texts. The study analyzed learning curves and class prediction errors for each classification algorithm to evaluate their performance in different training and validation stages. This analysis evaluated the model's robustness, generalizability, and areas for further improvement.

7.3.1 Experimental results

This subsection presents classification results from our extensive experimentation on the A3C3 dataset using different methods and configurations. The corpus includes various literary genres, such as drama, novella, poetry, short story, and novel, and a mixed scenario to examine the model's robustness.

We analyze texts at chapter and paragraph levels to evaluate the effectiveness of algorithms on classification accuracy for different text sizes. Our analysis uses eight distinct feature settings

to capture different textual characteristics. This approach provides an overview of our experimental findings across different levels, methodologies, and scenarios, contributing to a deep understanding of genre classification. The objective is to identify the minimal feature set for effective genre classification to simplify the process and maintain accuracy.

Literary works. The first set of experiments focuses on the literary text dataset. The results of these experiments are presented in Table 7.2. The TF-IDF model is the most promising for feature extraction when using language models. It achieves an F1 score of 1.00 using the OvO multi-class strategy and the Ridge classifier at the chapter level. Three algorithms, LR-CV, RC-CV, and LinSVC, demonstrate similar performance. The CountVectorizer and Word2Vec models also achieved comparable results. However, at the paragraph level, dropout rates increase, especially when using the Word2Vec model with different algorithms. This low-performance level is due to the limited ability for generalization to unseen instances in smaller datasets. The best results are obtained with the OvR multi-class strategy when using manually engineered features; the Ridge and LR algorithms show a significant F1 score of 0.987. The extensive set of features contributes substantially to effective genre classification.

Table 7.2: Comparing accuracy: Performance of ML-based algorithms on literary genres classification. Bold indicates the top-performing results in each experimental setup.

	One-versus-One					**One-versus-Rest**				
Level of analysis	**XGB**	**LR-CV**	**RC-CV**	**LinSVC**	**RF**	**XGB**	**LR-CV**	**RC-CV**	**LinSVC**	**RF**
Chapter-level										
TF-IDF	0.968	0.999	**1.000**	0.999	0.958	0.989	**1.000**	**1.000**	**1.000**	0.995
CountVec	0.972	0.994	0.999	0.996	0.993	0.988	0.997	0.999	0.997	0.999
Word2Vec	0.959	0.974	0.982	0.975	0.957	0.966	0.975	0.977	0.975	0.958
FE	0.964	**0.982**	0.949	0.923	0.930	0.977	**0.987**	**0.987**	0.964	0.935
Paragraph-level										
TF-IDF	0.935	0.994	**0.995**	0.994	0.907	0.955	0.996	**0.999**	0.996	0.955
CountVec	0.943	0.979	0.986	0.981	0.940	0.963	0.988	0.993	0.990	0.958
Word2Vec	0.912	0.945	0.945	0.938	0.899	0.917	0.948	0.931	0.936	0.901
FE	0.945	**0.966**	0.952	0.914	0.886	0.957	0.974	**0.976**	0.932	0.891

TF-IDF = Term Frequency-Inverse Document Frequency; **CountVec** = Count Vectorizer; **Word2Vec** = Word to Vector; **FE** = Feature Engineering (hand-crafted features); **XGB** = eXtreme Gradient Boosting; **LR-CV** = Logistic Regression Cross Validation; **RC-CV** = Ridge Classifier Cross Validation; **LinSVC** = Linear Support Vector Classification; **RF** = Random Forest.

Cross-domain. The next stage of the experiments extends our analysis to include all text data types. In this situation, the classification includes various texts in the linguistic context, such as literary works, newsroom columns, and social media posts. Table 7.3 outlines the results of this investigation. The study achieved an F1 score of 1.00 across multiple algorithms at the chapter level with different feature settings. This performance is consistent across model-generated and manually engineered attributes, with no significant decrease noticed at the paragraph level. Figure 7.10 illustrates the separation between data points, demonstrating the model's effectiveness in recognizing unique characteristics among diverse linguistic domains.

Table 7.3: Comparing accuracy: Performance of ML-based algorithms on cross-domain classification. Bold indicates the top-performing results in each experimental setup.

	One-versus-One					One-versus-Rest				
Level of analysis	**XGB**	**LR-CV**	**RC-CV**	**LinSVC**	**RF**	**XGB**	**LR-CV**	**RC-CV**	**LinSVC**	**RF**
Chapter-level										
TF-IDF	0.998	**1.000**	**1.000**	**1.000**	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**
CountVec	0.998	0.999	**1.000**	0.999	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**
Word2Vec	0.999	**1.000**	**1.000**	**1.000**	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**
FE	0.999	0.999	**1.000**	0.997	0.999	**1.000**	**1.000**	**1.000**	0.998	0.999
Paragraph-level										
TF-IDF	0.991	**0.999**	**0.999**	**0.999**	0.993	0.995	**0.999**	**0.999**	**0.999**	0.993
CountVec	0.992	0.998	0.996	0.998	0.994	0.995	**0.999**	0.997	**0.999**	0.994
Word2Vec	0.998	**0.999**	0.998	**0.999**	0.997	0.998	**0.999**	0.998	**0.999**	0.998
FE	0.996	0.998	**0.999**	0.993	0.991	0.998	**0.999**	**0.999**	0.993	0.991

TF-IDF = Term Frequency-Inverse Document Frequency; **CountVec** = Count Vectorizer; **Word2Vec** = Word to Vector; **FE** = Feature Engineering (hand-crafted features); **XGB** = eXtreme Gradient Boosting; **LR-CV** = Logistic Regression Cross Validation; **RC-CV** = Ridge Classifier Cross Validation; **LinSVC** = Linear Support Vector Classification; **RF** = Random Forest.

Table 7.4 presents the subsequent set of experiments. We evaluated the performance of DL-based methods on our dataset, using fastText, autoencoders, and transformers in Albanian genre classification. The difference we observed in classifying texts from various sources is also evident when using DL-based methods. It supports the identification of distinct regions between these data points.

The Ridge classifier is used for further analysis due to its consistent performance across various experimental scenarios. The OvR multi-class classification strategy is applied to investigate

Table 7.4: Comparing accuracy: Performance of DL-based algorithms on cross-domain classification. Bold indicates the top-performing results in each experimental setup.

Models	**fasttext**		**autoencoders**		**transformers**	
Level of analysis	**Chapter**	**Paragraph**	**Chapter**	**Paragraph**	**Chapter**	**Paragraph**
Literary genres	0.981	0.976	0.992	0.969	0.996	0.977
All writings	1.000	0.999	1.000	0.996	1.000	0.999

the model's efficacy in two scenarios at the chapter level.

We expand the analysis by excluding specific words that influence literary genre classification, including n-grams and capitalized words. These words could represent characters or personages in literary works. The aim was to evaluate their impact on the classification process, as detailed in Table 7.5.

Table 7.5: Comparing accuracy: Performance of the Ridge classifier on cross-domain classification.

Form of analysis	**no-ngrams**	**no-uppercase**		
Data	**FE**	**TF-IDF**	**CountVec**	**Word2Vec**
Literary works	0.978	0.974	0.948	0.950
All writings	1.000	1.000	0.999	1.000

The exclusion of n-grams from manually generated attributes did not significantly decrease the results. On the other hand, when using language model features for literary genre classification, the Ridge algorithm's performance dropped significantly, indicating the impact of reducing the presence of capitalized words in different literary genres.

The research findings reveal the effectiveness of different methods on various datasets and settings, demonstrating consistency in some methods and limitations in others across specific scenarios.

7.3.2 Learning curves

This section provides learning curves illustrating the relationship between training and cross-validated test scores. The Ridge estimator was used with the OvR strategy for different training sample sizes. Figure 7.1 and Figure 7.2 visualize the learning curves for two distinct scenarios in

the genre classification task. The figures show that the training score consistently outperforms the validation score. However, this gap decreases during learning, indicating improved generalization to unseen data instances. This phenomenon is due to the increased training data variety, which allows the model to capture more detailed and reliable patterns in the data distribution. The figures represent learning curves for literary genre classification and cross-domain scenarios.

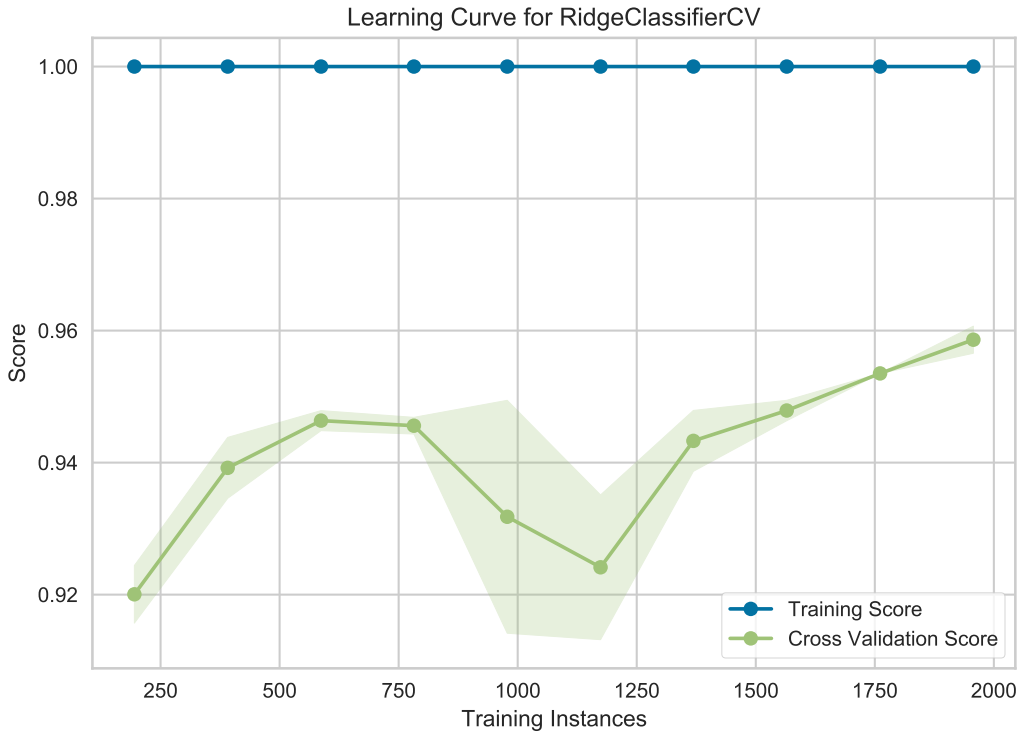


Figure 7.1: Learning curve for genre-related literary chapter-level classification using Ridge classifier and feature engineering.

7.3.3 Class prediction error analysis

This subsection analyzes class prediction errors to examine the effectiveness of ML-based models in accurately attributing genre to texts. The visualization of these errors provides insightful information about the model's performance, revealing misclassifications and identifying error-prone classes.

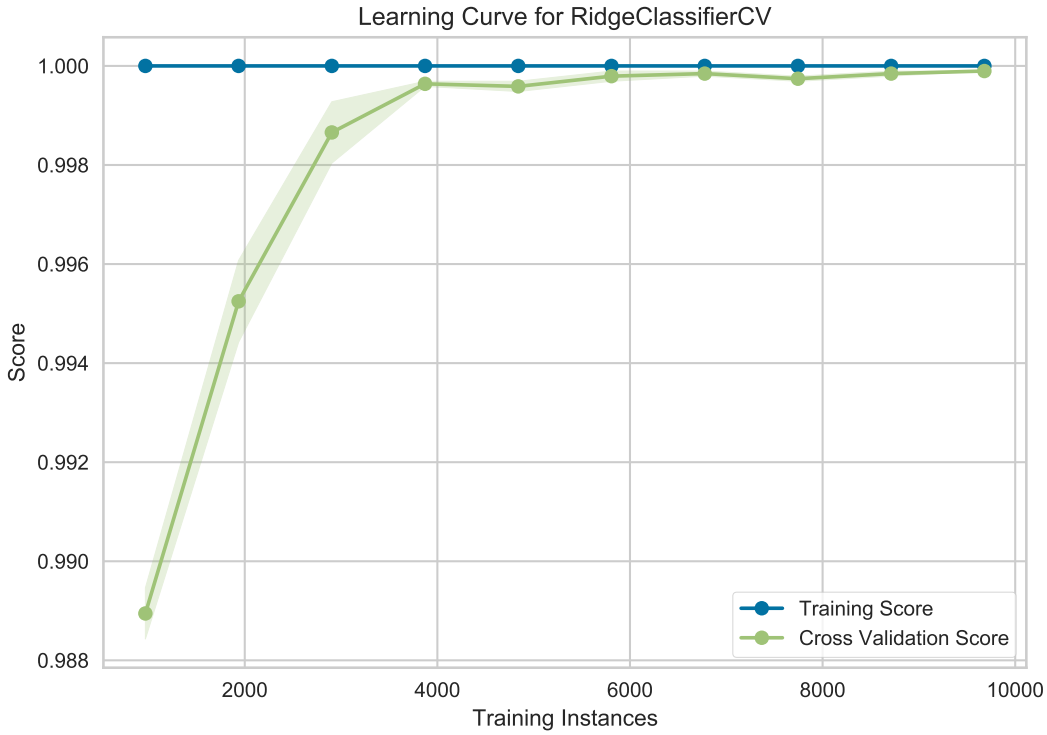


Figure 7.2: Learning curve for genre-related all writings chapter-level classification using Ridge classifier and feature engineering.

Literary works. The model demonstrates a low error rate in the classification of literary genres, as shown in Figure 7.3. The model accurately categorizes literary genres, demonstrating its effectiveness in various settings. However, there is constant confusion between novels and stories. The misclassification of novels with stories is complex due to the content, themes, and narrative style similarities. Novels may have specific characteristics that overlap with those found in stories, creating confusion in the classification process. This trend is illustrated in Figure C.17 in Appendix C.2.3. Genre classification can be challenging due to ambiguity in narrative style, content structure, and thematic elements, with some stories similar to novels.

Cross-domain. The Ridge algorithm achieved a perfect F1 score in cross-domain writings, with no significant confusion between different sources (Figure 7.4). The SHAP summary plot C.18 in Appendix C.2.3 illustrates the clear distinction between them, supporting this evidence.

The class prediction error analysis in genre classification provided significant information

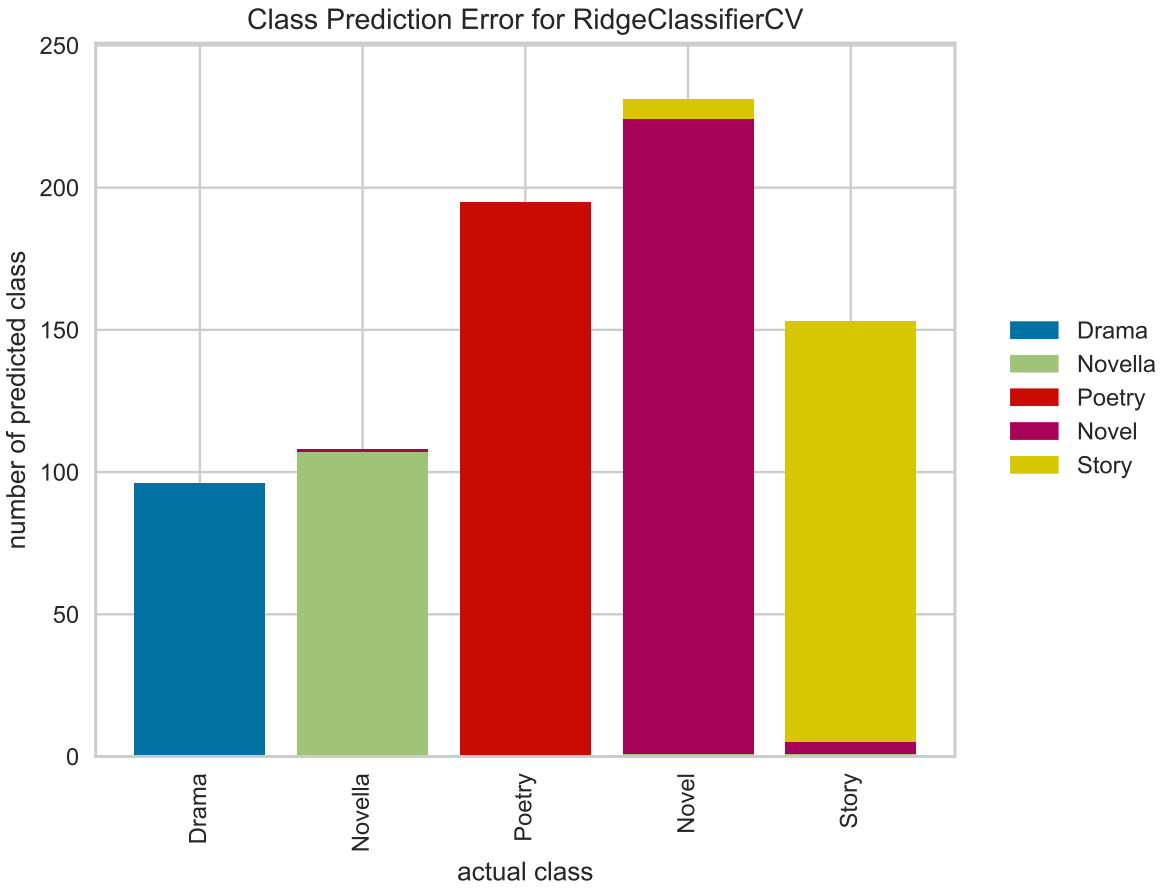


Figure 7.3: Class prediction error analysis for genre-related literary chapter-level classification using Ridge classifier and feature engineering.

for identifying areas where the models may have difficulty distinguishing between closely related genres. By analyzing these errors, we understand the challenges of genre classification in Albanian writings.

7.3.4 Feature analysis

This section provides a detailed analysis of features for classifying genres in Albanian texts. The study uses SHAP values and generates separate waterfall plots for each dataset to understand the importance of these features and their impact on the genre classification task. SHAP waterfall plots illustrate feature importance, with each bar corresponding to a feature and its length indicating its impact on the model's output for a given example.

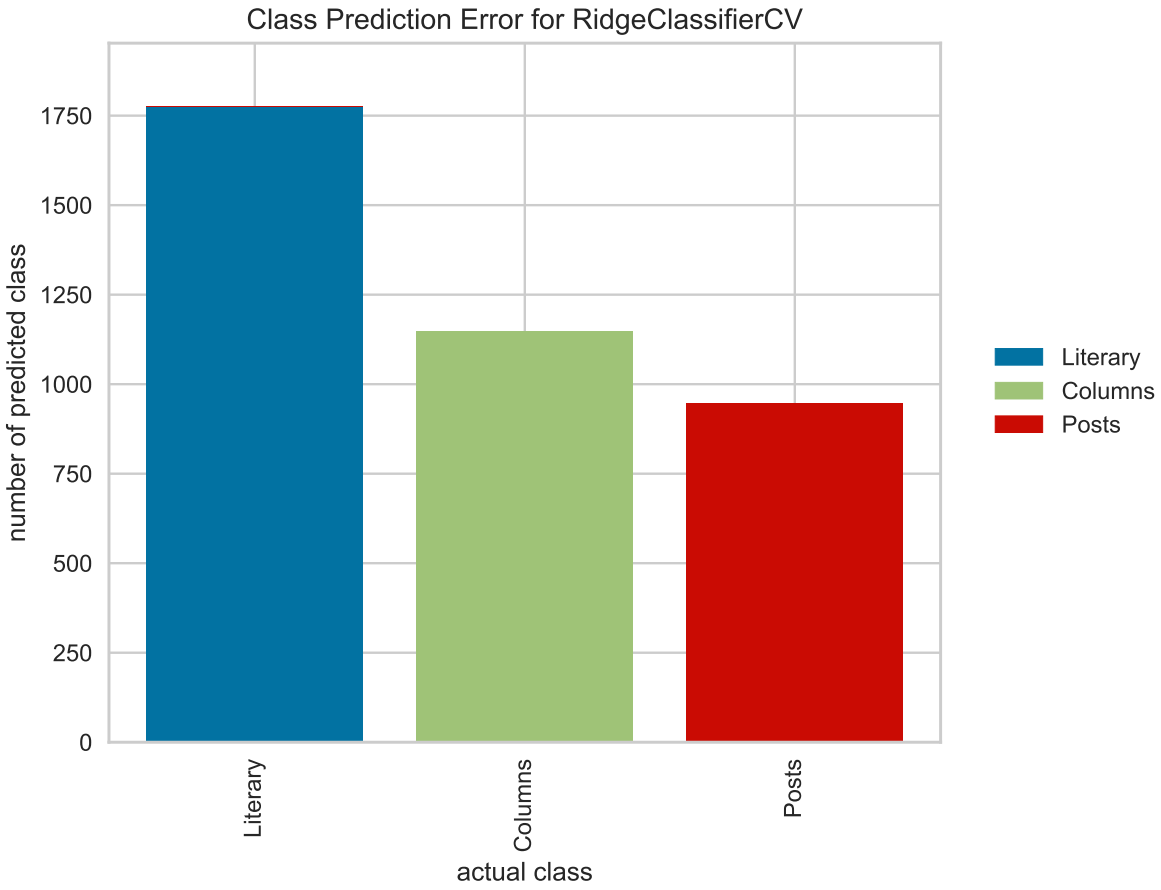


Figure 7.4: Class prediction error analysis for genre-related all writings chapter-level classification using Ridge classifier and feature engineering.

Literary works. The linguistic characteristics of literary texts and how those features relate to the genre can influence the classification of literary genres. Figure 7.5 presents the SHAP waterfall plot related to literary genres.

The 'average singular nouns per sentence' feature helps classify literary genres by indicating the average number of singular nouns per sentence. Literary works often include a rich vocabulary and diverse use of language, including singular nouns, to describe characters, locations, or objects. Literary genres with a higher average number of singular nouns per sentence may represent more complex and rich language, which helps with classification.

Proper nouns, like character names, place names, and titles, are frequently capitalized in literary works, indicating their presence in narrative pieces. However, their impact can vary based on the dataset, the classification algorithm, and the variety of literary genres in the dataset.

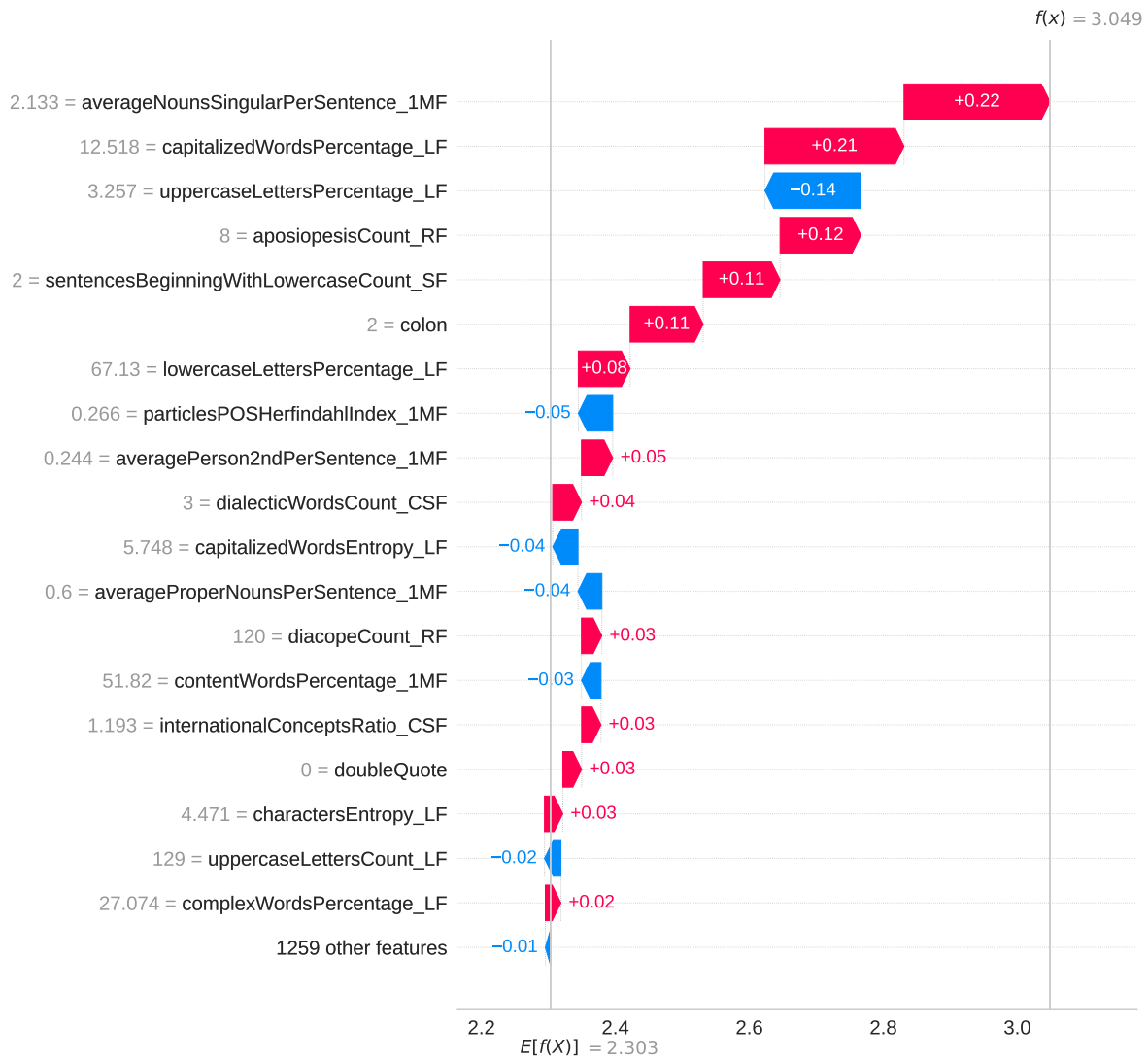


Figure 7.5: SHAP waterfall plot for genre-related literary chapter-level classification using XGBoost classifier and feature engineering.

The classification of literary genres is influenced by their unique linguistic and stylistic characteristics, as illustrated in Figure C.17 in Appendix C.2.3. *Poetry* is a distinctive genre with a unique structure, often characterized by short, fragmented sentences or lines. It is reflected in the 'sentence count' feature, distinguishing it from other genres with longer and more complex sentences. The *drama* genre frequently uses dialogue and dramatic devices, such as apsiopesis and colons, to create dramatic effects in plays and works. These devices help emphasize specific points and indicate dialogue within the narrative and dialogic structure. *Novellas* are

narratives that provide rich character development, often using dialectical words to indicate distinct speech patterns. *Short stories* often have a unique style with capitalized words for emphasis, which the higher count of these words can identify. *Novels* are longer works that frequently involve many characters, locations, or entities. Because of the extensive narrative, proper nouns, which are names of people, places, etc., are likely to be more frequent in this genre. These features represent distinctive linguistic and stylistic elements in genre texts, enabling accurate genre classification in written works.

Cross-domain. Specific linguistic and stylistic features are essential in classifying various text types. The 'sentences beginning with uppercase count' feature significantly affects the classification of different texts due to different writing styles and patterns (Figure 7.6).

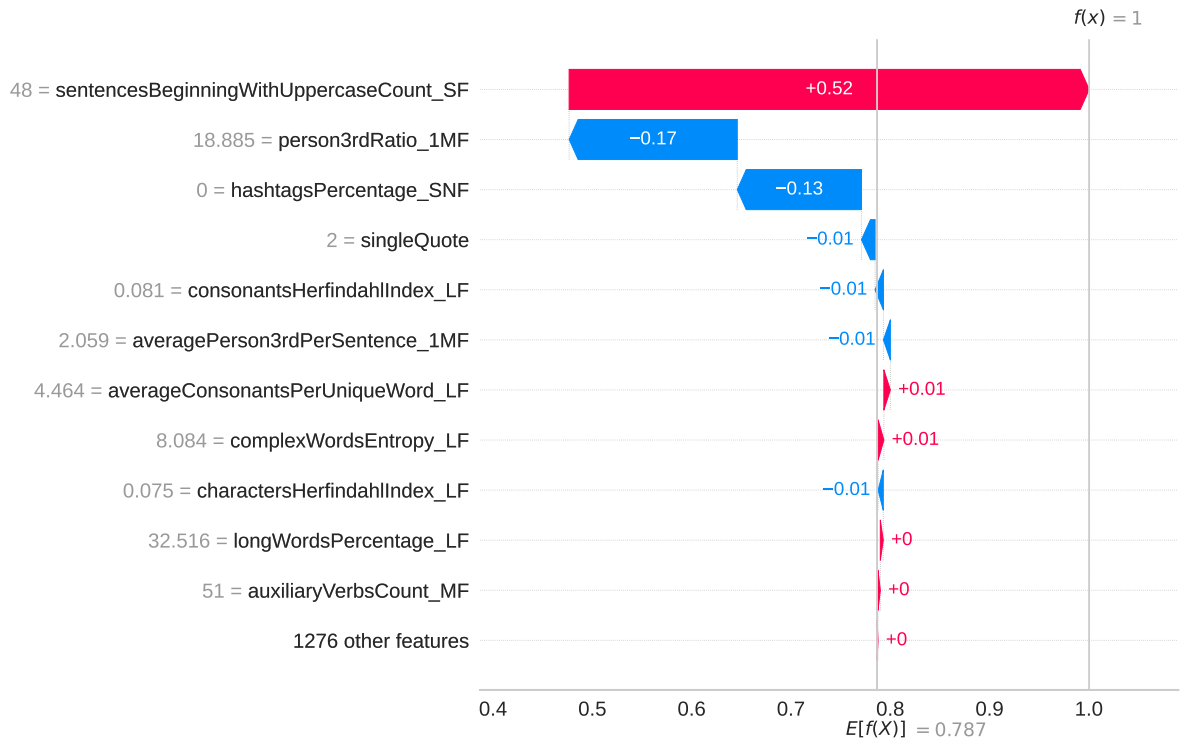


Figure 7.6: SHAP waterfall plot for genre-related all writings chapter-level classification using XGBoost classifier and feature engineering.

The SHAP summary plot shows that 'sentences beginning with uppercase count' significantly contribute to the classification of *literary texts*, as illustrated in the stacked feature importance visualization in Figure C.18, Appendix C.2.3. Sentence case is a standard convention in

well-written prose, following formal grammar rules. *Social media posts* often use hashtags to categorize content, indicating the topic's scope. The 'hashtags percentage' feature measures the percentage of text with hashtags, making it a relevant feature that distinguishes posts from other data points. *Newsroom columns* typically use a formal and objective tone to discuss public events, issues, or personalities. They often use third-person pronouns, particularly when discussing public figures or everyday situations.

Classification models consider an ensemble of features and their linear combinations; no single attribute can determine classification. The suitable combination of these features enables accurate genre classification in a mixed-domain corpus and contributes to a deeper understanding of genre-related features. Appendix C.2.3 provides detailed SHAP summary plots for each dataset, visualizing feature impact on the model's decision-making process. Appendix B.5 presents feature importance scores.

Some features have a minimal impact on the model's predictions. Excluding these less relevant attributes, which exhibit similar patterns across multiple genres, helps reduce model complexity without affecting the model's performance.

7.3.5 Feature selection

We performed feature selection analyses with literary works and the cross-domain corpus to improve the model's effectiveness for classifying genres. The goal was to identify the most informative features to optimize the model's performance using the Ridge classifier and RFE in a cross-validated scenario.

This section presents RFE curves for each dataset, detailing the iterative feature elimination process and its impact on model performance. The x-axis corresponds to feature count, while the y-axis represents the model's cross-validated score, providing insights into performance evolution with varying numbers of features.

These curves define two critical points: the exact number of features at which the Ridge model achieves its highest score and a threshold that indicates a performance drop of 0.02

from the peak, as marked by the dashed line. This threshold indicates the balance between feature inclusion and the stability of model performance.

Figure 7.7 illustrates the impact of different feature subsets on the Ridge model's performance in the literary domain. On the other hand, Figure 7.8 represents feature relevance in the cross-domain dataset and its impact on diverse text types.

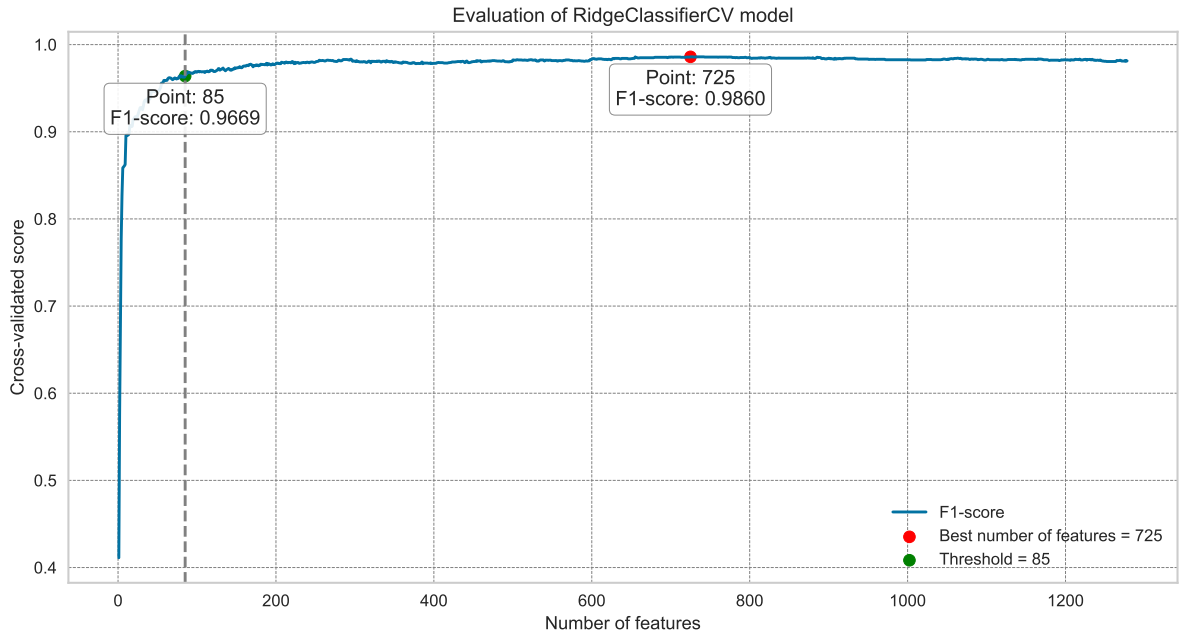


Figure 7.7: RFE for genre-related literary chapter-level classification using Ridge classifier and feature engineering.

The figures provide observations into the impact of feature selection on Ridge model performance, helping to identify the optimal number of features for superior predictive accuracy. Table 7.6 summarizes the results of our RFE-based feature selection process. This table presents the exact number of features for each domain with the highest F1 score and a threshold with a performance drop of 0.02.

Table 7.6: Feature selection outcomes using RFE in a cross-domain configuration.

Corpora	**Number of features**	**F1-score**	**Number of features (threshold)**	**F1-score (threshold)**
Literary genres	725	0.986	85	0.967
All writings	208	0.999	5	0.988

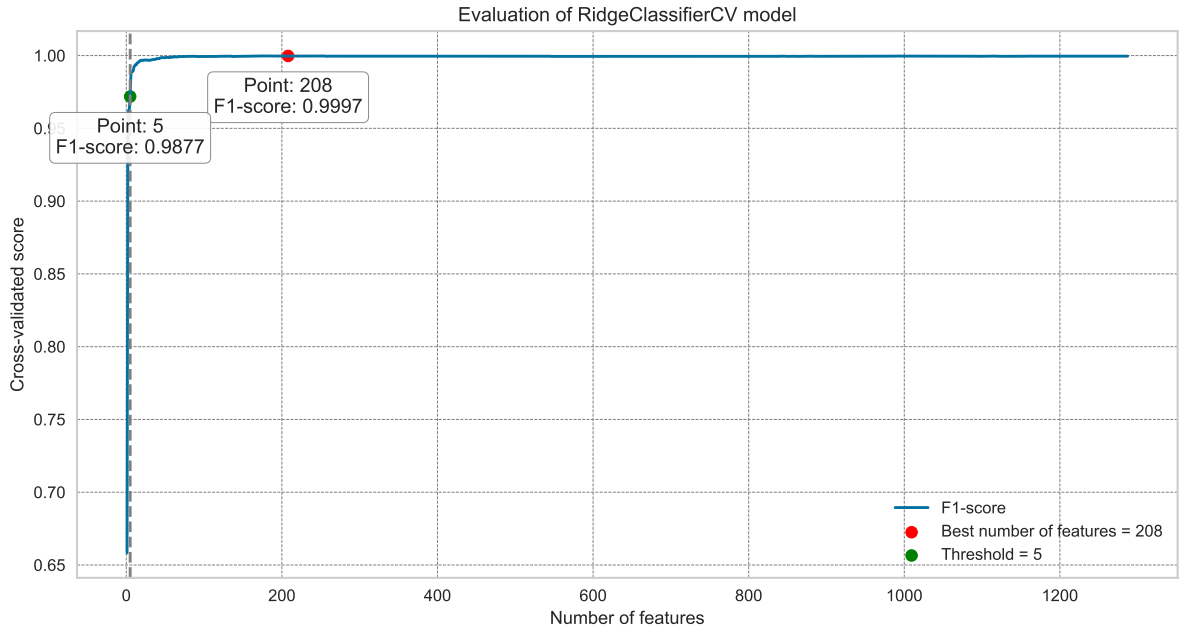


Figure 7.8: RFE for genre-related all writings chapter-level classification using Ridge classifier and feature engineering.

