GENDER-BASED ANALYSIS OF SLOVENE USER-GENERATED CONTENT

Iza Škrjanec

Master Thesis
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Prof. Dr. Nada Lavrač, Jožef Stefan Institute, Ljubljana, Slovenia, and Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
Working Supervisor: Dr. Senja Pollak, Jožef Stefan Institute, Ljubljana, Slovenia

Evaluation Board:
Assoc. Prof. Dr. Marko Robnik-Šikonja, Chair, Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia
Prof. Dr. Dunja Mladenić, Member, Jožef Stefan Institute, Ljubljana, Slovenia, and Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
Asst. Prof. Dr. Ana Zwitter Vitez, Member, Faculty of Humanities, University of Primorska, Koper, Slovenia, and Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Iza Škrjanec

GENDER-BASED ANALYSIS OF SLOVENE USER-GENERATED CONTENT

Master Thesis

ANALIZA SLOVENSKIH SPLETNIH UPORABNIŠKIH VSEBIN Z VIDIKA SPOLA

Magistrsko delo

Supervisor: Prof. Dr. Nada Lavrač

Working Supervisor: Dr. Senja Pollak

Ljubljana, Slovenia, August 2017

Acknowledgments

First and foremost I would like to thank my supervisor Nada Lavrač and my working supervisor Senja Pollak for their support, guidance, and patience during the two years I spent at the IPS. I believe our interdisciplinary collaboration has been most fruitful and can say I have acquired important skills and knowledge.

I would also like to thank the members of the committee, Marko Robnik-Šikonja, Dunja Mladenić, and Ana Zwitter Vitez, for their comments and suggestions. Thank you for helping me improve my research.

I am most grateful to the Janes national research project and Darja Fišer, the project leader, for allowing me to use the Twitter and blog data and for including me in the project tasks.

There are many people who have contributed to this thesis and I would like to express my gratitude to each and every one of you: Matej Martinc for helping me with the technical aspects of machine learning and kindly answering my n+1 emails; Luka Komidar for providing advice for the selection of statistical tests; Ben Verhoeven for his collaboration and the helpful discussions about the results; Vojko Gorjanc for the reflections on gender and language use; Jaka Čibej and Damjan Popič for the pep talks; and Gašper Pesek for the language check-up. Many thanks go to my friends and colleagues.

Finally, I am grateful for the support I received from my parents, my brother, and my sister. And thank you, Timotej, for encouraging me and standing by my side no matter what.

Abstract

Language as a social phenomenon is subject to variation and change. Among the social factors of variation, the gender of speakers has been studied by sociolinguists with regard to the linguistic practices displayed by women and men. Advances in natural language processing have enabled researchers to model linguistic variation in large text corpora using automated approaches.

In this Master Thesis we compare the language of women and men in Slovene user-generated content. User-generated content is a fairly new but increasingly popular domain and presents an important language resource due to its size, production in real time, and the informal nature of communication. In our analysis of gender-related variation, we rely on the approaches of computational stylometry, which aims to generalize the writing style into measurable patterns that can be compared between individual authors or groups of authors. We base the experiments on Slovene blog entries and Twitter messages (*tweets*), which were manually annotated for the authors' gender.

In the first step, we investigate the documents with machine learning techniques. Using a clustering algorithm, we construct topic ontologies of blogs to identify the predominant topics in entries by female and male bloggers. We then present automated methods that explore whether gender-related linguistic variation can provide predictive information to differentiate between women and men automatically. For this, we build two types of gender prediction models: a rule-based and a statistical classifier. The rule-based classification model takes into account the use of referential gender in verb phrases. It provides a baseline accuracy for more complex and time-consuming statistical models, for which we experiment with different features and learning algorithms. The best performing statistical models on tweets and blogs are further analyzed for the features ranked as most informative for female and male authors. Moreover, we present cross-genre experiments where the gender classifier was trained on one of the text genres and tested on the other.

In the second step, we analyze the linguistic choices of female and male authors that have less to do with topic and more to do with the writing style. Specifically, we contrast the use of words typical of a writing style (e.g., profane language, emoticons, hedges and intensifiers) by applying the Mann-Whitney test for testing the statistical significance of the differences, and the squared Pearson correlation coefficient as a measure of effect size.

The results of our experiments with gender prediction models show that rule-based classifiers work well, but noisy text might worsen their performance. For statistical models, the combination of Support Vector Machine as the learning algorithm and word unigrams as features achieves the best results on both tweets and blogs. The statistical analysis of stylistic variation further shows that the difference between female and male authors in the use of some discourses is statistically significant, though the effect size of gender on this variation is small. The results of cross-genre experiments with statistical models suggest that our tweet-trained model could be used for predicting gender in other Slovene user-generated content.

Povzetek

Jezik je družbeni pojav in je kot tak podvržen variaciji in spremembam. Med družbenimi dejavniki variacije sociolingvisti preučujejo tudi spol govorcev glede na jezikovne prakse, ki jih uporabljajo ženske in moški. Napredek v obdelavi naravnega jezika raziskovalcem omogoča, da modelirajo jezikovno variacijo v obsežnih besedilnih korpusih in z uporabo avtomatskih pristopov.

V tem magistrskem delu analiziramo jezik žensk in moških v slovenskih spletnih uporabniških vsebinah. Spletne uporabniške vsebine so dokaj nova, vendar vedno bolj priljubljena domena, zaradi obsežnosti, objavljanja v realnem času in neformalnega sporazumevanja pa predstavljajo pomemben jezikovni vir. V analizi jezikovne variacije z vidika spola se opiramo na pristope računalniške stilometrije, katere namen je posplošitev sloga pisanja na merljive vzorce, ki jih lahko nato primerjamo glede na posamezne avtorje ali skupine avtorjev. Naša analiza je osnovana na slovenskih blogovskih zapisih in sporočilih z družbenega omrežja Twitter (*tvitih*), ki so ročno označeni s podatkom o avtorjevem spolu.

V prvem delu naloge besedila analiziramo z metodami strojnega učenja. Z uporabo algoritma gručenja izdelamo tematske ontologije, da prepoznamo prevladujoče teme v blogovskih zapisih moških in žensk. Nato predstavimo avtomatske metode, s katerimi raziščemo, ali lahko s pomočjo razlik v jezikovni rabi avtomatsko prepoznamo spol avtorja besedil. V ta namen zgradimo dva tipa napovednih modelov za spol: model s klasifikacijskimi pravili in statistični model. Model s klasifikacijskimi pravili za razlikovanje med avtorji upošteva rabo referenčnega spola v glagolskih zvezah. Točnost modela s pravili služi kot osnova za primerjavo s kompleksnejšimi in časovno potratnejšimi statističnimi modeli, za katere smo eksperimentirali z različnimi značilkami in algoritmi strojnega učenja. Najboljši napovedni model za vsakega od besedilnih žanrov nato analiziramo na podlagi značilk, ocenjenih kot najbolj informativne za ženski ali moški razred. Poleg tega predstavimo čezžanrske poskuse, pri katerih napovedni model naučimo na enem žanru in testiramo na drugem.

V drugem delu analiziramo jezik moških in žensk glede na razlike, ki niso vezane na temo besedila, pač pa na slog pisanja. Natančneje avtorje primerjamo glede na rabo besed, značilnih za določen slog (npr. kletvice, emotikoni, omejevalci in ojačevalci diskurza), z uporabo Mann-Whitneyjevega testa za testiranje statistične pomembnosti razlik. Za merjenje velikosti učinka uporabimo kvadrirani Pearsonov korelacijski koeficient.

Rezultati poskusov so pokazali, da napovedni modeli s klasifikacijskimi pravili uspešno napovedujejo spol avtorja, vendar lahko šum v besedilih poslabša njihovo točnost. Statistični model, ki deluje na podlagi besednih unigramov kot značilk in metode podpornih vektorjev kot algoritma, se je izkazal kot najtočnejši tako na tvitih kot blogih. Statistična analiza slogovne variacije je pokazala, da so razlike v rabi nekaterih diskurzov statistično pomembne, vendar je velikost učinka spola na te razlike majhna. Glede na rezultate čezžanrskih eksperimentov lahko sklepamo, da je naš napovedni model, naučen za tvitih, primeren za napovedovanje spola v drugih žanrih slovenskih spletnih uporabniških vsebin.

Contents

List of Figures

List of Tables

Abbreviations

AP ... Author Profiling
API ... Application Programming Interface
LR ... Logistic Regression
NB ... Naïve Bayes
NLP ... Natural Language Processing
SVM ... Support Vector Machine
UGC ... User-Generated Content

Chapter 1

Introduction

The rise of digital media and social networks has strongly influenced the way we communicate and perceive communication. Luckily, the advancement in language technologies has brought methods and tools for collecting, processing, and analyzing the trends in digital interaction. Social media and the user-generated content published in social media enable users to express opinions and interact with others. Thus, the users actively participate in social processes and display their identities via their digital profiles. This makes user-generated content a relevant resource for studies in human behavior.

This Master Thesis is concerned with gender-related language variation in Slovene user-generated content (UGC). In this chapter, we provide an introduction to the subject of gender and language. We describe the motivation for comparing the language use of women and men in Slovene UGC. This chapter also states the hypotheses and goals of the thesis and presents its organization.

1.1 Gender and Language

Language is a social phenomenon and variation is inherent to its social nature (Nguyen, Dogruoz, Rose, & de Jong, 2016). With regard to gender and identity, language can be seen as central to the construction and reproduction of gendered selves, social structures, and relations (Gergen & Shotter, 1989). For scholars in language, the differences, variations, and contrasts in language reflect greater differences in society, such as socio-economic and political differences between individuals and groups. The relationship between gender and language constitutes an important question in sociolinguistics, a branch of linguistics that studies the social aspects of language, focusing on how members of groups and communities behave linguistically given their social variables (e.g., region, age, gender, ethnicity, level of education, etc.).

Traditionally, research on gender and language has been divided into two strands: the study of language form and the study of language function (Speer, 2005). Language form relates to how gender is represented in language: how particular words like pronouns, nouns, verbs or adjectives carry grammatical gender. From a more critical perspective, the studies of language form (see Goddard and Patterson (2000)) point also to sexist language in job titles or generic masculine pronouns. In this thesis, especially in Chapters 5 and 6, we take special notice of "referential gender", which is a matter of whom a certain linguistic form (e.g., personal pronouns or noun) refers to in a given context (Motschenbacher, 2010). Depending on the language in question, referential gender can be used as a female, male, or gender-indefinite reference (Weikert & Motschenbacher, 2015). In our analysis, we study to what extent the use of referential gender in first person relates to the gender of text authors in our corpus of user-generated content and whether the use of referential gender

can be indicative of the author's gender.

It this approach, we assume that the referential gender in self-referencing contexts implies the gender of the author, i.e. self-references in the feminine gender imply a female author, while masculine self-references point to a male author. However, it should be noted that the referential gender a person uses in self-references and this person's gender identity may not overlap (for transgender identities see Koletnik, Grm, and Gramc (2015), for the use of referential gender in non-heterosexual communities see Motschenbacher (2010)).

The study of how language is used by women and men investigates the language function, whereby it is hypothesized that the variation in language use between women and men originates in the way they are positioned in society and the roles they are expected to perform. Aside from the demographic variables of the speaker (gender, age, etc.), speaker goals and the audience are taken into consideration (Nguyen et al., 2016), whereby spoken interaction is usually analyzed. Typically, the goal of these studies is not to find as many differences between female and male speakers, but rather the differences that can tell us something about gender roles, hence the focus is mainly placed on the possible reasons behind the linguistic variation.

The communication styles of women and men have been interpreted by interactional sociolinguistics in three general frameworks in close connection to feminist studies. The work of Lakoff (1975) is viewed as pioneering for the "deficit" framework, where the differences between the genders and language are interpreted as a consequence of social restrictions concerning especially women and their language use. Shortly after, the "dominance" framework was developed – it is largely represented by Spender (1980), who claimed men are in a position of power and thus dominate and control the language form and function. The perspective was shifted by Tannen (1990) with the "difference" framework and the notion that women and men have contrasting communication expectations, leading to conflicts between them. Tannen (1990) explains that women perceive interaction a "rapport" and seek intimacy and connection, while men tend to prefer "report" talk to exchange impersonal information and maintain their status in the hierarchy. The author refers to these gender-based linguistic variations as "genderlects". All the three frameworks considered have been discussed and criticized in several publications (Speer, 2005; Goddard & Patterson, 2000), but the drawbacks of the frameworks will not be summarized in this thesis. Instead, we want to point out that while these interpretational frameworks contributed greatly to understanding and studying the communication of women and men, their claims are not supported by quantitative analyses of spoken interaction, but rather with manually collected material with a potential bias in choosing the examples that provide evidence for the researcher's claim (Widdowson, 2004).

The development of corpora and tools for their processing has brought to the field of gender and language the possibility of analyzing large collections of documents produced by women and men. As a result, descriptive analysis can be performed on large amounts of data leading to statistically relevant conclusions. Moreover, it can be enriched with the results of predictive tasks where the goal is to identify the gender of the authors based on their language use. Large amounts of gender-annotated data have a two-fold contribution: 1) they provide a quantitative complement to descriptive studies; 2) they enable further research in studies that involve large datasets with author gender annotation. Once a gender classifier is trained and tested, it can be applied to unknown instances or documents to predict the gender of their author. Corpora with author gender metadata can be used for further research, for example predicting the hierarchy between participants in e-mail communication (Prabhakaran, Reid, & Rambow, 2014), improving the detection of unwanted behavior displayed by cyberbullies or sexual predators (Dadvar & de Jong, 2012; Rangel, Rosso, Koppel, Stamatatos, & Inches, 2013), or observing the share of female

and male participants in a particular domain, e.g. academic writing (Vogel & Jurafsky, 2012).

Moreover, determining the profile of an anonymous text is applicable in text forensics (Chaski, 2001), e.g. in criminology. The field of marketing also benefits from author profiling, as it is in a company's interest to collect user feedback for their products and services and identify the most or least satisfied customers and their profiles.

1.2 Motivation

In this thesis, we study the gender-related linguistic variation in Slovene user-generated content (UGC), whereby we employ automated methods and statistical analysis upon large datasets of Slovene Twitter messages (*tweets*) and blog entries. We are interested in the differences between female and male authors with regard to topic, style, and referential grammatical gender, and whether these variations can be used to automatically identify the gender of the author.

The studies of the relationship between the document and the author's demographic and psychological profile usually involve collecting and processing data that is potentially interesting not only for natural language processing (NLP) but relevant and valuable as a resource for humanities and social science research. This is why resources, such as corpora with author metadata, should be studied with complementing methods. As Daelemans (2013) emphasizes, one of the problems of studies in computational stylometry is reflected in the lack of explanation. Namely, it occurs often that corpora are compiled with little regard to the balance of author gender, age, document topic or genre, but are nevertheless used for author profiling. The author profiling (AP) studies usually rely on quantitative performance evaluation; however, an explanation of the chosen features and an interpretation of results is crucial if we wish to make inferences about how the demographic and psychological identity of a person interacts with her or his writing style and how this can be adopted by machine learning models.

While extensive and multiple explanations for the variation between female and male speakers are provided in sociolinguistics, these studies strongly focus on spoken language (Nguyen et al., 2016). Their data is usually collected manually and the interpretation is based on either qualitative analysis in context or simple statistical comparison of frequencies. This has motivated us to try to employ a unified approach to the studied problem and include automated as well as qualitative methods for the analysis of linguistic variation between women and men.

In the thesis we analyze documents in the Slovene language that has so far been studied from a gender perspective in relation to the denomination of professions (Kunst Gnamuš, 1995) and gender representation in collocations that occur in the Slovene reference corpus (Gorjanc, 2007). Slovene user-generated content has been analyzed with regard to the author profile: Osrajnik, Fišer, and Popič (2015) conducted a corpus analysis of emoticons and expressive punctuation and found that female Twitter users generally include both categories of expressive language more often than male Twitter users. Ljubešić and Fišer (2016) built a classification model that distinguishes between corporate and personal Twitter accounts based on a variety of language-dependent as well as language-independent features.

1.3 Hypotheses and Goals

The main purpose of this thesis is to explore how gender-related linguistic variation can provide predictive information about the gender of a text author and how it can be used for

building gender classifiers for Twitter messages (*tweets*) and blog entries in Slovene. The gender-related linguistic variation is observed on three levels: referential gender, content, and style. For this, we employ the approaches used in author profiling and sociolinguistics. First, two types of gender classifiers are built and tested on tweets and blog entries. The use of referential gender is taken into account by a rule-based model with simple and manually constructed classification rules, while we experiment with features and machine learning algorithms with more time-consuming statistical models. We compare the performance of both classifier types in Slovene user-generated content. Additionally, we conduct experiments that include both genres of Slovene UGC (blog entries and tweets) by training a statistical model on one genre and testing it on the other.

On the other hand, we are interested in descriptive approaches to linguistic variation. We explore how the author's gender is reflected in the topical tendencies of documents, as well as stylistic choices and discourse (the use of profane language, emoticons, communication verbs and other).

In the thesis, the following hypotheses have been tested:

1. We hypothesize that for the prediction of author gender in Slovene, we can use the features related to textual content, referential gender, and style.

2. Our hypothesis is that features from predictive models can contribute to the sociolinguistic understanding of gender differences in writing.

3. We expect to find statistically significant differences between the language use of female and male authors in tweets and blogs when taking into account only the use of stylistic linguistic choices (such as profane language, intensifiers and hedges, emoticons, and others), which are neither topic-bound nor related to the referential gender.

4. Among the many factors that influence language use, the communication medium plays an important role. We hypothesize that models for gender prediction trained on a single UGC genre do no support generalization that goes beyond that genre. In the thesis we experiment with given datasets to test the models' reliance on a single UGC genre by testing the models on datasets of a different UGC genre.

The thesis aims to achieve the following goals:

1. Provide a critical overview of related work in the studies of gender-related linguistic variation in user-generated content.

2. Use machine learning to explore the differences in language use in Slovene UGC. More specifically, employ topic modelling to detect the predominant topics in entries by female and male bloggers. Furthermore, build two types of classification models for gender prediction (rule-based and statistical) and compare their performance.

3. Analyze the most informative features per class as ranked by the best performing statistical model to learn about the most differentiating features between female and male language use.

4. Explore sociolinguistic hypotheses on gender-based language variation and validate them on data from Slovene Twitter messages and blog entries.

1.4 Scientific Contributions

The thesis contributes to the fields of gender, language and author profiling. These contributions are as follows:

1. Construction of new rule-based and statistical models for gender prediction of Slovene Twitter messages and blog entries. Comparison of models for suggesting which method to use when faced with a gender prediction task in Slovene.

2. Analysis of the most informative features per class of the best performing statistical models for Twitter messages and blog entries.

3. Investigation of gender-related language variation beyond the topic and genre of the document by applying statistical analysis of writing style markers, as well as cross-genre gender prediction.

The results of the thesis and related research in gender prediction have been published in the following conference and workshop papers:

- Škrjanec, I. & Pollak, S. (2016). Topic ontologies of the Slovene blogosphere: A gender perspective. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities* (pp. 62–65). Ljubljana, Slovenia: Academic Publishing Division of the Faculty of Arts of the University of Ljubljana.

- Verhoeven, B., Škrjanec, I. & Pollak, S. (2017). Gender profiling for Slovene Twitter communication: The influence of gender marking, content and style. In *Proceedings of the EACL workshop, The 6th Workshop on Balto-Slavic Natural Language Processing* (pp. 119–125). Valencia, Spain: The Association for Computational Linguistics.

- Martinc, M., Škrjanec, I., Zupan, K. & Pollak, S. (2017). PAN 2017: Author Profiling – Gender and Language Variety Prediction. In L. Cappellato, N. Ferro, L. Goeuriot, & T. Mandl (Eds.), *CLEF 2017 Labs Working Notes*. Dublin, Ireland: CLEF and CEUR-WS.org.

1.5 Organization of the Thesis

The thesis is structured as follows. Chapter 2 provides an overview of related work from the fields of computational stylometry and author profiling. For the latter, we describe the state-of-the-art approaches for gender prediction in user-generated content for English and Balto-Slavic languages. We also present existing linguistic studies of Slovene user-generated content with regard to author profiling and gender.

Chapter 3 describes the Slovene Twitter and blog corpora that we used in our experiments. The methodology is explained in Chapter 4, where we first present the document representation for machine learning. Next, we present the methodology of gender prediction using statistical models and rule-based models. Then we present the methods and tools for topic modeling of blog entries. Finally, we describe the statistical tests applied to the analysis of writing styles in tweets and blog entries.

Chapters 5 and 6 present the experimental setting and results of gender prediction in tweets and blog entries, respectively. Chapter 6 also includes the setting and results of topic ontology construction. Furthermore, both chapters present the statistical analysis of writing styles, as well as its visualization.

Chapter 7 presents a comparison of gender prediction results from Chapters 5 and 6. We also present the setting and results of cross-genre experiments for gender prediction.

In Chapter 8, we conclude the thesis, list the lessons learned and propose suggestions for further work in gender prediction and analysis of gender-related linguistic variation.

Chapter 2

Background and Related Work

This chapter addresses the fields that study the relation between gender and language. We describe the methods and related work for predicting the demographic, psychological and health condition profiles in computational stylometry. The main focus is placed on gender-based analyses of user-generated content, whereby we describe the state-of-the-art approaches to gender prediction in user-generated content. We focus on the winners of the PAN shared tasks for author profiling. We also present related work in Balto-Slavic languages due to their similarity to Slovene, especially because both include referential gender markings. Finally, we describe the related tasks which have been addressed on Slovene user-generated content so far.

2.1 Computational Stylometry

Computational stylometry is a research field that focuses on the writing style in a textual document and how this style can be processed computationally. Its origin lies in literary studies of authorship (Stamatatos, 2009; Holmes, 1998). The basic premise of studies in computational stylometry is that style in a document can essentially be measured by generalizing it into patterns in order to compare the styles of various documents or to contrast authors by their styles. The methodology of computational stylometry involves scanning the text document to observe stylistic characteristics on various levels of language (morphological, grammatical, lexical, syntactical, and semantic), structural patterns (punctuation, use of capital letters), and text organization (number of sentences, word length). These characteristics are measured to construct a writing style of an author or group. Koppel, Schler, and Argamon (2008) have identified three di□erent tasks or problems that provide answers to the following questions: 1) what the text author is like, 2) which member of a closed group authored the text, 3) whether this member truly authored the text.

Problem 1 is called the “open case” (Rangel et al., 2013) and is generally referred to as *author profiling* (AP). Let us presume that a document or a collection of documents by the same author is provided and the identity of the author is unknown. If there exists no closed group of candidates, i.e. no potential authors of this document, it is still possible to identify the author’s demographic, psychological and health condition profile. In author profiling studies, the author is assigned to a particular population group based on her or his document(s), or rather based on the similarity of her or his documents and the documents of the assigned group.

In Problem 2 (the “closed case”), exists a finite group of potential authors or candidates and the task is to determine which one produced the given document. This problem is known as *authorship attribution* (Rangel et al., 2013), as we ascribe the authorship of a document to a single candidate. This usually includes the discovery of characteristics

typical of one author, but much less typical of other candidates.

We refer to Problem 3 as the task of *authorship verification*. The aim of this task is to determine whether or not a document was produced by a particular author, which means there is only one candidate. The process of verification relies on the similarity between the document in question and the author's other documents, which should be greater than the similarity with the documents produced by other authors.