We aimed to determine the most effective subsets of features in different domains (Appendix B.5), analyzing stylometric characteristics. We used RFE to systematically evaluate how various feature subsets affected the model's performance.

7.3.6 Dimensionality reduction

This section uses PCA and autoencoder models to reduce the dimensionality of the feature space. The following tables summarize the results obtained from these models. For the PCA model with results presented in Table 7.7, we first experimented to determine the number of components for 100% of the data variance. After that, we identify the number of components that give the optimal result, with a minimal drop in performance of 0.02 from the peak. Performance curves of the PCA model, showing its effectiveness on the number of components for each corpus, are detailed in Appendix C.2.1.

The autoencoder model employs a similar dimensionality reduction methodology (Table 7.8), using the complete feature set and gradually increasing the number by 50. A threshold

Table 7.7: Dimensionality reduction outcomes using PCA in a cross-domain configuration.

Corpora	Number of components - PCA cumulative variance=1.0	Number of components - Best score	F1-score - Best score	Number of components - Threshold score	F1-score - Threshold score
Literary genres	570	558	0.983	95	0.964
All writings	552	74	1.000	2	0.986

point of 0.02 is set for a reasonable reduction in predictive accuracy. Performance curves for each dataset are provided in Appendix C.2.2, with varying feature subsets.

Table 7.8: Dimensionality reduction outcomes using autoencoders in a cross-domain configuration.

Corpora	Best point	Best F1-score	Threshold point	Threshold F1-score	Epochs, early stopping (average)
Literary genres	650	0.996	150	0.983	13
All writings	400	0.999	50	0.999	16

The tables show that the autoencoder model outperforms the PCA model with minimal features and training epochs. However, even with two features, the PCA model achieved an impressive F1 score of 0.986 in cross-domain data. It demonstrates its reliability and effectiveness in identifying significant variations in text data, even with reduced feature space.

The dimensionality reduction technique of PCA generates data visualizations that facilitate the identification of patterns in the data's structure and relationships.

The PCA model visualization for literary genres in Figure 7.9 shows the distribution of data points in a two-dimensional space. The PCA visualization reveals distinct features in the poetry and drama genres, indicating differences from other literary genres. However, overlaps in the data clusters for novella, novel, and story genres indicate similarities or shared characteristics.

The cross-domain scenario illustrates three distinct clusters in regions without overlaps in data (Figure 7.10). This trend indicates substantial variations within each data cluster.

The subsection discussed the effectiveness of PCA and autoencoders in reducing feature dimensionality while preserving relevant information. It examined the impact of dimensionality

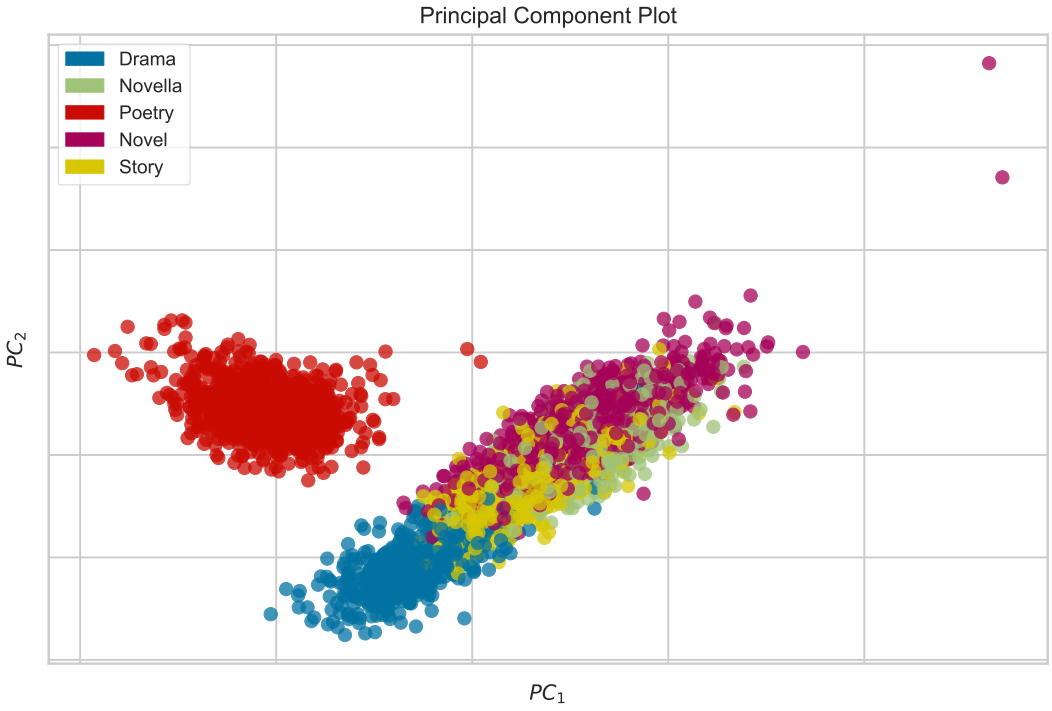


Figure 7.9: Illustration of literary genres data via PCA dimensionality reduction.

reduction techniques on model performance and provided significant information for further analysis.

7.4 Comparison with previous works

Our research methodology involves cross-validation with previous studies, as presented in Table 7.9. This approach compares our experimental results with the results reported in prior research. Cross-validation emphasizes the continuity of research and the importance of our findings in the authorship analysis domain.

7.5 Conclusion and future work

The research investigated genre classification in Albanian texts, examining the effectiveness of various methods. It highlights the importance of distinct linguistic features and demonstrates

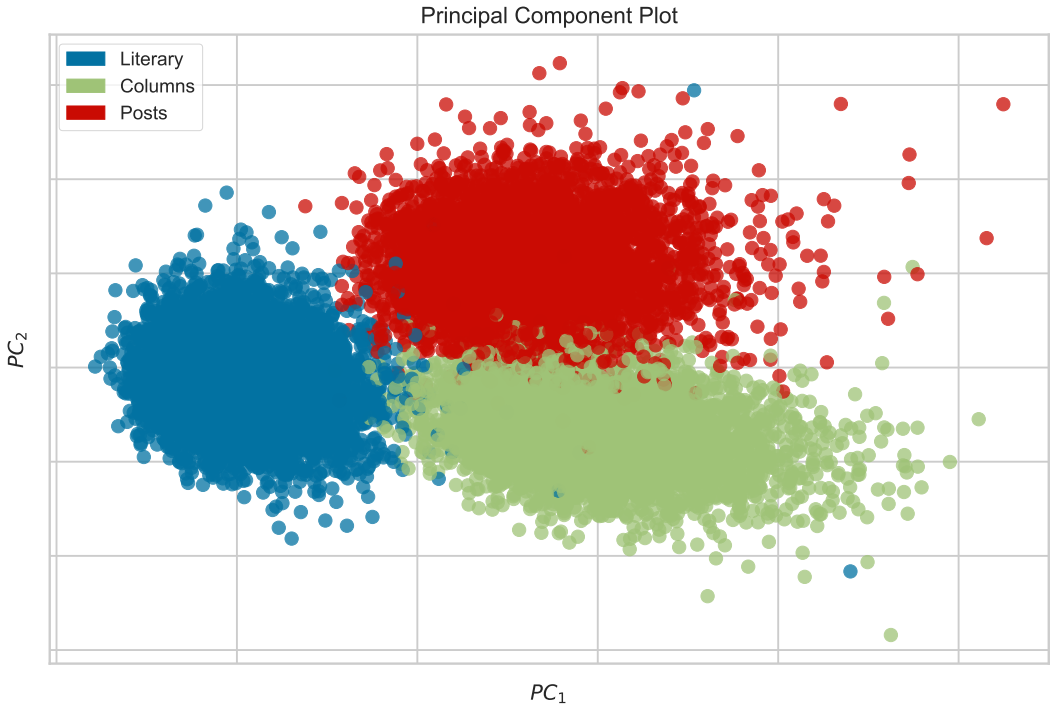


Figure 7.10: Illustration of all writings data via PCA dimensionality reduction.

the complexities of the genre classification task in Albanian. The findings contribute to the field of authorship analysis and provide a foundation for future progress in this field. We designed our experiments to address the following research questions:

1. *Can the genre be identified from an Albanian text?*

   The study investigated the genre classification of Albanian texts using different algorithms and explicitly designed language-related features. The research findings indicate that genre classification in Albanian texts is possible and validate the effectiveness of our approach.

2. *Do authors naturally use distinct classes of language styles for different genres?*

   This study explored writers' linguistic patterns while crafting texts in various genres. Careful analysis identifies genre-related characteristics such as capitalization preferences, sentence structure, vocabulary choices, and punctuation patterns. These findings demonstrate how authors adapt linguistic choices based on their writing genre, revealing unique

Table 7.9: Comparative analysis with prior studies in cross-domain classification settings (top-performing results).

Reference	Document type	Language	Number of genres	Number of samples	Features	Method	Performance
[115]	product reviews (book dataset)	English	3	2,121	CF (char and PoS n-grams, discriminative words frequency, AA features)	Random Subspace (Random Forest)	0.99 F-measure
[115]	product reviews (camera dataset)	English	3	1,638	CF (char and PoS n-grams, discriminative words frequency, AA features)	Bagging (Random Forest), Random Subspace (Random Forest)	0.98 F-measure
[86]	web pages	Telugu*	3	150	syllable extraction algorithm to extract char n-grams	SVM, RF	0.98 F-measure
our method	literary genres	Albanian	5	392	-	transformers	0.996 F1 score
our method	all writings (mixed)	Albanian	3	1,936	full set of features	OvR RC-CV, LR-CV	1.000 F1 score
our method	all writings (mixed)	Albanian	3	1,936	-	fasttext, autoencoders, transformers	1.000 F1 score

genre-specific attributes.

3. *What are the best methods for authorship analysis (genre classification) in Albanian?*

   An analysis was conducted to determine the most effective methods in the Albanian linguistic context. The research findings demonstrate the reliability of the Ridge model for genre classification in handling linguistic complexities. In addition to traditional machine learning, transformers capture complex textual patterns, demonstrating superior performance in the Albanian language scenario.

Analyzing language-related and stylometric features explicitly for each linguistic domain

contributes to understand different aspects of classifying genres. The examination of feature selection and dimensionality reduction methods enhances model efficiency while preserving critical information. Although the research focused on Albanian, the findings can be extended to other languages, resulting in cross-lingual progress in the field.

Future research in Albanian genre classification aims to expand the dataset, improve model reliability, and identify more contextual linguistic patterns. We also aim to refine feature engineering and explore unique domain-specific attributes in Albanian to increase model discriminatory effectiveness, contributing to natural language processing and text analysis.

Chapter 8

Gender identification task in Albanian

This chapter overviews the research's fundamental elements, including the new corpus, selected features, methods [120], and metrics used to evaluate the models' performance in authorship gender identification task in Section 8.2. Section 8.3 records experimental results and provides an in-depth feature analysis and selection [174]. A comparative analysis is conducted on prior research in Section 8.4, focusing on the relevance and advancement of our findings within the field of study. The chapter concludes with a review of research findings and potential directions for future research in Section 8.5.

8.1 Introduction

Authorship gender identification is a text analysis method that tries to identify an author's gender based on content-related attributes, linguistic structures, and context-specific characteristics in written texts. The gender identification task focuses on understanding the relationship between language use and gender personality. This task provides valuable insights into how language represents gender. It has practical applications in various fields, including sociolinguistic studies and forensic linguistics.

Statistical models, machine learning algorithms, and linguistic analysis are extensively utilized in this field to identify gender-specific patterns. This study focuses on adapting existing methodologies to improve the efficacy and model performance of gender identification experiments, specifically in the Albanian language scenario. We aim to contribute empirical findings

that expand the understanding of gender manifesting in language across different domains and genres, revealing gender-related linguistic characteristics.

The investigation aims to enhance our understanding of the complex relationship between gender and language through an extensive experimental approach. This chapter revisits the research questions from Chapter 1.

1. Can the gender be identified from an Albanian text?
2. Are there linguistic features that indicate gender?
3. What are the best methods for authorship analysis (gender identification) in Albanian?

This research presents diverse Albanian textual works for gender identification analyses and experiments. Cross-domain validation demonstrates the model's flexibility and applicability in various scenarios. The extensive experimentation and analysis significantly contribute to gender classification in Albanian texts by providing valuable insights into gender-related writing patterns. Our main contributions include the compilation of the novel corpus and the manual engineering of linguistic features crafted to represent distinctive gender characteristics. Using new visualization techniques, we identified the ensemble of features that significantly contribute to gender classification. The experiments reveal language-related characteristics that are essential for gender identification. Our research methodology involves cross-validation with previous studies to validate its relevance within the authorship analysis field.

8.2 Experimental setup

This section presents the novel dataset, outlines the selected features for analysis, and discusses the methodologies used to capture gender-specific characteristics. Also, we describe the metrics used to evaluate the performance of our models.

8.2.1 Materials

This subsection provides the specifications of our datasets, the hand-crafted features we've selected, and the language models for text data representation.

8.2.1.1 Data preparation

The A3C3 corpus is designed explicitly for authorship analysis and includes newsroom columns, social media posts, and literary texts. The corpus creation process is detailed in Chapter 3. Table 8.1 presents its specifications and statistics related to various NLP features. The table includes three aspects of language-related information: size, gender-related statistics, and POS data.

The A3C3 corpus is a diverse collection of 26,874,527 words and 339,080 unique tokens. The dataset is a valuable source for investigating gender identification in the Albanian language.

8.2.1.2 Combination of gender-related hand-crafted features

The gender classification task relies on a complete set of features detailed in Table 4.1 within Chapter 4. These features cover a text's lexical, morphological, and syntactic aspects of the language use. This combination of features is carefully engineered and selected, and it serves as the foundation for further experimental analyses.

8.2.1.3 Language models

This section provides the language models used for effectively representing textual data. These models transform raw textual content into numerical matrices for practical analysis. Our experiments use three well-known models for text representation, including TF-IDF, CountVetorizer, and Word2Vec.

Table 8.1: The corpus size and gender-related statistics of the A3C3 corpus.

Level / Features	Literary	Columns	Posts	Writings
Corpus size				
total number of samples	338	654	879	1,239
average number of words	5,840.57	8,426.85	7,095.80	10,841.70
average number of unique words	306.05	206.20	202.74	192.35
average word length (in characters)	5.72	5.89	6.16	5.89
average sentence length (in words)	16.09	22.40	19.97	20.74
average unique words per sentence	0.84	0.55	0.57	0.37
average number of sentences	363.06	376.18	355.34	522.86
average syllables per word	1.88	2.04	2.13	2.02
max length word	51	66	98	66
Authorship-related task statistics				
total number of gendres	2	2	2	2
average number of samples per gender	169	327	439.5	619.5
average number of tokens per gender	2,920.29	4,213.43	3,547.90	5,420.85
average number of unique words per gender	153.02	103.1	101.37	96.17
POS-related information				
lexical density (proportion of content words)	49.46	49.03	50.39	49.34
average number of function words (TBN)	47.77	4.76	45.58	47.18
average number of verbs (TBN)	12.80	1.09	9.74	11.01
average number of nouns (TBN)	18.38	1.90	20.22	19.05
average number of adjectives (TBN)	5.28	6.24	6.55	6.16
average number of adverbs (TBN)	5.01	4.26	3.23	4.20
average number of pronouns (TBN)	9.07	4.19	7.46	8.46

TBN = Token-Based Normalization.

8.2.2 Classification methods and parameter settings

The study uses various ML- and DL-based methods selected after extensive evaluation to ensure their effectiveness within our experimental approach. This selection includes XGBoost and Random Forest ensemble techniques, Logistic Regression and Ridge Classifier linear models, and Linear Support Vector Classifier. The research uses state-of-the-art DL-based techniques, such as fastText, autoencoders, and transformers, to explore AA in Albanian texts. These techniques reveal intricate patterns and relationships, enabling the identification of linguistic structures and gender-specific markers.

8.2.3 Evaluation metrics

The experiments used a 5-fold cross-validation method to evaluate the performance of the methods. The evaluation uses the F1 score, a precision and recall metric. Due to class imbalance, weighted averaging of the F1 score was applied, with weights defined in proportion to example counts for each class.

8.3 Gender identification results

In this section, we report the experiments performed to evaluate the practicality and effectiveness of the A3C3 corpus. In addition to the experimental results, we analyzed the learning curves and class prediction errors. This investigation evaluated the algorithms' performance in various training and validation phases. It assessed the model's reliability and generalizability and identified potential areas for improvement. This extensive examination contributes to improving gender classification techniques, especially in the less studied area of the Albanian language.

8.3.1 Experimental results

This section investigates gender attribution using various methods and configurations. The A3C3 dataset, created from diverse sources like literary works, newsroom columns, and social media posts, is used to analyze gender-related characteristics. A mixed scenario with all text types validates the method's robustness, providing a deep understanding of the classification results.

We analyze text at two different size levels: the chapter level and the paragraph level. The aim is to evaluate the impact of different levels of text size on classification performance. The analysis uses eight different feature settings to represent distinct features of written language. This methodology provides an overview of our empirical findings across various dimensions,

methodologies, and scenarios and contributes to a deeper understanding of gender classification in Albanian written works. The focus is on identifying the minimal feature set for effective gender identification.

Literary works. The experiment sequence begins with the literary works dataset. Table 8.2 displays the results of different classification algorithms and feature settings in identifying gender. The TF-IDF model demonstrates superior performance in feature extraction, with impressive results achieved at the chapter level using LR-CV, RC-CV, and LinSVC. Comparable performance scores are obtained with CountVectorizer and Word2Vec. However, an increased dropout rate is observed at the paragraph level when using the Word2Vec model with different algorithms. It is evident in smaller datasets, where word embeddings need more support to capture data variation adequately. As a result, the limited capacity to generalize to unseen data results in reduced performance. The Ridge algorithm achieves an impressive F1 score of 0.991 when applying the complete set of hand-crafted features at the chapter level. This feature set enables accurate data classification and analysis for effective gender identification.

Table 8.2: Comparing accuracy: Performance of ML-based algorithms on literary works classification. Bold indicates the top-performing results in each experimental setup.

Level of analysis	Set of features	XGB	LR-CV	RC-CV	LinSVC	RF
Chapter-level	TF-IDF	0.988	**1.000**	**1.000**	**1.000**	0.996
	CountVec	0.989	0.999	0.999	**1.000**	0.998
	Word2Vec	0.989	0.997	0.999	0.996	0.990
	FE	0.986	**0.995**	0.993	0.980	0.962
Paragraph-level	TF-IDF	0.970	**0.998**	**0.998**	**0.998**	0.965
	CountVec	0.972	0.993	0.996	0.994	0.971
	Word2Vec	0.955	0.978	0.975	0.969	0.936
	FE	0.964	0.982	**0.985**	0.961	0.914

TF-IDF = Term Frequency-Inverse Document Frequency; **CountVec** = Count Vectorizer; **Word2Vec** = Word to Vector; **FE** = Feature Engineering (hand-crafted features); **XGB** = eXtreme Gradient Boosting; **LR-CV** = Logistic Regression Cross Validation; **RC-CV** = Ridge Classifier Cross Validation; **LinSVC** = Linear Support Vector Classification; **RF** = Random Forest.

Newsroom columns. Table 8.3 presents experimental results focusing on newsroom columns. We observe a consistent results trend within this context, similar to previous tables. The Ridge

classifier's performance decreases significantly at the paragraph level, primarily due to the columns' shorter and fragmented structure. The limited paragraph content impacts the model's ability to capture sufficient information, reducing prediction accuracy.

Table 8.3: Comparing accuracy: Performance of ML-based algorithms on newsroom columns classification. Bold indicates the top-performing results in each experimental setup.

Level of analysis	**Set of features**	**XGB**	**LR-CV**	**RC-CV**	**LinSVC**	**RF**
Chapter-level	TF-IDF	0.975	**0.999**	**0.999**	**0.999**	0.979
	CountVec	0.976	0.996	0.997	0.995	0.988
	Word2Vec	0.981	0.986	0.989	0.986	0.977
	FE	0.974	0.983	**0.991**	0.964	0.940
Paragraph-level	TF-IDF	0.942	0.988	**0.991**	0.988	0.942
	CountVec	0.947	0.979	0.976	0.978	0.949
	Word2Vec	0.934	0.949	0.947	0.948	0.922
	FE	0.934	0.956	**0.961**	0.857	0.874

TF-IDF = Term Frequency-Inverse Document Frequency; **CountVec** = Count Vectorizer; **Word2Vec** = Word to Vector; **FE** = Feature Engineering (hand-crafted features); **XGB** = eXtreme Gradient Boosting; **LR-CV** = Logistic Regression Cross Validation; **RC-CV** = Ridge Classifier Cross Validation; **LinSVC** = Linear Support Vector Classification; **RF** = Random Forest.

Social media posts. The following experiment sequence analyzes the dataset of social media posts, with results consistent with previous findings. However, due to the concise nature of social media posts, the results of these experimental series showed a slight performance drop compared to the other settings, as shown in Table 8.4.

Cross-domain. The study expands the examination to include the entire corpus. The classification analysis utilizes texts from various domains, such as literary works, newsroom columns, and social media posts, with the results reported in Table 8.5. The experimental findings in gender classification show that performance is generally lower when considering all types of texts together compared to individual textual data. The data variation exhibited by numerous linguistic attributes, word usage, writing habits, and speech formality influences the model's performance in detecting gender-related patterns. Domain knowledge inclusion and additional feature engineering may be needed to enhance the model's effectiveness.

Table 8.4: Comparing accuracy: Performance of ML-based algorithms on social media posts classification. Bold indicates the top-performing results in each experimental setup.

Level of analysis	Set of features	XGB	LR-CV	RC-CV	LinSVC	RF
Chapter-level	TF-IDF	0.967	0.994	**0.995**	0.994	0.925
	CountVec	0.969	0.989	0.989	0.990	0.969
	Word2Vec	0.943	0.964	0.956	0.964	0.925
	FE	0.950	**0.973**	0.972	0.900	0.890
Paragraph-level	TF-IDF	0.924	**0.973**	0.972	0.972	0.894
	CountVec	0.927	0.961	0.952	0.960	0.915
	Word2Vec	0.873	0.895	0.884	0.895	0.853
	FE	0.896	0.906	**0.923**	0.814	0.819

TF-IDF = Term Frequency-Inverse Document Frequency; **CountVec** = Count Vectorizer; **Word2Vec** = Word to Vector; **FE** = Feature Engineering (hand-crafted features); **XGB** = eXtreme Gradient Boosting; **LR-CV** = Logistic Regression Cross Validation; **RC-CV** = Ridge Classifier Cross Validation; **LinSVC** = Linear Support Vector Classification; **RF** = Random Forest.

Table 8.5: Comparing accuracy: Performance of ML-based algorithms on cross-domain classification. Bold indicates the top-performing results in each experimental setup.

Level of analysis	Set of features	XGB	LR-CV	RC-CV	LinSVC	RF
Chapter-level	TF-IDF	0.959	0.986	**0.989**	0.984	0.883
	CountVec	0.960	0.980	0.983	0.979	0.912
	Word2Vec	0.937	0.954	0.950	0.953	0.918
	FE	0.939	**0.960**	0.959	0.915	0.871
Paragraph-level	TF-IDF	0.900	**0.945**	0.944	**0.945**	0.821
	CountVec	0.904	0.936	0.912	0.934	0.814
	Word2Vec	0.864	0.864	0.860	0.864	0.817
	FE	0.871	0.893	**0.909**	0.788	0.804

TF-IDF = Term Frequency-Inverse Document Frequency; **CountVec** = Count Vectorizer; **Word2Vec** = Word to Vector; **FE** = Feature Engineering (hand-crafted features); **XGB** = eXtreme Gradient Boosting; **LR-CV** = Logistic Regression Cross Validation; **RC-CV** = Ridge Classifier Cross Validation; **LinSVC** = Linear Support Vector Classification; **RF** = Random Forest.

We used DL-based methods to evaluate the corpus in the following experiments, as presented in Table 8.6. We experimented with fastText, autoencoders, and transformers to demonstrate their effectiveness in handling gender identification difficulties. The performance scores obtained with these methods were comparable to those achieved with model-generated features and manually crafted attributes for different text types. However, in the cross-domain

setting, they demonstrated lower F1 scores. As noted in the previous chapter, the transformer model's effectiveness is evident for tasks with fewer classes, as the data volume within each class significantly impacts model performance. The fastText model reliably performed across various linguistic contexts in the entire corpus, demonstrating its ability to capture and generalize within different scenarios. However, the autoencoders model has lower F1 scores, possibly due to architectural limitations or hyperparameter tuning.

Table 8.6: Comparing accuracy: Performance of DL-based algorithms on cross-domain classification. Bold indicates the top-performing results in each experimental setup.

Models	**fasttext**		**autoencoders**		**transformers**	
Level of analysis	**Chapter**	**Paragraph**	**Chapter**	**Paragraph**	**Chapter**	**Paragraph**
Literary works	0.994	0.987	0.997	0.978	1.000	0.997
Newsroom columns	0.992	0.981	0.988	0.944	1.000	0.981
Social media posts	0.976	0.965	0.967	0.896	0.992	0.968
Writings	0.952	0.920	0.953	0.887	0.976	0.937

The Ridge classifier has been selected as the primary classifier for further investigations due to its robust performance in various experimental settings, particularly in chapter-level text analysis across four dataset settings.

The study aimed to evaluate the impact of certain words, such as n-grams and capitalized words, on gender classification by excluding them from hand-crafted features and language models. These words generally represent artistic characters or public figures in various written texts. Table 8.7 presents the results of this examination.

Table 8.7: Comparing accuracy: Performance of the Ridge classifier on cross-domain classification.

Form of analysis	**no-ngrams**	**no-uppercase**		
Data	**FE**	**TF-IDF**	**CountVec**	**Word2Vec**
Literary works	0.988	0.991	0.976	0.987
Newsroom columns	0.972	0.987	0.972	0.973
Social media posts	0.942	0.945	0.937	0.931
All writings	0.944	0.924	0.921	0.889

The investigation found that removing n-grams from manually engineered features did not significantly impact results. However, we observed a performance drop in the Ridge algorithm

when classifying gender using model-generated features. It highlights the importance of capitalized words across various textual sources. The literary works dataset consistently provided the most reliable performance across different scenarios and configurations.

The experimental results provide valuable insights into the performance of different methods across different textual data and configurations. Some methods demonstrated robustness in different scenarios, and others revealed context-specific limitations, highlighting the need for customized approaches in gender classification tasks.

8.3.2 Learning curves

This section presents learning curves to illustrate the relationship between training and cross-validated scores. The Ridge algorithm's performance is evaluated across different training sample sizes. Figure 8.1 and Figure 8.2 show that the training score consistently outperforms the validation score, with the performance gap decreasing over the learning process. This trend indicates an improvement in the model's generalization ability for previously unseen data instances. This improvement is due to the increasing variety of training data, which enables the model to capture more meaningful patterns in the distribution of data points. Learning curves provide valuable insights into the model's generalization and learning process, particularly in the cross-domain scenario, revealing the data variations in gender classification.

8.3.3 Class prediction error analysis

This subsection analyzes class prediction errors to examine ML-based models' performance predicting the writer's gender from text data. It illustrates incorrect classifications made by the model, which improves our understanding of its performance by revealing error-prone classes.

The study visualizes class prediction errors within the cross-domain scenario, as the error rate for misclassifying gender remains consistent across other subsets of the A3C3 corpus. The model constantly shows a low error rate in gender identification for various writing types, as presented in Figure 8.3. This linearity indicates the model's robustness and effectiveness in classifying different text data examples.

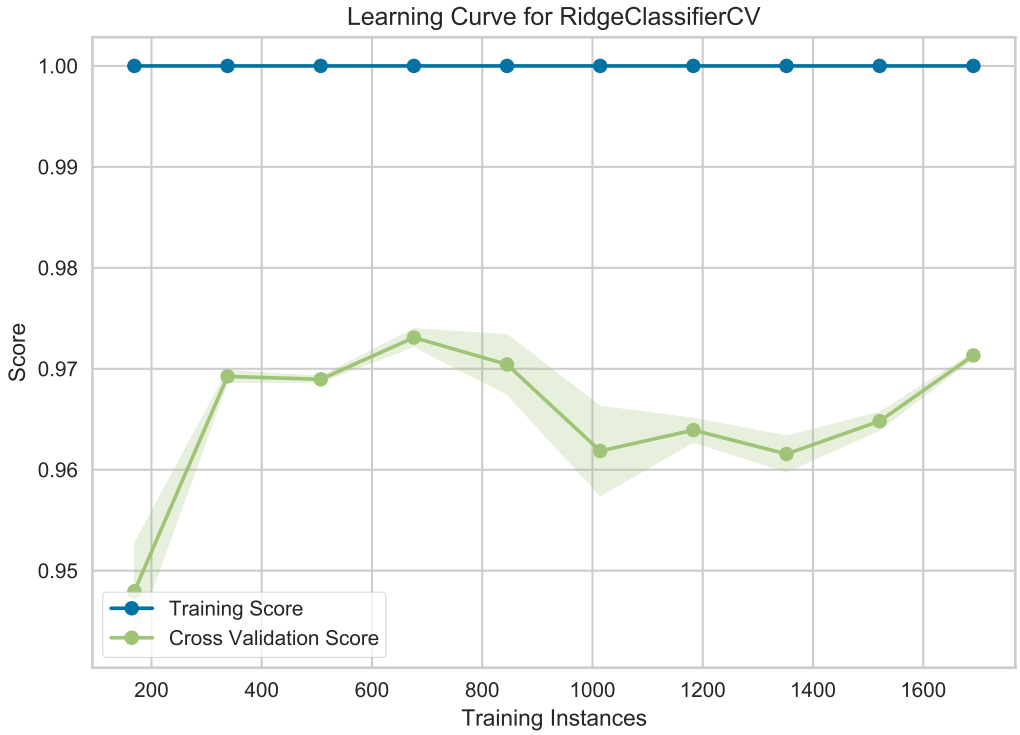


Figure 8.1: Learning curve for gender-related literary chapter-level classification using Ridge classifier and feature engineering.

Misclassifications in the author's gender identification task arise from the complexity of accurately predicting an author's gender based on text. Factors contributing to these confusions include variations in writing styles, gender-neutral language, diverse linguistic contexts, and model limitations.

The class prediction error analysis in gender classification provides valuable insights into the model's abilities and potential challenges, expanding our understanding of the task's variations and complex nature.

8.3.4 Feature analysis

This section examines the features utilized in the gender identification of Albanian texts, as defined in Chapter 4. We aim to understand the importance and impact of these features on this

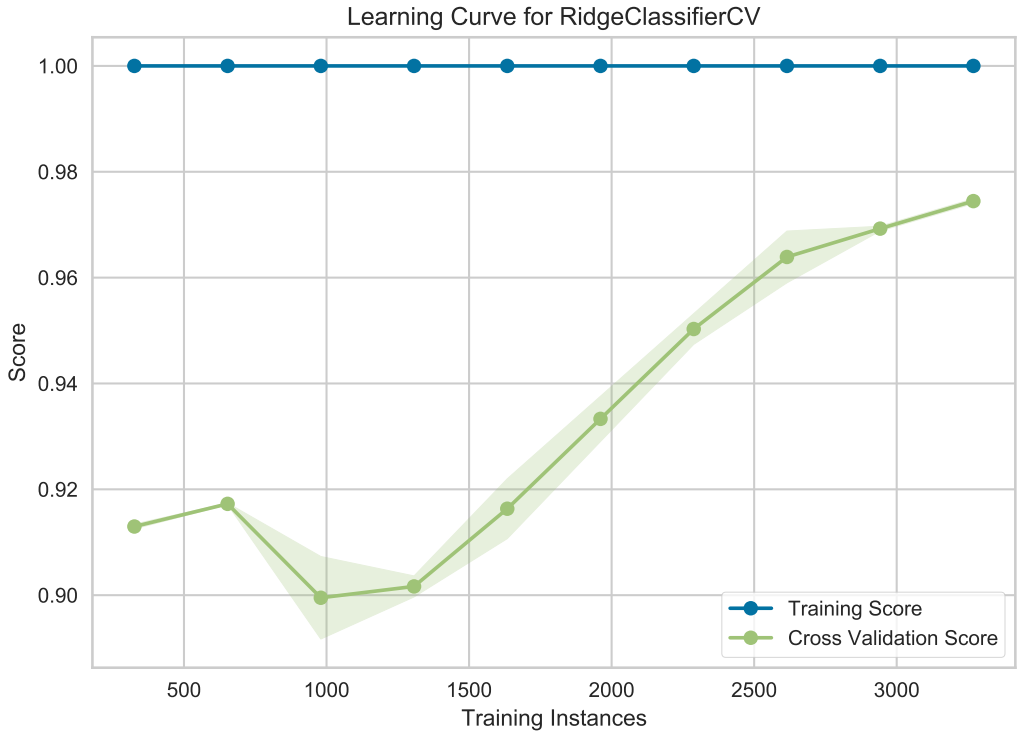


Figure 8.2: Learning curve for gender-related newsroom columns chapter-level classification using Ridge classifier and feature engineering.

classification task. The study uses SHAP scores and waterfall plots to evaluate the significance of individual features in gender identification across various textual domains.

SHAP waterfall plots visually illustrate the importance of features, with each bar representing a specific feature and its influence on the model's predictions. These plots reveal distinctive features with a significant positive or negative impact on the classification output.

Literary works. The analysis reveals that linguistic attributes, such as parenthesis, punctuation, and consonants, are crucial for gender identification in literary works. The SHAP waterfall plot, as shown in Figure 8.4, illustrates the effect of these features.

Literary works exhibit unique narrative patterns, with some authors using parentheses more often. Authors use parentheses to express comments or insights in their writing, which differs between males and females. Literary texts include artistic and figurative language, resulting in varying punctuation preferences. Female authors may use more punctuation to express their

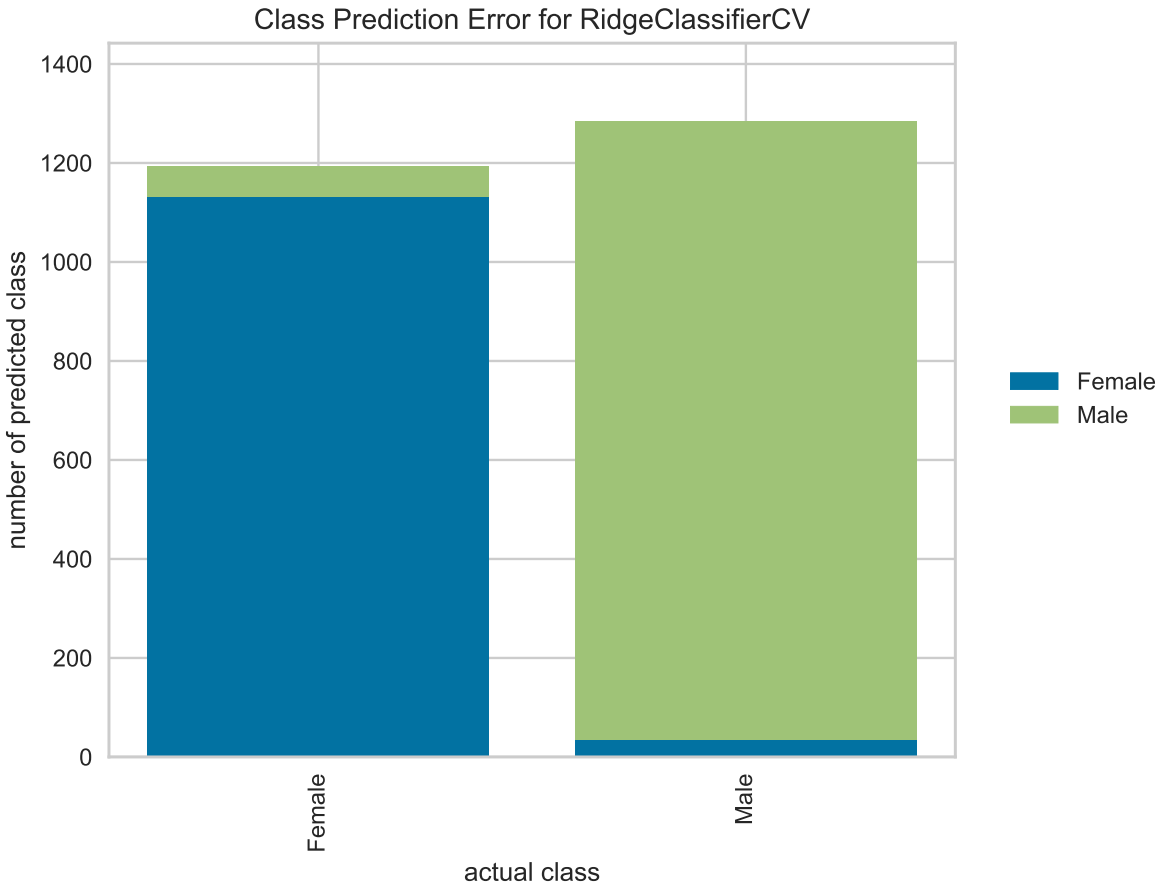


Figure 8.3: Class prediction error analysis for gender-related all writings chapter-level classification using Ridge classifier and feature engineering.

emotions, while male authors use punctuation differently to communicate a specific attitude or rhythm. The 'average number of punctuation marks per sentence' feature represents these stylistic variations. This feature helps identify patterns in punctuation usage by male and female authors and contributes to accurate gender classification in literary texts.

Word usage and sound distribution in writing texts can vary depending on gender. Male and female authors use consonants differently to create a specific tone or feeling in their creations. The 'average consonants per word' feature can identify a stylistic preference for male authors by using more complex consonant words.

Newsroom columns. The frequency and pattern of capitalization contribute to gender-specific writing patterns in newsroom columns. Capitalized words correspond to a distinct writing convention, as shown in the SHAP waterfall plot in Figure 8.5. Male and female writers

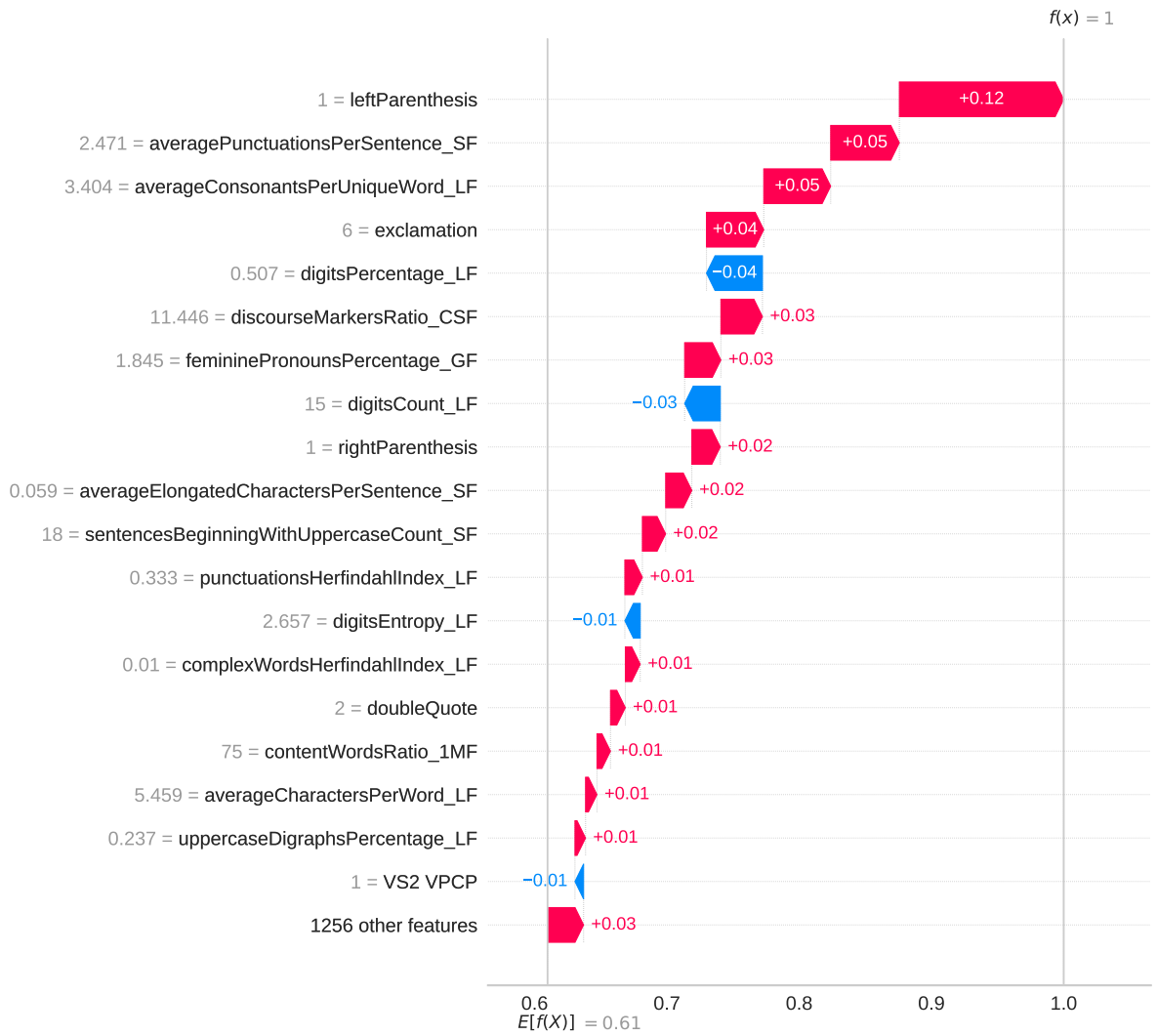


Figure 8.4: SHAP waterfall plot for gender-related literary chapter-level classification using XGBoost classifier and feature engineering.

emphasize different aspects of a particular topic, resulting in a varying frequency of capitalized words.

The 'syntactic phrases Herfindahl index' feature reveals the various and lengthy sentence structures that different authors use to express their ideas. It can be attributed to the author's gender, highlighting the unique preferences of each gender in sentence structure.

Social media posts. The SHAP waterfall plot in Figure 8.6 for social media posts analyzes feature importance based on gender-based variations in various linguistic characteristics.

Single quotes in written speech can represent a specific voice or expression, often used by

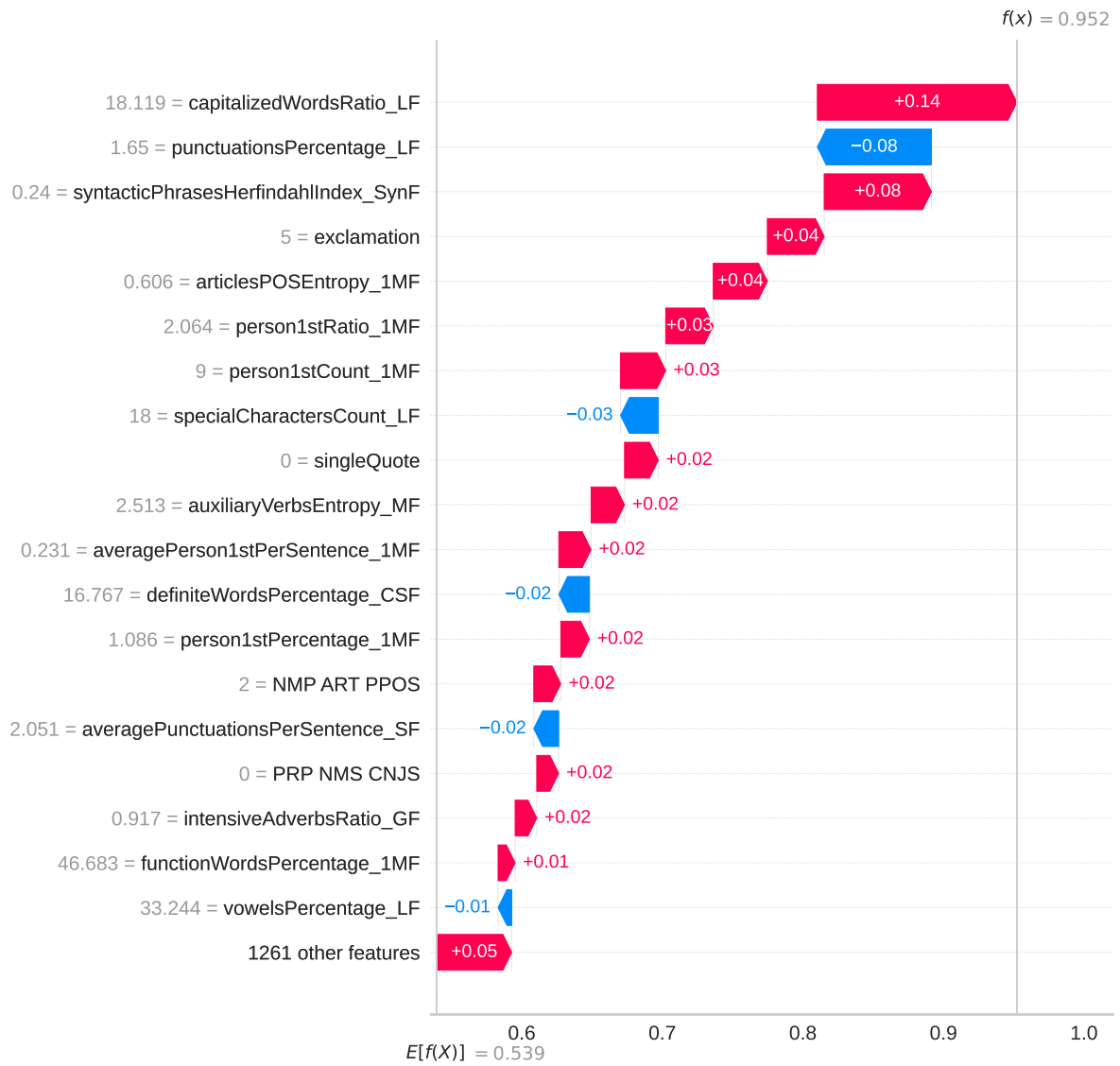


Figure 8.5: SHAP waterfall plot for gender-related columns chapter-level classification using XGBoost classifier and feature engineering.

social media users to share and quote content from various sources. Some authors may use single quotes to express sarcasm or irony or highlight specific words. The frequency of single quotes contributes to distinctive linguistic characteristics and gender-specific writing patterns that help identify the gender of social media users.

People frequently use language as a form of individual expression, intentionally selecting nouns and adjectives that reflect their personality or gender identity. The 'masculine adjective

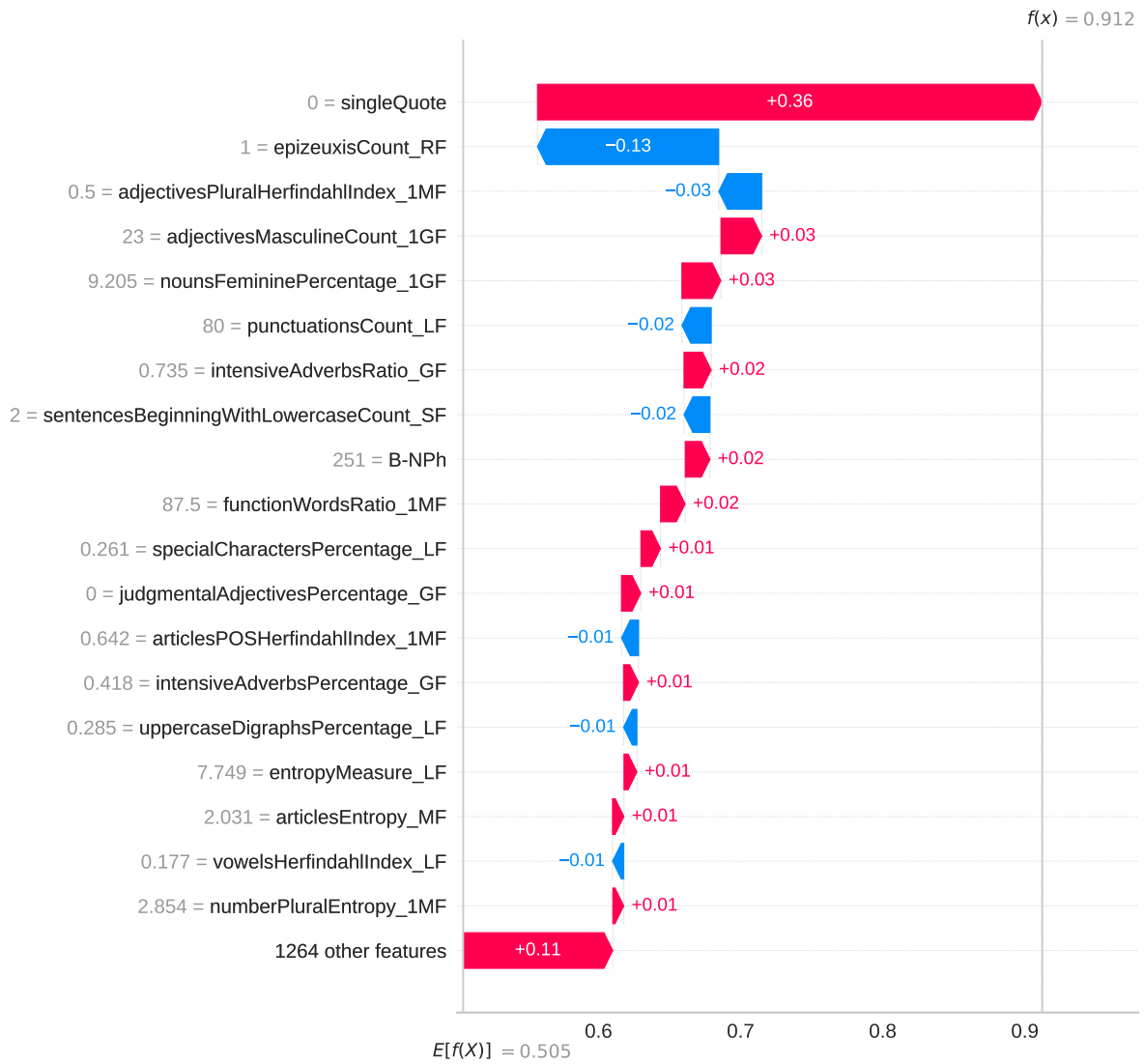


Figure 8.6: SHAP waterfall plot for gender-related posts chapter-level classification using XGBoost classifier and feature engineering.

count' and 'feminine noun percentage' features significantly influence gender identification in social media posts relating to language usage patterns and gender-specific content representation. Male authors may use more adjectives associated with determination or strength, while female writers may use adjectives associated with affection or empathy. These specific markers help define the author's gender.

Cross-domain. Figure 8.7 illustrates the SHAP waterfall plot, analyzing feature importance in the cross-domain scenario.

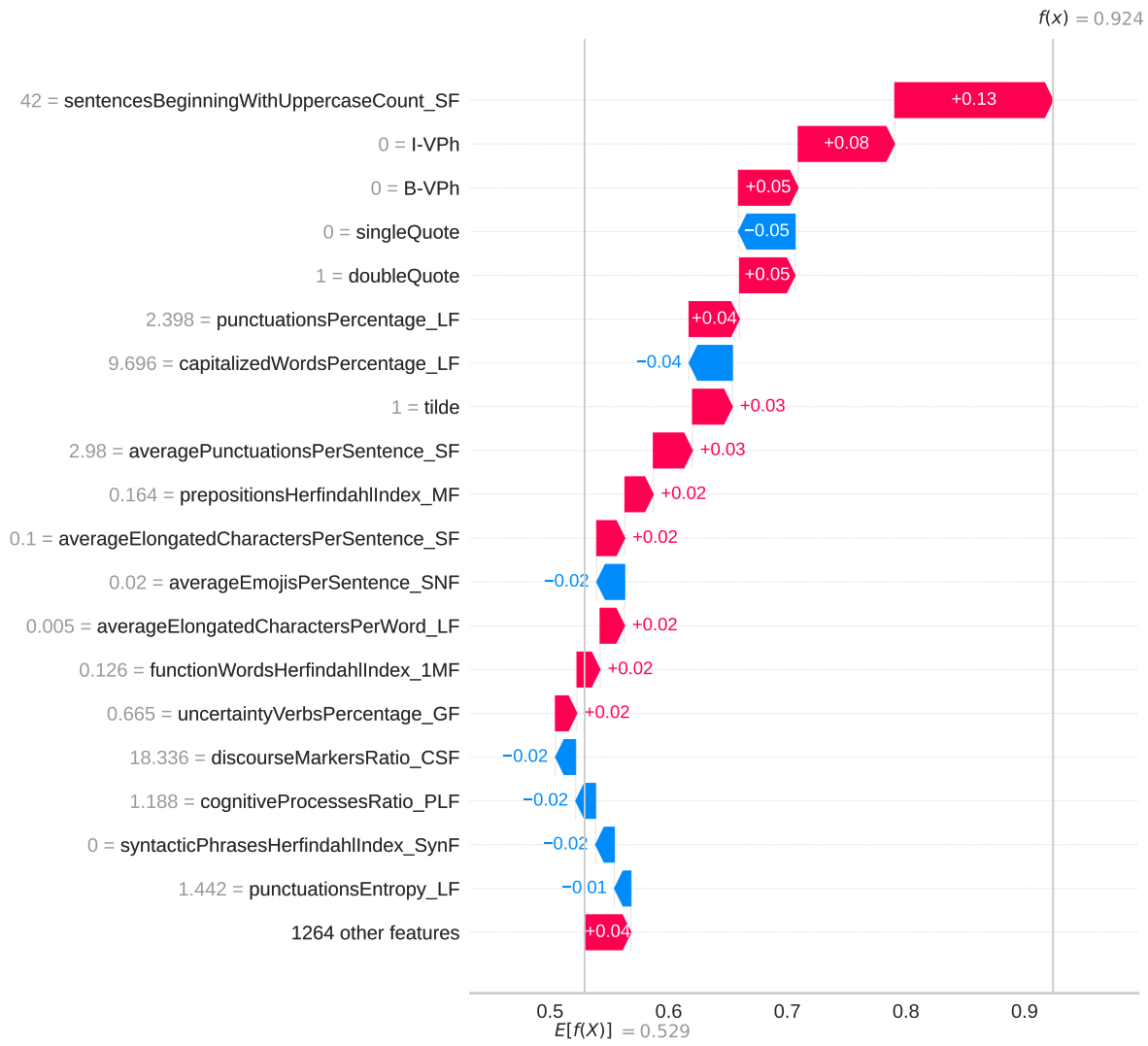


Figure 8.7: SHAP waterfall plot for gender-related all writings chapter-level classification using XGBoost classifier and feature engineering.

The sentence structure and capitalization patterns are other writing characteristics that vary between male and female writers. Capitalized words at the beginning of sentences are related to formal writing and grammatical rules, with choices determined by educational or creation standards.

Gender-related writing habits can influence verb phrase usage, with male and female authors having distinct preferences for communicating actions, emotions, or ideas. One individual might prefer short verb phrases, while the other might use more adverbs or modifiers for more

detailed expressions. The structure of verb phrases and the frequency of active or passive voice in the writing can affect the author's narrative perspective, indicating gender differences.

These characteristics in Albanian texts suggest that they capture different characteristics of an author's gender in various contexts. Understanding the importance of these features increases the model's ability to classify gender in different textual sources correctly.

The SHAP summary plots for each dataset in Appendix C.3.3 present a detailed visualization of feature contributions to the model's predictions. Appendix B.5 provides a deeper understanding of each feature's significance.

Some features have a minimal impact on the model's predictions, providing a recurring pattern across different classes. Excluding these features helps reduce computational complexity while maintaining the algorithm's efficacy.

8.3.5 Feature selection

This section discusses the feature selection approach to determine the most informative feature subset and enhance the model's effectiveness. We used the Ridge algorithm and handcrafted features with RFE in a cross-validated configuration.

The figures in this section represent an RFE curve for a specific subset of the original corpus, illustrating the impact of the iterative feature elimination process on the model's performance. The x-axis corresponds to the number of features, while the y-axis represents the model's cross-validated score with different features.

These curves define two critical points: the peak performance number of features and a threshold number of features with a performance drop of 0.02 from the peak. This threshold indicates the balance between feature inclusion and model stability, providing information on feature significance.

Figure 8.8 illustrates the Ridge model's performance for different feature subsets in the literary domain. Similarly, Figure 8.9 provides insight into the impact of feature selection on model performance in the newsroom columns dataset. Figure 8.10 focuses on social media posts and represents the interaction between features and the model's performance. Finally,

Figure 8.11 visualizes the features' importance and effect on model performance with different text types.

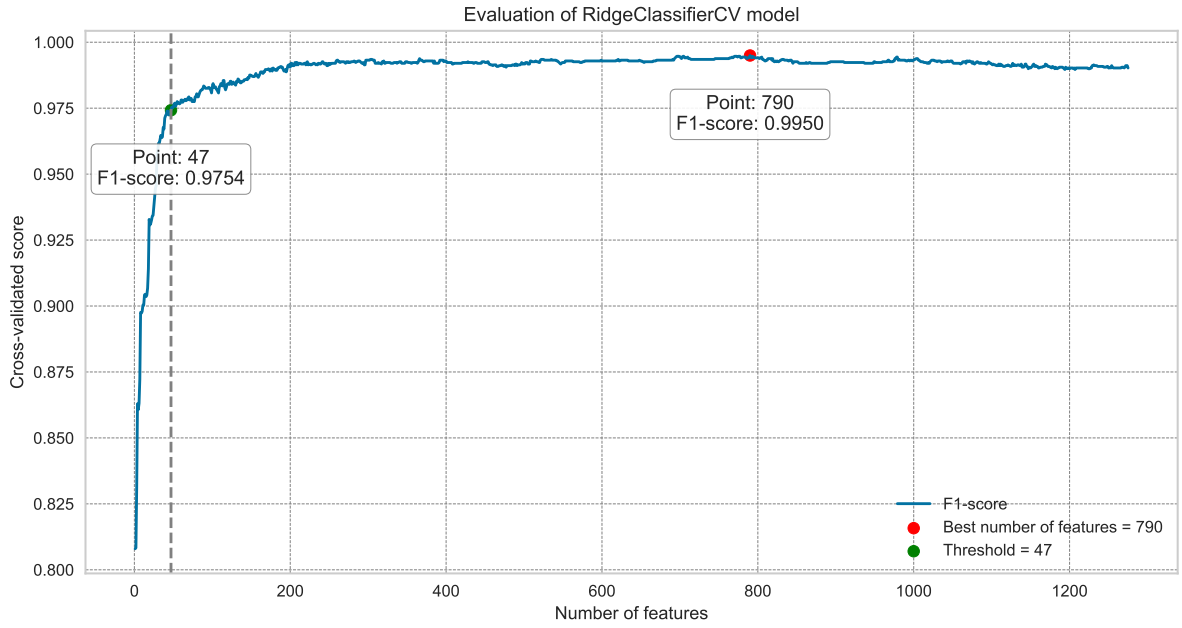


Figure 8.8: RFE for gender-related literary chapter-level classification using Ridge classifier and feature engineering.

The RFE-based feature selection process yields the optimal feature count with the highest F1 score for each dataset and the required number of features for a threshold F1 score with a performance drop of 0.02. Table 8.8 summarizes these outcomes.

Table 8.8: Feature selection outcomes using RFE in a cross-domain configuration.

Corpora	Number of features	F1-score	Number of features (threshold)	F1-score (threshold)
Literary works	790	0.995	47	0.975
Newsroom columns	445	0.976	102	0.957
Social media posts	947	0.950	98	0.930
All writings	1262	0.949	177	0.929

The study aimed to identify the most informative feature combinations in various text data, considering their distinctive language-related and stylometric attributes. This extensive analysis provides insights into optimal feature configurations for each domain and contributes to

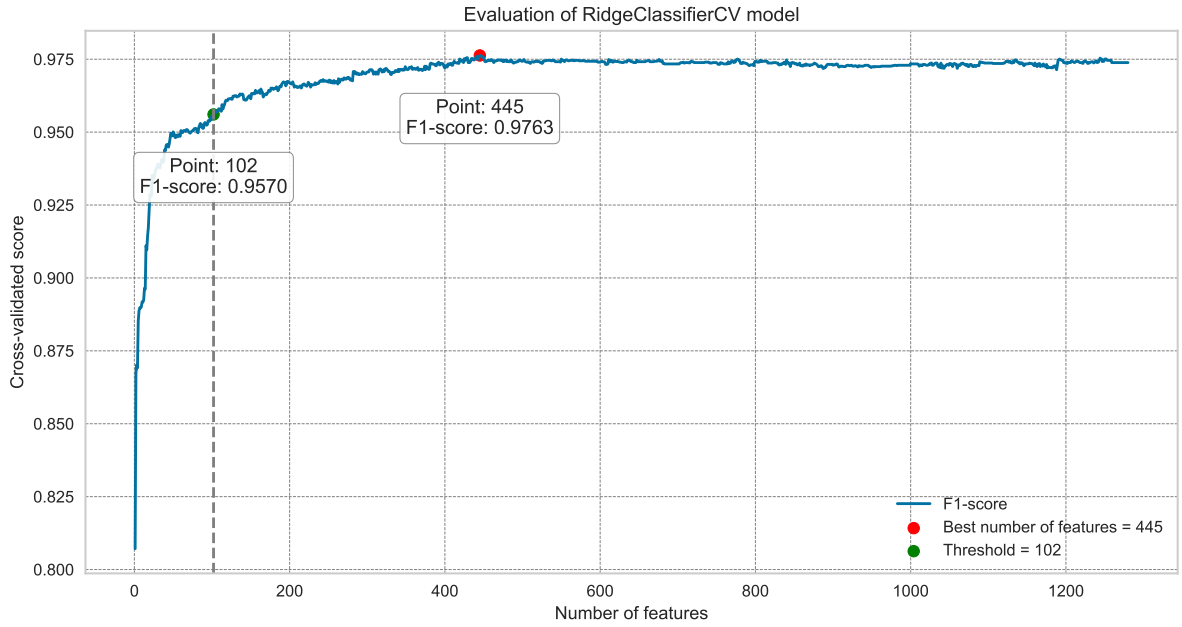


Figure 8.9: RFE for gender-related columns chapter-level classification using Ridge classifier and feature engineering.

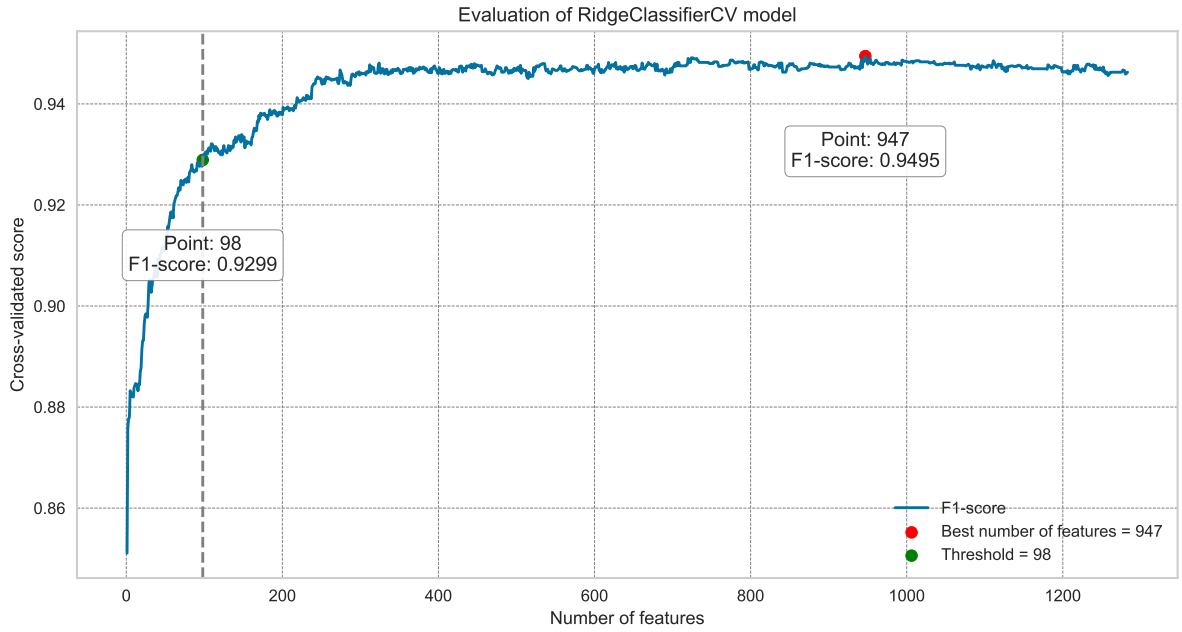


Figure 8.10: RFE for gender-related posts chapter-level classification using Ridge classifier and feature engineering.

the robustness of our gender identification model across various textual sources.

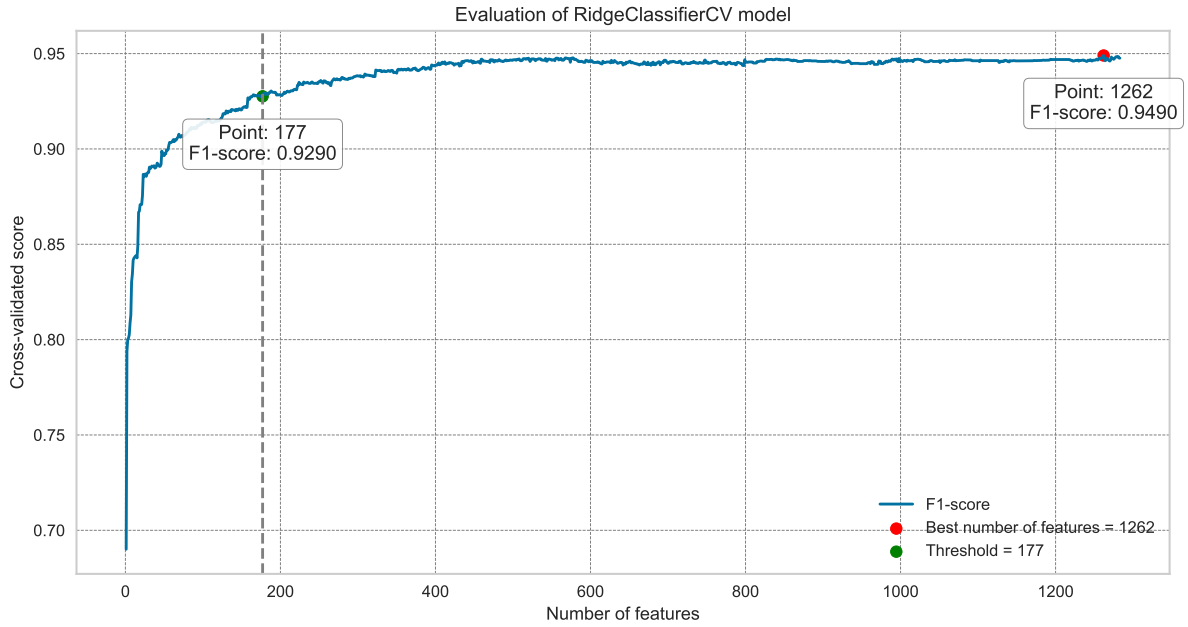


Figure 8.11: RFE for gender-related all writings chapter-level classification using Ridge classifier and feature engineering.

8.3.6 Dimensionality reduction

The section explores the use of PCA and autoencoder models to reduce the dimensionality of the feature space. The following tables present the summarized results.

For the PCA model with results presented in Table 8.9, we conducted experiments to determine the number of components to retain 100% of the data variance. Next, we identify the component count that results in the highest F1 score and the number of features for a performance drop of 0.02 from the peak. The performance curves in Appendix C.3.1 illustrate the effectiveness of the PCA model corresponding to the number of components for each dataset.

The autoencoders model is additionally used for dimensionality reduction (Table 8.10), experimenting with the complete feature array, and increasing the feature count by 50. A threshold point is set to achieve a tolerable reduction while maintaining an informative subset of features. Appendix C.3.2 provides performance curves for the autoencoder-based model.

The results presented in previous tables indicate that the autoencoder model yields impressive performance scores with minimal features and reduced training epochs.

Table 8.9: Dimensionality reduction outcomes using PCA in a cross-domain configuration.

Corpora	Number of components - PCA cumulative variance=1.0	Number of components - Best score	F1-score - Best score	Number of components - Threshold score	F1-score - Threshold score
Literary works	540	417	0.990	36	0.972
Newsroom columns	542	537	0.970	68	0.952
Social media posts	550	524	0.948	189	0.928
All writings	544	520	0.932	117	0.913

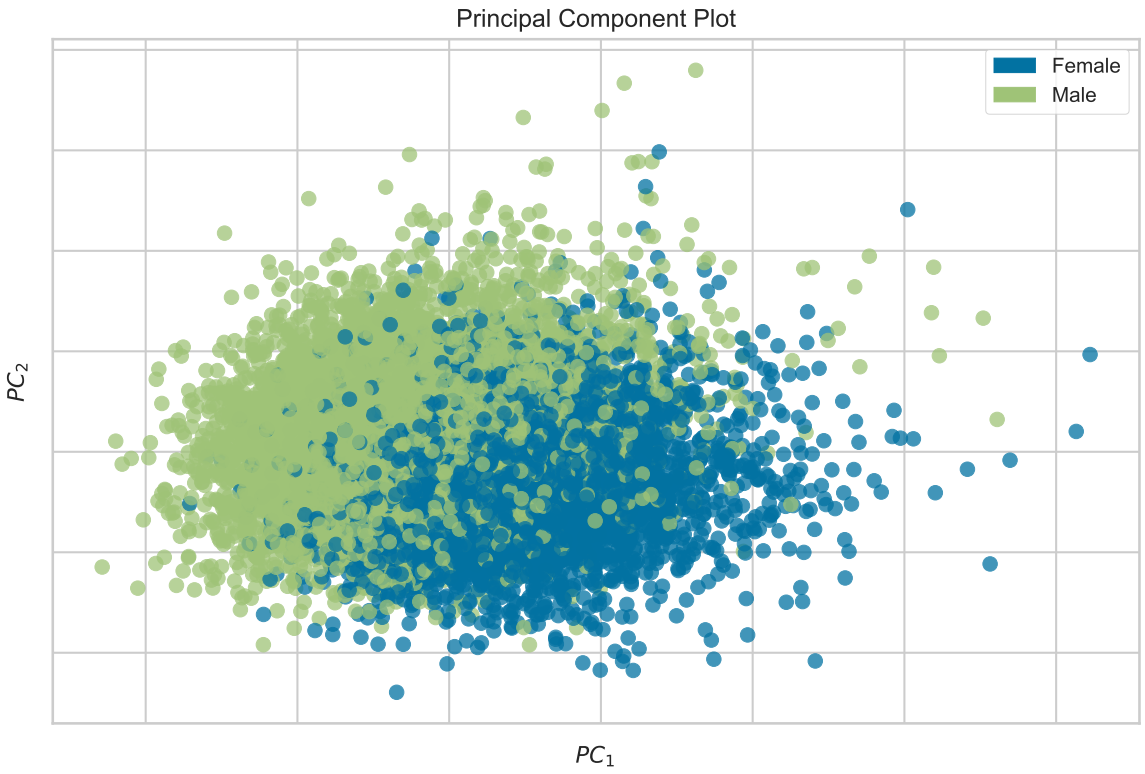
Table 8.10: Dimensionality reduction outcomes using autoencoders in a cross-domain configuration.

Level of analysis	Best point	Best F1-score	Threshold point	Threshold F1-score	Epochs, early stopping (average)
Literary works	600	0.998	50	0.994	24
Newsroom columns	550	0.990	50	0.972	26
Social media posts	1250	0.975	150	0.956	41
All writings	850	0.965	150	0.945	33

The study provides a PCA visualization for the cross-domain scenario due to the consistent pattern in data structure across four distinct corpora. Visualizing data points in this context helps understand data clusters and their relationships.

Figure 8.12 illustrates the distribution and relationships of data points in a two-dimensional space. The distribution of data points provides valuable information about potential gender-related clusters and distinct stylometric patterns.