In this thesis, we employ the methods of author profiling (Problem 1) to predict the gender profile of Slovene Twitter and blog users based on the documents they produced. Because of this, the next sections focus on related work and the state-of-the-art approaches to predicting the author gender from text.

2.2 Gender Profiling

Gender profiling or the automated prediction of a person's gender is an interesting problem that has been on the rise since the early 2000s and has established itself an important task in computational stylometry and natural language processing (NLP). The workflow of gender profiling involves large textual datasets with manually labeled author gender. Based on this dataset, we make generalizations about the relationship between the gender of the author and the language used in the document. This is typically achieved with techniques of machine learning, where a predictive model for gender is constructed.

Many researchers have addressed gender profiling based on text documents, so various approaches and algorithms are described in the literature. An overview of studies in automated gender prediction shows that the task has been applied to several text genres, such as fiction, non-fiction, and emails, while user-generated content (UGC) remains the central source of data for gender profiling: Twitter, Facebook, blogs, forums, YouTube etc. Because domains di□er in characteristics, preprocessing requirements and author behavior, and because this thesis is concerned with Slovene UGC, we focus on state-of-the-art methods for gender prediction in UGC.

As Daelemans (2013) points out, the question of feature construction and feature selection is of crucial importance in the field of author profiling, as successful features should be genre- and topic-independent, and resilient against individual style change and conscious manipulation. Stamatatos (2009) provides an overview of feature types for tasks in authorship attribution, but they are frequently applied in author profiling as well: lexical (word n-grams, word frequencies, errors, and vocabulary richness), character (character n-grams, letters or digits), syntactic (parts-of-speech (POS), chunks, and sentence or phrase structure), semantic (synonyms and syntactic dependencies), and application-based.

Koppel, Argamon, and Shimoni (2002) built a gender prediction model using the Balanced Winnow algorithm on the fiction and non-fiction data from the British National Corpus. To avoid the topic bias, function words and POS-tags were applied as features. The classification models performed with an accuracy of 79.5% ($\pm$1.1%) for fiction, and 82.6 ($\pm$ 0.99) for non-fiction, whereby both datasets were gender-balanced. Function words were proven as useful features by Schler, Koppel, Argamon, and Pennebaker (2006), who performed gender and age classification on English blogs. They applied the multi-class real Winnow algorithm on a balanced dataset and ran experiments with stylistic features (function words, selected POS, UGC-specific neologisms) and content features (1,000 word unigrams with the highest information gain). The authors report that while models based solely on stylistic features perform better than those with content features, the best model includes both feature types and performs with an accuracy of 80.1%. Content-wise, they found that female bloggers focus more on their personal lives, while male bloggers write more about politics, technology and money.

Mukherjee and Liu (2010) improved the existing methods for the gender classification of blog authors and proposed a new feature selection algorithm that uses an ensemble of feature selection criteria. They used a combination of four features and one newly constructed feature. The F-measure[1] feature is a distinction between the contextuality and formality of a document and is based on the frequency of the POS usage. A lower F-measure score indicates contextuality (implicitness) of the text, marked by a greater use of pronouns, verbs, adverbs, and interjections. In contrast, a higher F-measure value means that the text is more explicit (indicated formality), as it contains more nouns, adjectives, prepositions, and articles. The second feature used by Mukherjee and Liu (2010) is what they call "stylistic features", which are actually words and blog-specific words, i.e. abbreviations (e.g., *lol*), words mimicking spoken discourse (e.g., *hmmm*) or symbols typical of UGC (e.g., emoticons). Their next feature is a list of word endings that indicate the use of emotionally intense adverbs and adjectives, which are often ascribed to female language. They also used a word list of twenty factors from Argamon, Koppel, Pennebaker, and Schler (2007); factors are groups of related words that tend to occur in similar documents and are mostly topic-related (*Family*, *Work*, *Location*, *Poetic*, *Swearing*, *Romance*, etc.). Mukherjee and Liu (2010) added three new "word classes" (of mostly adverbs and adjectives) according to their connotation, which is either emotional (e.g, *careful*, *puzzled*), positive (e.g., *cool*, *wow*) or negative (e.g., *stupid*, *hopeless*). Their newly constructed features are POS sequence patterns which differ from normal POS-tags and n-grams in that the sequences are not of fixed lengths (a maximum of 7 consecutive tags). All the mined POS sequence patterns were used as features. They applied their feature selection algorithm to reduce the vector space and carried out some experiments using different classification (Naïve Bayes, SVM) and regression (SVM regression) methods, with different feature weighting (boolean and term frequency). They determined that the highest accuracy (88.56%) was achieved when only the POS sequence feature was used together with their feature selection algorithm.

Rao, Yarowsky, Shreevats, and Gupta (2010) carried out an interesting comparison of three SVM models for gender prediction of Twitter users. Based on sociolinguistic research of spoken discourse (Macaulay, 2005) they constructed a list of what they call sociolinguistic features, as the prosodic cues from spoken discourse are absent in Twitter. The list included smileys, ellipses, possessive bigrams (*my_XX*, *our_XX*), references to self, markers of agreement, affection, excitement etc. Word uni- and bigrams were also used as features in the second SVM model. The third model was stacked and its features were predictions from the n-gram feature and sociolinguistic models along with their prediction weights. The sociolinguistic model performed better (71.76%) than the n-gram model (68.70%), while the stacked model performed best (72.22%). They also determined that emoticons, ellipses, character repetition, repeated exclamation, puzzled punctuation, and the abbreviation *OMG* ("oh my god") were more typical of tweets by women. The authors also report that the sociolinguistic model improved when bigrams with the possessive pronoun "my" from the bigram model were added to it.

The above reported performances of statistical approaches to gender identification were achieved by models trained and tested on the same UGC genre – either blog entries or tweets. Some studies (Sarawgi, Gajulapalli, & Choi, 2011; Rangel et al., 2013) indicate a genre bias and show that the classifiers trained on one UGC genre and tested on another perform worse than classifiers trained and tested on the same genre, suggesting that the existing achievements of gender identification are limited to each UGC genre individually. Sarawgi et al. (2011) also point out a topic bias in gender, with the help of which the above listed models perform with a high accuracy. In order to find stronger evidence supporting gender-specific styles of language and to achieve gender identification model robustness

[1] Note that here F-measure does not refer to the F-score or F-measure used for performance evaluation.

against change in topic and genre, Sarawgi et al. (2011) define their gender identification task as cross-topic and cross-genre. Their dataset consisted of blog entries from seven distinctive topics (education, travel, spirituality, history, book reviews, entertainment, and politics) and scientific papers from the NLP community. Several experiments were conducted using three types of statistical language models: probabilistic context-free grammar that learns deep long-distance syntactic patterns; token-level language models that learn shallow lexico-semantic patterns, and character-level models that learn morphological patterns. Finally, they also used the bag-of-words (BOW) representation with a maximum entropy classifier. On topic-balanced blog data, the character-level model performed best (71.3%). In the cross-topic experiment (trained on 6 topics, tested on the remaining one topic), the character-level model performed best (68.3%). In the next setting, the models were evaluated across topic and genre, as they were trained on the blog dataset and tested on scientific papers; the best performing model was the BOW approach with the accuracy of 61.5%, while the character-level model achieved 58.5%. When the cross-topic approach was applied to scientific papers, which use formal language and do not give lexical or topical cues, the best performing models were the probabilistic context-free grammar and the character-level model (both 76.0%), suggesting that deep syntactic patterns play a much greater role in detecting gender-specific styles.

While the above mentioned related studies performed gender prediction for documents in one language, important advances have been made on preprocessing and predicting the gender in multilingual corpora. Verhoeven, Daelemans, and Plank (2016) present the TwiSty corpus, which is a collection of Dutch, German, Italian, French, and Portuguese tweets, containing gender and personality annotation of the tweet authors. For each language, a gender and personality classification model was built comprising n-gram features on both word (of length 1 and 2) and character (of length 3 and 4) level. The proposed methodology was applied to the corpus of Slovene tweets (see Section 2.5).

The TwiSty corpus was used by Ljubešić, Fišer, and Erjavec (2017) for testing whether an SVM model with language-independent features (such as the share of emojis, punctuation, background color, or posting frequency) outperforms gender prediction models that are based on BOW representations (such as n-grams). They found that in most cases where the training and testing language di□er, the model with language-independent features performs better than the BOW models. Furthermore, they report on features that are associated with the female class across the six languages included: the percentage of emojis, the mean retweet count, the reddish background color, a higher number of tweets per day, and the number of favorites (tweets marked by the user with a heart symbol). In contrast, the features related to male users are the following: percentage of tweets with URLs, percentage of tweets sent from an iOS device, percentage of emoticons, question marks, and hashtags, as well as location sharing.

Aside from n-grams and other data-driven features pre-prepared (a priori) lexica have also been applied to the study of gender-related variation. One such system involving predefined categories is Linguistic Inquiry and Word Count (LIWC) (Tausczik & Pennebaker, 2010), which performs extensive linguistic analysis of input documents. It is comprised of pre-prepared dictionaries for 74 word categories, which are either more grammatical (function words, references to self, etc.) or topical (vocabulary on particular emotions, work, leisure, etc.)[2].

The LIWC system was used for gender-related linguistic analysis in two interesting studies. Newman, Groom, Handelman, and Pennebaker (2008) combined several gender-annotated datasets of mostly written, but also spoken samples from other studies. The multivariate analysis of variance (MANOVA) has shown that women use more pronouns, social

[2]For full list, see Tausczik and Pennebaker (2010)

words, references to psychological processes, verbs, references to home, and positive as well as negative emotions. In contrast, men outdid women in the following: word length, use of numbers, articles, prepositions, swear words and references to current concerns. Their findings have largely confirmed the results of studies that used smaller datasets. Similarly, Schwartz et al. (2013) found a correlation between the messages of female Facebook users and emotional words, references to self, and psychological and social processes, while male Facebook users employed more swear words and object references in their messages.

Gender is a complex social phenomenon not limited to binarism (Butler, 1990), which has been explored in Dutch tweets by Nguyen et al. (2014). They compared the results of automated gender prediction with gender labels given by human annotators and found that Twitter users express their gender in varying degrees and in di□erent ways. Namely, Twitter users who use language that deviates from what is expected of their biological sex are labelled incorrectly by human annotators as well as predictive models. The authors emphasize that models training on a binary class data (female, male) result in a stereotypical predictive model that works well for most users, but also o□ers a simplistic view on linguistic variation. Thus, Nguyen et al. (2014) suggest computational approaches could enable complex modelling of variation not only between, but also within speakers, especially in the ever increasing data on social media.

A similar exploratory analysis was conducted by Bamman, Eisenstein, and Schnoebelen (2014) who performed clustering of Twitter users based on 10,000 most frequent words and found that users with similar styles and topical interests are clustered together. They also trained a Logistic Regression gender prediction model using the same features as for clustering. A closer observation of incorrectly classified individuals shows that their social network comprises significantly fewer same-gender contacts, which suggests that same-gender language markers are correlated with homophily.

2.3 The PAN Shared Tasks on Author Profiling

The author profiling community has organized a series of scientific events and shared tasks on digital text forensic called PAN (Uncovering Plagiarism, Authorship, and Social Software Misuse)[3]. One of the tasks involves author profiling (AP). Since they called the first task in 2013, they have addressed age and gender profiling together with personality or language variety identification for several languages and genres of UGC. The participants of PAN AP 2013 (Rangel et al., 2013) had to predict the age and gender of English and Spanish Netlog users. Meina et al. (2013) achieved the highest overall accuracy for English by using a random forest classifier and a set of various features (number of conversations and paragraphs; POS-tags, POS-sequences, readability measure, number of errors, emoticons, abbreviations, bad words, emotion words, topic-specific features, and an n-gram model). Santosh, Bansal, Shekhar, and Varma (2013) predicted the profile of Spanish users most successfully by applying the most di□erentiating n-grams of the female and male class, punctuation-related features, POS-tags and, topic modelling to a decision tree algorithm.

The PAN AP for 2014 (Rangel et al., 2014) again addressed gender and age prediction on English and Spanish tweets and social media with the addition of English hotel reviews. The most successful gender identification model for English and Spanish was built using a LibLinear classifier (López-Monroy, Montes-y-Gómez, Escalante, & Villaseñor-Pineda, 2014). The winning team used second order attributes, but also considered the information among documents belonging to the same class, i.e., the same profile (females, males). The approach found more specific subgroups of authors (male employees, female teenagers, etc.). By using intra-class relationships inside target profiles, they automatically generated few,

[3] PAN: http://pan.webis.de/index.html

but more detailed attributes, which improved the classification rates, achieving 64.76% accuracy.

In the PAN 2015 author profiling task (Rangel et al., 2015), the participants had to identify the gender, age, and five personality traits (extroverted, stable, agreeable, conscientious, and open) of Twitter users for English, Dutch, Italian, and Spanish. The best achieving team (Álvarez-Carmona, López-Monroy, Montes-y-Gómez, Villaseñor-Pineda, & Escalante, 2015) built the most discriminant and descriptive features using second order attributes and latent semantic analysis techniques jointly in a LibLinear classifier.

The task of PAN 2016 (Rangel et al., 2016) for author profiling involved age and gender prediction in a cross-genre setting, as the training set was comprised of English, Spanish, and Dutch tweets. The early bird set for English and Spanish was collected from social media, and the test set from blogs. The early bird and test set for Dutch was collected from reviews. The best joint performance on the test set for gender and age profiling for all three languages was achieved by Busger op Vollenbroek et al. (2016), who employed a combination of stylistic features (function words, POS-tags, emoticons, punctuation symbols) along with second order representation in an SVM model.

The task for PAN author profiling for 2017 (Rangel, Rosso, Potthast, & Stein, 2017) addressed gender and language variety prediction based on tweets in English, Spanish, Portuguese, and Arabic. Our contribution on the shared task is presented in the paper by Martinc, Škrjanec, Zupan, and Pollak (2017), where the majority of the work was completed by the first author. The task was to build a classification model for predicting the gender of the author (female or male) and the language variety of the tweet. The following languages (and their varieties) were included in the dataset: English (Canadian, Irish, North American, Australian, New Zealand, and British), Spanish (Argentine, Colombian, Venezuelan, European Spanish, Chilean, Mexican, and Peruvian), Portuguese (Brazilian and European Portuguese), and Arabic (Egyptian, Maghrebi, Gulf, and Levantine). The preprocessing of tweets involved several steps, some of which were language-specific, (such as tweet reversal for the Arabic dataset, and removal of tweets with a large number of errors detected by the spell checker).

Other preprocessing steps were dependent on feature construction: POS-tagging and removal of punctuation and stop words were used for English, Spanish, and Portuguese. All the datasets underwent the substitution of hashtags, user mentions, and URLs with constant strings (e.g., "@USERMENTION" for every user mention). Next, all tweets authored by the same user were concatenated into one document. A number of features were used for training and developing the classification model and they can be divided into three groups: 1) features related to the writing style and grammar, 2) sentiment-based features, 3) vocabulary. The stylistic and grammatical features included punctuation trigrams, suffix character tetragrams, POS trigrams, character flooding, and word lists with spelling or vocabulary specific to a language variety (used only for data in English). The vocabulary features were word uni- and bigrams, as well as word bound character tetragrams. We counted the emojis used by each author and added a feature with document sentiment based on the sum of emoji sentiment. Several classification models and ensemble classifiers were tested. For the final evaluation, the Logistic Regression algorithm was used.

Our model performed with a joint classification accuracy (see Rangel et al. (2017)) of 82.85% for gender and language variety identification on the entire test set (involving all four languages) and was scored as second best in the global ranking. Among all participants, our approach was the most successful for predicting author gender in Arabic tweets.

2.4 Gender Profiling in Balto-Slavic Languages

While the task of gender prediction and automated approaches to studying gender poses a new problem for Slovene, related studies in other Balto-Slavic languages can already be found. In particular, research of Lithuanian and Russian documents focuses on the relation between the author gender and speech or writing style; however, no work on user-generated content has been published in English for an international readership.

Kapočiūtė-Dzikienė, Šarkutė, and Utka (2014) used a balanced corpus of parliamentary speech transcripts in Lithuanian to perform gender prediction. They tested several grammatical and lexical features, such as character and word n-grams, POS-tags, and compounds consisting of token or lemma and POS-tag or extended POS-tags, which included information on morphological categories. In their experiments, they compared the performance of Naïve Bayes and Support Vector Machine (SVM), and found that the highest classification accuracy (74.61%) was achieved by SVM when used over lemmatized word bigrams.

The same corpus of Lithuanian parliamentary speeches was used to test whether female and male legislators can be distinguished based on the use of most frequent words. For this, Mandravickaitė and Krilaviius (2017) performed unsupervised learning (hierarchical clustering), whereby the distances were computed according to the frequency of over 7,000 most frequent words. The authors successfully constructed two separate clusters for each gender and thus showed that variation exists between female and male speakers based on the use of most frequent words, which are in turn mostly function words.

Kapočiūtė-Dzikienė, Utka, and Šarkutė (2015) predicted the author gender in Lithuanian literary texts. When the SVM algorithm was applied to the feature vector comprising of lemma and POS-tag compounds, the model predicted the gender of 89.0% authors correctly. However, the performance with word lemmas as features did not significantly di□er from this optimal setting.

The studies in author profiling for Russian draws from the RusPersonality corpus, which includes students' essays, picture descriptions, and personal letters, as well as a range of author metadata (gender, age, native language, personality traits from psychometric tests, etc.). Sboev, Litvinova, Gudovskikh, Rybka, and Moloshnikov (2016) built a model with a convolutional neural network that predicts the author gender with an accuracy of 86% ($\pm 3\%$) based on grammatical features (POS-tags, noun case, verb form, gender, and number). They also experimented with other topic-independent features (syntactic relations, number of punctuation marks, emotive dictionaries) and algorithms (Gradient Boosting Classifier, Adaptive Boosting Classifier, ExtraTrees, PNN, Random Forest, SVM).

The RusPersonality corpus was also used for computing the correlation between the author profile (gender and personality) and POS-sequences by Litvinova, Seredin, and Litvinova (2015). The POS-bigram comprising of a preposition and a noun was found to have a weak, but statistically significant correlation with gender, and it positively correlated with male authors. The regression model based on this feature performed with an accuracy of 65% on a gender-balanced test set.

2.5 Analysis of Slovene User-Generated Content

The Slovene language is well-equipped with language resources and technologies for standard written language, such as part-of-speech taggers (Grčar & Krek, 2012; Ljubešić & Erjavec, 2016), lemmatizers (Juršič, Mozetič, Erjavec, & Lavrač, 2010), syntactic parsers (Dobrovoljc, Krek, & Rupnik, 2012), and the Gigafida referential corpus of standard Slovene (Logar Berginc et al., 2012). However, not much attention has been given to computational

stylometry. To the best of our knowledge, the first real-life application of authorship attribution is presented by Zwitter Vitez (2011), who identified the most probable author of a provocative document published under a pseudonym. Zwitter Vitez used a range of lexical and readability features to compare the anonymous text with the documents of potential authors.

While focused studies about Slovene computer-mediated communication and user-generated content (UGC) brought the first results of the Slovene language online, the Slovene UGC began to be analyzed on a larger scale within the scope of the Janes[4] national research project. The most notable contributions of the Janes project include the Janes corpus of Slovene UGC and the tools for collecting, preprocessing, and analyzing UGC (Fišer, Erjavec, and Ljubešić, 2017). The Janes corpus comprises five UGC genres (tweets, blog entries, forum posts, online news and their comments, and Wikipedia talk pages). The subcorpora of tweets and blog entries are enriched with manual annotation about the authors' account type and gender, thus they make suitable resources for analysis in user behaviour and author profiling.

A corpus study by Osrajnik et al. (2015) used the Janes tweet subcorpus to compare female and male language use with regard to non-verbal expressiveness. Their analysis of linguistically substandard tweets has shown that female users tend to include more emoticons (e.g., *:))*, *:/*, *:-))))*) than male users. In turn, male users use more expressive punctuation symbols (e.g., *...*, *!?*, *????*) in their tweets.

The study by Ljubešić and Fišer (2016) presents the experiments for automatically discriminating between private and corporate Twitter accounts based on the Janes tweet corpus and the manual annotations. Their approach involves various features that are either language-independent or language-dependent. Two types of the language-dependent features were used: 1) a BOW model, and 2) morphosyntactic tags and surface forms. The language-independent features relate to tweet text (e.g., the average number of tweets containing URLs, the mean of number of posting hour, variance of posting weekday) or user metadata (e.g. the number of followers the user has). For each of the three feature types, an SVM classifier was built. While the BOW classifier performed best among the three, an ensemble classifier that uses the output of all three feature-type classifiers performs best, achieving an overall weighted F-measure of 94.31%, and an F-measure for private users of over 96%.

A preliminary study of gender profiling of Slovene Twitter users was presented by Verhoeven, Škrjanec, and Pollak (2017). For the experiments, the tweets from the Janes Twitter corpus (Fišer, Erjavec, & Ljubešić, 2016) were imported into the previously built classification system for author profiling by Verhoeven et al. (2016), who performed gender and personality prediction on tweets in six languages: Portuguese, Spanish, French, Italian, Dutch, and German (see Section 2.2). For the classification of Slovene Twitter users, the setting from experiments with other languages was preserved: the SVM algorithm was learned based on token word uni- and bigrams, and character three- and tetra-grams with TF-IDF weighting. Altogether, 3,490 users (68.5% were male, and 31.5% female) were drawn into analysis, with each of them represented by 200 tweets in Slovene. In 10-fold cross validation, the model achieved 92.6% accuracy, which is more than the accuracy scores on tweets in other languages tested imported into the classification system (Verhoeven et al., 2017).

This Thesis provides a broader study of Slovene UGC, which has not yet been analyzed with machine learning methods and statistical analysis with regard to the author's gender. In the analysis of gender-related linguistic variation, we address the differences between female and male authors that occur on the level of referential gender, document topic,

[4] Janes is a national research project funded by the Slovenian Research Agency: http://nl.ijs.si/janes/

and the writing style. For this, we employ methods that have been used for the study of variation as described in this chapter; namely, we employ the methods of gender profiling, topic modeling, and stylometric analysis, and apply them to Slovene tweets and blog entries.

Chapter 3

Corpus Description

This chapter presents the Slovene Twitter and blog data used for the study of gender-related linguistic variation. Both corpora were collected as a part of the Janes[1] national research project which was funded by the Slovenian Research Agency between July 1, 2014, and June 30, 2017. The project focuses on the analysis of nonstandard Slovene language and its results include the Janes corpus of Slovene user-generated content (UGC). The Janes corpus (Fišer, Erjavec, & Ljubešić, 2016) comprises five genres of UGC: Twitter messages (*tweets*), forum posts, blog entries and their comments, news comments, and user and page talks from Wikipedia. This chapter describes the Twitter and blog corpora with regard to their collection and preprocessing. In the experiments, version 0.4 of the corpus was used.

3.1 Data Collection

This section presents the process of collecting the Twitter and blog subcorpora. Because the document structure of each genre di□ers, special extractor tools were developed for each domain.

3.1.1 The Twitter Corpus

Slovene tweets were collected with a tool called TweetCat developed by Ljubešić, Fišer, and Erjavec (2014). The tool specializes in collecting tweets in smaller languages and has so far been used for building Slovene, Croatian, and Serbian Twitter corpora. TweetCat searches the Twitter Application Programming Interface[2] (API) to find the tweets of users who use the desired language. We present the process of harvesting Slovene tweets as described in Fišer, Erjavec, and Ljubešić (2016).