In the cross-domain scenario, the data clusters representing male and female authors overlap, demonstrating similarities in their written content. However, despite the overlap, clear clusters are evident, demonstrating the existence of linguistic patterns specific to each gender. The trend observed in data groups highlights the complexity of author gender classification, as writing styles may share similar characteristics but also contain specific gender-specific features.



1

Figure 8.12: Illustration of all writings data via PCA dimensionality reduction.

The section demonstrates the effectiveness of PCA and autoencoders in reducing feature dimensionality while maintaining significant information and highlights the impact of the dimensionality reduction technique on model performance.

8.4 Comparison with previous works

The research methodology involves cross-validation with prior work to examine our experimental findings. The comparative analysis is outlined in Table 8.11. This analysis provides insights into the research's contributions, including the robustness of our results, new findings, and improvements. Cross-validation emphasizes the continuing aspect of research and the significance of our work in authorship analysis.

Table 8.11: Comparative analysis with prior studies in cross-domain classification settings (top-performing results).

Reference	Document type	Language	Number of authors	Number of samples	Features	Method	Performance
[20]	tweets	Greek	463	45,848	word-based features (TF-IDF encoding)	SVM (linear)	0.70 accuracy
[85]	chat messages	Turkish*	1,616	218,742	keywords, misspelled, slang and emoticon words, discriminating words	SVM (linear)	0.835 accuracy
[81]	short messages (SMS)	English	-	55,835	word-based TF-IDF (n-gram)	Decision Tree (J48)	0.791 accuracy
our method	literary works	Albanian	30	338	-	transformers	1.000 F1 score
our method	newsroom columns	Albanian	52	654	-	transformers	1.000 F1 score
our method	social media posts	Albanian	44	879	-	transformers	0.992 F1 score
our method	all writings (mixed)	Albanian	122	1,239	-	transformers	0.976 F1 score

8.5 Conclusion and future work

This study explored the gender identification task in Albanian texts using a range of ML- and DL-based techniques. It demonstrates the effectiveness of different configurations and feature sets and identifies an ensemble of linguistic elements significantly impacting gender classification. The research findings contribute to authorship analysis and deepen our understanding of the gender identification task within the Albanian linguistic scenario. The results have provided valuable information from our experiments to address the following research questions:

1. *Can the gender be identified from an Albanian text?*

   The study investigates the relationship between linguistic characteristics and gender identification in Albanian texts, analyzing various linguistic features and methods to identify

gender-related patterns. The experimental outcomes indicate the potential to identify an author's gender across different sources and textual contexts.

2. *Are there linguistic features that indicate gender?*

   The study examines numerous language-related features in Albanian as indicators of an author's gender, contributing to understanding gender-specific language use. It identifies patterns related to gender in punctuation preferences, capitalization standards, syntactic structures, specific adjectives, and nouns. These features reveal consistent gender patterns in various texts and demonstrate the relevance of gender-related linguistic analysis.

3. *What are the best methods for authorship analysis (gender identification) in Albanian?*

   The research identified effective methods for capturing gender-related markers in Albanian authorship analysis. The Ridge linear model can handle diverse feature sets. As state-of-the-art models, transformers show promising results in Albanian texts while identifying the author's gender. The findings affirm the research question and indicate exciting possibilities in Albanian authorship analysis.

The experimental findings contribute valuable information to the Albanian gender identification task and provide potential directions for further research in other textual domains. This research highlights the importance of linguistic analyses to examine different aspects of gender identification in Albanian and establish a foundation for future investigations.

Future research in gender identification of Albanian texts aims to expand the dataset to include texts from different sources and contexts, increasing the robustness of the model. Another aspect is the refinement of feature engineering and the exploration of domain-specific characteristics unique to the language. Adapting the methodology to a multi-lingual scenario offers cross-lingual pattern investigation, aiming to advance gender identification in the Albanian language and beyond with valuable insights and significant contributions.

Chapter 9

Final discussion

This investigation presented significant findings with substantial contributions to the field. Learning curves revealed that no fixed universal data threshold ensures the recognition of authorial uniqueness. The trends in learning curves demonstrated significant improvements in generalization over time, with a decreasing difference between the training and validation scores.

Class prediction error analysis exposed different aspects of model performance in tasks related to authorship. Specifically, authors of the same genre exhibit common linguistic patterns, narrative styles, and thematic focuses in identifying authorship, leading to misclassifications within the dataset. In genre classification, we observed a constant confusion between novels and short stories due to their similarities in content, themes, and narrative styles. However, there is a clear difference in mixed-text settings. Gender identification demonstrated the existence of writing styles not strictly related to gender by revealing style diversity, gender neutrality, and distinctive domain features.

Feature analysis identified distinct linguistic elements contributing to the model's effectiveness in different tasks. The model performance and feature importance analyses demonstrated the complexity of different authorship analysis tasks in Albanian.

The experiments were designed to validate the hypotheses presented in the study, providing a robust experimental approach that enhances the relevance of the research's contributions to the field.

1. *The novel Albanian authorship analysis corpus significantly impacts Albanian authorship-related tasks.*

Our findings strongly support the hypothesis that the novel AAA corpus significantly impacts authorship analysis in Albanian. The novel corpus expands our analysis with diverse linguistic styles, genres, and themes. This dataset enables models to learn from diverse linguistic patterns and styles in Albanian texts. The models generalize effectively and make reliable predictions across various text data, contributing to the effectiveness of Albanian authorship analysis.

2. *There are highly relevant feature groups to be used in training effective and efficient machine learning models for each subtask of authorship analysis.*

   The research indicates that specific linguistic and stylometric features significantly improve the models' ability to identify unique patterns in tasks related to authorship. The feature analysis supports this hypothesis, highlighting the impact of various features in meeting task specifications for optimal results. In authorship identification within literary works, consonants, syntactic phrases, exclamation marks, and commas capture phonetic distinctions, sentence structures, and punctuation preferences. Newsroom columns use exclamation marks, uppercase digraphs, stopwords, and complex words to characterize unique stylistic patterns and content choices. Social media posts use third-person words, negative particles, and short sentences to represent narrative perspective, author tone or sentiment, and concise communication styles. In cross-domain writings, the distribution of characters and long sentences reflects text composition and narrative complexity. For genre classification in the literary corpus, poetry's sentence count signifies its structure, stories' capitalized words reflect narrative style, colons in dramas indicate dialogue or specific points, dialectical words in novellas reflect distinct speech patterns, and novels use proper nouns for rich character development. In the cross-domain corpus, capitalization patterns reflect formal and stylistic conventions in literary works; columns use the third-person narrative style for an objective tone; and social media posts use hashtags to label content. In gender identification within literary texts, parentheses, punctuations, and consonants reflect unique narrative styles, emotional nuances, and gender-specific moods. Newsroom columns use capitalized words for emphasis, syntactic phrases for

linguistic structures, and exclamation marks to denote emotions. In social media posts, single quotes represent specific tones, while masculine adjectives and feminine nouns contribute to gender-specific expressions. In the cross-domain scenario, sentence case signifies formal writing and grammar, and verb phrases reflect gender-related communication preferences in expressing actions, emotions, or ideas. These features capture authorial voices and styles in different linguistic contexts.

3. *Hand-crafted, task-adapted features-based traditional machine learning models will outperform automatic feature extraction-based deep learning models.*

   This research provides documentation of outcomes in evaluating the performance of traditional ML- and DL-based methods in the authorship analysis domain. Hand-crafted features capture different aspects of language patterns and stylometric elements for various tasks related to authorship. On the other hand, DL-based methods automatically learn representations and characteristics from text data without explicit manual engineering. Research findings indicate that ML-based algorithms with task-adapted features perform better on specific authorship analysis tasks and configurations. The study showcases the efficacy of manual-generated features in providing linguistic and stylistic data for authorship analysis, aiding the model in identifying text data patterns.

In the study of Albanian AA, we experienced situations that presented limitations, expanding our understanding of linguistic complexity. The complex morphological structure and dialectal variations of the Albanian language result in significant difficulties in identifying standardized language-related features. This phenomenon affected the performance of the models. The diverse genres provided diversity, requiring genre-specific analyses and feature engineering. The lack of ready-made datasets limited the depth of the study, impacting model training and evaluation. These limitations indicate the complexities of authorship analysis in Albanian and suggest further research and exploration to address these challenges.

Chapter 10

Conclusion and future work

10.1 Conclusion

Authorship analysis in low-resource languages, such as Albanian, has presented various difficulties. To address this, we created the first Albanian corpus adapted to authorship challenges with different types of texts. This dataset provided a solid foundation for further analyses and investigations.

Our analysis used various ML and DL-based methods to evaluate the effectiveness of different feature settings and configurations. A significant contribution is feature engineering, which represents different characteristics and patterns in authors' writing patterns and styles. The research on feature selection and dimensionality reduction techniques demonstrated their potential to increase model effectiveness while preserving relevant information. This exploration of authorship analysis in different linguistic domains in Albanian provided essential information into the challenges and possibilities of the field.

This study experimentally presents the potential to identify authorship, genre, and gender in Albanian written works. Our research provides promising results on a diverse dataset, demonstrating the methods' reliability and generalization ability. The existence of distinctive patterns and characteristics in language use helps with effective authorship identification, understanding of gender-related features, and genre differentiation.

The study identified authorial patterns, gender indicators, and genre attributes through a detailed analysis of hand-crafted features. The analysis reveals elements that define author

writing habits in Albanian, ranging from sound distribution, structural elements, punctuation choices, and POS tags. Our findings indicate that no universally specific feature set exists for authorship analysis. Instead, the features' impact on the model's performance differs based on context, domain, and text characteristics. The experimental results demonstrate the reliability of the linear Ridge model, considering linguistic variations. This study employed cutting-edge methods such as autoencoders, transformers, and FastText. Impressive results in cross-domain data demonstrate the potential of our research and contribute to authorship analysis in Albanian texts.

This research demonstrates the implicit complexities of the Albanian language within its context and suggests further investigation and advancement in this area.

10.2 Future work

This research contributes to a deep understanding of authorship styles and provides a robust foundation for future research. However, some directions are given for enhancing the effectiveness of methods for specific tasks related to authorship, with explicit emphasis on authorship identification, genre classification, and gender identification.

The dataset's continued development and extension aim to maintain its significance and include texts from various genres, linguistic styles, and periods. Furthermore, it supports the investigation of cross-domain authorship analysis and the promotion of multi-lingual research.

The expansion of domain-specific and psycho-linguistic features for specific tasks related to authorship provides more adaptive feature representations. Research into cross-domain and cross-linguistic feature extraction provides new opportunities for authorship analysis. Furthermore, transfer learning techniques can help adapt models trained in other languages or domains to Albanian authorship analysis tasks.

The findings from this research have provided a foundation for authorship analysis tasks. Future efforts may include tuning model settings, integrating advanced model architectures, and examining the significance of domain knowledge.

Appendix A

Sample texts from the corpus

A.1 Author-related data

Table A.1: Representative excerpts from the corpus of literary works.

Literary works
Dhe Bajrami doli nga dera, i bindur për ato që kish menduar dhe i kënaqur për atë që kishte bërë. - Oh, unazë! Madje, sfidimi i tij na qe kthyer në njëfarë zbavitjeje. Sytë i qenë zgurdulluar tej masës dhe qe gati t'i hidhej plakës; sepse i dukej që ajo me ato ushtrimet e saj po ja. Xhexhoja ngjiti shkallët. Në këtë mënyrë tregimi Hijet në kuvertë krijon një klithje deni në skaj, një klithje për humanitetin që ka arritur në kufirin e zhdukjes. Era ulërinte si mallkim i tokës spanjolle dhe tërë batalioni Ish strukur ndër shtrojera e guva, ledhe të pështjella me gjethe.

Table A.2: Representative excerpts from the corpus of newsroom columns.

Newsroom columns
Do të duhet kohë për ta rindërtuar stereotipin, i cili i lidhte racat, religjionet, apo etnitë e caktuara me krimin dhe i amnistonte disa të tjera kjo. Kryesorja, është se kohët e fundit, kanë ardhur duke i parafruar si kurrë më parë qëndrimet e tyre në skenën politike mbarëkombëtare. E pikërisht, këtu ku duket se ndodhet pika më e fuqishme që i bashkon, është edhe pika më e dobët e kufirit që i ndan ata jo thjeshtë si programe, si filozofi, si kahje në spektrin politik, si qasje, si frymë apo si gërmë. Eshtë formë e pastër korrupsioni, madje edhe një tufë lule, kur të tjerë nxënës të varfër nuk e përballojnë dot ta blejnë, kthehet në tmerr dhe ofendim për ta, korrupsion që vazhdon prej vitesh, denonconi të gjithë ata që abuzojnë me emrin e lartë të mësuesit dhe ju kërkojnë dhurata dhe ju rrjepin.

Table A.3: Representative excerpts from the corpus of social media posts.

Social media posts
Kah me ja mbajt tash ne që kujtuam se shpëtuam?Jam e mllefosur, sepse veç në Albinin dhe në Vjosën kam pas shpresë. E fundit ne Europe. Mbrëmë, ka nisur shtrimi me zhavorr i rrugës, në mënyrë që kalimi i këmbësorëve dhe lëvizja e veturave të jetë më e lehtë. Asnjë qindarkë për shuarjen e qindra zjarreve në gjithë Shqipërinë duke shkaktuar vdekje, djegie pyjesh, kullotash, dhe duke vënë në rrezik fshatra të tërë, etj ! Qoftë kjo festë gëzim për të gjithë! Integrimi i të gjithë pjesëtarëve të shoqërisë pa dallime është rrënjësor për zhvillim kolektiv, prandaj i lumtë Down Syndrome Kosova që po ua mundëson të rinjëve me sindromën Down një jetë të tillë.

A.2 Genre-related data

Table A.4: Representative excerpts from the corpus of drama genre.

Drama
MIKELI: Deri këtu arrin të shkojë “krenaria” jote, Petrit?
Sulejmani: Bejtash aga, çka don prej meje.
SADIKU: Ç’lidhje ka vdekja e Markusit me fejesën tonë?
Rrushja: Tashti do ta heq unë. Kam sjellë lajme.
Bardhoku: Kah po ia mban kështu?
Më ndihmon, Xhafer, t’i çoj këto plaçka në akshihane.
Mirëpo jam nanë, o Murrash.
Tuçi: Kadale, kadale, mos u nxej!

Table A.5: Representative excerpts from the corpus of poetry genre.

Poetry
Edhe unë dashurova dhe në jetë pata fat,
lexoj para popullit.
As mua nuk më njohin të tjerët
Sa herë që ia behën
Jo përkundrazi
Të varreve të stërgjyshërve
në zot hetuan gjarpërin
Mbeta lakuriq rrugëve
viset e tua kreshnike nuk i shkeli dot njeri.

Table A.6: Representative excerpts from the corpus of novella genre.

Novella
Zaltën, shumë prej jush keni zbritur nga skena dhe jeni mplakur, prandaj flisni kështu". Seanca mbyllet, tha ai. - Jo dhe aq, ikim!- tha ajo. Ato dy ditë kaluan si në ëndërr mes kënaqësisë së shoqërisë së Mirit që ish më i ëmbël e më i dashur se kurrë dhe etheve të kërkimit të ndonjë shenje që të vërtetonte dyshimet e saj. Pirgu, siç na e përshkroi ai, është jo fort i lartë, sa dy shtate njeriu, dhe nga ana ushtarake s'ka kurrfarë vlere përballë Murit tonë, por tmerri që dërgon është sa i njëqind kështjellave bashkë. Kjo gjë e preku dhe e mallëngjeu. - Me kujdes, me kujdes, - u dëgjua zëri i infermieres, që kishte shërbimin atë ditë. Fjalët e saj, si gugitje pëllumbeshe, ishin jo vetëm të dashura, por edhe të zgjuara, të matura, ajo ishte e mençur, e ditur dhe shpirtmadhe.

Table A.7: Representative excerpts from the corpus of novel genre.

Novel
Ngaqë gjatë një pjese të shtegtimit vëmendja i shpërqendrohej sa andej këndej, përherë e më tepër Gjorgut po i dukej se udhët që përshkonte ai ishin pa vazhdimësi, plot boshllëqe apo kapërcime të mëdha. Shpërblimi për veprat e mira, do të jetë lumturia e përhershme, pa vuajtje; kurse ferri, si diçka e kundërt. Tradhtarë i ndyrë! Se sa e çmojnë banorët e këtushëm tokën ku mund të mbjellësh çkado qoftë që rritet dhe bulëzon, dëshmojnë edhe varrezat që nuk janë të vendosura në një vend, por nga të katër anët e fshatit, nëpër vendet përplot gurishte ku vetëm ata e dinë sa e vështirë është të hapësh varr për të ndjerin.

Table A.8: Representative excerpts from the corpus of story genre.

Story
Dredhoi kah shkallët. Dhe kur zu të shkundet si ai mani i nunit, buzëqesha. - Ka nga ata që digjen në zjarr ditë e natë. Po kjo velje e tij ne nuk na fyeu aspak. - Sa vesh të mirë keni, - kishte thënë me admirim. Hë, mo, e ke parë? Të pranishmit të tërë qeshën bile njëri e pyeti: - A është e vërtetë xha Bit?! Këtu, për një kohë, mendimi iu pre. Edhe kur takoheshim në qytet (ai banonte në qytet se unë tërë jetës kam banuar në fshat), ne takoheshim në klub dhe, kur qëllonim të qemë të dy kokë për kokë, digjnim vajguri me hallet e botës.

A.3 Gender-related data

Table A.9: Representative excerpts from the corpus of male gender.

Male
Çfarë bëra?
Për shembull hakerat iranianë u përpoqën të vendosin një farë disipline me skandalet.
Shqipëria dhe Kosova bashkëpunojnë ngushtë.
As që bëhet fjalë!
Kështu, rimbushet me tym e mjegull Lugina e Betontë e Kosovës.
Si risim propozoi që fill mbas gjimnastikës të tërhiqej secili në vete dhe të lexonte dy-tri faqe të zgjedhuna ndër libra autorësh të pazhigla e ideologji të caktueme të shekujve nandëmbëdhjet e njëzet.
Mjafton të marrësh në analizë (pa u ndalur shumë) drejtuesit politikë në qarqe, për të konstatuar sa e ka telendisur Partinë Demokratike, amatori politik dhe aktori Basha.

Table A.10: Representative excerpts from the corpus of female gender.

Female
Por ajo, e kujdesshme dhe e vetëdijshme për rolin që kishte, nuk zuri në gojë asgjë të tillë.
Këtu dua të ndalem tek pozita shumë inferiore e gjuhës shqipe në përdorimin zyrtar të saj.
Por, ka vetëm një regjion, për të cilin nuk diskutohet.
- Në këtë orë?
E djaloshi, me sytë e mbyllur shtrëngon me të dy duart timonin e autobusit dhe shkel përplot vetëbesim dhe krenari papuçen e gazit.
Ato janë krijesa të thjeshta, formojnë natyrshëm ekosistemin e tyre, i cili ka një ndjeshmëri shumë të lartë.

Appendix B

Detailed description of the engineered feature sets

B.1 Linguistic features

B.1.1 Morphological features

Table B.1: List of particles included in the morphological feature set.

Particles									
edhe	duke	vetëm	bash	tepër	more	qoftë	madje	sesa	mbase
deri	gjer	sikur	gati	ndoshta	sidomos	afro	pikërisht	vallë	qysh

Table B.2: List of prepositions included in the morphological feature set.

Prepositions										
në	me	për	më	nga	pos	deri	gjatë	gjer	larg	nën
sipas	midis	tek	ndër	mes	ndaj	rreth	mbas	kundër	jashtë	pas
para	mbi	nëpër	prej	drejt	pranë	sipër	përpara			

Table B.3: List of conjuctions included in the morphological feature set.

Conjuctions											
edhe	kur	por	pra	apo	vetëm	prej	ose	nëse	sikur	ku	në
ndaj	siç	kurse	sepse	sesi	pasi	pranë	andaj	përpara	meqë	si	as
dhe	nga	se	po	që	sa						

Table B.4: List of interjections included in the morphological feature set.

Interjections												
a!?	he..he..	he	pa	ish	ej	ou	shët	oj	more	bis	bah	bam
or	ua	ua!?	urra	hë	oh	ptu	ee	eh	oo	vaj	aha	mori
oi	eu	ah	ori	ih	hëm	au	of	uf	ysh	ehu	psht	pis
uj	ore	uh	hej	ura	jau	mor	ehe	oho	bërr	ohu		

Table B.5: List of stopwords included in the morphological feature set.

Stopwords									
edhe	kishte	është	ishte	duke	gjatë	kështu	ashtu	bërë	tani
shumë	këtë	prej	vetëm	tyre	sikur	pastaj	deri	këtu	gjitha
mund	janë	parë	herë	gjithë	vetë	qenë	madhe	ndonjë	sepse
kanë	mirë	kishin	ishin	nëpër	marrë	jetë	kemi	kësaj	tonë
tjetër	para	disa	këto	duhet	fund	cili	atje	kurrë	diçka

B.2 Domain-specific features

B.2.1 Content-specific features

Table B.6: List of abstract concepts included in the content-specific feature set.

Abstract concepts										
jetë	punë	kohë	qetësi	fjalë	histori	dashuri	drejtë	art	liri	trim
frikë	uratë	tmerr	faj	egër	vdekje	frymë	zemër	gaz		

Table B.7: List of archaisms included in the content-specific feature set.

Archaisms									
ves	mëri	brat	ulk	oskan	delme	ulka	brisa	mrikë	oleum
satem	kjaj	rinos	famulli	klaj	franë	familia	bërrakë	gluha	zagorje

Table B.8: List of international concepts included in the content-specific feature set.

International concepts								
histori	kimi	kesh	rimë	person	art	bar	krim	komunist
letrare	gjeneral	organ	graf	tiran	limit	grup	kafe	ballkan
familje	profesor	test	nato	parti	tekst	shkollë	univers	magji

Table B.9: List of discourse markers included in the content-specific feature set.

Discourse markers									
dhe	për	se	po	që	ashtu	kaq	kështu	pastaj	veç
edhe	pas	por	pra	para	kurse	sepse	pasi	atëherë	përpara
vetëm	ende	tjetër	nëse	tepër	sipas	meqë	ndërsa	prandaj	sesa

B.2.2 Gender features

Table B.10: List of affective adjectives included in the gender feature set.

Affective adjectives						
mirë	qetë	bukur	qartë	lumtur	këndshme	mrekullueshme
ëmbël	paqe	shenjtë	gëzuar	kënaqur	gatshëm	gjakftohtë
pushtuar	hyjnor	freskët	tronditur			

Table B.11: List of assertion words included in the gender feature set.

Assertion words						
qartë	pikërisht	patjetër	sigurisht	plotësisht	konkretisht	besueshëm
natyrisht	njëherësh	tërësisht				

Table B.12: List of independence-related words included in the gender feature set.

Independence-related words					
liri	pavarësi	çlirim	personalitet	sovran	vetëqeverisje
autonomi	individualitet	sundues	vetëvendosje	sovranitet	decentralizim
vetëbesim	emancipim	suprem			

Table B.13: List of judgmental adjectives included in the gender feature set.

Judgmental adjectives					
mirë	bukur	habitshme	hijshme	çuditshëm	mrekullueshëm
trishtuar	imagjinar	këndshëm	tmerrshëm	frikshëm	jashtëzakonshëm
padurueshme	tronditës	urryer	shkëlqyer	mërzitshme	mesatare
fantastike	mesatar				

B.2.3 Social-network features

Table B.14: List of social media terms included in the social-network feature set.

Social media lexicons									
kërko	tag	lajm	lidhje	grup	link	live	kërkim	ndjekur	vlerësim
ndjek	post	like	fotografi	ndiq	koment	mesazh	storie	status	etiketë

B.3 Psycho-linguistic features

Table B.15: List of cognitive processes included in the psycho-linguistic feature set.

Cognitive processes							
punë	mendim	shkak	çudi	kujdes	bisedë	kritikë	kujtesë
kuptim	dyshim	lidhje	kujtim	problem	ndërmend	vetëdije	mësim
besim	çështje	tregim	menduar	studim	vëmendje	zbulim	bazë
përfundim	dëshmi	shprehje	kuptuar	rrëfim	nxë		

Table B.16: List of neutral emotions included in the psycho-linguistic feature set.

Neutral emotions						
drejtë	asnjënës	sistematik	faktik	thjeshtë	mesatarisht	zakonshëm
neutral	moderuar	paanshëm	rëndom	normal	tipik	barabartë
përgjithshëm	objektiv	natyror	formal	rëndomtë	indiferent	

Table B.17: List of time concepts included in the psycho-linguistic feature set.

Time concepts									
erë	dje	orë	ditë	kohë	atëherë	herët	kaluar	nesër	vjetër
tash	tani	natë	vjet	pastaj	shkuar	fillim	mëngjes	lindje	shekull
sot	çast	shpejt	viti	mot	menjëherë	mbrëmje	muaj	nesërm	stinë

Table B.18: List of positive emotions included in the psycho-linguistic feature set.

Positive emotions							
gaz	liri	trim	dashuri	shpirti	qëndrim	qeshje	përpjekje
fuqi	zemër	drejtësi	ëndje	pasuri	respekt	kënaqësi	paqe
rëndësi	dije	qëllim	lidhje	kujtim	vullnet	shpirtërore	qetësim
vlerë	shëndet	rregull	besim	magji	vëmendje	thellësi	tërësi
nur	gëzim	intelekt	bukuri	dëshirë	pavarësi	prani	forcë
durim	zhvillim	guxim	shpresë	kujtesë	ndihmë		

Table B.19: List of negative emotions included in the psycho-linguistic feature set.

Negative emotions							
zi	mjerë	krim	keq	hidhur	pikëllim	sulm	dhembje
mani	zët	turp	mëri	zor	mërzi	shkretë	trishtuar
thatë	ligë	frikë	shuar	tmerr	trishtim	fatkeqësi	shqetësim
ulur	dyshim	rrezik	vetmi	vrarë	ndotë	mundim	vetmuar
gabim	varur	dhimbje	bosh	varfër	tmerrshme	dënim	ankth

Table B.20: List of social processes included in the psycho-linguistic feature set.

Social processes							
punë	nder	liri	luftë	urim	mësim	marrëveshje	gabim
mendim	shtim	bind	frikë	hapur	pushim	forcë	përpjekje
veprim	kuptim	kënaq	lidhje	besim	shembull	durim	zhvillim
drejtim	besoj	krijim	gëzim	mbledhje	ndërtim	shqetësim	zgjedhje
dëshirë	guxim	bisedë	shpresë	bashkim	çlirim	kalim	rrëfim
ndihmë	ndryshim	marrje	qëndrim	udhëtim	respekt	shprehje	

B.4 Calculation of the feature set

Code Listing B.1: Calculating the complete set of extracted features.

```
functions = globals()

user_defined_functions_LF = [name for name,
                         func in functions.items()
                         if not name.startswith("__")
                            and inspect.isfunction(func)
                            and name.endswith("_LF")]

for feature_name in user_defined_functions_LF:
    features_names.append(feature_name)
for feature_names_freq in user_defined_functions_LF_freq:
    returnValue = functions[feature_names_freq](str(""), str(folder))
    keys = list(returnValue.keys())
    for e in keys:
        features_names.append(e)

for txt, label, pos in zip(X, y, POS):
    feature_res_l = []

    for name in user_defined_functions_LF:
        feature_res_l.append(functions[name](str(txt)))

    for name in user_defined_functions_LF_freq:
        returnValue = functions[name](str(txt), folder)
        values = list(returnValue.values())
        for e in values:
            feature_res_l.append(e)
```

B.5 Feature importance scores

Table B.21: Feature importance scores for author-related literary works in the chapter level.

Rank	Feature	Score	Rank	Feature	Score
1	graveAccent	0.128634	31	repeatedConsonantsCount_RF	0.006208
2	underscore	0.076144	32	comma	0.006091
3	adjectivesCount_1MF	0.073273	33	numberSingularRatio_1MF	0.0059
4	genderFemininеCount_1MF	0.034916	34	PRTV PCL1 VS3	0.005838
5	averageVerbsPluralPerSentence_1MF	0.033483	35	averageSpecialCharactersPerSentence_SF	0.005798
6	averageDigitsPerSentence_SF	0.023245	36	pastParticipleVerbsCount_1MF	0.005782
7	longSentencesCount_SF	0.023134	37	numberPluralPercentage_1MF	0.005659
8	complexWordsEntropy_LF	0.02276	38	averagePunctuationsPerSentence_SF	0.0053
9	PREL	0.017668	39	averagePerson3rdPerSentence_1MF	0.005127
10	averageIndefiniteWordsPerSentence_CSF	0.017091	40	uppercaseWordsEntropy_LF	0.004876
11	elongationCharactersCount_LF	0.016603	41	capitalizedWordsRatio_LF	0.004759
12	articlesPOSHerfindahlIndex_1MF	0.015061	42	averageNounsSingularPerSentence_1MF	0.004615
13	averageAbstractConceptsPerSentence_CSF	0.012973	43	adjectivesRatio_1MF	0.004566
14	conjunctionsPOSPercentage_1MF	0.012798	44	assertionWordsPercentage_GF	0.004562
15	pastParticipleVerbsPercentage_1MF	0.012453	45	lowercaseLettersPercentage_LF	0.004302
16	averageConjunctionsPOSPerSentence_1MF	0.01118	46	syllablesPercentage_LF	0.004108
17	NFS PPOS	0.01107	47	syntacticPhrasesCount_SynF	0.003975
18	exclamation	0.010938	48	complexWordsHerfindahlIndex_LF	0.003787
19	nounsCount_1MF	0.010727	49	averageAssertionWordsPerSentence_GF	0.003674
20	complexWordsPercentage_LF	0.010425	50	averageArticlesPerSentence_MF	0.003405
21	PDEM NMS	0.010178	51	socialProcessesPercentage_PLF	0.003261
22	subWordsRatio_1MF	0.009326	52	averageSubWordsPerSentence_1MF	0.00321
23	averageCharactersPerWord_LF	0.008539	53	shortWordsRatio_LF	0.003154
24	averageConsonantsPerWord_LF	0.008155	54	definiteWordsRatio_CSF	0.003106
25	NP CNJC	0.007858	55	auxiliaryVerbsHerfindahlIndex_MF	0.003064
26	punctuationsPercentage_LF	0.007624	56	VS3	0.002958
27	auxiliaryVerbsSingularPercentage_1MF	0.00732	57	posTypeHerfindahlIndex_1MF	0.002909
28	specialCharactersEntropy_LF	0.006961	58	nounsFemininеRatio_1GF	0.002802
29	properNounsCount_1MF	0.00635	59	I-NPh	0.002767
30	sentencesBeginningWithLowercaseCount_SF	0.006246	60	+ 156 other features	0.17327542

Table B.22: Feature importance scores for author-related newsroom columns in the chapter level.

Rank	Feature	Score	Rank	Feature	Score
1	singleQuote	0.031193	31	averageSpecialCharactersPerSentence_SF	0.006569
2	functionWordsRatio_1MF	0.023907	32	questionMark	0.006418
3	CNJS	0.019347	33	masculineSuffixWordsCount_GF	0.006349
4	fullStop	0.018233	34	abbreviationsRatio_1MF	0.005974
5	averageProperNounsPerSentence_1MF	0.018045	35	syllablesRatio_LF	0.005751
6	ART NMP	0.017371	36	adverbsRatio_1MF	0.005739
7	punctuationsCount_LF	0.017057	37	sichelsS_LF	0.005579
8	allVerbsSingularRatio_1MF	0.016318	38	averageShortWordsPerSentence_SF	0.005454
9	VS1	0.014792	39	averageDigitsPerSentence_SF	0.005431
10	PRTS VS3	0.014724	40	particlesPOSRatio_1MF	0.005243
11	capitalizedWordsCount_LF	0.013273	41	punctuationsPercentage_LF	0.005198
12	complexWordsRatio_LF	0.01153	42	internationalConceptsRatio_CSF	0.005146
13	wordLength3	0.01148	43	abbreviationsCount_1MF	0.005082
14	averageNounsSingularPerSentence_1MF	0.010484	44	averageVowelsPerWord_LF	0.00496
15	longWordsCount_LF	0.010003	45	numberPluralCount_1MF	0.004724
16	abstractConceptsHerfindahlIndex_CSF	0.009976	46	punctuationsEntropy_LF	0.004703
17	capitalizedWordsHerfindahlIndex_LF	0.009498	47	averageCaptalizedWordsPerSentence_SF	0.004669
18	mentionsPercentage_SNF	0.009399	48	exclamation	0.004623
19	averageWhiteSpacesPerSentence_SF	0.009167	49	sentencesBeginningWithLowercaseCount_SF	0.004594
20	person3rdPercentage_1MF	0.009079	50	pronounsPercentage_1MF	0.004536
21	averageInternationalConceptsPerSentence_CSF	0.008858	51	doubleQuote	0.004504
22	averageArticlesPOSPerSentence_1MF	0.008217	52	specialCharactersPercentage_LF	0.004493
23	averageEmojisPerSentence_SNF	0.008011	53	yulesKMeasure_LF	0.004453
24	lowercaseLettersPercentage_LF	0.007936	54	ADV	0.004423
25	averageConsonantsPerWord_LF	0.007794	55	entropyMeasure_LF	0.004231
26	averageParticlesPOSPerSentence_1MF	0.007442	56	longWordsPercentage_LF	0.004186
27	capitalizedWordsRatio_LF	0.007279	57	syllablesPercentage_LF	0.004159
28	properNounsPercentage_1MF	0.007161	58	VAUXP3	0.004092
29	punctuationsHerfindahlIndex_LF	0.006698	59	averageArticlesPerSentence_MF	0.003998
30	minLengthSentence_SF	0.006605	60	+ 259 other featrues	0.372242281

Table B.23: Feature importance scores for author-related social media posts in the chapter level.

Rank	Feature	Score	Rank	Feature	Score
1	wordLength39	0.089035	31	adverbsRatio_1MF	0.005526
2	averageVerbsPerSentence_1MF	0.028468	32	articlesCount_MF	0.0052
3	adjectivesSingularCount_1MF	0.02772	33	fleschKincaidGradeLevel_SF	0.005155
4	shortWordsCount_LF	0.020888	34	uppercaseLettersCount_LF	0.005122
5	tilde	0.020143	35	charactersEntropy_LF	0.005083
6	punctuationsPercentage_LF	0.019391	36	prepositionsCount_MF	0.00489
7	definiteWordsRatio_CSF	0.015748	37	particlesHerfindahlIndex_MF	0.004595
8	graveAccent	0.014045	38	stopwordsHerfindahlIndex_MF	0.004552
9	conjunctionsPOSCount_1MF	0.01294	39	person3rdCount_1MF	0.00447
10	wordLength16	0.012882	40	exclamation	0.004468
11	hashtagsRatio_SNF	0.012875	41	wordLength21	0.004461
12	nounsMasculineCount_1GF	0.012689	42	averageAllVerbsSingularPerSentence_1MF	0.004454
13	lowercaseDigraphsCount_LF	0.012437	43	averageGenderFemininePerSentence_1MF	0.004425
14	wordLength23	0.011535	44	complexWordsRatio_LF	0.004415
15	dialecticWordsCount_CSF	0.01119	45	nounsCount_1MF	0.004296
16	questionMark	0.011033	46	averageNonSpaceCharactersPerSentence_SF	0.004226
17	averageAdjectivesPerSentence_1MF	0.010821	47	adjectivesFeminineCount_1GF	0.004194
18	verticalBar	0.009902	48	shortWordsEntropy_LF	0.004186
19	socialProcessesCount_PLF	0.007663	49	PINT	0.004144
20	verbsEntropy_1MF	0.00759	50	PPOS	0.004133
21	epanalepsisCount_RF	0.007312	51	uppercaseDigraphsCount_LF	0.004086
22	shortWordsRatio_LF	0.007008	52	adjectivesPluralPercentage_1MF	0.004042
23	averageFunctionWordsPerSentence_1MF	0.006465	53	stopwordsRatio_MF	0.003966
24	stopwordsEntropy_1MF	0.006434	54	PRTC	0.003906
25	averageStopwordsPerSentence_MF	0.006165	55	averageLongWordsPerSentence_SF	0.003804
26	capitalizedWordsPercentage_LF	0.006041	56	definiteWordsHerfindahlIndex_CSF	0.003793
27	averageNounsPerSentence_1MF	0.005854	57	entropyMeasure_LF	0.003677
28	averagePunctuationsPerSentence_SF	0.00579	58	posTypeHerfindahlIndex_1MF	0.003663
29	wordLength17	0.005743	59	punctuationsEntropy_LF	0.003598
30	allVerbsSingularRatio_1MF	0.005561	60	+ 277 other features	0.323559263

Table B.24: Feature importance scores for author-related all writings in the chapter level.