The tool uses a list of seed terms to find the tweets in the chosen language. Seed terms are highly frequent words that do not overlap with the vocabulary of other languages. For Slovene, 20 seed terms were used including *še* [yet], *kaj* [what], *mogoče* [maybe], *vendar* [but], etc. On the one hand, the tool identifies new users by checking if seed words occur in their tweets, in which case it inspects 200 of the user's tweets with a list of additional 40 seed words in Slovene. On the other hand, the tool includes the new tweets of previously confirmed users. It also carries out language identification with seed terms on tweets by the followers and friends of confirmed users.

Together with the tweet text, the tool extracts the handle (username), time stamp of the creation and harvesting of the tweet, and the number of retweets and likes. The tweets are saved as an output with an XML structure. The Janes corpus 0.4 includes 7,503,199

[1] Janes project web page: http://nl.ijs.si/janes/
[2] Twitter API: https://dev.twitter.com/overview/api

tweets by 8,749 users published between June 2013 and January 2016. In the following sections, we describe how the tweets were preprocessed and annotated with metadata.

3.1.2 The Blog Corpus

The Janes corpus contains blog entries from two Slovene blog portals: PublishWall[3] and rtvslo.si[4]. For the selection of these portals, two criteria that allowed for a larger harvesting of blog entries and their comments were taken into consideration: the unified document structure in each portal, and the popularity among Slovene bloggers (Fišer, Erjavec, & Ljubešić, 2016). The collection of blog data from each portal was carried out by specialized extraction tools. The tools saved the blog entries and their comments as an XML format including the blogger's and commentator's username, as well as the date of posting and harvesting. The Janes blog subcorpus comprises 23,515 blog entries by 243 bloggers from rtvslo.post and 18,515 blog entries by 615 bloggers from PublishWall. The blog entries in the corpus were published between October 2006 and January 2016.

3.2 Preprocessing and Linguistic Annotation

This section presents the automated preprocessing steps and linguistic annotation of the Janes tweet and blog corpora.

The documents underwent tokenization and sentence segmentation using the tool developed by Ljubešić and Erjavec (2016) for processing non-standard documents in Slovene, Croatian and Serbian. In computer-mediated communication, the diacritics are often omitted, which results in poorer performance of NLP tools. The diacritics in the Janes corpus were restored using the tool for automatic rediacritization by Ljubešić, Erjavec, and Fišer (2016).

In the following preprocessing steps, the documents underwent normalization, which means that words written in non-standard orthography were assigned the standard spelling (e.g., *tut* and *tud* are normalized as *tudi* [also]). The normalization was done based on a manually normalized training corpus of tweets (Fišer, Erjavec, & Ljubešić, 2016). In the final step of preprocessing, the corpus was automatically tagged with part-of-speech tags and lemmatized using the tool by Ljubešić and Erjavec (2016). The POS-tagger performs with an accuracy of 94.3% on the standard Slovene test set. The POS-tagset was expanded with the POS-tags for elements that typically occur in UGC: URL and email addresses, emoticons and emojis, hashtags, and user mentions.

3.3 Metadata

The Janes corpus is enriched with metadata on the text and user level, which were obtained either automatically or manually. Given the multilingual context of UGC, each document underwent language identification. For this the langid.py program (Lui & Baldwin, 2012) was used. The identified language tags were additionally corrected with heuristics resulting in four possible tags for the entire corpus: *Slovene*, *English*, *Serbian/Croatian/Bosnian*, and *undefined* (Fišer, Erjavec, & Ljubešić, 2016). UGC documents display various levels of standardness in terms of capitalization, grammar, and vocabulary. Using an automated method developed by Ljubešić et al. (2015), the documents in the corpus were annotated on a three-level scale for their linguistic and technical standardness, whereby the linguistic level takes into account deviations from grammatical and syntactic norms, while technical

[3] PublishWall: http://www.publishwall.si/
[4] rtv.slo.si/blog: http://www.rtvslo.si/blog

standardness refers to the use of capitalization and punctuation, as well as typographical errors. The documents in the corpus are annotated with sentiment information that is either positive, negative, or neutral (Smailović, Grčar, Lavrač, & Žnidarčič, 2014; Fišer, Smailović, Erjavec, Mozetič, & Grčar, 2016).

The metadata on the user level include the user gender and account type. The metadata on region is available only for geotagged tweets (Čibej, 2016). Gender and account type were first annotated automatically. The gender tag (*male*, *female*, or *neutral*) was assigned according to the use of referential gender in verb phrases (Fišer, Erjavec, & Ljubešić, 2016). A program was built to find auxiliary verb forms in first person singular (*sem* [am], *nisem* [am not], *bom* [I will]) and the l-participles, which have explicit gender markings. Each such participle contributed one point to the female or male gender indicators. The program compared the indicators and if the ratio exceeded 0.7, while at least 1% of the documents of this user contained gender indicators, the user was annotated with the corresponding tag. If these criteria were not met, the user was tagged as *neutral*. The program annotated users with the *female* and *male* gender tag as private accounts (tag *private*), i.e. accounts used by individuals who publish in their free time. The users who were assigned the *neutral* gender tag were categorized as corporate users. The *corporate* tag applies to accounts of companies and other o□cial profiles managed by professionals or paid users.

The automatic annotation of gender and account type in tweets and blog entries underwent manual revision, whereby the username, profile picture, and documents were checked to either confirm or change the automatic tag. The manual revision has shown that the program correctly classified the gender of roughly 76% Twitter users (Fišer, Erjavec, & Ljubešić, 2016) and almost 78% of bloggers.

Chapter 4

Methodology

In this chapter, we present the methodological background for the gender-based analysis of Slovene user-generated content. First, the focus is placed on the document representation in the tasks of text mining. We then describe the machine learning approaches applied to the Twitter and blog data: we use unsupervised machine learning to explore the topical variation between female and male bloggers, and supervised machine learning to build classification models for gender prediction. Furthermore, we present the compilation of stylistic word lists and the statistical methods for analyzing the writing style of female and male authors.

4.1 Document Representation for Machine Learning

The input data in text mining are more or less structured natural language documents. Because most machine learning and data mining algorithms are generally not designed to work on textual data, the feature construction is an important step in text mining, thus obtaining a feature-based representation (Brank, Mladenić, & Grobelnik, 2010). A commonly used approach to feature extraction is the construction of a bag-of-words (BOW) representation. In the BOW representation, each document is transformed into a vector that contains the words from the entire corpus by recording the frequency of each word in each document.

For document representation, words can be used as terms for indexing, but also phrases or word and character n-grams (Sebastiani, 2002; Daelemans, 2013). Word n-grams are units made of an n-number of consecutive words or tokens in general. For example, from the tweet "To pa je bila top sprostitev."[1] the following word unigrams (n-grams of length 1) are generated: "To", "pa", "je", "bila", "top", "sprostitev", ".". The same tweet could produce the following word bigrams (n-grams of length 2): "To pa", "pa je", "je bila", "bila top", "top sprostitev", "sprostitev.". In the same way that sentences and documents are split into word n-grams, words can be split into character n-grams of various lengths. The word "bila" generates four character unigrams: "b", "i", "l", and "a", or three character bigrams: "bi", "il", and "la".

In order to compare documents, the aim is to find units that are typical of some documents, but rare in others. For this, several weighting schemes can be applied, such as term frequency (the frequency of a word in the given document), binary weights (records presence or absence of a word in the document), or term frequency-inverse document frequency (TF-IDF). The TF-IDF weighting scheme is presented in Equation 4.1 (Kobayashi & Aono, 2007), whereby w_{ij} stands for the weight of the word i in document j; tf_{ij} is the

[1] English: *That was quite a relaxation.*

term frequency of the word i in document j (the number of times this word appears in the given document); N stands for the total number of documents; and df_i is the number of documents containing the word i. As can be seen from the equation, the words with a high frequency across the entire corpus are given lower weights.

$$w_{ij} = tf_{ij} * log(\frac{N}{df_i}) \tag{4.1}$$

Document feature vectors tend to be large in size. Feature selection is often applied to reduce the feature space dimensionality and thus save computer resources and preprocessing time as well as possibly increase the machine learning model's performance (Dhillon, Kogan, & Nicholas, 2004). While a number of feature selection methods exist (see Sebastiani (2002)), we briefly describe the SelectFromModel method implemented in the Scikit-learn software (Pedregosa et al., 2011), which we used in our classification experiments. From the feature space, the SelectFromModel method[2] removes the features that score below a certain threshold which is left in the default settings or set by the software user. Namely, in the model training phase, each feature is assigned a coe□cient attribute, i.e. the weight assigned by the learning algorithm. The coe□cient in models based on Naïve Bayes is the empirical log probability of features given a class. Logistic Regression models use the coe□cient of the features from the decision function, and similarly, models based on the Support Vector Machine use the weights assigned to the features by the decision function.

When the SelectFromModel method is applied, features which have a coe□cient lower than the threshold are removed from the feature space. In the default settings, the threshold equals the mean of coe□cients. For our experiments in author gender prediction with statistical models, we used the SelectFromModel with the threshold set to default.

4.2 Gender Prediction with Statistical Models

In this section we present the methodology of gender prediction with statistical models, which we approach as a supervised machine learning task. Section 4.2.1 provides the description of three algorithms we use: Support Vector Machine, Naïve Bayes, and Logistic Regression. For the model evaluation, we apply the cross validation method, which is presented in Section 4.2.2.1. We analyze the most informative features of statistical models using the method described in 4.2.2.2.

4.2.1 Machine Learning Algorithms

In this section we present the algorithms we employed for building the gender prediction models: the Support Vector Machine, Naïve Bayes, and Logistic Regression. For model construction and evaluation the Scikit-learn library was used (Pedregosa et al., 2011).

- **Support Vector Machine**

  The Support Vector Machine (SVM) algorithm (Cortes & Vapnik, 1995) is one of the most widely used approaches to text mining and gender prediction tasks (see Section 3). A linear SVM algorithm maps the training examples from the original feature space into a multidimensional feature space and uses a discriminant function to separate between the classes as widely as possible with a margin hyperplane (Witten & Frank, 2005). The training samples that lie closest to the hyperplane are referred

[2] http://scikit-learn.org/stable/modules/feature_selection.html

to as “support vectors” and are used to build the discriminant (decision) function. Unseen examples are then mapped into the multidimensional space and classified based on their position relative to the hyperplane. In the linear SVM algorithm, the following formula for the discriminant function is used (Guyon, Weston, Barnhill, & Vapnik, 2002):

$$D(x) = x * w + b \tag{4.2}$$

In Equation 4.2, *x* is the feature vector of the unseen example, *w* is the SVM weight vector, and *b* is the bias value of the hyperplane. If the discriminant function returns a positive value, the assigned class is positive, and if it returns a negative value, the assigned class is negative.

- **Naïve Bayes**

  Classifiers using the Naïve Bayes algorithm are probabilistic classifiers that rely on the Naïve Bayes theorem. The theorem takes into account the prior and conditional probability of an event (Bramer, 2013). An event refers to class membership, so the Naïve Bayes algorithm computes the probability of each class ($P(c_i)$) for an unseen instance. Prior probability is the class probability when no information is available, except for the probability of each class in the training set. In contrast, conditional (or posterior) probability is the probability of a class after additional information about the conditions (feature values) has been obtained, so $P(c_i \mid f_1 = v_1)$ is the probability of class c_i given that the value of feature f_1 is v_1. This algorithm performs under the “naïve” assumption that the features are independent of each other.

  In text classification tasks, a document is typically represented as a feature vector of binary of weighted terms from the document. Document d is represented by the vector $\vec{d_j} =< w_{ij}...w_{|T|j} >$, where w is a word from the bag of words and T is the size of the feature space (Sebastiani, 2002). The probability that the document d with the vector representation $\vec{d_j}$ belongs to class c_i is computed using the following formula Sebastiani (2002):

$$P(c_i|\vec{d_j}) = \frac{P(c_i) * P(\vec{d_j}|c_i)}{P(\vec{d_j})} \tag{4.3}$$

  The document is classified into the class with the highest probability. Despite the drawback of naïve assumption, the Naïve Bayes classifier works very well, especially when combined with the feature selection procedures that exclude redundant and thus non-independent features (Witten & Frank, 2005).

- **Logistic Regression**

  The Logistic Regression algorithm (Cox, 1958) builds a linear classifier. It estimates the probability of class y given the document x ($P(y|x)$) by extracting features from the input and combining them linearly (Jurafsky & Martin, 2009).

  Unlike Linear Regression, the Logistic Regression algorithm builds a linear model based on a transformed target variable which is obtained with the logit transformation function (Witten & Frank, 2005). The formula in Equation 4.4 (Cox, 1958; Jurafsky & Martin, 2009) is used to compute the probability of each class; the class with the highest probability is then chosen::

$$P(c|x) = \frac{exp(\sum_{i=1}^{N} w_i f_i(c,x))}{\sum_{c' \epsilon C} exp(\sum_{i=1}^{N} w_i f_i(c'x))} \quad (4.4)$$

In Equation 4.4 we compute the probability of class c for a given observation x, where w_i represents the weights of document i, f_i represents the features of document i, c' is the estimate of the correct class, and C is the set of all classes. The exp notation stands for the exponent function $exp(x) = e^x$.

The weights are learned with conditional maximum likelihood estimation, which means the parameters w are chosen in a way that they maximize the probability of the classes in the training data given the observations x (Jurafsky & Martin, 2009).

4.2.2 Model Evaluation Methods

In this section, we present the methods for evaluating the performance of statistical gender prediction models. To compare the model performance in di□erent settings, we use the classification accuracy. In a binary prediction task, the classification accuracy is defined as the sum of the number of true positives and true negatives, divided by the number of all instances. In Section 4.3.4.1, we present cross validation and the train–test split. Section 4.3.4.2 proposes a qualitative approach to model evaluation by analyzing the features that provide most information about class membership.

4.2.2.1 Cross Validation and Separate Train and Test data

When training a classification model, we wish to estimate the model's performance on instances that were not used in the training process. A frequently used evaluation technique called cross validation provides an estimation of how well the model will generalize over unseen instances. In cross validation, the data is divided into a k number of approximately equal parts or folds. In each iteration, one fold is used for testing, and the rest $k-1$ folds are used for training (Witten & Frank, 2005). In total, the learning process is carried out k times and a number of k performance measures are obtained. In the experiments described in Sections 5.1.3 and 6.2.3 we build statistical gender prediction models and evaluate their performance using 10-fold cross validation. We report on the mean and standard deviation of the 10 classification accuracy scores obtained from training each model.

Many machine learning methods are used to build models where the train and test data share the same feature space and distribution, but in real-world applications the unlabeled dataset might di□er strongly from the train set. In such cases it is recommended to transfer knowledge between task domains, i.e. employ transfer learning, as it can greatly improve the performance of learning as well as avoid manual labeling of unseen data (Pan & Yang, 2009). This thesis does not apply the methods of transfer learning to automated gender prediction; however, we perform cross-genre model evaluation, where a statistical model for gender prediction is trained on one UGC genre (e.g., tweets) and tested on the other (e.g., blog entries). We perform cross-genre experiments to test the classifier's robustness and potential application to other UGC genres.

4.2.2.2 Most Informative Features per Class

Aside from classification accuracy, we discuss the performance of statistical models for gender prediction by observing the features with the largest weights. We refer to these features as "the most informative features".

Linear models such as Support Vector Machine, Naïve Bayes, and Logistic Regression assign a weight to each feature to perform classification. The weight vector is learned

from the training instances by applying a function that calculates the optimal weights to predict a class or the probability of a class (Jurafsky & Martin, 2009). For each of the three algorithms that we used in our experiments the process of weight learning is described in Section 4.2.1. Feature weights provide insight into gender prediction models as we can infer which features are strongly associated with female authors, but not with their male counterparts, and vice versa.

We use a simple Python function that takes a classification model as the input and then outputs a list of features with the largest weights. In the function we specify the number of features per class, so the function returns a number of features that have the largest weight, i.e. carry the most predictive information for classifying an instance into this class. The function outputs features together with their weights. In the thesis, this function is used to analyze the 1,000 most informative features for the female and male class in the best performing statistical models for gender prediction in tweets and blog entries.

4.3 Gender Prediction with Classification Rules

In this section we present the rule-based model for gender prediction. This model uses manually constructed classification rules to assign the class to authors based on the occurrence of referential gender in the text (see Section 1.1), more specifically they rely on the grammatical gender of verb l-participles in self-referencing context. The classification rules are written manually in as a Python program that processes non-lemmatized documents. We perform gender prediction using the profile-based approach, which means the texts of a single author are concatenated into one instance (Stamatatos, 2009).

The rule-based model is built under the simplified assumption that the referential gender use in self-referencing context indicates the gender of the author. Thus, the use of feminine referential gender in self-references implies the text was written by a female author, while the masculine referential gender points to a possible male author[3].

The rule-based model searches for first person auxiliary verb forms (*sem* [am], *nisem* [am not], and *bom* [I will]) and non-standard spellings of the first two forms (*sm* [am], and *nism* [am not]). These first person auxiliary verb forms serve as node words, i.e. when such a form is found, the preceding and following words of the node word are observed. The rules consider two words before and two words after the node word, whereby the endings of these words are checked for gender marking. An indicator for the female and male class is calculated based on the use of feminine or masculine gender of l-participles so that the ending "-la" signals the feminine grammatical gender of l-participles, while "-al", "-il" and "-el" indicate the masculine grammatical gender. If the majority (70%) of all gender indicators belongs to one class, the rules assign the author to the corresponding class. We additionally experiment with the minimum count of indicators (3 or 5) per author. If the indicator minimum or the indicator majority conditions are not satisfied, the rules classify a user as "undefined".

4.4 Topic Ontologies of Blog Entries with the OntoGen Editor

Knowledge representation is an important aspect when dealing with large collections of textual data. Semantic knowledge can be explored with topic modelling, i.e. building

[3] It should be noted that gender identities are not limited to this gender binarism (the female and male gender), but rather form a range of separate or overlapping categories. However, the question of individual gender identities of text authors in our corpus goes beyond the scope of this thesis.

models that automatically discover topics from a large collection of documents (Yang, Pan, Downey, & Zhang, 2014). A frequent approach to topic modelling of textual documents is the automated construction of topic ontologies, which are defined as sets of topics connected via various relations, whereby each topic includes a set of related documents (Fortuna, Grobelnik, & Mladenić, 2005). This section presents the methods of constructing topic ontologies by employing the k-means clustering algorithm using the OntoGen tool (Fortuna, Grobelnik, & Mladenić, 2007).

The OntoGen tool[4] is a semi-automatic data-driven ontology editor that combines text mining techniques with a fairly simple user interface (Fortuna et al., 2007). It constructs and visualizes topic ontologies by performing document clustering, meaning it groups similar documents together. In the OntoGen tool, the documents are represented as a bag-of-words (BOW) with TF-IDF weights (see Section 4.1). In order to cluster together topic-wise related documents, a similarity measure is applied. There are several measures for comparing. Cosine similarity is widely used and implemented in OntoGen as well. It equals the cosine of the angle ϕ between the vectors. The cosine is computed as the dot product between two document or word vectors (i and j) using the following formula (Senellart & Blondel, 2007):

$$cos(\phi) = \frac{i \cdot j}{\sqrt{i \cdot i \times j \cdot j}} \tag{4.5}$$

In Equation 4.5 we define the angle ϕ between the vectors i and j. If the documents comprise of a similar word distribution, the angle between their vectors is small, thus resulting in a cosine close to 1. In turn, documents that share a small number of words have a larger angle between their vectors, so the cosine similarity lies close to zero.

The cosine similarities between documents are used for the formation of document clusters using the k-means clustering algorithm. The k-means employs iterative distance-based clustering and is an exclusive clustering algorithm, as every instance is assigned to a single cluster (Witten & Frank, 2005; Bramer, 2013). First, the number of clusters is specified: this is the parameter k. Among all instances, k instances are randomly chosen as cluster centers (centroids). Each instance is assigned to the centroid closest to it given the chosen distance measure (e.g., cosine similarity) and thus a k number of clusters is formed. Next, the centroids of the k-clusters are re-calculated and these centroids are to be the new cluster centers. The process is repeated with these new centroids and the iteration continues until the centroids have stabilized (Witten & Frank, 2005).

In the OntoGen tool, the k is provided by the user. The tool then suggests a k number of document clusters – topics. The user then decides whether to add the clusters to the ontology. Each topic is represented with a set of keywords and the user can freely rename the topic. The user can also manually move the documents. Additionally, if the input documents are pre-categorized, a method for grouping the instances according to the labels is also supported. Another view is gained by inspecting SVM keywords, which are the words most distinctive for the selected concept with regard to its sibling concepts in the hierarchy (e.g., words contrasting male and female entries categorized in a selected topic). Figures 4.1 and 4.2 show the OntoGen interface with the menu on the left-hand side and the ontology visualization on the right-hand side.

[4]OntoGen is freely available at http://ontogen.ijs.si/

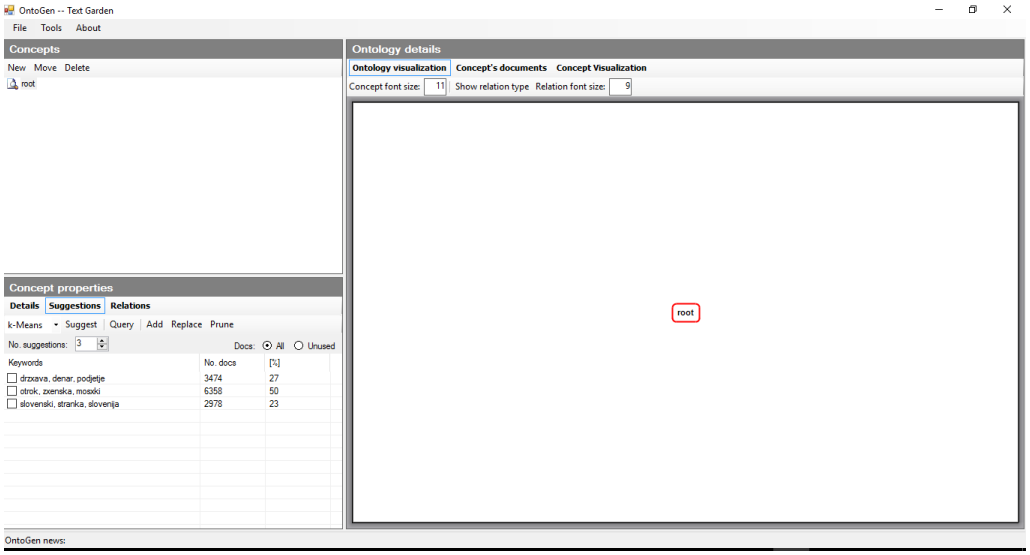

Figure 4.1: The OntoGen user interface showing the import of the corpus and the menu on the left.

4.5 Discursive Features and Statistical Methods for Their Analysis

In this section, we explain the methodology of used for statistical analysis for writing style. One of the tasks of this thesis is to compare the writing styles of female and male users in tweets and blog entries. In our approach, we measure the features for quantifying the writing style of each author and use statistical methods to compare the scores between female and male authors. For each author, we compute the frequency of words from pre-prepared lists, where each word list represents a particular writing style. This section presents the content and size of each word list.