Rank	Feature	Score	Rank	Feature	Score
1	wordLength39	0.089035	31	wordLength17	0.005288
2	averageVerbsPerSentence_1MF	0.028468	32	averageVerbsPerSentence_1MF	0.005232
3	adjectivesSingularCount_1MF	0.02772	33	hashSign	0.004595
4	shortWordsCount_LF	0.020888	34	articlesPOSHerfindahlIndex_1MF	0.004297
5	tilde	0.020143	35	hashtagsRatio_SNF	0.004119
6	punctuationsPercentage_LF	0.019391	36	punctuationsPercentage_LF	0.003868
7	definiteWordsRatio_CSF	0.015748	37	averageNounsSingularPerSentence_1MF	0.003849
8	graveAccent	0.014045	38	uppercaseLettersPercentage_LF	0.003711
9	conjunctionsPOSCount_1MF	0.01294	39	posTypesToUniqueWordsRatio_1MF	0.003687
10	wordLength16	0.012882	40	capitalizedWordsCount_LF	0.003661
11	hashtagsRatio_SNF	0.012875	41	conjunctionsPercentage_MF	0.003594
12	nounsMasculineCount_1GF	0.012689	42	functionWordsRatio_1MF	0.003573
13	lowercaseDigraphsCount_LF	0.012437	43	sentenceLength2	0.0035
14	wordLength23	0.011535	44	averageConsonantsPerUniqueWord_LF	0.003479
15	dialecticWordsCount_CSF	0.01119	45	shortWordsPercentage_LF	0.003266
16	questionMark	0.011033	46	averageElongatedCharactersPerWord_LF	0.003226
17	averageAdjectivesPerSentence_1MF	0.010821	47	internationalConceptsPercentage_CSF	0.003221
18	verticalBar	0.009902	48	verbsPercentage_1MF	0.003093
19	socialProcessesCount_PLF	0.007663	49	sentencesBeginningWithLowercaseCount_SF	0.003078
20	verbsEntropy_1MF	0.00759	50	fullStop	0.003072
21	epanalepsisCount_RF	0.007312	51	I-NPh	0.003069
22	shortWordsRatio_LF	0.007008	52	consonantsEntropy_LF	0.003058
23	averageFunctionWordsPerSentence_1MF	0.006465	53	sentencesBeginningWithUppercaseCount_SF	0.00301
24	stopwordsEntropy_1MF	0.006434	54	genderFeminineRatio_1MF	0.003004
25	averageStopwordsPerSentence_MF	0.006165	55	NFS ART NP	0.002974
26	capitalizedWordsPercentage_LF	0.006041	56	emojisPercentage_SNF	0.002951
27	averageNounsPerSentence_1MF	0.005854	57	adverbsCount_1MF	0.00294
28	averagePunctuationsPerSentence_SF	0.00579	58	VS3	0.002859
29	wordLength17	0.005743	59	PRTS VS3	0.002825
30	allVerbsSingularRatio_1MF	0.005561	60	+ 494 other features	0.302529782

Table B.25: Feature importance scores for genre-related literary works in the chapter level.

Rank	Feature	Score	Rank	Feature	Score
1	sentencesCount_SF	0.135703	31	adjectivesFemininеCount_1GF	0.003777
2	properNounsCount_1MF	0.065557	32	verbsSingularRatio_1MF	0.003692
3	dialecticWordsCount_CSF	0.061706	33	verbsSingularHerfindahlIndex_1MF	0.003658
4	capitalizedWordsCount_LF	0.055506	34	asterisk	0.003639
5	lowercaseLettersPercentage_LF	0.028965	35	vowelsPercentage_LF	0.003542
6	aposiopesisCount_RF	0.02877	36	PCL1 VS3 NFS	0.003501
7	uppercaseWordsRatio_LF	0.018591	37	slash	0.003423
8	capitalizedWordsRatio_LF	0.014639	38	diacriticsCount_LF	0.003404
9	averageVowelsPerUniqueWord_LF	0.011197	39	wordLength2	0.003382
10	sentencesBeginningWithLowercaseCount_SF	0.007731	40	particlesPOSEntropy_1MF	0.003367
11	vowelsEntropy_LF	0.007517	41	averageElongatedCharactersPerUniqueWord_LF	0.003317
12	discourseMarkersCount_CSF	0.007445	42	rightParenthesis	0.003259
13	syntacticPhrasesCount_SynF	0.007278	43	punctuationsCount_LF	0.003206
14	PRP NFS	0.007092	44	auxiliaryVerbsSingularRatio_1MF	0.003166
15	auxiliaryVerbsSingularCount_1MF	0.00667	45	averagePastParticipleVerbsPerSentence_1MF	0.003165
16	complexWordsHerfindahlIndex_LF	0.006636	46	VS3 CNJS	0.003132
17	uniqueWordsCount_LF	0.006424	47	averageCaptalizedWordsPerSentence_SF	0.003106
18	hapaxLegomenaPercentage_LF	0.006261	48	averageProperNounsPerSentence_1MF	0.00309
19	singleQuote	0.005948	49	capitalizedWordsPercentage_LF	0.003073
20	averageLongWordsPerSentence_SF	0.005769	50	graveAccent	0.00303
21	person3rdEntropy_1MF	0.005651	51	genderMasculineCount_1MF	0.002998
22	repeatedConsonantsCount_RF	0.00541	52	consonantsEntropy_LF	0.002992
23	PRTN VAUXS3 VPCP	0.005332	53	syllablesCount_LF	0.002924
24	PRP NFS PRP	0.00532	54	NFS ADJFS	0.002888
25	doubleQuote	0.005044	55	charactersEntropy_LF	0.002799
26	PRTC ADV	0.004903	56	complexWordsEntropy_LF	0.002798
27	uppercaseLettersPercentage_LF	0.004544	57	abstractConceptsHerfindahlIndex_CSF	0.002746
28	imperativeVerbsPercentage_1MF	0.004359	58	exclamation	0.002724
29	nounsPluralPercentage_1MF	0.004152	59	stopwordsCount_MF	0.002724
30	uppercaseWordsPercentage_LF	0.004017	60	+ 26 other features	0.05814571

Table B.26: Feature importance scores for genre-related all writings in the chapter level.

Rank	Feature	Score	Rank	Feature	Score
1	sentencesBeginningWithUppercaseCount_SF	0.311804	4	averageHashtagsPerSentence_SNF	0.036279
2	person3rdCount_1MF	0.246745	5	person3rdPercentage_1MF	0.024564
3	hashtagsPercentage_SNF	0.235696			

Table B.27: Feature importance scores for gender-related literary works in the chapter level.

Rank	Feature	Score	Rank	Feature	Score
1	leftParenthesis	0.184222	25	intensiveAdverbsPercentage_GF	0.007616
2	rightParenthesis	0.038852	26	averageFemininePronounsPerSentence_GF	0.007534
3	numberSingularRatio_1MF	0.025002	27	abstractConceptsPercentage_CSF	0.007375
4	averageConsonantsPerUniqueWord_LF	0.018378	28	auxiliaryVerbsHerfindahlIndex_MF	0.00735
5	auxiliaryVerbsSingularHerfindahlIndex_1MF	0.017897	29	digitsPercentage_LF	0.007218
6	nounsFeminineHerfindahlIndex_1GF	0.017462	30	averageHapaxLegomenaPerSentence_SF	0.006446
7	digitsCount_LF	0.017306	31	VS2 VPCP	0.006132
8	averageElongatedCharactersPerSentence_SF	0.015951	32	averageSpecialCharactersPerSentence_SF	0.006128
9	consonantsHerfindahlIndex_LF	0.014483	33	internationalConceptsPercentage_CSF	0.006119
10	interjectionsHerfindahlIndex_MF	0.014169	34	digitsEntropy_LF	0.006096
11	socialMediaLexiconsPercentage_SNF	0.011948	35	punctuationsPercentage_LF	0.006084
12	averageElongatedCharactersPerWord_LF	0.011946	36	ADV NFS	0.006079
13	averageConsonantsPerWord_LF	0.01138	37	doubleQuote	0.006063
14	sentencesCount_SF	0.010524	38	particlesHerfindahlIndex_MF	0.005914
15	definiteWordsEntropy_CSF	0.010112	39	stopwordsPercentage_MF	0.005832
16	PCL1 VAUXS3 VPCP	0.009875	40	internationalConceptsCount_CSF	0.00561
17	lowercaseLettersPercentage_LF	0.009453	41	femininePronounsPercentage_GF	0.005513
18	averagePunctuationsPerSentence_SF	0.009163	42	complexWordsRatio_LF	0.005461
19	exclamation	0.009005	43	positiveEmotionsCount_PLF	0.00534
20	slangWordsHerfindahlIndex_IF	0.008887	44	averageAuxiliaryVerbsPOSPerSentence_1MF	0.005151
21	charactersHerfindahlIndex_LF	0.008535	45	discourseMarkersPercentage_CSF	0.00515
22	elongatedWordsHerfindahlIndex_LF	0.008296	46	discourseMarkersRatio_CSF	0.004893
23	sentencesBeginningWithUppercaseCount_SF	0.008119	47	averageAuxiliaryVerbsSingularPerSentence_1MF	0.004845
24	sentenceLength27	0.007751			

Table B.28: Feature importance scores for gender-related newsroom columns in the chapter level.

Rank	Feature	Score	Rank	Feature	Score
1	capitalizedWordsRatio_LF	0.102061	31	averageSpecialCharactersPerSentence_SF	0.005003
2	person1stRatio_1MF	0.041366	32	averageConsonantsPerUniqueWord_LF	0.004894
3	person1stCount_1MF	0.032513	33	socialMediaLexiconsCount_SNF	0.004875
4	averagePerson1stPerSentence_1MF	0.021309	34	doubleQuote	0.004852
5	singleQuote	0.019695	35	articlesRatio_MF	0.004822
6	atSign	0.014871	36	comma	0.004805
7	properNounsPercentage_1MF	0.012462	37	exclamation	0.00479
8	mentionsPercentage_SNF	0.012168	38	auxiliaryVerbsPOSHerfindahlIndex_1MF	0.00476
9	averageAllVerbsSingularPerSentence_1MF	0.010965	39	auxiliaryVerbsEntropy_MF	0.004672
10	ART NP ART	0.01036	40	cardinalNumbersRatio_1MF	0.004577
11	averageNumberSingularPerSentence_1MF	0.01021	41	articlesPOSCount_1MF	0.00434
12	graveAccent	0.00919	42	NFS PIND	0.004329
13	PRTV PRTS VMOD	0.008738	43	posTypeEntropy_1MF	0.004295
14	elongatedWordsPercentage_LF	0.008462	44	averageDialecticWordsPerSentence_CSF	0.004289
15	punctuationsPercentage_LF	0.008096	45	asterisk	0.00427
16	entropyMeasure_LF	0.00792	46	masculinePronounsPercentage_GF	0.004228
17	specialCharactersCount_LF	0.007534	47	PPER	0.0042
18	person1stPercentage_1MF	0.007405	48	PREL PCL1	0.004188
19	B-VPh	0.007352	49	ADJMS	0.00415
20	articlesPOSEntropy_1MF	0.007198	50	averagePunctuationsPerSentence_SF	0.004083
21	auxiliaryVerbsPluralCount_1MF	0.007078	51	verbsEntropy_1MF	0.004024
22	properNounsRatio_1MF	0.006923	52	NMS ART NFS	0.003913
23	syntacticPhrasesHerfindahlIndex_SynF	0.006868	53	NP NP	0.003884
24	functionWordsCount_1MF	0.006275	54	shortWordsHerfindahlIndex_LF	0.003786
25	lowercaseDigraphsCount_LF	0.006058	55	I-VPh	0.003743
26	functionWordsPercentage_1MF	0.005882	56	genderFemininеCount_1MF	0.003642
27	person1stHerfindahlIndex_1MF	0.005304	57	interjectionsPOSRatio_1MF	0.003592
28	particlesEntropy_MF	0.005189	58	articlesPOSHerfindahlIndex_1MF	0.003489
29	nounsSingularCount_1MF	0.005098	59	functionWordsHerfindahlIndex_1MF	0.003426
30	NMP ART ADJMP	0.00507	60	+ 43 other features	0.123215741

Table B.29: Feature importance scores for gender-related social media posts in the chapter level.

Rank	Feature	Score	Rank	Feature	Score
1	singleQuote	0.066766	31	stopwordsEntropy_1MF	0.004386
2	definiteWordsCount_CSF	0.034267	32	allVerbsSingularPercentage_1MF	0.004266
3	CNJC	0.020873	33	averagePronounsPerSentence_1MF	0.004259
4	conjunctionsPOSCount_1MF	0.011326	34	syntacticPhrasesCount_SynF	0.00399
5	masculinePronounsCount_GF	0.010623	35	comma	0.003967
6	femininePronounsCount_GF	0.00942	36	NMP ART NMS	0.003911
7	conjunctionsCount_MF	0.008989	37	particlesPOSCount_1MF	0.003858
8	mentionsPercentage_SNF	0.008571	38	averageConsonantsPerSentence_SF	0.00377
9	mentionsRatio_SNF	0.006765	39	adjectivesFemininePercentage_1GF	0.003728
10	adjectivesCount_1MF	0.00676	40	PRT	0.003718
11	properNounsCount_1MF	0.006578	41	averageGenderMasculinePerSentence_1MF	0.003665
12	sentenceLength3	0.006134	42	PDEM NMS	0.003664
13	shortWordsEntropy_LF	0.006128	43	pronounsCount_1MF	0.003645
14	assonanceCount_RF	0.005976	44	nounsPluralPercentage_1MF	0.003621
15	hashtagsRatio_SNF	0.005567	45	punctuationsCount_LF	0.003611
16	nounsPluralCount_1MF	0.005453	46	uniqueWordsPercentage_LF	0.003602
17	timeConceptsEntropy_PLF	0.005449	47	cognitiveProcessesHerfindahlIndex_PLF	0.003601
18	epizeuxisCount_RF	0.005443	48	negativeEmotionsPercentage_PLF	0.003583
19	ADJMS	0.005201	49	diacriticsPercentage_LF	0.003449
20	B-NPh	0.005181	50	PRP NP NP	0.003432
21	averageNonAlphabeticCharactersPerSentence_SF	0.00515	51	stopwordsRatio_MF	0.003275
22	VPCP PRP	0.004932	52	nounsFemininePercentage_1GF	0.00323
23	doubleQuote	0.004915	53	wordLength39	0.003222
24	specialCharactersPercentage_LF	0.004796	54	genderFeminineRatio_1MF	0.003159
25	genderMasculineCount_1MF	0.004775	55	underscore	0.00309
26	adjectivesSingularCount_1MF	0.004722	56	longWordsCount_LF	0.003053
27	wordLength1	0.004647	57	averageHapaxDislegomenaPerSentence_SF	0.003035
28	PCL1 VAUXS3	0.004513	58	NFP NFS	0.003019
29	averageFunctionWordsPerSentence_1MF	0.004467	59	wordLength3	0.003003
30	hashtagsPercentage_SNF	0.004391	60	+ 39 other features	0.104951247

Table B.30: Feature importance scores for gender-related all writings in the chapter level.

Rank	Feature	Score	Rank	Feature	Score
1	capitalizedWordsPercentage_LF	0.048862	31	asterisk	0.004443
2	sentencesBeginningWithUppercaseCount_SF	0.037454	32	punctuationsEntropy_LF	0.00432
3	stopwordsRatio_MF	0.019972	33	ART ADV	0.004161
4	hashtagsPercentage_SNF	0.012024	34	punctuationsPercentage_LF	0.004094
5	hashSign	0.011683	35	charactersEntropy_LF	0.003986
6	doubleQuote	0.008983	36	ART ADJFP CNJC	0.003864
7	discourseMarkersRatio_CSF	0.008892	37	particlesPOSCount_1MF	0.003815
8	I-VPh	0.008694	38	lowercaseLettersPercentage_LF	0.003801
9	person3rdRatio_1MF	0.008515	39	underscore	0.003792
10	stopwordsPercentage_MF	0.008264	40	averageCaptalizedWordsPerSentence_SF	0.00372
11	averageElongatedCharactersPerWord_LF	0.008249	41	prepositionsPOSPercentage_1MF	0.003703
12	B-VPh	0.007868	42	NFS ART NP	0.003623
13	averageHashtagsPerSentence_SNF	0.007709	43	specialCharactersHerfindahlIndex_LF	0.003601
14	hashtagsRatio_SNF	0.007082	44	nounsEntropy_1MF	0.003567
15	functionWordsHerfindahlIndex_1MF	0.006923	45	averageElongatedCharactersPerSentence_SF	0.003529
16	mentionsPercentage_SNF	0.006901	46	properNounsPercentage_1MF	0.003433
17	leftParenthesis	0.006735	47	PPOS	0.003359
18	B-AdjPh	0.006119	48	uncertaintyVerbsCount_GF	0.003351
19	singleQuote	0.006054	49	longSentencesCount_SF	0.00335
20	rightParenthesis	0.005835	50	averageDiscourseMarkersPerSentence_CSF	0.003265
21	averageMentionsPerSentence_SNF	0.005767	51	NP ART	0.00325
22	elongationCharactersCount_LF	0.005709	52	punctuationsCount_LF	0.003233
23	exclamation	0.005605	53	indefiniteWordsRatio_CSF	0.003095
24	atSign	0.005549	54	sentenceLength23	0.003027
25	stopwordsCount_MF	0.005408	55	averageAuxiliaryVerbsPerSentence_MF	0.00298
26	elongationCharactersPercentage_LF	0.005171	56	capitalizedWordsRatio_LF	0.002969
27	person3rdPercentage_1MF	0.005122	57	adverbsRatio_1MF	0.00295
28	tilde	0.004994	58	averageConsonantsPerSentence_SF	0.00292
29	particlesCount_MF	0.004646	59	uncertaintyVerbsRatio_GF	0.002902
30	averageElongatedCharactersPerUniqueWord_LF	0.004482	60	+ 118 other features	0.225559992

Appendix C

Feature analysis and model performance visualizations

C.1 Author identification task

C.1.1 Dimensionality reduction using PCA

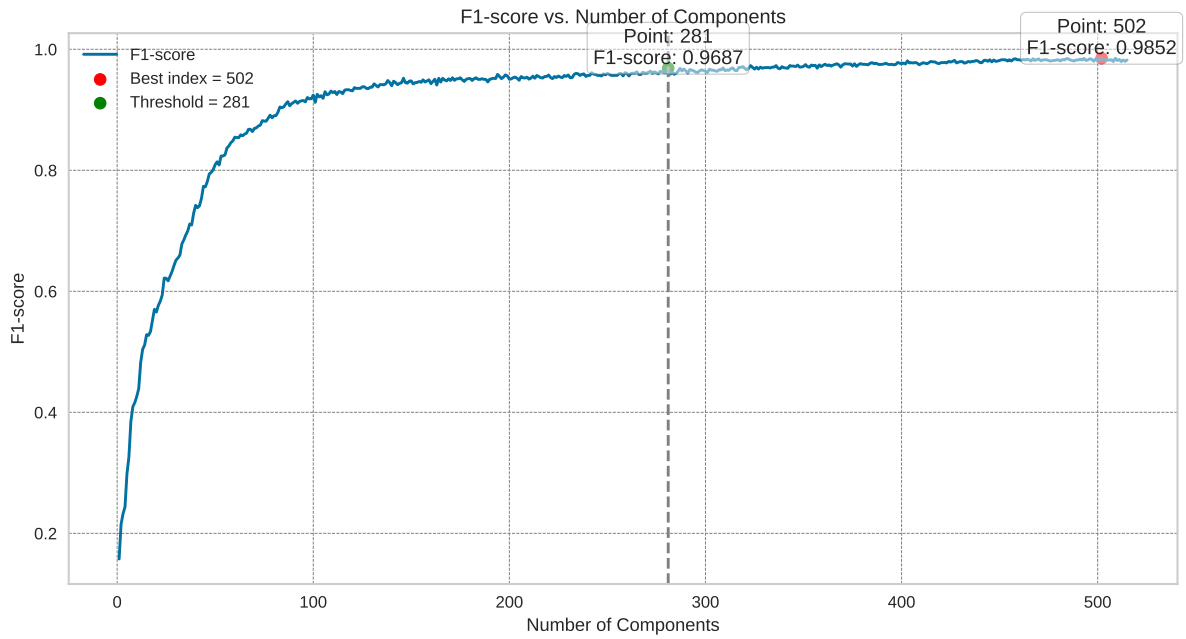


Figure C.1: PCA for author-related literary chapter-level classification using Ridge classifier and feature engineering.

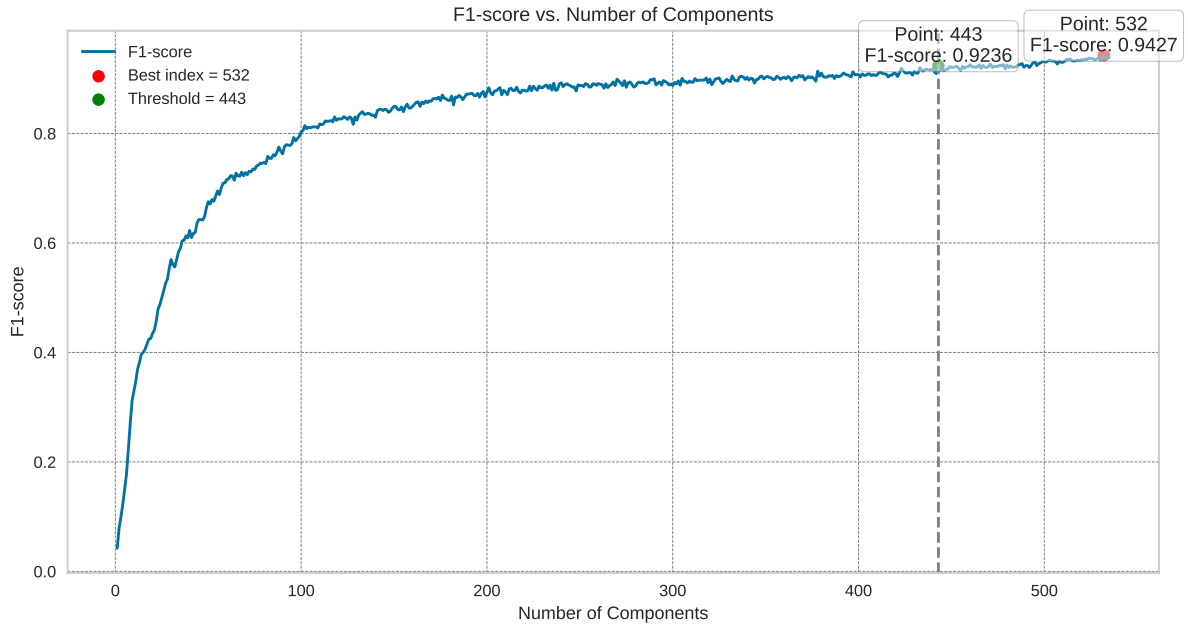


Figure C.2: PCA for author-related columns chapter-level classification using Ridge classifier and feature engineering.

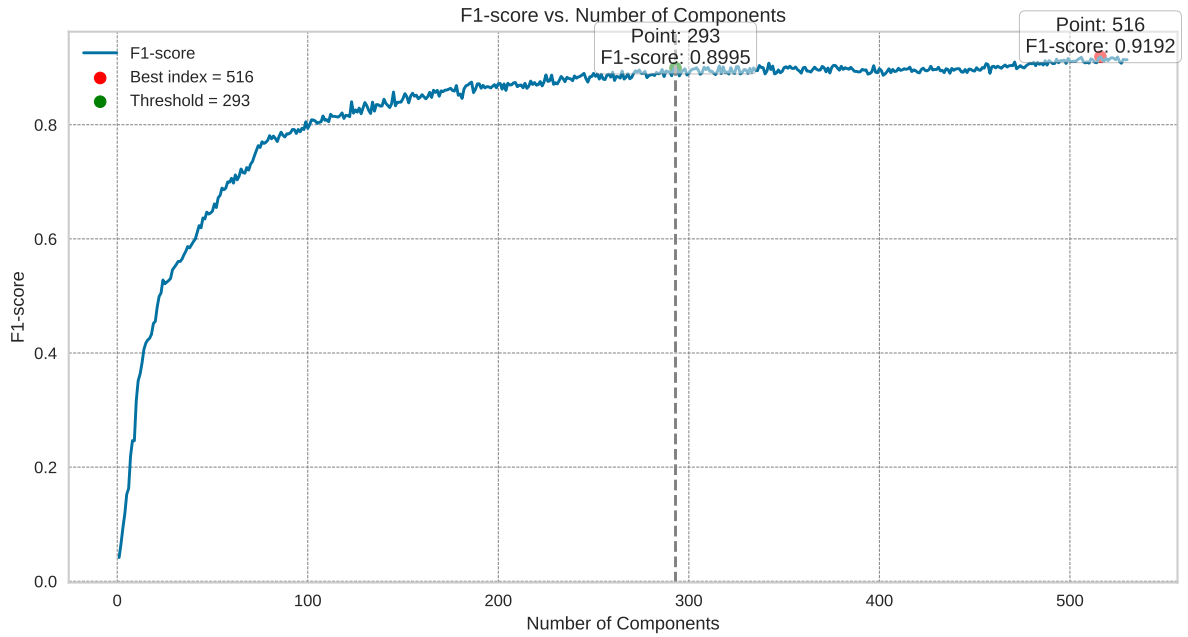


Figure C.3: PCA for author-related posts chapter-level classification using Ridge classifier and feature engineering.

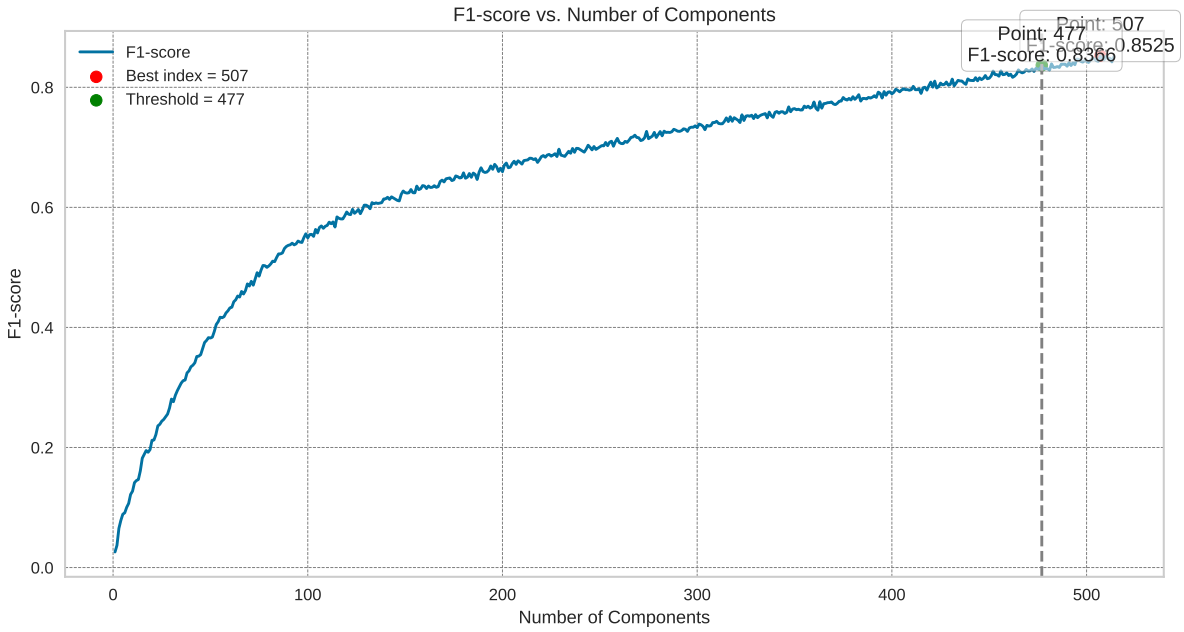


Figure C.4: PCA for author-related all writings chapter-level classification using Ridge classifier and feature engineering.

C.1.2 Dimensionality reduction using autoencoder-based model

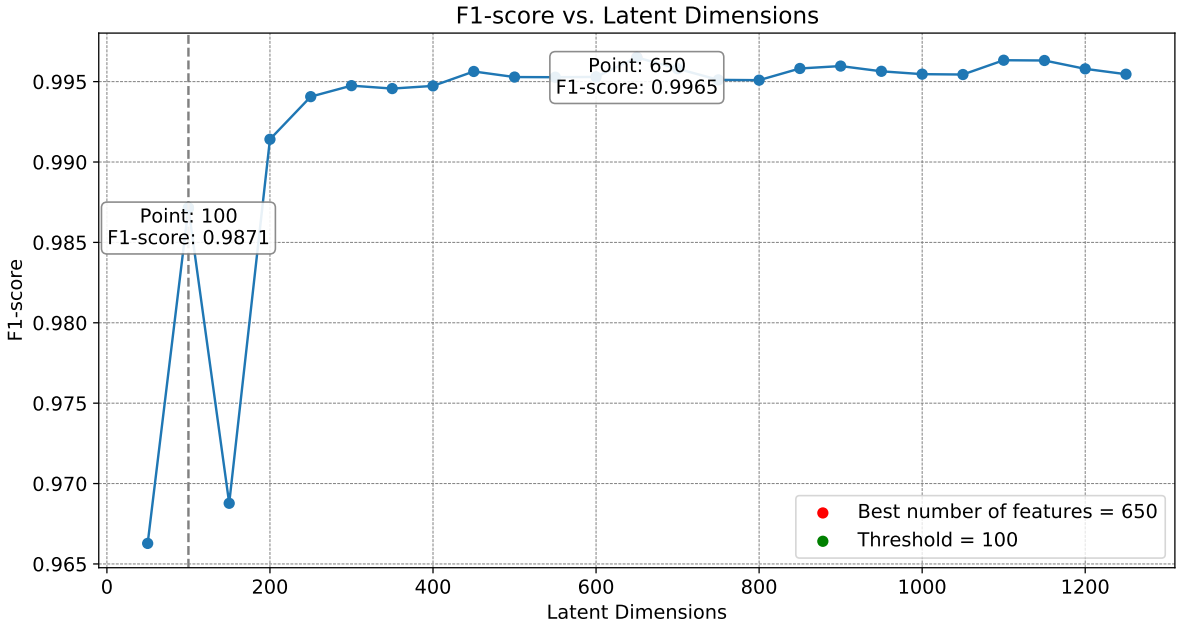


Figure C.5: Autoencoder-based model for author-related literary chapter-level classification using Ridge classifier and feature engineering.

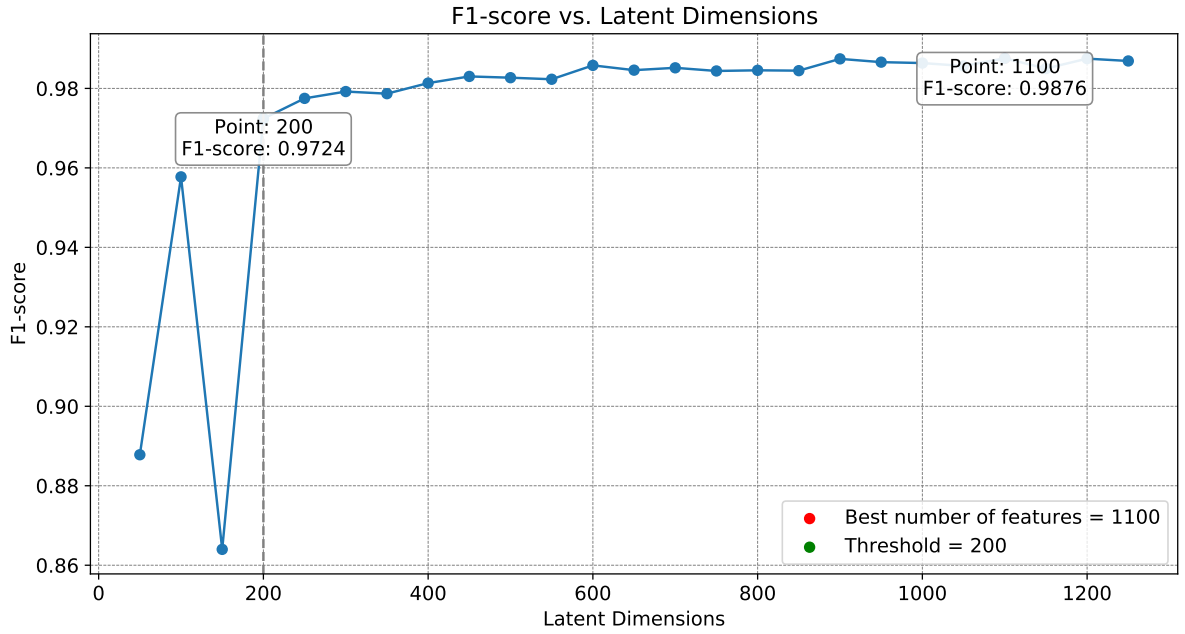


Figure C.6: Autoencoder-based model for author-related columns chapter-level classification using Ridge classifier and feature engineering.

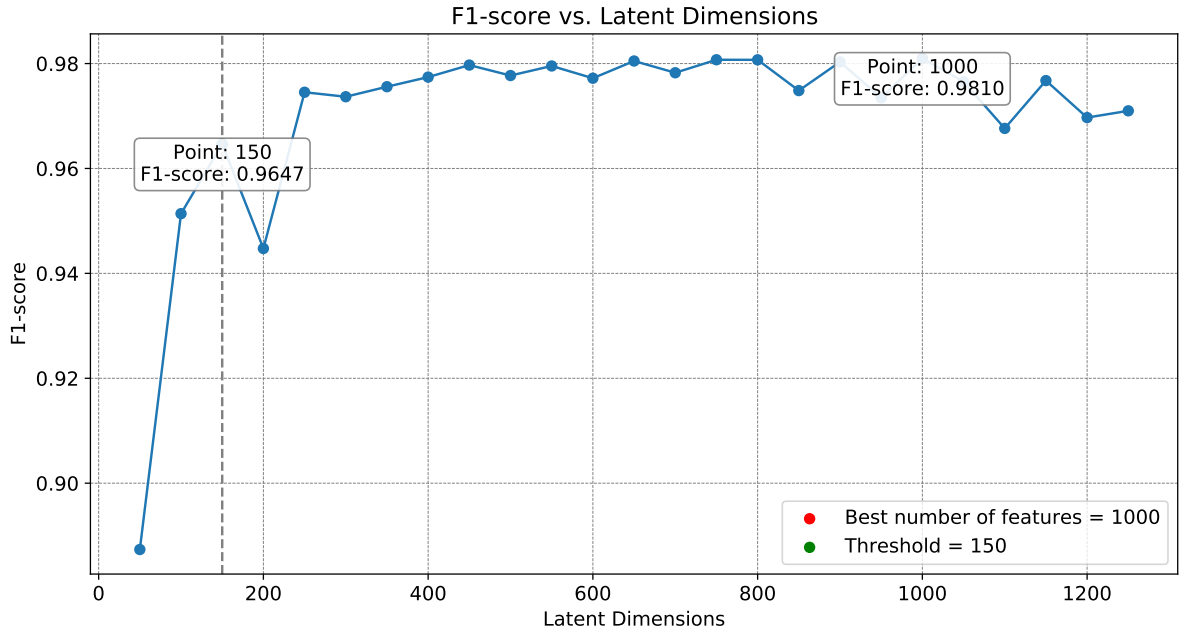


Figure C.7: Autoencoder-based model for author-related posts chapter-level classification using Ridge classifier and feature engineering.