We understand the use of words from lists as a stylistic choice and expect to see discrepancies in the choices made by female and male authors. For this we apply the Mann-Whitney *U*-test to test which writing styles di□er in a statistically significant manner between women and men. Next, we use Pearson's coe□cient to compute the correlation between each style and the gender, whereby we focus on observing the size and sign of Pearson's coe□cient. In order to estimate the e□ect of gender on the writing style, we calculate the square of Pearson's coe□cient, which tells us how much variance in the writing style can be explained by the gender variable. In hypothesis testing, we use the alpha level of 0.5. All statistical testing was carried out in the SciPy Python library[5].

4.5.1 Writing Style Presented as Word Lists

We approach the analysis of writing style by counting the frequency of particular words in the documents of a single author. We use 18 word lists containing words that are associated with the same style to count the word occurrences. The word lists can be generally divided into four categories based on the style they describe: expressive language and symbols (*Emoticons*, *Emojis*, *Emoticons and emojis*, *Positive words*, *Negative words*, *Emotional words*), modality markers (*Intensifiers*, *Hedge words*, *Modal verbs*), grammatical choices (*Function words*, *Negation*), non-standard or new vocabulary (*Non-standard words*,

[5]SciPy: https://www.scipy.org/

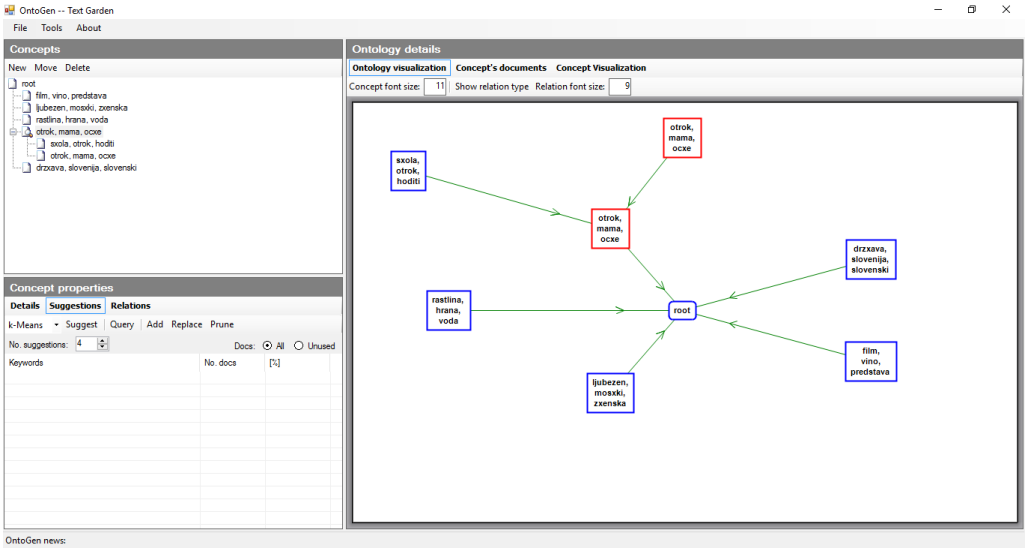

Figure 4.2: The construction of the topic ontology and its hierarchical visualization on the right-hand side.

New words, *Janes glossary*, *Profanity*), and social and cognitive processes (*Communication verbs*, *Cognition verbs*, *Social words*).

Each list contains the lemma forms of words. Table 4.1 presents examples and the total size (number of words) of each list.

The *Emoticons* list contains emoticons, which are sequences of punctuation symbols or numbers that resemble a human face or other non-verbal symbols. The *Emoticons* list was collected from the Wikipedia[6] web page.

The *Emojis* list comprises of "emojis" which are small images representing faces or other motifs. The list was extracted from the Janes corpus using the SketchEngine concordance tool (Kilgarri et al., 2014).

The *Positive words* and *Negative words* lists are a publicly available[7] sentiment lexicon for Slovene constructed by Kadunc and Robnik-Šikonja (2016), who translated an English sentiment lexicon by Hu and Liu (2004) into Slovene and used it for sentiment analysis. For our study, the Slovene *Positive words* and *Negative words* were combined into the *Emotional words* list.

Intensifying adverbs and adjectives are collected into the *Intensifiers* list, while the *Hedge words* list comprises of adverbs used for hedging statements. Both lists are based on the examples of intensifiers and hedges provided in **penebaker2013** Biber (2007) and Hyland (2005). The English examples were translated manually into Slovene.

The *Modal verbs* list contains the infinitive forms of five Slovene modal verbs. Adverbs and pronouns that express negation are collected in the *Negation* list.

Function words are often referred to as "stop words" in text mining studies and are understood as words with a high frequency which carry little meaning. The *Function words* list was constructed by extracting the lemma forms of prepositions, conjunctions, pronouns, and particles from the balanced Kres corpus of standard written Slovene (Logar Berginc et al., 2012). For the extraction, we used the SketchEngine (Kilgarri et al., 2014) concordance tool.

The word lists with marked vocabulary (*Non-standard words*, *New words*, *Janes glos-*

[6] https://en.wikipedia.org/wiki/List_of_emoticons

[7] The lexicon is available on the Clarin repository: http://hdl.handle.net/11356/1097

sary) were taken from the glossary prepared for the construction of the dictionary of Slovene Twitterese (Gantar, Škrjanec, Fišer, & Erjavec, 2016). The glossary was extracted from the Janes corpus and each word in the glossary was categorized as either non-standard, new, or as an abbreviated form. For the purpose of writing style analysis, the words categorized as non-standard are grouped into the *Non-standard words* list, the new vocabulary is collected into the *New words* list, while the *Janes glossary* contains the entire glossary. Vulgar expressions and swear words are compiled into the *Profanity* list and are partially based on the list of o□ensive terms by Luis von Ahn[8], while some additional vocabulary was taken from the collaborative dictionary named Razvezani Jezik[9] ("The Unleashed Tongue"), which can be edited and contributed to by any visitor.

The lists of specific verbs (*Communication verbs* and *Cognition verbs*) are largely based on the annotation and analysis of semantic roles in the Slovene training corpus (Krek, Gantar, Dobrovoljc, & Škrjanec, 2016). The words denoting social interaction (*Social words*) were translated from examples by Pennebaker (2013). Furthermore, the Thesaurus function in the SketchEngine concordance tool was used to extract synonyms or similar words from the Kres corpus.

Table 4.1: Examples of words from lists representing writing styles.

Word list	Examples	List length
Emoticons	:-), :o, <3	154
Emojis	👆, 🍺, 🌺	550
Emoticons and emojis	:D, 💕, 🍋	704
Positive words	*ugodje* [pleasure], *barvit* [colorful], *zmagati* [to win]	2,647
Negative words	*izguba* [loss], *nepravičen* [unfair], *odlašati* [to delay]	6,690
Emotional words	*zlo* [evil], *všeč* [to like], *brutalen* [brutal]	9,336
Intensifiers	*zares* [really], *grozno* [awfully], *bistveno* [essentially]	64
Hedge words	*malce* [a bit], *deloma* [partially], *skoraj* [almost]	27
Modal verbs	*morati* [to must], *želeti* [to wish], *hoteti* [to want]	5
Function words	*če* [if], *za* [for], *naš* [our]	311
Negation	*nič* [nothing], *nihče* [nobody], *nikjer* [nowhere]	9
Non-standard words	*kofe* [co□ee], *šansa* [chance], *psihič* [psycho]	300
New words	*blender* [blender], *hejter* [hater], *zloadati* [to load]	510
Janes glossary	*biznis* [business], *čaga* [party], *mejbi* [maybe]	906
Profanity	*fak* [fuck], *kuzla* [bitch], *scati* [to piss]	180
Communication verbs	*vprašati* [to ask], *govoriti* [to talk], *trditi* [to claim]	72
Cognition verbs	*upati* [to hope], *vedeti* [to know], *meniti* [to think]	35
Social words	*prijatelj* [friend], *darilo* [gift], *vesel* [happy]	125

We measure the writing style of each author in the tweet and blog corpora by observing the occurrence of words from the word lists described above. For each author, we compute the share of these target words in their word total, meaning 18 numeric values are produced for each author using the following formula:

$$share = 100 * \frac{\#\,target\,words}{\#\,word\,total} \tag{4.6}$$

The numeric score obtained from Equation 4.6 is used as the input for statistical testing with the Mann-Whitney *U*-test and Pearson's correlation. Moreover, the shares are used in

[8] The list of o□ensive words is available at https://www.cs.cmu.edu/ biglou/resources/bad-words.txt
[9] http://razvezanijezik.org/

the visualization with parallel coordinates that are plotted using the Matplotlib library[10].

4.5.2 Mann-Whitney *U*-test

For each of the word lists presented in Section 4.3.1, we compare the scores of female and male authors; to do this, we apply the Mann-Whitney *U*-test (Mann & Whitney, 1947). The Mann-Whitney test is a non-parametric test that evaluates the difference between two groups of scores from an independent-measures design. Unlike its parametric counterpart, the *t*-test, it does not assume a normal distribution of the dependent variable. The Mann-Whitney *U*-test compares two samples by first ranking individual scores. The null hypothesis states that there are no differences between the two samples; therefore, there is no tendency for the ranks of one sample to be systematically higher (or lower) than the ranks of the other sample (Gravetter & Wallnau, 2013). The alternative hypothesis is that a difference exists between the two samples, so the scores of one sample have systematically higher (or lower) ranks than those of the other sample. For each of the samples, the *U*-statistics is computed using the following equation:

$$U_1 = n_1 * n_2 + \frac{n_1 * (n_1 + 1)}{2} - \Sigma R_1 \tag{4.7}$$

In Equation 4.7, n_1 is the size of sample 1, n_2 is the size of sample 2, and ΣR_1 is the sum of ranks in sample 1. The *U*-score for the second sample (U_2) is computed the same way with appropriate alternation: n_2 is used instead of n_1 in the fraction, and the sum of ranks ΣR_2 is used as the sum of ranks in sample 2. The smaller value of U_1 and U_2 is used when consulting significance tables. If the sample data produce a *U*-score that is less than or equal to the table value, the null hypothesis is rejected and the difference between the samples is estimated to be statistically significant (Gravetter & Wallnau, 2013).

To decide between parametric and non-parametric tests, we examined the distributions of target word occurrence in the blog and tweet corpus by visualizing the distribution in histograms. For each target list, a histogram for each class was created based on absolute frequencies of target words in the documents. Most of the word lists displayed a skewed distribution. Thus, we decided to use the Mann-Whitney *U*-test for all the word lists to ensure comparability of results.

4.5.3 Pearson's Correlation Coefficient and Its Squared Value

Pearson's correlation coefficient (*r*) measures the degree and direction of the linear relationship between two numeric variables (Gravetter & Wallnau, 2013). Because our data contain a dichotomous independent variable (gender with *female* and *male* as possible values) and a numerical dependent variable (share of particular words in all words by one author), we use the point-biserial correlation coefficient (r_{pb}), which is a version of Pearson's *r* used for measuring the relationship between a numerical and dichotomous variable. We compute the r_{pb} score for every word list using the following formula:

$$r_{pb} = \frac{M_1 - M_2}{s_{(n-1)}} * \sqrt{\frac{n_1 * n_2}{n * (n-1)}} \tag{4.8}$$

In Equation 4.8, M_1 and n_1 represent the mean and size of sample 1, respectively, and M_2 and n_2 are the mean and size of sample 2, respectively. The standard error is s_n, and n represents the total sample size. The values of r_{pb} range from -1 to +1. A positive r_{pb} value indicates a positive correlation with one of the dichotomous values, while a negative r_{pb} signals a correlation with the other dichotomous value. An absolute r_{pb} value of 0.1

[10] Matplotlib: https://matplotlib.org/

indicates a low correlation, an absolute value around 0.3 means the correlation is moderate, while an absolute r_{pb} value of 0.5 signals a large correlation (Cohen, 1988).