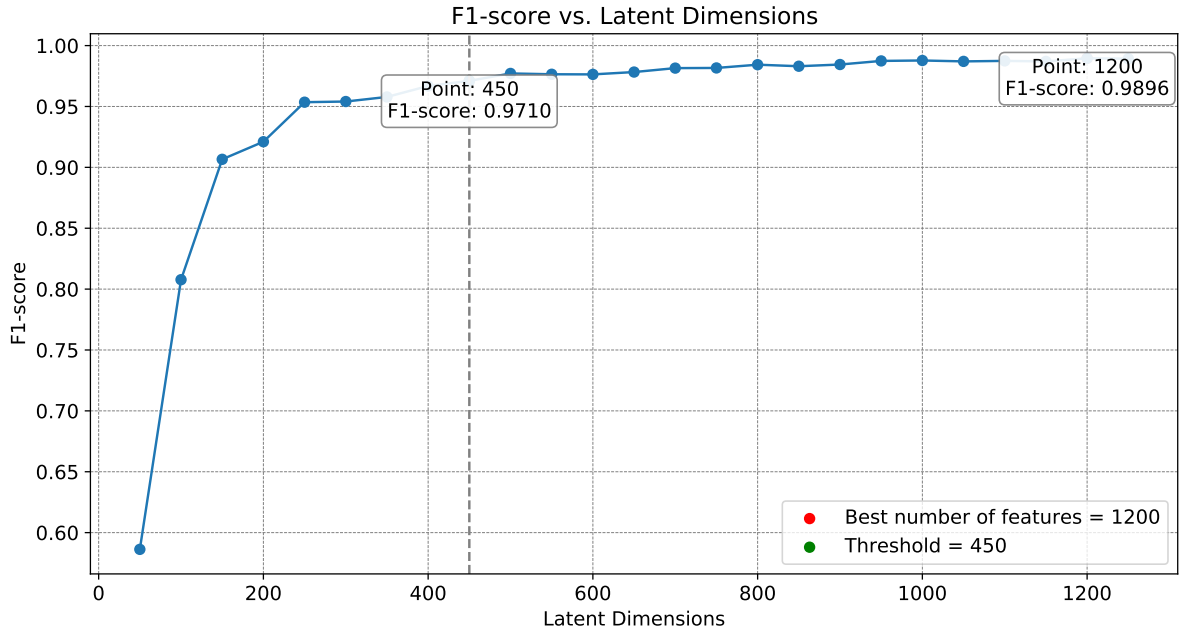


Figure C.8: Autoencoder-based model for author-related all writings chapter-level classification using Ridge classifier and feature engineering.

C.1.3 Feature analysis using SHAP

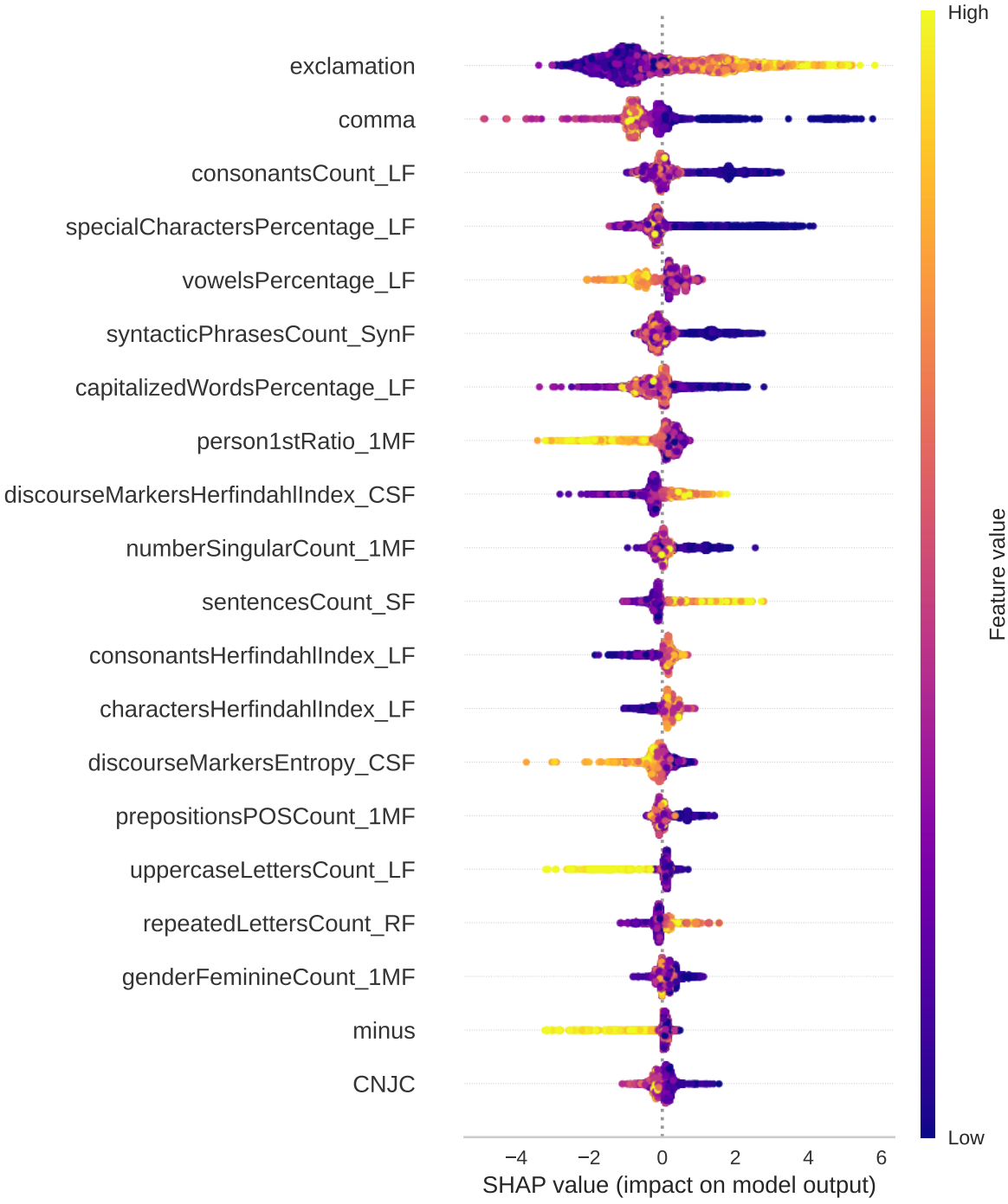


Figure C.9: SHAP summary plot for author-based literary classification using XGBoost classifier and feature engineering.

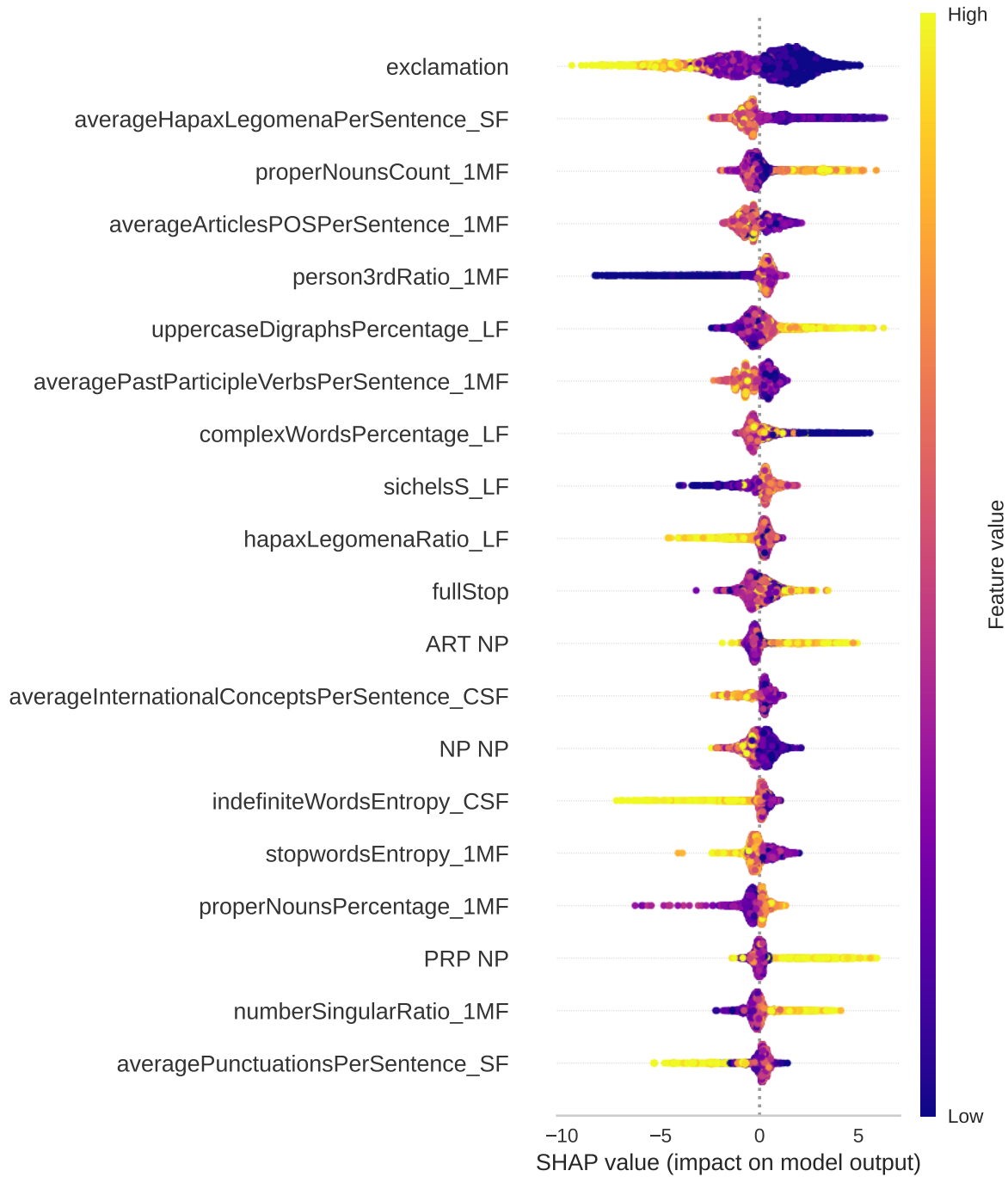


Figure C.10: SHAP summary plot for author-based columns classification using XGBoost classifier and feature engineering.

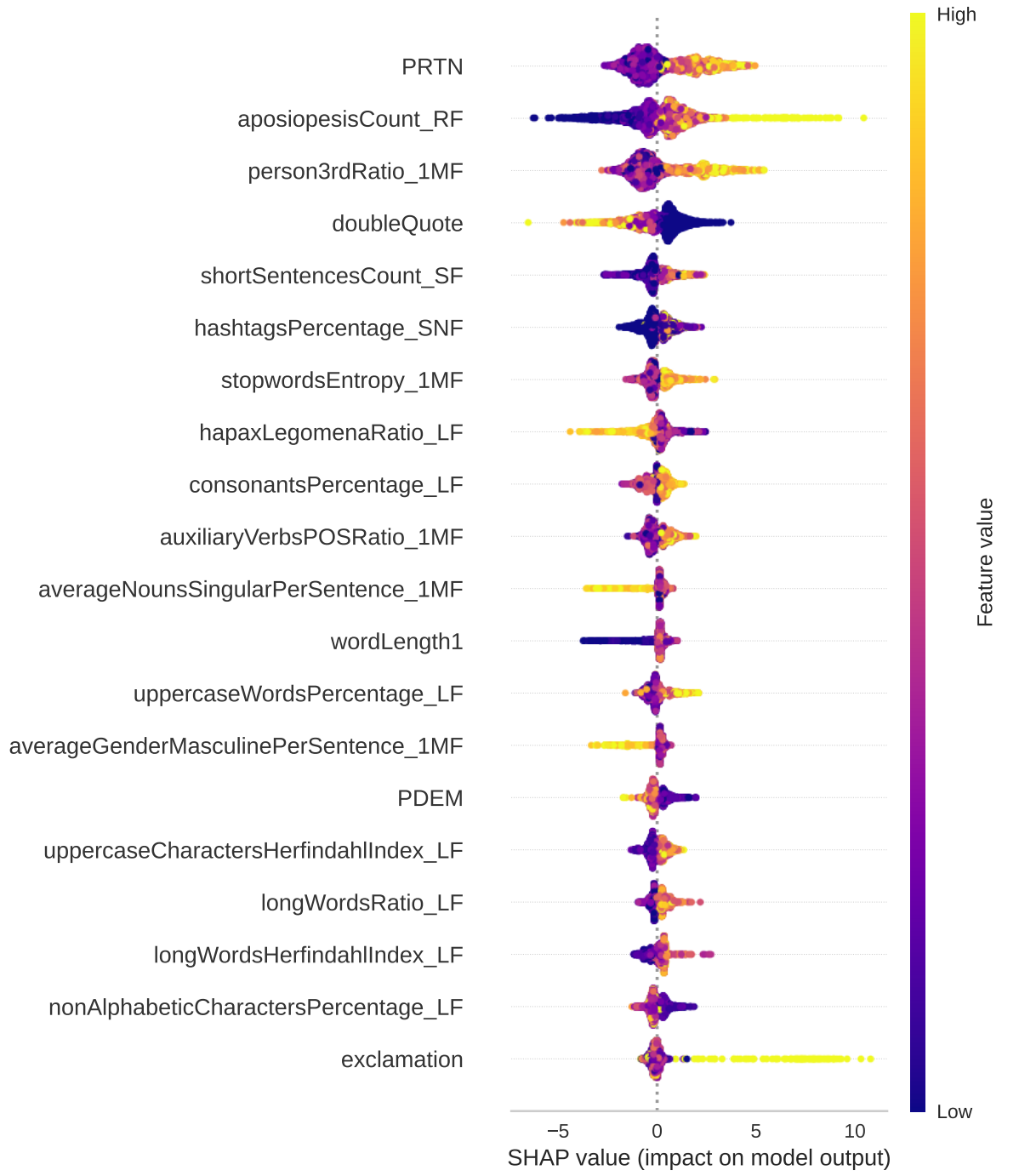


Figure C.11: SHAP summary plot for author-based posts classification using XGBoost classifier and feature engineering.

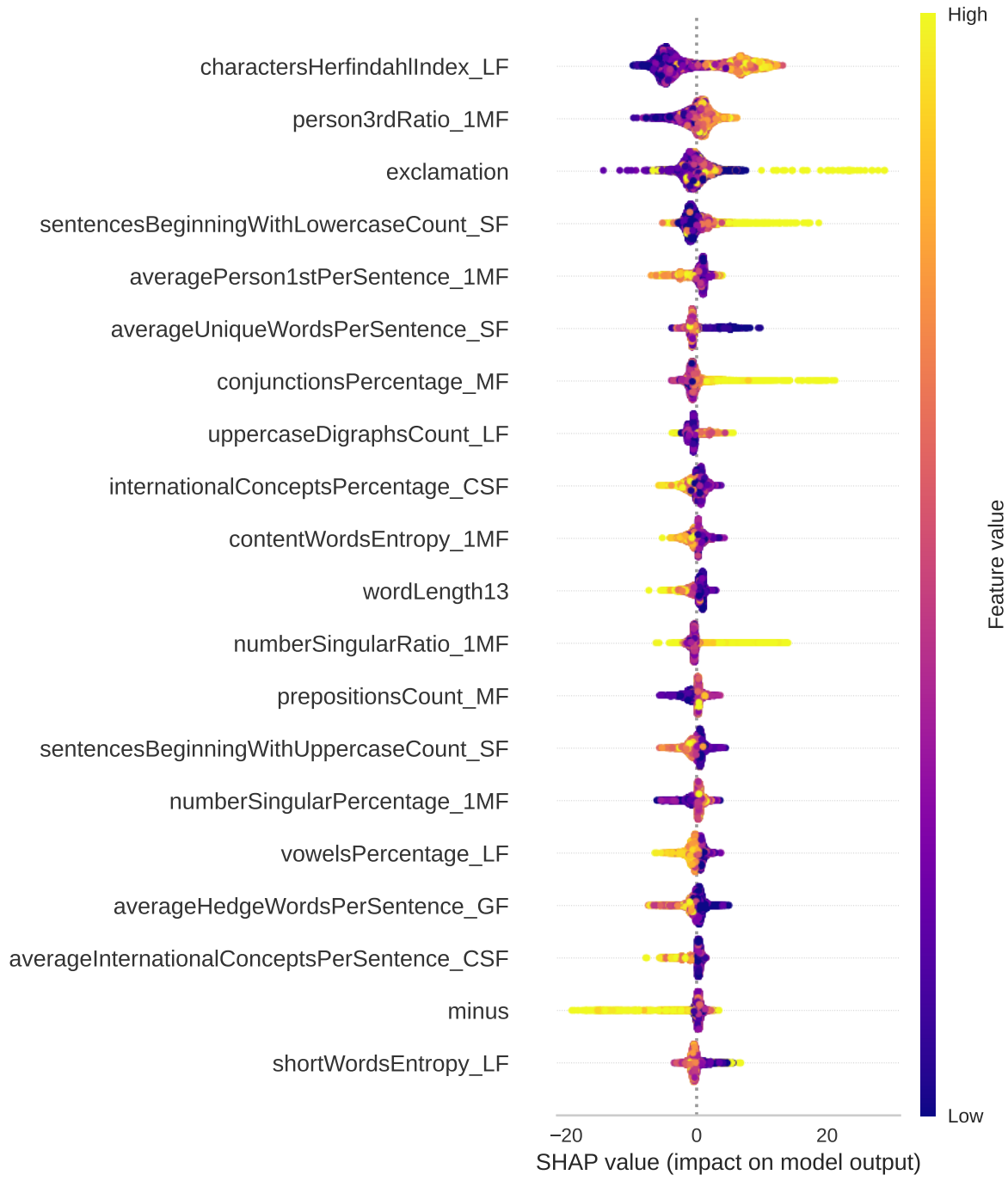


Figure C.12: SHAP summary plot for author-based all writings classification using XGBoost classifier and feature engineering.

C.2 Genre identification task

C.2.1 Dimensionality reduction using PCA

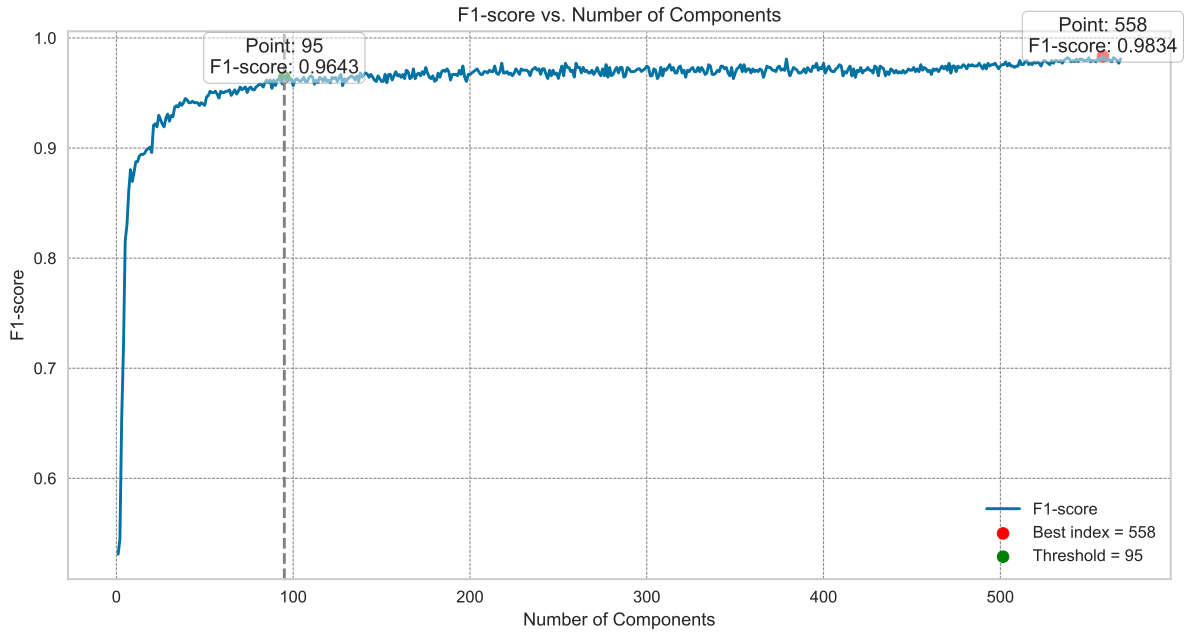


Figure C.13: PCA for genre-related literary chapter-level classification using Ridge classifier and feature engineering.

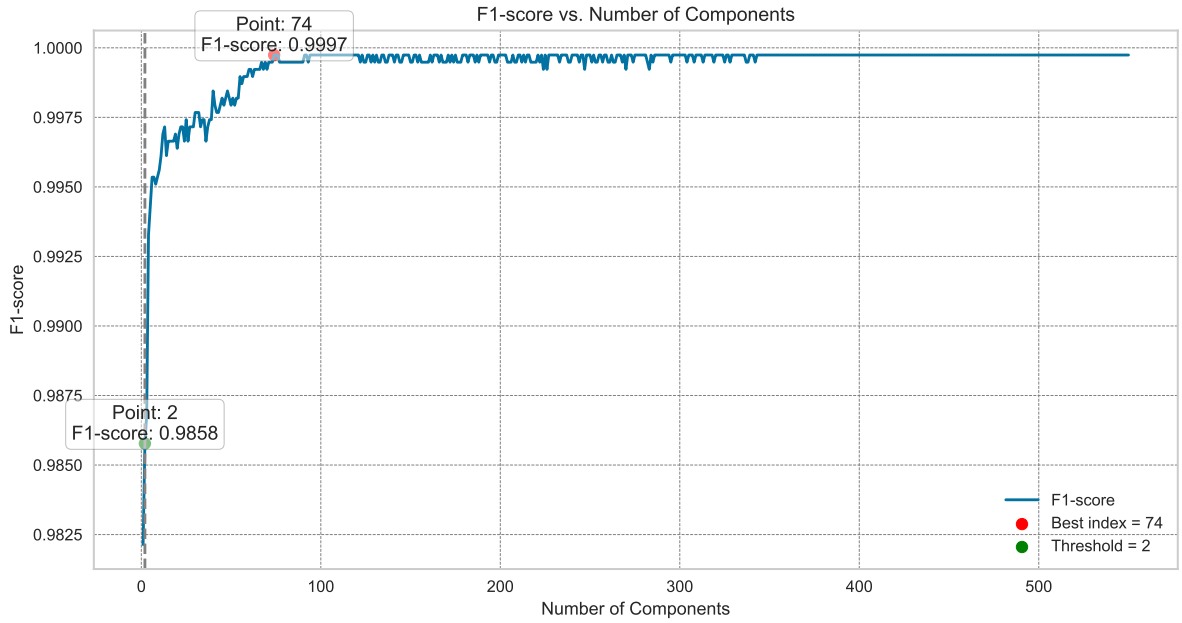


Figure C.14: PCA for genre-related all writings chapter-level classification using Ridge classifier and feature engineering.

C.2.2 Dimensionality reduction using autoencoder-based model

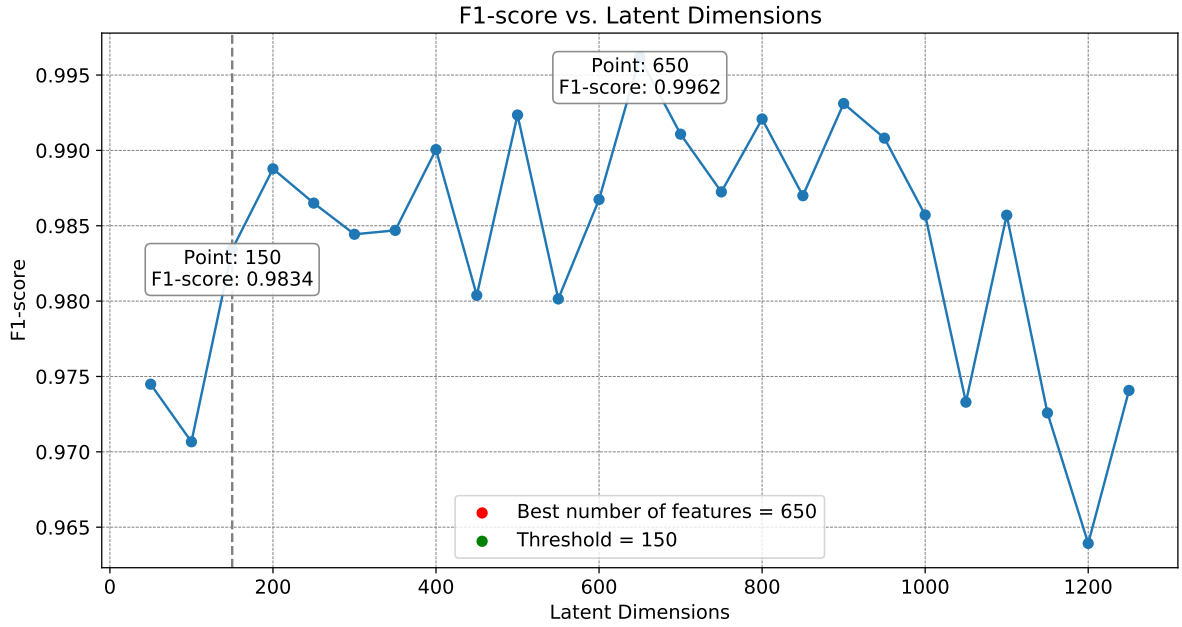


Figure C.15: Autoencoder-based model for genre-related literary chapter-level classification using Ridge classifier and feature engineering.

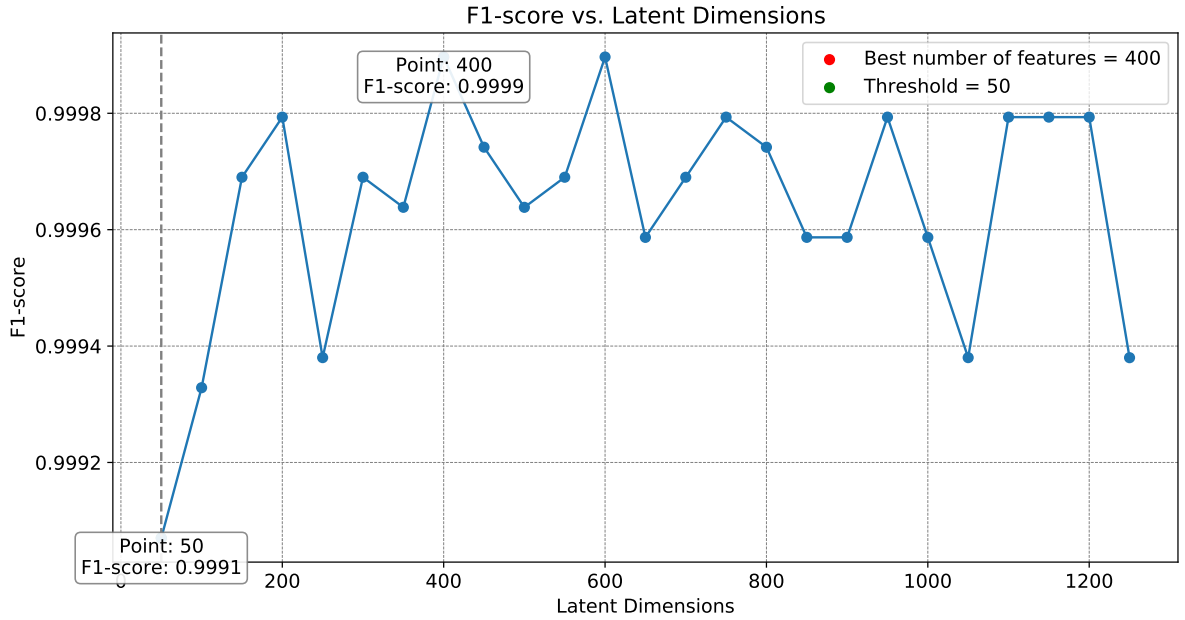


Figure C.16: Autoencoder-based model for genre-related all writings chapter-level classification using Ridge classifier and feature engineering.

C.2.3 Feature analysis using SHAP

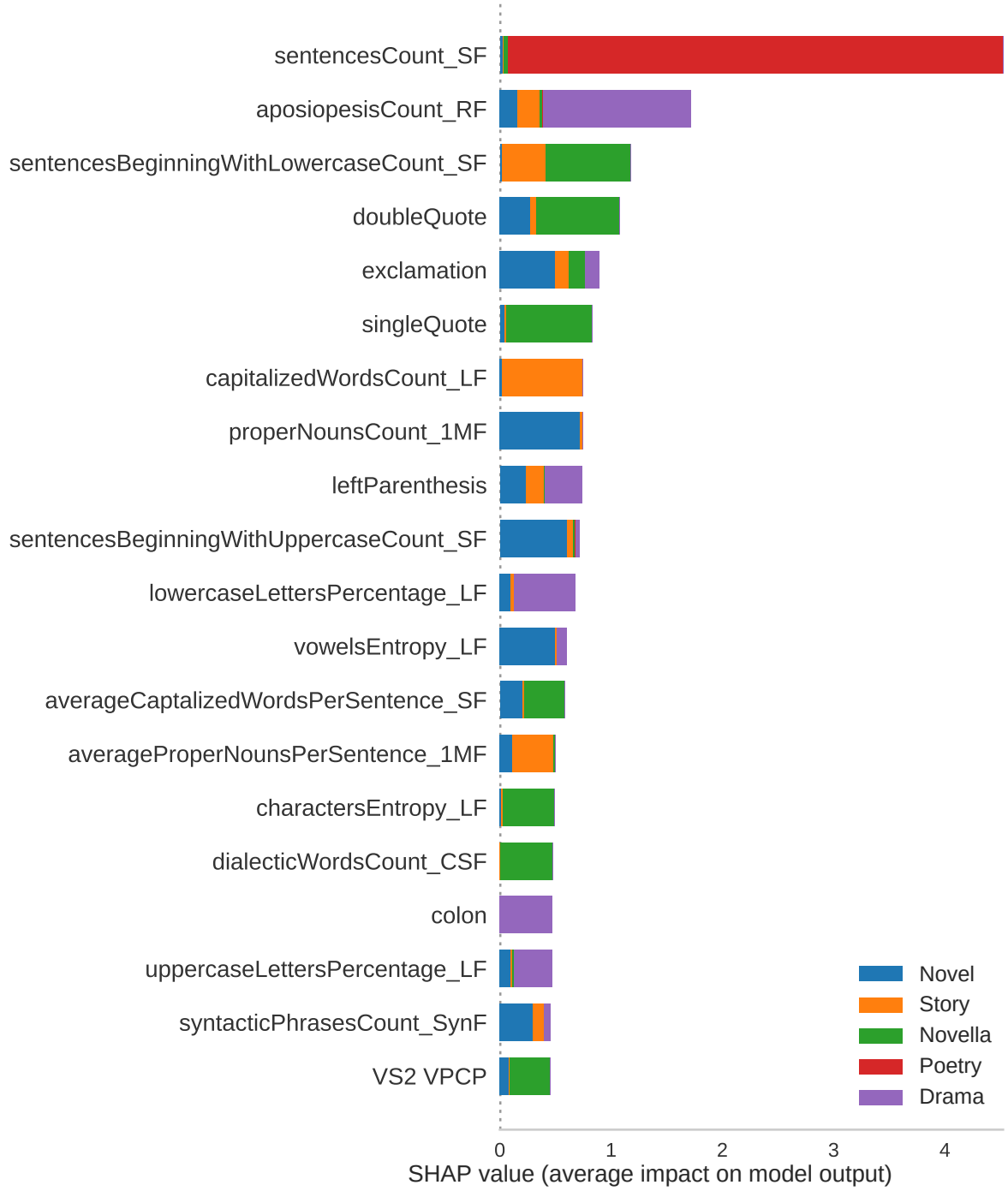


Figure C.17: SHAP summary plot for genre-related literary classification using XGBoost classifier and feature engineering.

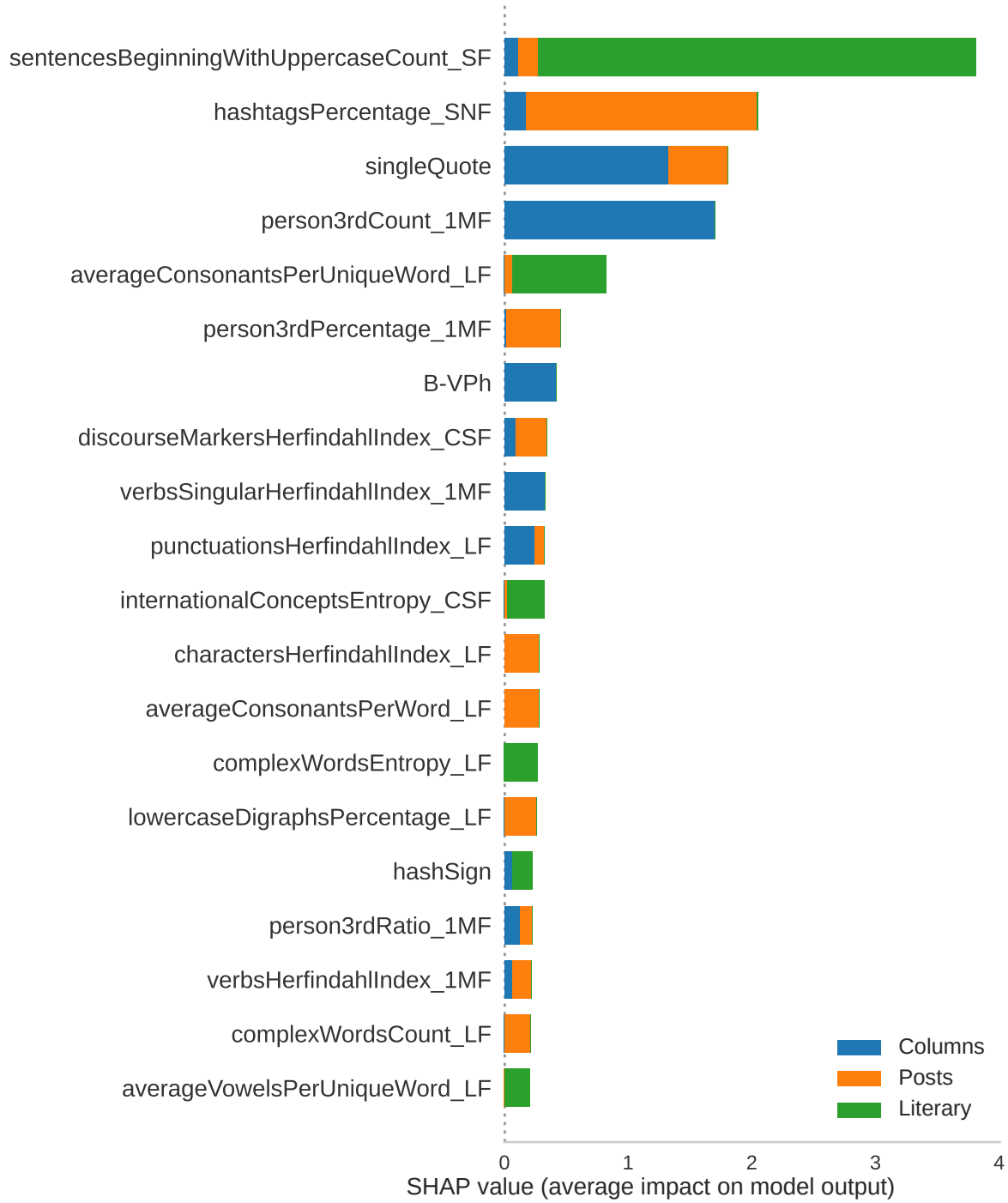


Figure C.18: SHAP summary plot for genre-related all writings classification using XGBoost classifier and feature engineering.

C.3 Gender identification task

C.3.1 Dimensionality reduction using PCA

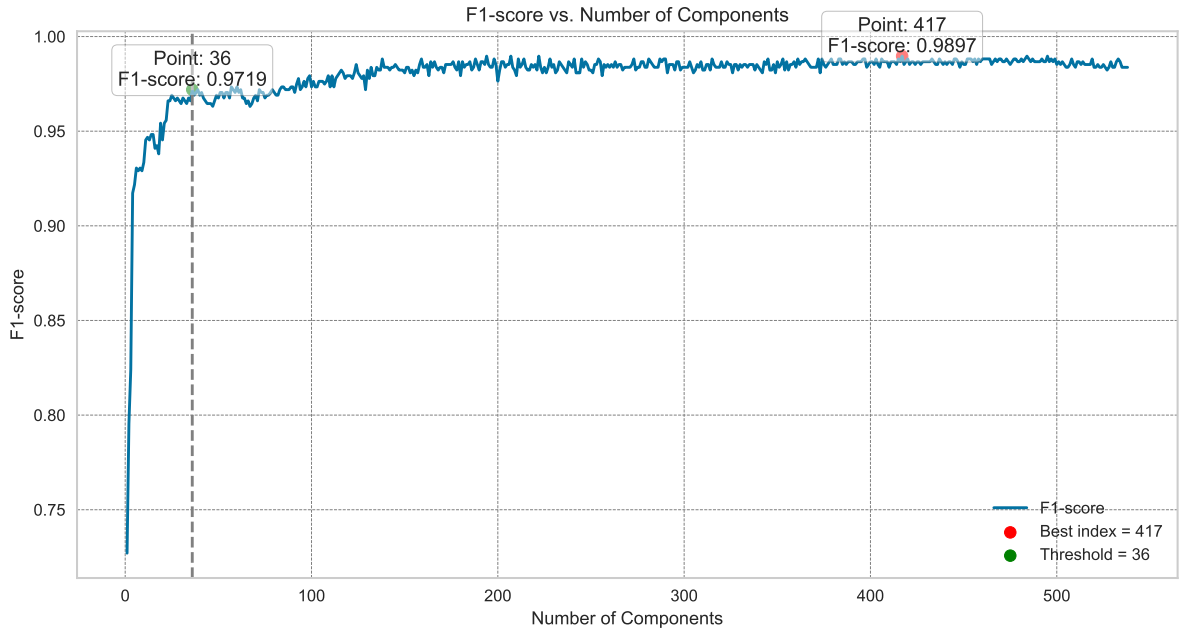


Figure C.19: PCA for gender-related literary chapter-level classification using Ridge classifier and feature engineering.

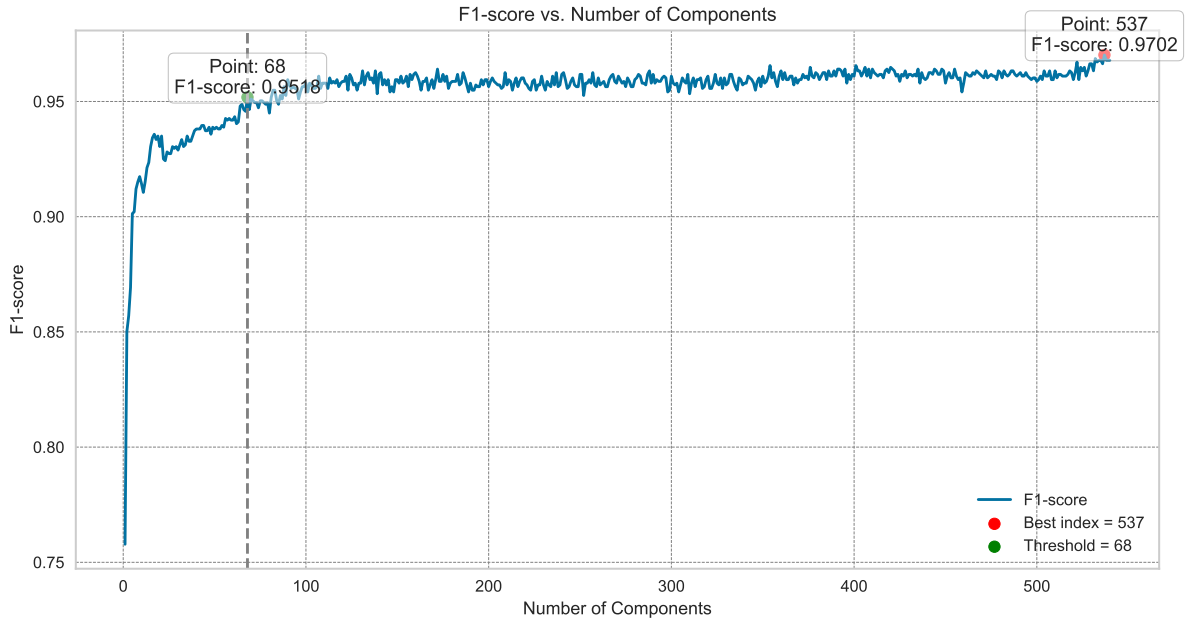


Figure C.20: PCA for gender-related columns chapter-level classification using Ridge classifier and feature engineering.

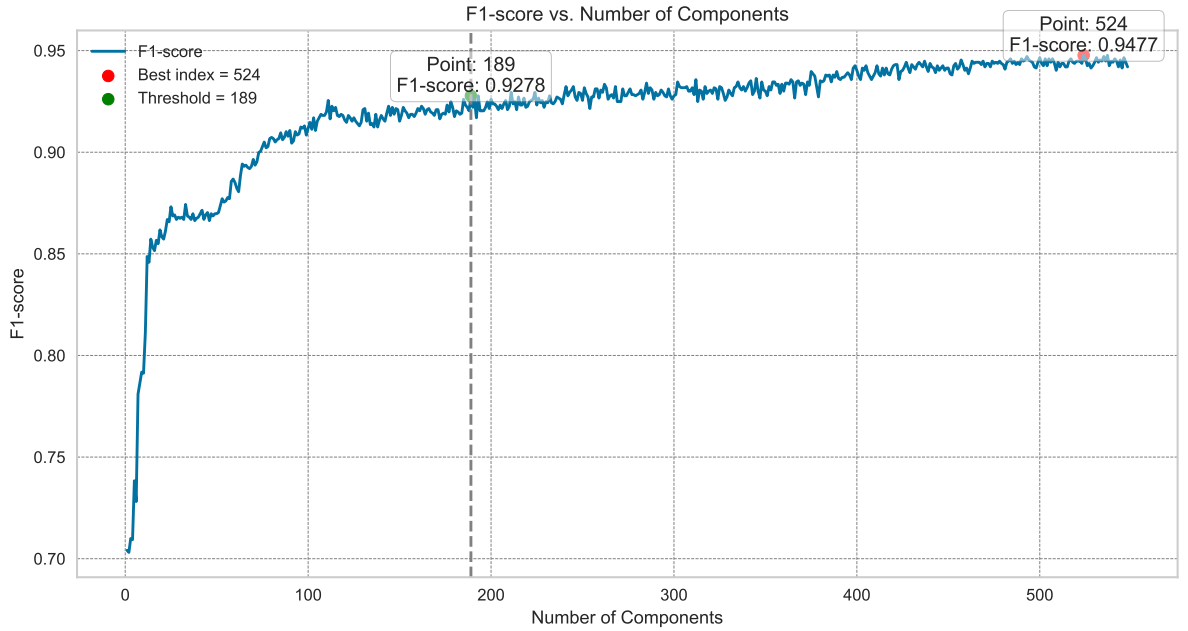


Figure C.21: PCA for gender-related posts chapter-level classification using Ridge classifier and feature engineering.

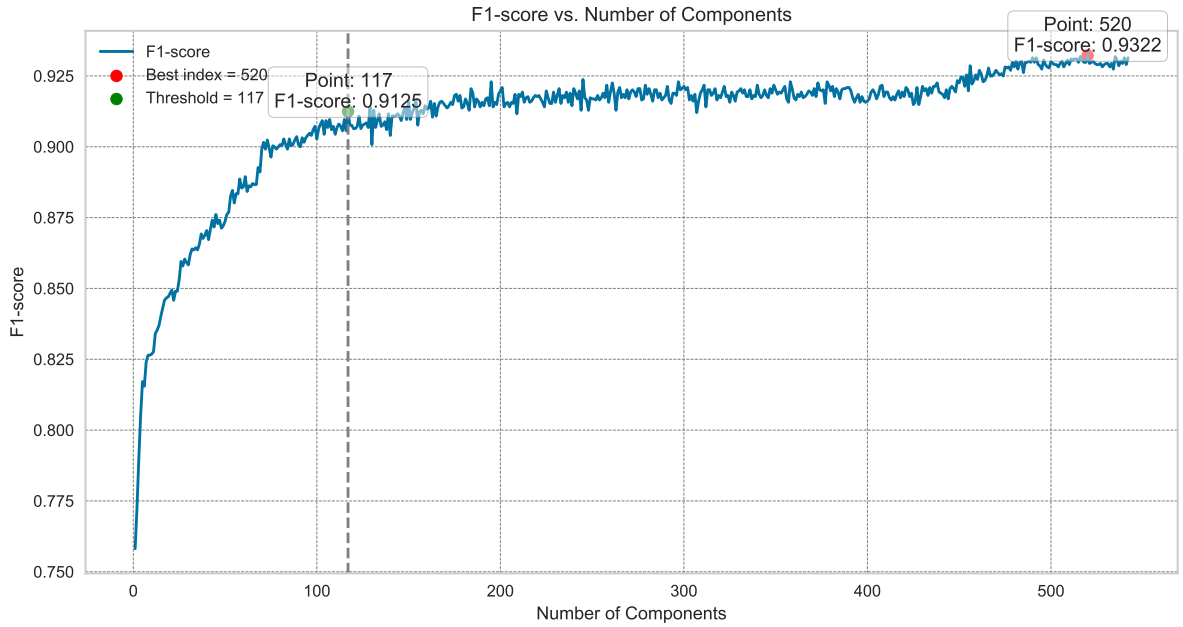


Figure C.22: PCA for gender-related all writings chapter-level classification using Ridge classifier and feature engineering.

C.3.2 Dimensionality reduction using autoencoder-based model

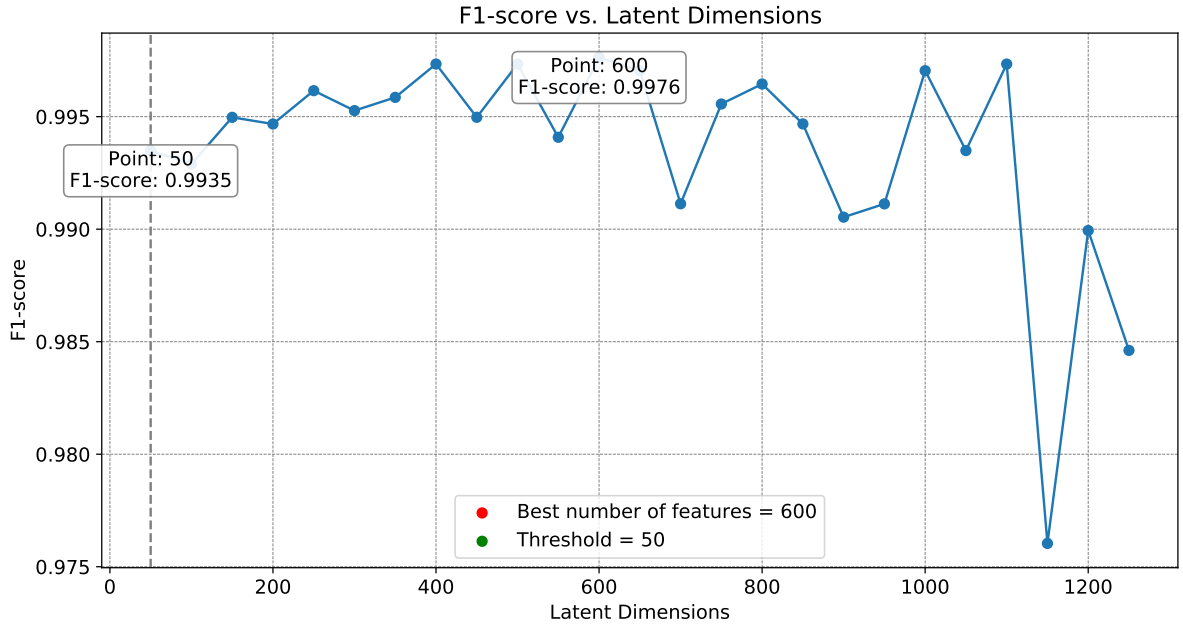


Figure C.23: Autoencoder-based model for gender-related literary chapter-level classification using Ridge classifier and feature engineering.

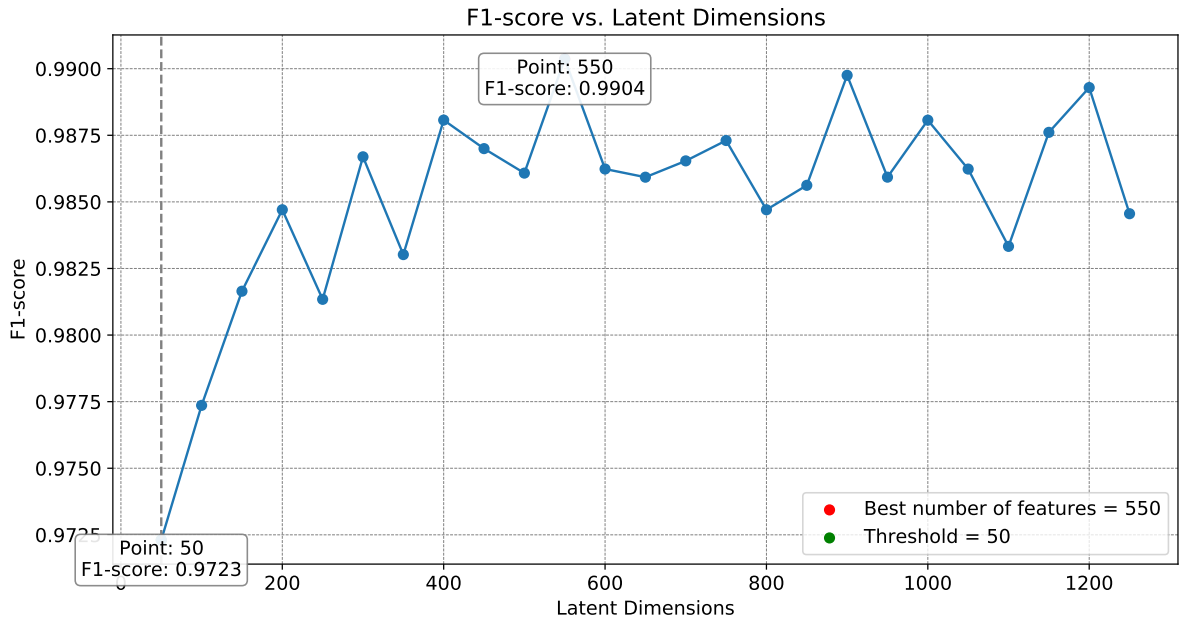


Figure C.24: Autoencoder-based model for gender-related columns chapter-level classification using Ridge classifier and feature engineering.

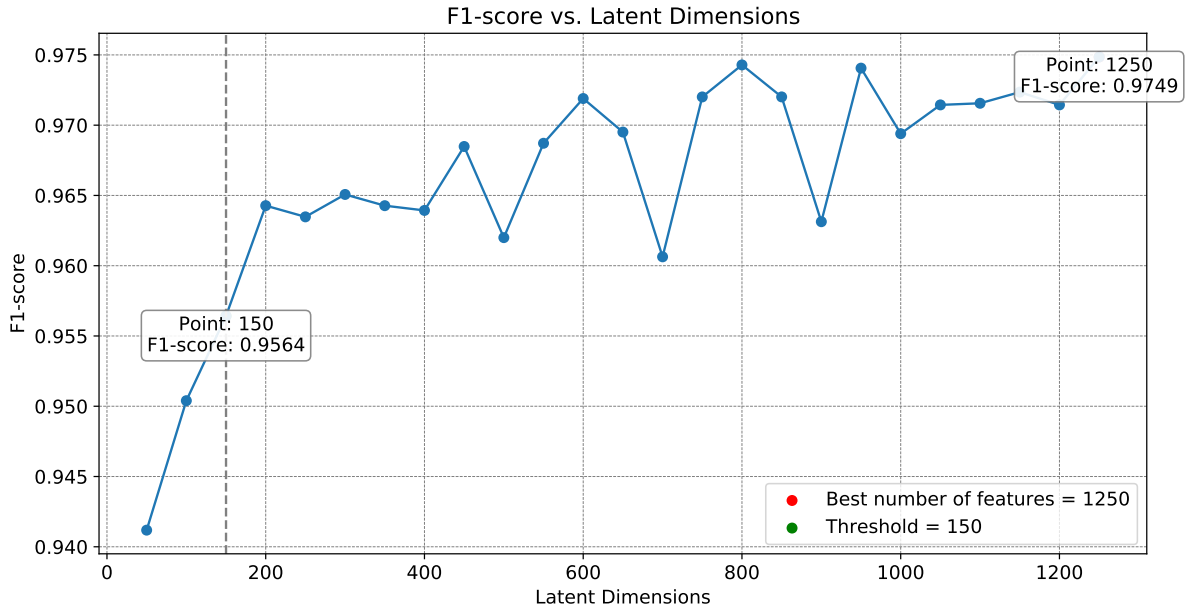


Figure C.25: Autoencoder-based model for gender-related posts chapter-level classification using Ridge classifier and feature engineering.

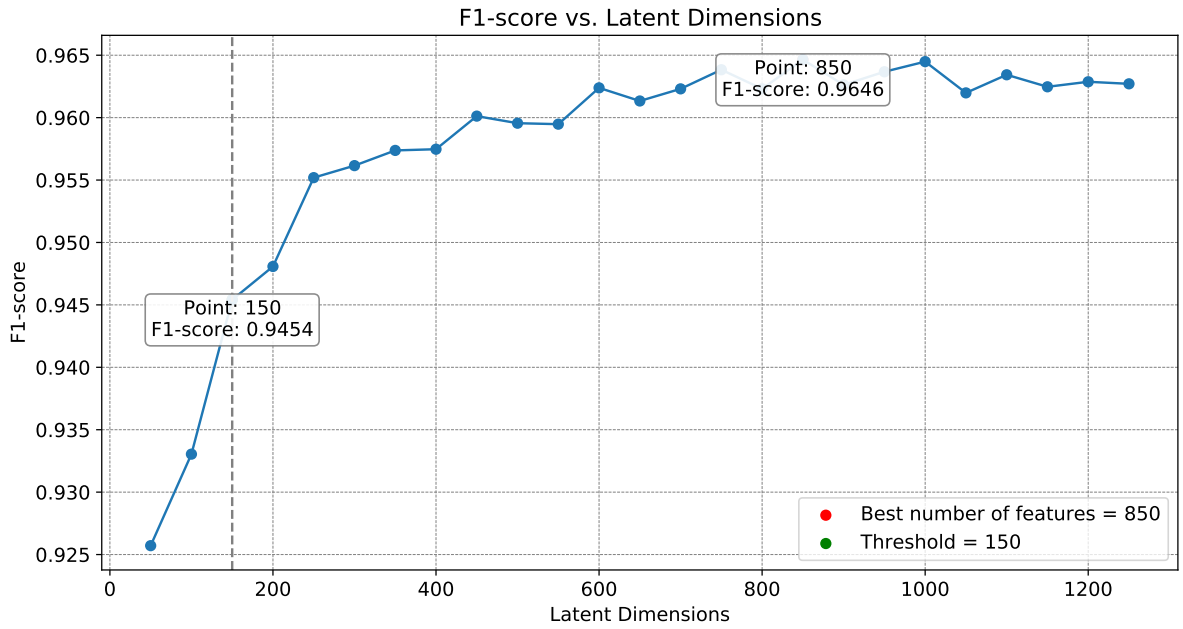


Figure C.26: Autoencoder-based model for gender-related all writings chapter-level classification using Ridge classifier and feature engineering.

C.3.3 Feature analysis using SHAP

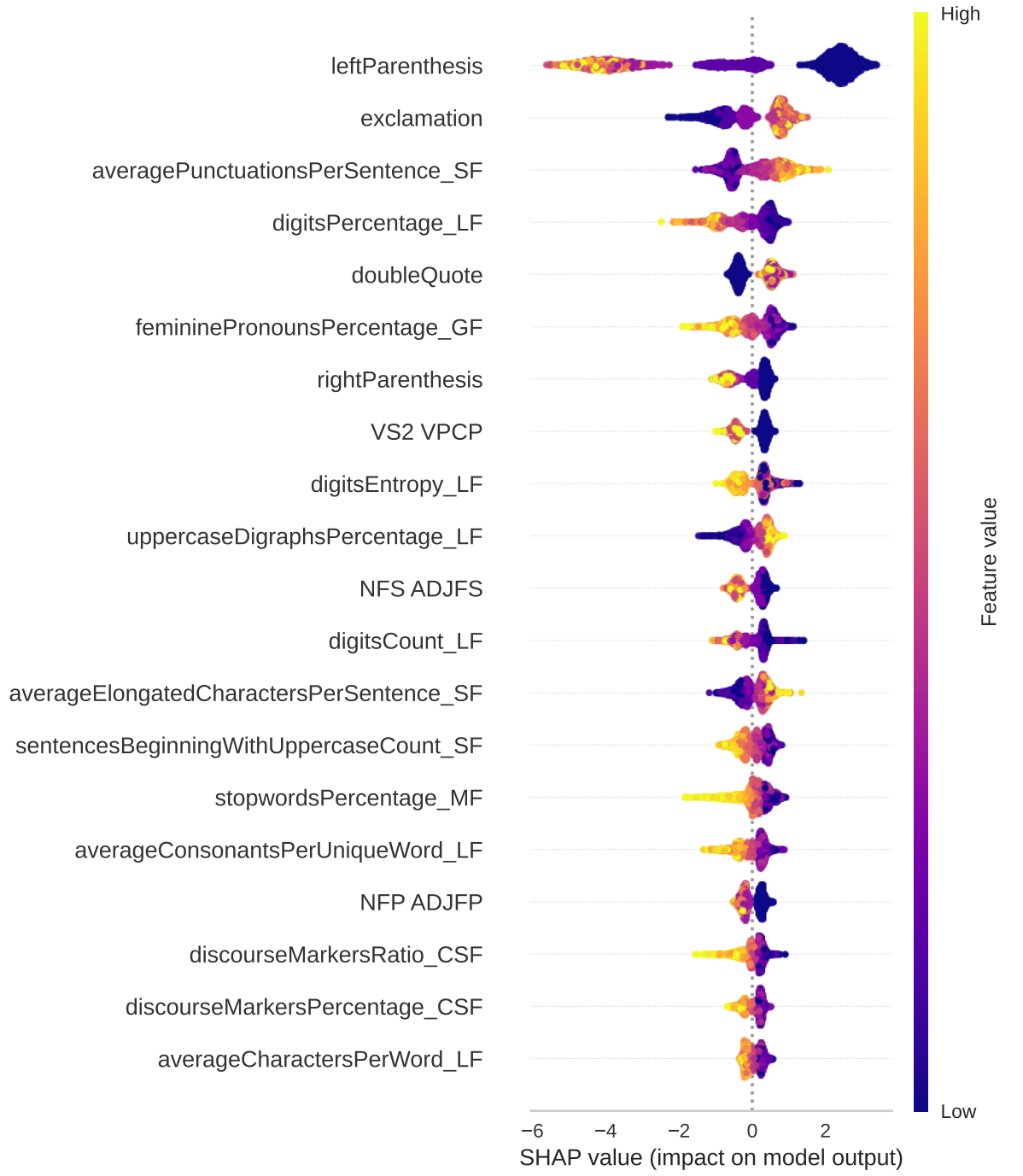


Figure C.27: SHAP summary plot for gender-related literary classification using XGBoost classifier and feature engineering.

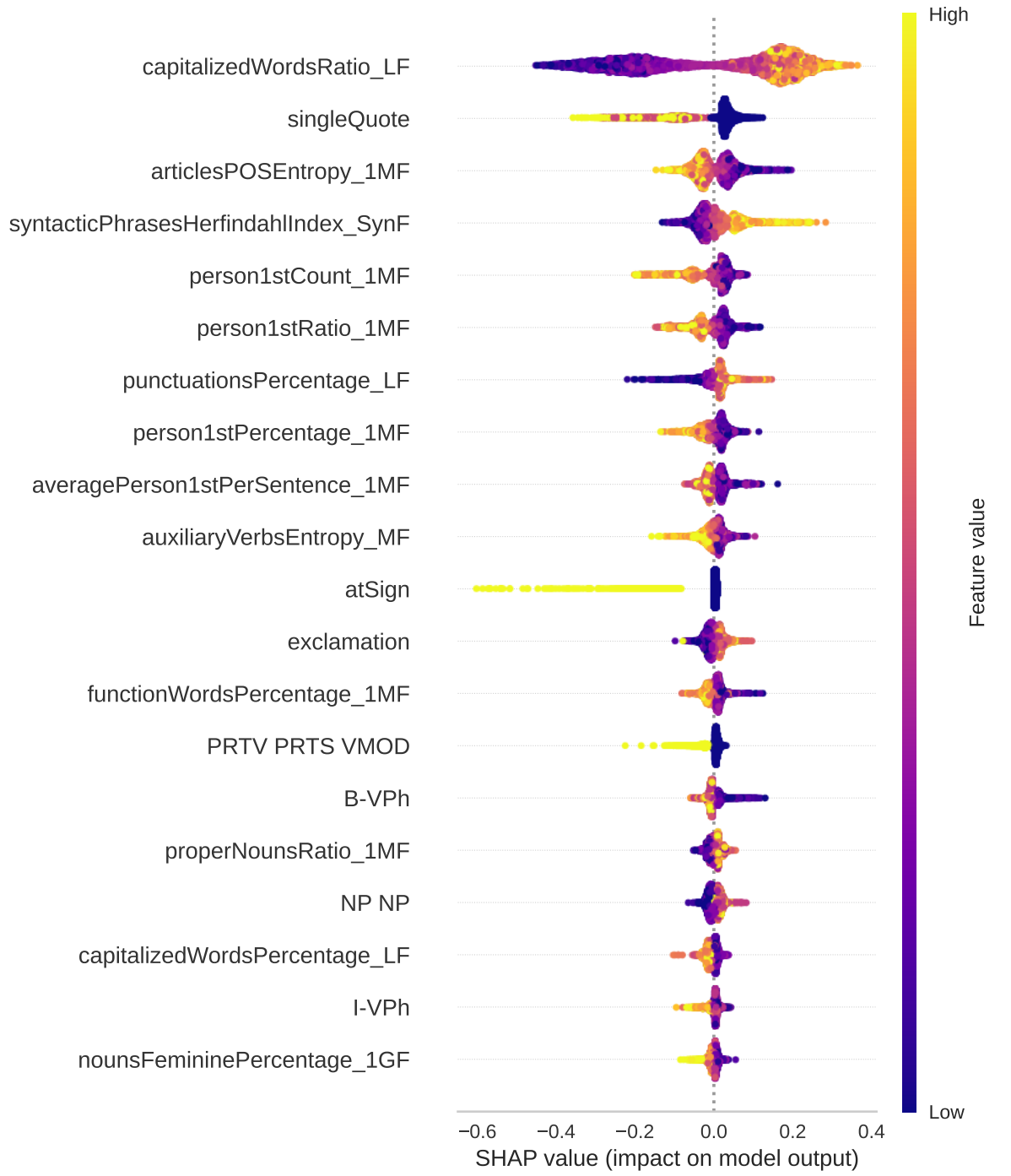


Figure C.28: SHAP summary plot for gender-related columns classification using XGBoost classifier and feature engineering.

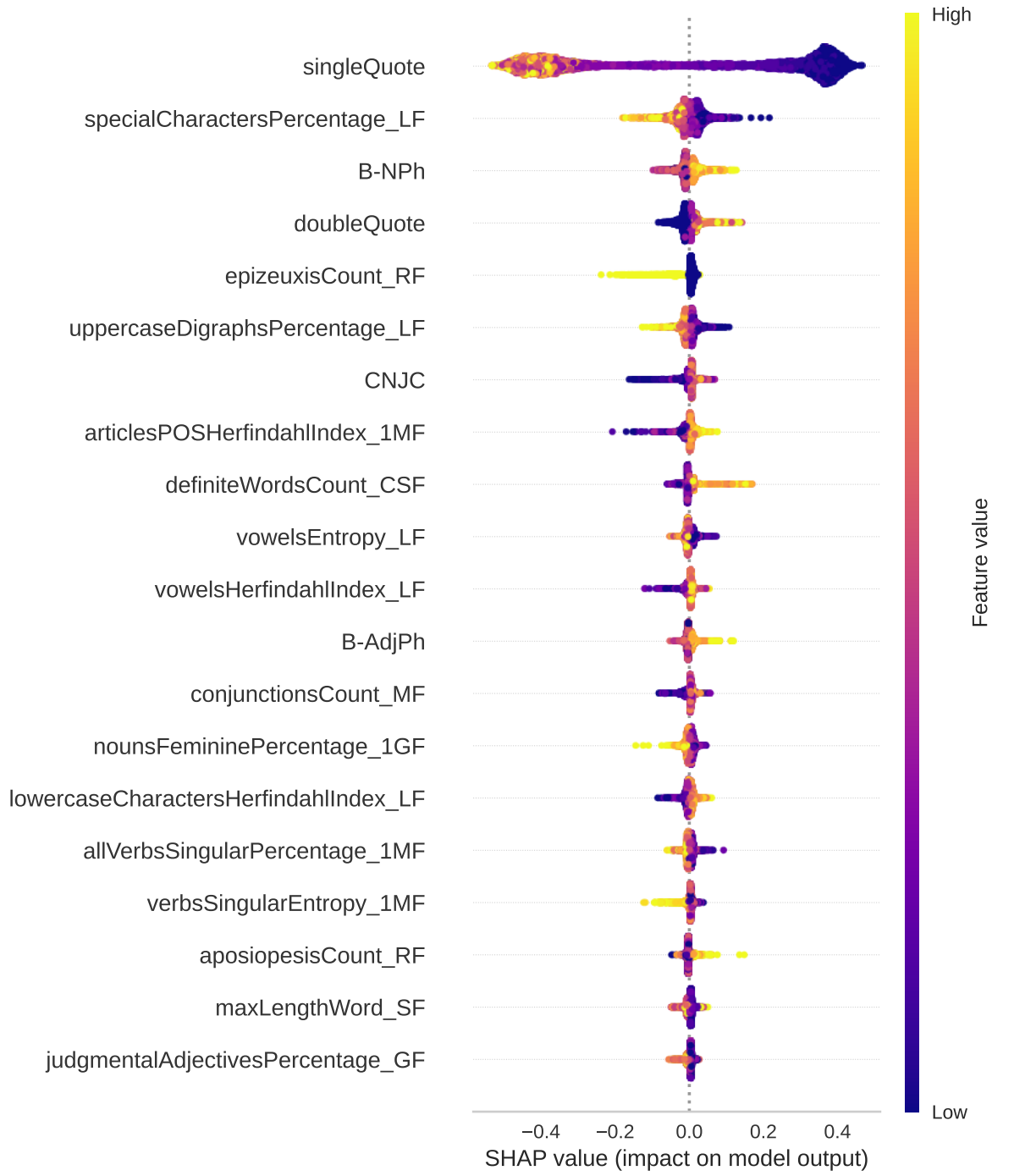


Figure C.29: SHAP summary plot for gender-related posts classification using XGBoost classifier and feature engineering.

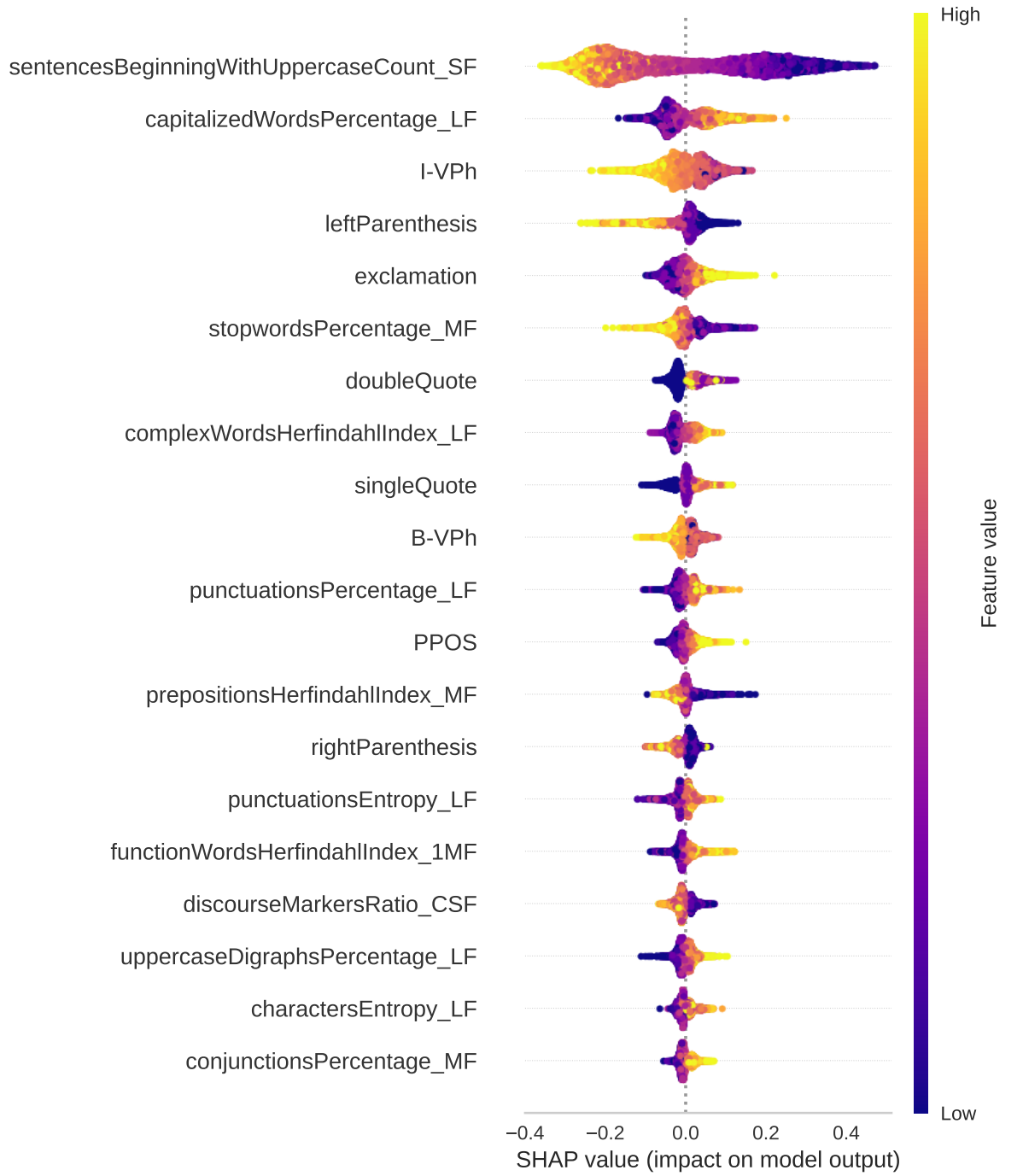


Figure C.30: SHAP summary plot for gender-related all writings classification using XGBoost classifier and feature engineering.

References

[1] Hervé Abdi and Lynne J Williams. “Principal component analysis”. In: *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010), pp. 433–459.

[2] M. Abuhamad et al. “Large-scale and Robust Code Authorship Identification with Deep Feature Learning”. en. In: *ACM Trans. Priv. Secur* 24.4 (2021). doi: 10.1145/3461666.

[3] Y. Abuhammad et al. “Authorship attribution of modern standard Arabic short texts”. en. In: *ACM Int. Conf. Proceeding Ser* 1.1 (2021), pp. 1–7, doi: 10.1145/3485557.3485563.

[4] H.V. Agun, S. Yilmazel, and O. Yilmazel. “Effects of language processing in Turkish authorship attribution”. en. In: *Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017* 1 (2018-01), pp. 1876–1881, doi: 10.1109/BigData.2017.8258132.

[5] M. Al-Ayyoub et al. “Feature extraction and selection for Arabic tweets authorship authentication”. en. In: *J. Ambient Intell. Humaniz. Comput* 8.3 (2017), pp. 383–393, doi: 10.1007/s12652-017-0452-1.

[6] M. Al-Sarem et al. “Combination of stylo-based features and frequency-based features for identifying the author of short Arabic text”. en. In: *ACM Int. Conf. Proceeding Ser* (2018). doi: 10.1145/3289402.3289500.

[7] S.A. Alanazi. “Toward identifying features for automatic gender detection: A corpus creation and analysis”. en. In: *IEEE Access* 7.Ii (2019), pp. 111931–111943, doi: 10.1109/ACCESS.2019.2932026.

[8] F. Alonso-Fernandez et al. “Writer Identification Using Microblogging Texts for Social Media Forensics”. en. In: *IEEE Trans. Biometrics, Behav. Identity Sci* 3.3 (2020). Available: pp. 1–22, url: http://arxiv.org/abs/2008.01533.

[9] F. Alonso-fernandez et al. “Writer Identification Using Microblogging Texts for Social Media Forensics”. en. In: XX (2020), pp. 1–22,

[10] H. Alshaher and J. Xu. “A New Term Weight Scheme and Ensemble Technique for Authorship Identification”. en. In: *ACM Int. Conf. Proceeding Ser* (2020), pp. 123–130, doi: 10.1145/3388142.3388159.

[11] K. Alsmearat et al. “Author gender identification from Arabic text”. en. In: *J. Inf. Secur. Appl* 35 (2017), pp. 85–95, doi: 10.1016/j.jisa.2017.06.003.

[12] K. Alsmearat et al. “Emotion analysis of Arabic articles and its impact on identifying the author’s gender”. en. In: *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA* November (2016). doi: 10.1109/AICCSA.2015.7507196.

[13] A.S. Altheneyan and M.E.B. Menai. “Naïve Bayes classifiers for authorship attribution of Arabic texts”. en. In: *J. King Saud Univ. - Comput. Inf. Sci* 26.4 (2014), pp. 473–484, doi: 10.1016/j.jksuci.2014.06.006.

[14] R. Ameri and H. Beigy. “Authorship identification from unstructured texts : A stylometric approach”. en. In: *CSI J. Comput. Sci. Eng* 17.2 (2020), pp. 34–45,

[15] W. Anwar et al. “An empirical study on forensic analysis of Urdu text using LDA-based authorship attribution”. en. In: *IEEE Access* 7 (2019), pp. 3224–3234, doi: 10.1109/ACCESS.2018.2885011.

[16] K.A. Apoorva and S. Sangeetha. “Deep neural network and model-based clustering technique for forensic electronic mail author attribution”. en. In: *SN Appl. Sci* 3.3 (2021), pp. 1–12, doi: 10.1007/s42452-020-04127-6.

[17] S. Aykent and G. Dozier. “Author Identification via a Distributed Neural-Evolutionary Hybrid (DiNEH)”. en. In: *Conf. Proc. - IEEE SOUTHEASTCON* (2020). doi: 10.1109/SoutheastCon44009.2020.9249707.

[18] M.G. Babenko, O.K. Strashkova, and I.A. Babenko. “Development of the neural network method for analyzing the literary text from the point of view of genre identification”. en. In: *Proc. 2017 Int. Conf. "Quality Manag. Transp. Inf. Secur. Inf. Technol. IT QM IS 2017*. 2017, pp. 162–165, doi: 10.1109/ITMQIS.2017.8085789.

[19] S. Balint et al. “Classifying Written Texts Through Rhythmic Features Citation”. en. In: *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. 2016, pp. 121–129, doi: 10.1007/978-3-319-44748-3.

[20] S. Baxevanakis et al. “A machine learning approach for gender identification of Greek tweet authors”. en. In: *ACM Int. Conf. Proceeding Ser* (2020), pp. 419–422, doi: 10.1145/3389189.3397992.

[21] N.M.Sharon Belvisi, N. Muhammad, and F. Alonso-Fernandez. “Forensic Authorship Analysis of Microblogging Texts Using N-Grams and Stylometric Features”. en. In: *Work. Biometrics Forensics, IWBF* (2020), pp. –, 1–6, doi: 10.1109/IWBF49977.2020.9107953.

[22] N.E. Benzebouchi et al. “Authors’ Writing Styles Based Authorship Identification System Using the Text Representation Vector”. en. In: *16th Int. Multi-Conference Syst. Signals Devices, SSD* (2019), pp. 371–376, doi: 10.1109/SSD.2019.8894872.

[23] Joseph Berkson. “A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function”. In: *Journal of the American Statistical Association* 48.263 (1953), pp. 565–599.

[24] M. Bhargava, P. Mehndiratta, and K. Asawa. “Stylometric analysis for authorship attribution on twitter”. en. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics* 8302 LNCS (2013), pp. 37–47, doi: 10.1007/978-3-319-03689-2_3.

[25] Piotr Bojanowski et al. “Enriching Word Vectors with Subword Information”. In: *arXiv preprint arXiv:1607.04606* (2016).

[26] D. Boughaci, M. Benmesbah, and A. Zebiri. “An improved N-grams based model for authorship attribution”. en. In: *Conf. Comput. Inf. Sci. ICCIS* (2019), pp. 1–6, doi: 10.1109/ICCISci.2019.8716391.

[27] M.L. Brocardo and I. Traore. “Continuous authentication using micro-messages”. en. In: *Conf. Privacy, Secur. Trust. PST* (2014), pp. 179–188, doi: 10.1109/PST.2014.6890938.

[28] M.L. Brocardo et al. “Authorship verification for short messages using stylometry”. en. In: *Conf. Comput. Inf. Telecommun. Syst. CITS* (2013). doi: 10.1109/CITS.2013.6705711.

[29] J Brownlee. *Master Machine Learning Algorithms Discover How They Work and Implement Them From Scratch i Master Machine Learning Algorithms*. Tech. rep. Technical report, 2016.

[30] Jason Brownlee. *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery, 2020.

[31] Jason Brownlee. *Machine learning mastery with Python: understand your data, create accurate models, and work projects end-to-end*. Machine Learning Mastery, 2016.

[32] Gertjan van den Burg. *Algorithms for multiclass classification and regularized regression*. Tech. rep. 2018.

[33] P. Canbay, E.A. Sezer, and H. Sever. "Authorship modelling approach for authorship verification on the Turkish texts". en. In: *26th IEEE Signal Process. Commun. Appl. Conf. SIU 2018*. Available: 2018, pp. 1–4, doi: 10.1109/SIU.2018.8404436.

[34] Ercan Canhasi, Rexhep Shijaku, and Erblin Berisha. "Albanian fake news detection". In: *ACM Transactions on Asian and Low-Resource Language Information Processing* February (2022), pp. 0–24. issn: 2375-4699. doi: 10.1145/3487288. url: https://doi.org/10.1145/3487288.

[35] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[36] Xue-wen Chen and Jong Cheol Jeong. "Enhanced recursive feature elimination". In: *Sixth international conference on machine learning and applications (ICMLA 2007)*. IEEE. 2007, pp. 429–435.

[37] N. Cheng, R. Chandramouli, and K.P. Subbalakshmi. "Author gender identification from text". en. In: *Digit. Investig* 8.1 (2011), pp. 78–88, doi: 10.1016/j.diin.2011.04.002.

[38] H.Ahmed Chowdhury, M.A.Haque Imon, and M.S. Islam. "A Comparative Analysis of Word Embedding Representations in Authorship Attribution of Bengali Literature". en. In: *Conf. Comput. Inf. Technol. ICCIT* December (2018). doi: 10.1109/ICCITECHN.2018.8631977.

[39] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20 (1995), pp. 273–297.

[40] J.E. Custódio and I. Paraboni. “Stacked authorship attribution of digital texts”. en. In: *Expert Syst. Appl* 176.July 2020 (2021). doi: 10.1016/j.eswa.2021.114866.

[41] Z. Damiran and K. Altangerel. “Author identification: An experiment based on Mongolian literature using decision trees”. en. In: *Proc. - 2014 7th Int*. Conf. Ubi-Media Comput. Work. U-MEDIA, 2014, pp. 186–189, doi: 10.1109/U-MEDIA.2014.22.

[42] S. Das and P. Mitra. “Author identification in bengali literary works”. en. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics* 6744 LNCS (2011), pp. 220–226, doi: 10.1007/978-3-642-21786-9_37.

[43] T.K. Dugar, S. Gowtham, and U.K. Chakraborty. “Hyperparameter tuning for enhanced authorship identification using deep neural networks”. en. In: *Proc. Int. Conf. Trends Electron. Informatics, ICOEI 2019* Icoei (2019-04), pp. 206–211, doi: 10.1109/ICOEI.2019.8862631. url: https://doi.org/10.1109/ICOEI.2019.8862631.

[44] H.J. Escalante, T. Solorio, and M. Montes-Y-Gómez. “Local histograms of character n-grams for authorship attribution”. en. In: *ACL-HLT* 1 (2011), pp. 288–298,

[45] Anastasia Fedotova et al. “Authorship attribution of social media and literary russian-language texts using machine learning methods and feature selection”. In: *Future Internet* 14.1 (2022). issn: 19995903. doi: 10.3390/fi14010004.

[46] Mark Fenner. *Machine learning with Python for everyone*. Addison-Wesley Professional, 2019.

[47] O. Fourkioti, S. Symeonidis, and A. Arampatzis. “Language models and fusion for authorship attribution”. en. In: *Inf. Process. Manag* 56.6 (2019), pp. 102061, doi: 10.1016/j.ipm.2019.102061.

[48] J.A. García-Díaz, R. Colomo-Palacios, and R. Valencia-García. “Psychographic traits identification based on political ideology: An author analysis study on Spanish politicians’ tweets posted in 2020”. en. In: *Futur. Gener. Comput. Syst* 130 (2022), pp. 59–74, doi: 10.1016/j.future.2021.12.011.

[49] J. Gaston et al. “Authorship Attribution via Evolutionary Hybridization of Sentiment Analysis, LIWC, and Topic Model Features”. en. In: *2018 IEEE Symposium Series on Computational Intelligence (SSCI*. 2018, pp. 933–940.

[50] S.T.P. Gupta, J.K. Sahoo, and R.K. Roul. “Authorship identification using recurrent neural networks”. en. In: *ACM Int. Conf. Proceeding Ser* (2019), pp. 133–137, doi: 10.1145/3325917.3325935. url: https://doi.org/10.1145/3325917.3325935.

[51] H. Gómez-Adorno et al. “Document embeddings learned on various types of n-grams for cross-topic authorship attribution”. en. In: *Computing* 100.7 (2018), pp. 741–756, doi: 10.1007/s00607-018-0587-8.

[52] M. Hajja, A. Yahya, and A. Yahya. “Authorship Attribution of Arabic Articles”. en. In: *Commun. Comput. Inf. Sci* 1108.Aiccsa (2019), pp. 194–208, doi: 10.1007/978-3-030-32959-4_14.

[53] D. Harman. “Overview of the first TREC conference”. en. In: *Proc. Annu. Int. ACM SIGIR Conf. Res. Dev. Infofmation Retr* Figure 1 (1993), pp. 36–47, doi: 10.1145/160688.160692.

[54] D. Harman. “Overview of the Second Text REtrievalConference (TREC-2”. en. In: *Proc. - Natl. Online Meet.* 1993, pp. 181–187,

[55] Zellig S Harris. “Distributional structure”. In: *Word* 10.2-3 (1954), pp. 146–162.

[56] S.U. Hassan et al. “Deep stylometry and lexical & syntactic features based author attribution on PLoS digital repository”. en. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics* 10647 LNCS (2017), pp. 119–127, doi: 10.1007/978-3-319-70232-2_10.

[57] Geoffrey E Hinton and Richard Zemel. “Autoencoders, minimum description length and Helmholtz free energy”. In: *Advances in neural information processing systems* 6 (1993).

[58] Tin Kam Ho. “Random decision forests”. In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

[59] Arthur E Hoerl and Robert W Kennard. “Ridge regression: Biased estimation for nonorthogonal problems”. In: *Technometrics* 12.1 (1970), pp. 55–67.

[60] D.I. Holmes and R.S. Forsyth. “The Federalist revisited: New directions in authorship attribution”. en. In: *Lit. Linguist. Comput* 10.2 (1995), pp. 111–127, doi: 10.1093/llc/10.2.111.

[61] A.S. Hossain, N. Akter, and M.D.Saiful Islam. “A stylometric approach for author attribution system using neural network and machine learning classifiers”. en. In: *ACM Int. Conf. Proceeding Ser* (2020). doi: 10.1145/3377049.3377079. url: https://doi.org/10.1145/3377049.3377079.

[62] M.R. Hossain et al. “Authorship classification in a resource constraint language using convolutional neural networks”. en. In: *IEEE Access* 9 (2021), pp. 100319–100338, doi: 10.1109/ACCESS.2021.3095967.

[63] Eduard Hovy and Julia Lavid. “Towards a ‘ Science ’ of Corpus Annotation : A New Methodological Challenge for Corpus Linguistics”. In: *International Journal of Translation* 22.1 (2010), p. 25.

[64] S. Hriez and A. Awajan. “Authorship Identification for Arabic Texts Using Logistic Model Tree Classification”. en. In: *Advances in Intelligent Systems and Computing*. Vol. 1229. 2020, pp. 656–666, doi: 10.1007/978-3-030-52246-9_48.

[65] J. Hurtado, N. Taweewitchakreeya, and X. Zhu. “Who wrote this paper? Learning for authorship de-identification using stylometric featuress”. en. In: *Proc. 2014 IEEE 15th Int. Conf. Inf. Reuse Integr. IEEE IRI*. 2014, pp. 859–862, doi: 10.1109/IRI.2014.7051981.

[66] M.A. Islam et al. “Authorship attribution on Bengali literature using stylometric features and neural network”. en. In: *4th Int. Conf. Electr. Eng. Inf. Commun. Technol. iCEEiCT* (2018), pp. 360–363, doi: 10.1109/CEEICT.2018.8628106.

[67] F. Jafariakinabad and K.A. Hua. “A Self-Supervised Representation Learning of Sentence Structure for Authorship Attribution”. en. In: *ACM Trans. Knowl. Discov. Data* 16.4 (2022), pp. 1–16, doi: 10.1145/3491203.

[68] F. Jafariakinabad and K.A. Hua. “Style-aware neural model with application in authorship attribution”. en. In: *Proc. - 18th IEEE Int. Conf. Mach. Learn. Appl. ICMLA*, 2019, pp. 325–328, doi: 10.1109/ICMLA.2019.00061.

[69] F. Jafariakinabad, S. Tarnpradab, and K.A. Hua. "Syntactic Recurrent Neural Network for Authorship Attribution". en. In: (2019). Available: url: http://arxiv.org/abs/1902.09723.

[70] Armand Joulin et al. "Bag of Tricks for Efficient Text Classification". In: *arXiv preprint arXiv:1607.01759* (2016).

[71] Armand Joulin et al. "FastText.zip: Compressing text classification models". In: *arXiv preprint arXiv:1612.03651* (2016).

[72] P. Juola. "An Overview of the Traditional Authorship Attribution Subtask". en. In: *CLEF* (2012). Available: pp. 37–41, url: http://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-Juola2012.pdf.

[73] Patrick Juola. *Authorship Attribution*. now Publishers Inc., 2006, pp. 1–115. isbn: 978-1-60198-118-9.

[74] Besim Kabashi and Thomas Proisl. "Albanian part-of-speech tagging: Gold standard and evaluation". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018. url: https://aclanthology.org/L18-1412.

[75] Arbana Kadriu. "NLTK tagger for Albanian using iterative approach". In: *Proceedings of the ITI 2013 35th International Conference on Information Technology Interfaces*. IEEE. 2013, pp. 283–288. url: https://ieeexplore.ieee.org/document/6649039.

[76] Arbana Kadriu and Lejla Abazi. "A Comparison of Algorithms for Text Classification of Albanian News Articles". In: *ENTRENOVA-ENTerprise REsearch InNOVAtion* 3.1 (2017), pp. 62–68. url: https://hrcak.srce.hr/251111.

[77] Arbana Kadriu, Lejla Abazi, and Hyrije Abazi. "Albanian Text Classification: Bag of Words Model and Word Analogies". In: *Business Systems Research* 10.1 (2019), pp. 74–87. issn: 18479375. doi: 10.2478/bsrj-2019-0006.

[78] R. Kaur, S. Singh, and H. Kumar. *Authorship Analysis of Online Social Media Content*. en. Vol. 46. Springer Singapore, 2019.

[79] A. Khatun et al. “Authorship Attribution in Bangla Literature (AABL) via Transfer Learning using ULMFiT”. en. In: *ACM Trans. Asian Low-Resource Lang. Inf. Process* (2022). doi: 10.1145/3530691.

[80] A. Khatun et al. “Authorship attribution in bangla literature using character-level CNN”. en. In: *Conf. Comput. Inf. Technol. ICCIT* (2019), pp. 18–20, doi: 10.1109/ICCIT48885.2019.9038560.

[81] A.J. Khdr and C. Varol. “Age and Gender Identification by SMS Text Messages”. en. In: *Conf. Artif. Intell. Data Process. IDAP* (2018), pp. 1–5, doi: 10.1109/IDAP.2018.8620780.

[82] I. Khomytska, V. Teslyuk, and L. Bordyuk. “The Kolmogorov-Smirnov’s Test for Authorship Attribution on the Phonological Level”. en. In: *Int. Sci. Tech. Conf. Comput. Sci. Inf. Technol* 1.May (2020), pp. 259–262, doi: 10.1109/CSIT49958.2020.9322042.

[83] Nelda Kote et al. “Morphological tagging and lemmatization of Albanian: A manually annotated corpus and neural models”. In: *arXiv preprint arXiv:1912.00991* (2019). url: https://arxiv.org/abs/1912.00991.

[84] Mark A Kramer. “Nonlinear principal component analysis using autoassociative neural networks”. In: *AIChE journal* 37.2 (1991), pp. 233–243.

[85] T. Kucukyilmaz, A. Deniz, and H.E. Kiziloz. “Boosting gender identification using author preference”. en. In: *Pattern Recognit. Lett* 140 (2020), pp. 245–251, doi: 10.1016/j.patrec.2020.10.002.

[86] K.P. Kumari and A.V. Reddy. “Syllable n-gram approach for identification and classification of genres in Telugu language”. en. In: *1st Int. Conf. Networks Soft Comput. ICNSC 2014 - Proc.* 2014, pp. 125–129, doi: 10.1109/CNSC.2014.6906646.

[87] A.O. Kusakci. “Authorship attribution using committee machines with k-nearest neighbors rated voting,” 11th Symp. Neural Netw”. en. In: *Appl. Electr. Eng* April (2012), pp. 161–166, doi: 10.1109/NEUREL.2012.6419997.

[88] I. Lagutina et al. “Automatic Extraction of Rhythm Figures and Analysis of Their Dynamics in Prose of 19th-21st Centuries”. en. In: *2020 26th Conference of Open Innovations Association (FRUCT*. 2020, pp. 247–255.

[89] K. Lagutina. “A Survey on Stylometric Text Features”. en. In: *Conf. Open Innov. Assoc. Fruct* (2019), pp. 184–195, doi: 10.23919/FRUCT48121.2019.8981504.

[90] K. Lagutina and N. Lagutina. “A Survey of Models for Constructing Text Features to Classify Texts in Natural Language”. en. In: *Conf. Open Innov. Assoc. Fruct* Section IV (2021), pp. 222–233, doi: 10.23919/FRUCT52173.2021.9435512.

[91] K. Lagutina et al. “Authorship verification of literary texts with rhythm features”. en. In: *Conf. Open Innov. Assoc. Fruct* 2021-Janua (2021). doi: 10.23919/FRUCT50888.2021.9347649.

[92] O. Litvinova et al. “Ruspersonality: a Russian Corpus for Authorship Porfiling and Deception Detection”. en. In: *2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT* May (2016), pp. 1–7.

[93] T. Litvinova, O. Litvinova, and P. Panicheva. “Authorship attribution of Russian forum posts with different types of N-gram features”. en. In: *ACM Int. Conf. Proceeding Ser* (2019), pp. 9–14, doi: 10.1145/3342827.3342834.

[94] L.M. Lopez-Santamaria et al. “Age and gender identification in unbalanced social media”. en. In: *CONIELECOMP* March (2019), pp. 74–80, doi: 10.1109/CONIELECOMP.2019.8673125.

[95] Harold Love. *Attributing Authorship*. Cambridge University Press, 2002, pp. 1–281. isbn: 978-0-521-78339-2.

[96] D. Lowe and R. Matthews. “Shakespeare Vs. Fletcher: A stylometric analysis by radial basis functions”. en. In: *Comput. Hum* 29.6 (1995), pp. 449–461, doi: 10.1007/BF01829876.

[97] L. Lundmark et al. “The Applicability of Authorship Verification to Swedish Discussion Forums”. en. In: *Proc. - 2020 IEEE Int.* Conf. Big Data, Big Data, 2020, pp. 1805–1812, doi: 10.1109/BigData50022.2020.9378373.

[98] I. Markov, E. Stamatatos, and G. Sidorov. “Improving cross-topic authorship attribution: The role of pre-processing”. en. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics* Cic (2018), pp. 289–302, doi: 10.1007/978-3-319-77116-8_21.

[99] R. Marukatat et al. “Authorship attribution analysis of thai online messages”. en. In: *ICISA* (2014). doi: 10.1109/ICISA.2014.6847369.

[100] R.K. Maurya, M.R. Saxena, and N. Akhil. “Authorship Attribution Using Stylometry and Machine Learning Techniques”. fr. In: *Adv. Intell. Syst. Comput* 384 (2016-01), pp. 247–257, doi: 10.1007/978-3-319-23036-8.

[101] Tony McEnery and Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. 2012.

[102] A. Mehri, A.H. Darooneh, and A. Shariati. “The complex networks approach for authorship attribution of books”. en. In: *Phys. A Stat. Mech. its Appl* 391.7 (2012), pp. 2429–2437, doi: 10.1016/j.physa.2011.12.011.

[103] Tomas Mikolov et al. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems* 26 (2013).

[104] Arta Misini, Arbana Kadriu, and Ercan Canhasi. “A survey on authorship analysis tasks and techniques”. In: *SEEU Review* 17.2 (2022), pp. 153–167. doi: 10.2478/seeur-2022-0100.

[105] Arta Misini, Arbana Kadriu, and Ercan Canhasi. “A3C: Albanian Authorship Attribution Corpus”. In: *Economic Recovery, Consolidation, and Sustainable Growth*. Cham: Springer Nature Switzerland, 2023, pp. 755–763. isbn: 978-3-031-42511-0. doi: 10.1007/978-3-031-42511-0_49.

[106] Arta Misini, Arbana Kadriu, and Ercan Canhasi. “Albanian Authorship Attribution Model”. In: *2023 12th Mediterranean Conference on Embedded Computing (MECO)*. 2023, pp. 1–5. doi: 10.1109/MECO58584.2023.10155046.

[107] Arta Misini, Arbana Kadriu, and Ercan Canhasi. “Authorship classification techniques: Bridging textual domains and languages”. In: *International Journal on Information Technologies and Security* 16.1 (2024), pp. 27–38. doi: 10.59035/UKBE1226.

[108] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[109] S. Nagaprasad et al. “Empirical evaluations using character and word N-grams on authorship attribution for Telugu text”. en. In: *Adv. Intell. Syst. Comput* 343 (2015), pp. 613–623, doi: 10.1007/978-81-322-2268-2_62.

[110] Z. Nazir et al. “Authorship Attribution for a Resource Poor Language - Urdu”. en. In: *ACM Trans. Asian Low-Resource Lang. Inf. Process* 21.3 (2022), pp. 1–23, doi: 10.1145/3487061.

[111] T. Neal et al. “Surveying stylometry techniques and applications”. fr. In: *ACM Comput. Surv* 50.6 (2017). doi: 10.1145/3132039.

[112] A. Neocleous and A. Loizides. “Machine Learning and Feature Selection for Authorship Attribution: The Case of Mill, Taylor Mill and Taylor, in the Nineteenth Century”. en. In: *IEEE Access* 9 (2021), pp. 7143–7151, doi: 10.1109/ACCESS.2020.3047583.

[113] S. Nirkhi, R.V. Dharaskar, and V.M. Thakare. “Authorship Verification of Online Messages for Forensic Investigation”. en. In: *Phys. Procedia* 78 (2016), pp. 640–645, doi: 10.1016/j.procs.2016.02.111.

[114] S. Okuno, H. Asai, and H. Yamana. “A challenge of authorship identification for ten-thousand-scale microblog users”. it. In: *Proc. - 2014 IEEE Int*. Conf. Big Data, IEEE Big Data, 2014, pp. 52–54, doi: 10.1109/BigData.2014.7004491.

[115] A. Onan. “An ensemble scheme based on language function analysis and feature engineering for text genre classification”. en. In: *J. Inf. Sci* 44.1 (2018), pp. 28–47, doi: 10.1177/0165551516677911.

[116] S. Ouamour and H. Sayoud. “Authorship attribution of short historical arabic texts based on lexical features”. en. In: *2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discov. CyberC, 2013, pp. 144–147, doi: 10.1109/CyberC.2013.31.

[117] S. P. and S.P.Hoshiladevi Ramnial. “Authorship Attribution Using Stylometry and Machine Learning Techniques”. fr. In: *Adv. Intell. Syst. Comput* 384.August (2016), pp. 247–257, doi: 10.1007/978-3-319-23036-8.

[118] H. Paci et al. “Author identification in Albanian language”. en. In: *2011 14th International Conference on Network-Based Information Systems*. Conf. Network-Based Inf. Syst. NBiS, 2011, pp. 425–430, doi: 10.1109/NBiS.2011.71.

[119] P. Panicheva and T. Litvinova. “Authorship Attribution in Russian in Real-World Forensics Scenario”. en. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell.*

Lect. Notes Bioinformatics 11816 LNAI (2019), pp. 299–310, doi: 10.1007/978-3-030-31372-2_25.

[120] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[121] J. Petrik and D. Chuda. "Source code authorship approaches natural language processing". en. In: *ACM Int. Conf. Proceeding Ser* September (2018), pp. 58–61, doi: 10.1145/3274005.3274031.

[122] S. Phani, S. Lahiri, and A. Biswas. "A supervised learning approach for Authorship attribution of Bengali literary texts". en. In: *ACM Trans. Asian Low-Resource Lang. Inf. Process* 16.4 (2017), pp. 15, doi: 10.1145/3099473.

[123] J.P. Posadas-Durán et al. "Application of the distributed document representation in the authorship attribution task for small corpora". en. In: *Soft Comput* 21.3 (2017), pp. 627–639, doi: 10.1007/s00500-016-2446-x.

[124] N. Potha and E. Stamatatos. "A profile-based method for authorship verification". en. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics* 8445 LNCS (2014), pp. 313–326, doi: 10.1007/978-3-319-07064-3_25.

[125] N. Potha and E. Stamatatos. "Improved algorithms for extrinsic author verification". en. In: *Knowl. Inf. Syst* 62.5 (2020), pp. 1903–1921, doi: 10.1007/s10115-019-01408-4.

[126] N. Potha and E. Stamatatos. "Intrinsic author verification using topic modeling". en. In: *ACM Int. Conf. Proceeding Ser* (2018). doi: 10.1145/3200947.3201013.

[127] Noun Project Press. *Noun Project*. https://thenounproject.com/. Accessed: 2023-03-04. 2010.

[128] M.A. Raafat, R.A.F. El-Wakil, and A. Atia. "Comparative study for Stylometric analysis techniques for authorship attribution". fr. In: *Conf. MIUCC* (2021), pp. 176–181, doi: 10.1109/MIUCC52538.2021.9447600.

[129] R. Ramezani. "A language-independent authorship attribution approach for author identification of text documents". en. In: *Expert Syst. Appl* 180 (2020-05), pp. 115139, doi: 10.1016/j.eswa.2021.115139.

[130] R. Ramezani, N. Sheydaei, and M. Kahani. “Evaluating the effects of textual features on authorship attribution accuracy”. en. In: *Proc* (2013), pp. 108–113, doi: 10.1109/ICCKE.2013.6682828.

[131] T.R. Reddy, B.V. Vardhan, and P.V. Reddy. “N-gram approach for gender prediction”. en. In: *Proc. - 7th IEEE Int. Adv. Comput. Conf. IACC 2017*. 2017, pp. 860–865, doi: 10.1109/IACC.2017.0176.

[132] A. Rexha et al. “Authorship identification of documents with high content similarity”. en. In: *Scientometrics* 115.1 (2018), pp. 223–237, doi: 10.1007/s11192-018-2661-6.

[133] A. Rocha. “Authorship Attribution for Social Media Forensics”. en. In: *IEEE Trans. Inf. Forensics Secur* 12.1 (2017), pp. 5–33, doi: 10.1109/TIFS.2016.2603960.

[134] A. Romanov et al. “Authorship identification of a russian-language text using support vector machine and deep neural networks”. en. In: *Futur. Internet* 13.1 (2021), pp. 1–16, doi: 10.3390/fi13010003.

[135] T. Rose, M. Stevenson, and M. Whitehead. “The reuters corpus volume 1 - From yesterday’s news to tomorrow’s language resources”. en. In: *Proc* 1 (2002), pp. 827–833, url: http://www.lrec-conf.org/proceedings/lrec2002/pdf/80.pdf.

[136] Alvin E Roth. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

[137] C. Saedi and M. Dras. “Siamese networks for large-scale author identification”. en. In: *Comput. Speech Lang* 70 (2021), pp. 101241, doi: 10.1016/j.csl.2021.101241.

[138] R. Sarwar and S.-U. Hassan. “UrduAI: Writeprints for Urdu Authorship Identification”. en. In: *ACM Trans. Asian Low-Resource Lang. Inf. Process* 21.2 (2022), pp. 1–18, doi: 10.1145/3476467.

[139] R. Sarwar et al. “StyloThai: A scalable framework for stylometric authorship identification of Thai documents”. en. In: *ACM Trans. Asian Low-Resource Lang. Inf. Process* 19.3 (2020). doi: 10.1145/3365832.

[140] J. Savoy. “Authorship attribution based on specific vocabulary”. en. In: *ACM Trans. Inf. Syst* 30.2 (2012). doi: 10.1145/2180868.2180874.

[141] N.S. Saygili et al. “Taking advantage of Turkish characteristic features to achieve authorship attribution problems for Turkish”. en. In: 2017, pp. 1–4, doi: 10.1109/siu.2017.7960438.

[142] N.S. Saygili et al. “Taking Advantage of Turkish Characteristic Features to Tackle with Authorship Attribution Problems for Turkish”. en. In: *Commun. Appl. Conf* November (2017), pp. 1–4, doi: 10.1109/siu.2017.7960438.

[143] A. Sboev et al. “Deep Learning Network Models to Categorize Texts According to Author’s Gender and to Identify Text Sentiment”. en. In: *Proc* (2016), pp. 1101–1106, doi: 10.1109/CSCI.2016.0210.

[144] A. Sboev et al. “Machine Learning Models of Text Categorization by Author Gender Using Topic-independent Features”. en. In: *Procedia Comput. Sci* 101 (2016), pp. 135–142, doi: 10.1016/j.procs.2016.11.017.

[145] J. Schler et al. “Effects of age and gender on blogging”. nl. In: *AAAI Spring Symp. - Tech. Rep* SS-06-03 (2006), pp. 191–197,

[146] S.E. Seker, K. Al-Naami, and L. Khan. “Author attribution on streaming data”. en. In: *Proc. 2013 IEEE 14th Int. Conf. Inf. Reuse Integr. IEEE IRI 2013*. 2013, pp. 497–503, doi: 10.1109/IRI.2013.6642511.

[147] Y. Seroussi, I. Zukerman, and F. Bohnert. “Authorship Attribution with Topic Models”. en. In: *Assoc. Comput. Linguist* 70.8 (2013), pp. 269–310, doi: 10.1162/COLI.

[148] E. Sezerer, O. Polatbilek, and S. Tekir. “A Turkish Dataset for Gender Identification of Twitter Users”. en. In: 2019, pp. 203–207, doi: 10.18653/v1/w19-4023.

[149] S. Shao et al. “An Ensemble of Ensembles Approach to Author Attribution for Internet Relay Chat Forensics”. en. In: *ACM Trans. Manag. Inf. Syst* 11.4 (2020). doi: 10.1145/3409455.

[150] Lloyd S Shapley. “Notes on the n-person game—ii: The value of an n-person game”. In: (1951).

[151] P.K. Singh, K.S. Vivek, and S. Kodimala. “Stylometric analysis of E-mail content for author identification”. en. In: *ACM Int. Conf. Proceeding Ser* (2017). doi: 10.1145/3109761.3109770.

[152] J. Soler-Company and L. Wanner. "On the relevance of syntactic and discourse features for author profiling and identification". en. In: *15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf.* Vol. 2. 2017, pp. 681–687, doi: 10.18653/v1/e17-2108.

[153] J. Soler-Company and L. Wanner. "On the role of syntactic dependencies and discourse relations for author and gender identification". en. In: *Pattern Recognit. Lett* 105 (2018), pp. 87–95, doi: 10.1016/j.patrec.2017.12.006.

[154] Karen Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of documentation* 28.1 (1972), pp. 11–21.

[155] L. Srinivasan and C. Nalini. "An improved framework for authorship identification in online messages". en. In: *Cluster Comput* 22 (2019), pp. 12101–12110, doi: 10.1007/s10586-017-1563-3.

[156] E. Stamatatos. "Author identification: Using text sampling to handle the class imbalance problem". en. In: *Inf. Process. Manag* 44.2 (2008), pp. 790–799, doi: 10.1016/j.ipm.2007.05.012.

[157] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. "Automatic text categorization in Terms of Genre and Author". en. In: *Comput. Linguist* 26.4 (2001). Available: pp. 471–495, url: http://casa.disi.unitn.it/~moschitt/books/Book2005-reduced.pdf.

[158] K. Sundararajan and D.L. Woodard. "What constitutes ' style ' in authorship attribution?" en. In: *27th Int. Conf. Comput. Linguist* (2018), pp. 2814–2822,

[159] X. Tang, S. Liang, and Z. Liu. "Authorship attribution of the golden lotus based on text classification methods". en. In: *ACM Int. Conf. Proceeding Ser* Part F1481 (2019), pp. 69–72, doi: 10.1145/3319921.3319958.

[160] M.J.G. Ucelay. "Profile-based approach for age and gender identification". en. In: *CEUR Workshop Proc.* Vol. 1609. 2016, pp. 864–873,

[161] P. Varela, E. Justino, and L.S. Oliveira. "Selecting syntactic attributes for authorship attribution". en. In: *Proc. Int. Jt. Conf. Neural Networks* (2011), pp. 167–172, doi: 10.1109/IJCNN.2011.6033217.

[162] P. Varela et al. “A computational approach for authorship attribution of literary texts using sintatic features”. en. In: *Proc. Int. Jt. Conf. Neural Networks* (2016-10), pp. 4835–4842, doi: 10.1109/IJCNN.2016.7727835.

[163] P.J. Varela et al. “A Computational Approach for Authorship Attribution on Multiple Languages”. en. In: *Proc. Int. Jt. Conf. Neural Networks* July (2018-07). doi: 10.1109/IJCNN.2018.8489704.

[164] P.J. Varela et al. “Authorship Attribution in Latin Languages using Stylometry”. en. In: *IEEE Lat. Am. Trans* 18.4 (2020), pp. 729–735, doi: 10.1109/TLA.2020.9082216.

[165] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[166] A. Venckauskas et al. “Open class authorship attribution of lithuanian internet comments using one-class classifier”. en. In: *Proc* 11 (2017), pp. 373–382, doi: 10.15439/2017F461.

[167] B. Verhoeven and W. Daelemans. “CLiPS stylometry investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text”. en. In: *Proc* (2014), pp. 3081–3085,

[168] B. Verhoeven, W. Daelemans, and B. Plank. “Twisty: A multilingual Twitter stylometry corpus for gender and personality profiling”. en. In: 2012, pp. 1632–1637,

[169] B. Verhoeven, I. Škrjanec, and S. Pollak. “Gender profiling for slovene Twitter communication: The influence of gender marking, content and style,” BSNLP 2017 - 6th Work”. en. In: *Balto-Slavic Nat. Lang. Process* April (2017), pp. 119–125, doi: 10.18653/v1/w17-1418.

[170] B. Vijayakumar and M.M.M. Fuad. “A new method to identify short-text authors using combinations of machine learning and natural language processing techniques”. fr. In: *Procedia Comput. Sci* 159 (2019), pp. 428–436, doi: 10.1016/j.procs.2019.09.197.

[171] V. Vysotska, O. Kanishcheva, and Y. Hlavcheva. “Authorship Identification of the Scientific Text in Ukrainian with Using the Lingvometry Methods”. en. In: *Int. Sci. Tech. Conf. Comput. Sci. Inf. Technol* 2 (2018), pp. 34–38, doi: 10.1109/STC-CSIT.2018.8526735.

[172] D. Wright. “Using word n-grams to identify authors and idiolects”. en. In: *Int. J. Corpus Linguist* 22.2 (2017), pp. 212–241, doi: 10.1075/ijcl.22.2.03wri.

[173] H. Wu, Z. Zhang, and Q. Wu. “Exploring syntactic and semantic features for authorship attribution”. en. In: *Appl. Soft Comput* 111 (2021), pp. 107815, doi: 10.1016/j.asoc.2021.107815.

[174] yellowbrick. *Yellowbrick: Machine Learning Visualization*. https://www.scikit-yb.org/en/latest/. Accessed: 2023-04-05. 2016.

[175] Y. Z. and J. Zobel. “Effective and Scalable Authorship Attribution Using Function Words”. en. In: *Airs* Lncs 3689 (2005). Available: pp. 174–189, doi: 10.1007/11562382_14.

[176] C. Zhang et al. “Authorship identification from unstructured texts”. en. In: *Knowledge-Based Syst* 66 (2014), pp. 99–111, doi: 10.1016/j.knosys.2014.04.025.

[177] Y. Zhao and J. Zobel. “Effective and Scalable Authorship Attribution Using Function Words”. en. In: *Asia Inf. Retr. Symp* 3689 (2005), pp. 174–189,

[178] A. Zhou, Y. Zhang, and M. Lu. “C-Transformer Model in Chinese Poetry Authorship Attribution”. en. In: *Int. J. Innov. Comput. Inf. Control* 18.3 (2022), pp. 901–916, doi: 10.24507/ijicic.18.03.901.