Additionally, a measure of effect size was computed to estimate the size of the effect the author's gender has on the measured writing style. For this, the coefficient of determination r^2 was computed. It measures the variability of one variable that can be determined from the relationship with the other variable (Gravetter & Wallnau, 2013). In our task, we measure how much of the variance in the use of writing styles is accounted for by the author's gender. The r^2 can be calculated by squaring Pearson's coefficient or, in our case, the point-biserial coefficient r_{pb}. Cohen (1988) provides guidelines for interpreting the r^2 score: a value of 0.01 indicates a small effect, a value of 0.09 is a medium effect, while a large effect starts at 0.25 or more. When turning the r^2 score into percentages by multiplying it with 100, the score tells us the share of variance accounted for by gender.

Chapter 5

Analysis of Twitter Messages

In this chapter, we explore the relationship between the language used in documents and the gender of their author. More specifically, we analyze Slovene Twitter messages (*tweets*) to compare the language of women and men to learn about the variation in language use. We examine these linguistic variations on three levels: use of grammatical gender (as referential gender), general vocabulary, and particular vocabulary indicating a gender-related discourse style (genderlect), to learn whether and how they can provide predictive information on the author's gender. Section 5.1 describes our approach to automated prediction of Twitter user gender: we present the creation of the Twitter subcorpus, and the setting and the performance results of rule-based and statistical machine learning models. For the best performing statistical model, we analyze features that help distinguish between female and male authors most successfully. Section 5.2 presents the comparison of female and male Twitter users in terms of their writing style. We explore how the authors differ in the use of words that represent a particular style, whereby these variations are tested statistically with the Mann-Whitney U-test, and Pearson's correlation coefficient as a measure of effect size. The writing style traits with a proven statistically significant difference are visualized as parallel coordinates, which enables the intra- and inter-class comparison of authors.

5.1 Models for Automated Gender Prediction

The aim of this section is to examine which patterns in language can be used to distinguish automatically between female and male Twitter users. For this task, two types of classification models for gender prediction are built. This section provides the description of the model construction and performance. We also present the subcorpus created for machine learning models and statistical testing. First we discuss rule-based classification models that consider the use of grammatical gender when it is realized as referential gender of first person singular references. The second type of models are statistical models that employ machine learning algorithms to learn to distinguish between the classes based on various features. Several models are built and compared given the algorithm and features. Furthermore, we focus on the best performing model by observing the features ranked as most informative for the female and male class of Twitter users.

5.1.1 Experimental Setting

For the purpose of automated gender prediction, a subcorpus of the original Janes corpus of tweets (see Section 3) was built. This section describes the construction of the tweet subcorpus. Moreover, we present the setting of the rule-based and statistical classification.

The original Janes corpus of tweets was modified for the task of gender prediction (see Section 3 for the description and preprocessing of the original corpus). A tweet subcorpus was created comprising tweets by private user accounts of female or male gender that have at least 10 tweets in Slovene in the original Janes corpus. The subcorpus statistics are presented in Table 5.1. For the subsequent experiments with gender prediction models, the profile-based approach (Stamatatos, 2009) was used, meaning that the tweets of a single user were concatenated into one document.

Table 5.1: Tweet subcorpus statistics: female and male private users.

	Users	Tweets	Tokens
Female	1,985	1,779,066	20,816,523
Male	4,218	3,564,378	43,244,116

For the rule-based classification, the non-lemmatized version of the subcorpus was used, whereas both the non-lemmatized and the lemmatized versions of text were experimented with for statistical models. Furthermore, special attention in terms of preprocessing was given to the following elements specific to Twitter messages:

- Hashtags: Hashtags are symbols that are comprised of the hash sign "#" and a phrase. Hashtags often indicate the topic of a tweet. They can function as sentence elements or can be added to the tweet text without syntactic dependency.
- User mentions: Users can mention each other in tweets by linking the at sign "@" and the handle (username) of another user. This often occurs in conversations when particular users are addressed.
- Web links: Addresses of web pages (URLs) are often included in tweets.

These Twitter-specific communication elements can generally be handled in three ways: complete removal, substitution by a constant string, or they are left intact in the tweet text. When substituted, the constant strings serve as replacements, e.g.: "username" for user mentions, "#hashtag" for hashtags, and "URL" for web links. Based on preliminary experiments with preprocessing of these Twitter-specific elements, we opted for their complete removal in the construction of gender prediction models.

5.1.2 Rule-Based Models

The tweet corpus in its tokenized and non-lemmatized version was used, as the classification model relies on the occurrence of verb participles, which is a non-finite verb form. See Section 4.3 for the methodology of the rule-based model.

The results of the classification are presented in Table 5.2. The table provides the classification accuracy scores achieved on tweets given the minimum number of gender indicators and the node words used.

The best classification accuracy (68.56%) is achieved when a minimum of 3 gender indicators occur per user and the non-standard spellings are included in the node words. However, this optimal setting outperforms the majority class vote by only 0.57% or around 35 users from the corpus of 6,203 users. To compare the performance for the female and male class, detailed results of the optimal setting are presented as a confusion matrix in Table 5.3.

From Table 5.3 several observations can be made. The majority of female (68.82%) and male (68.44%) Twitter users were correctly classified by the rule-based model, thus

Table 5.2: Results of author gender prediction by classification rules on tweets.

Indicator minimum	Node words	Classification accuracy
3	sem, nisem, bom	66.94%
3	sem, nisem, bom, sm, nism	**68.56%**
5	sem, nisem, bom	57.65%
5	sem, nisem, bom, sm, nism	59.65%
Majority class		67.99%

Table 5.3: Confusion matrix for the optimal setting of the rule-based model on tweets.

Predicted	Actual	
	female	male
female	1,366	12
male	26	2,887
undefined	593	1,319
Total	1,985	4,218

achieving a high recall. Interestingly, only 0.2% of male users were given the female label, while 1.3% of female users were classified as male. The model made most erroneous classifications in labeling about a third of each class as undefined: 31.27% of male users and 29.87% of female users. We are interested in the possible reasons for incorrect classification into the opposite or undefined class. First, the incorrectly classified male users were analyzed. A small number (12 or 0.28%) were assigned the female class. For these users, well over 200 tweets were included in the corpus. The verb phrases in their tweets consisting of an auxiliary verb and an l-participle were observed in context. A closer analysis shows interesting patterns in the text. One of these erroneously classified users consistently uses the feminine referential gender when referring to self (*sem mislila* [I thought], *sem videla* [I saw], *sem ji dala* [I gave it to her]), and the username may also suggest that the user might be female, meaning the manual annotation is probably incorrect. Two other users from the female class classified to the male class have consistently used the female referential gender for self-references. From the tweets, it is evident that one is referred to with nouns in feminine forms (*mami* [mommy]). For the other user, the profile was searched for on the Twitter web page. In this user's short description or "bio", the user is described with a noun in female form *bralka* [reader]. Among the users classified as female, 4 seem to use very few gender markings in verb phrases, but an overview of the descriptions in their profiles shows that they refer to themselves with feminine nouns (*oma* [grandma], *avtorica* [author], *Evropejka* [European]). A single male user who was assigned the female class has few tweets (25) displays a feminine referential gender; however, a closer reading of the tweets shows that the gender markings are used in citations of female speakers (e.g., of female patients, a female hotel receptionist, and a mother). The username suggests that the user is male (a Slovene male name).

Over a third of male users (1,319 or 31.27%) were classified by the rules as "undefined". For a great majority of them (1,336), the rules found very few gender markings for both the female and male class. However, a few users stand out, as they refer to themselves with verb phrases containing an auxiliary verb, but the spelling of the verb l-participle is nonstandard (*sm ji odgovoriu* [I replied to her], *kloniro se bom* [I will have myself cloned]), which is not detected by the classification rules. By reading these users' tweets, it is evident that they indeed refer to themselves with masculine nouns (*nism super-junak* [I'm not a

superhero]) and adjectives (*bom star* [I will be old]).

Among the 1,985 female users in the corpus, 26 or 1.31% were classified as male. Even though the number of tweets per each of these incorrectly classified users is high, a greater number of male rather than female indicators are found in their automatically generated tweets, which link their Twitter accounts to other social media, such as Facebook or YouTube. In the automatically generated tweets, the masculine referential gender of verb l-participles is used as neutral (*Objavil sem novo sliko na Facebooku* [I posted a new photo on Facebook] and *Na seznam predvajanja @YouTube URL sem dodal videoposnetek* [I added a video to a playlist @YouTube URL]). For the users with over 50 male indicators, a closer reading of user-generated tweets revealed that they refer to themselves in feminine forms of the verb l-participles (*sem se kopala* [I bathed]) and adjectives (*sem prav žalostna* [I'm really sad]), but the number of automatically generated tweets is greater.

The classification model assigned the "undefined" label to almost a third of female users (593 or 29.87%). A small group (15) of these users displays a high number (12 to 65) of female and male indicators, but the ratio between them is lower than 0.7. In their tweets, the masculine referential gender in verb phrases is again used in automatically generated messages linked to social media. In some tweets, the masculine referential gender occurs in quotations of male speakers mostly within headlines of news articles, whereby a male politician or some other male public personality is cited and the quotation is followed by an URL (e.g., *Janša: Kot predsednik SDS se bom boril za to, da stranka zmaga na volitvah URL* [Janša: As president of SDS I will fight for the party to win the election URL]). Furthermore, we observed that the tweets by female users were classified as undefined when few verbal gender indicators were found by the rules. They display few references to self with verb l-participles, but express gender in feminine forms of nouns (*kot da sem ena kriminalka* [as if I'm a criminal]), adjectives (*sem ogorčena* [I'm outraged]), or the l-participles outside the window (*sem ga že 2x gledala* [I have already watched it 2x]). A closer observation of verb phrases shows that sometimes only a part of the verb phrase occurs, as the auxiliary verb is omitted (*pripravila mapo ...* [prepared the folder...]), but the reference to self can be assumed from the context.

5.1.3 Statistical Models

For the task of automated gender prediction with statistical models, we tested several settings with regard to text form (non-lemmatized and lemmatized). Furthermore, three machine learning algorithms were compared based on their classification accuracy: Support Vector Machine (SVM), Logistic Regression (LR), and Naïve Bayes (NB). Various types of features were applied:

1. Word n-grams: uni- and bigrams;
2. Character n-grams: (2–4)-grams;
3. External word lists: for each list a feature was constructed, whereby the share of the words from the list occurring in the author's document was inserted as the feature value using Equation 4.6;
4. Feature union: experiments included single feature types and feature unions.

For the word and character n-gram feature vectors, the TF-IDF weighting scheme was applied. The feature space was reduced by first setting the minimum document frequency to 5, and the maximum document frequency to 80% (i.e., occurring in tweets a of minimum 5 and a maximum of 4,962 users[1]). The SelectFromModel (see Section 4.1) feature selection

[1] These parameters were set after preliminary experiments with no document frequency restrictions.

was also applied. Previous experiments with the preprocessing of Twitter-specific elements have shown that the deletion of these elements results in better classifier performance than substitution with a constant string or no preprocessing.

The average classification accuracy results, together with the standard deviation from 10-fold cross validation, are shown in Table 5.4. The majority vote classifier was built using the DummyClassifier within Scikit-learn.

Table 5.4: Classification accuracy ± standard deviation scores obtained from 10-fold cross validation in gender prediction experiments with statistical models using various features, text forms, and three algorithms: Support Vector Machine (SVM), Logistic Regression (LR), and Naïve Bayes (NB). The majority vote classifier is provided as a baseline.

Feature	Form	SVM(%)	LR(%)	NB(%)
word unigram	token	**90.26±1.13**	86.31±1.35	69.43±1.08
word uni- and bigram	token	89.63±0.72	84.67±1.27	68.63±1.92
character (2-4)-gram	token	89.31±1.22	83.94±1.33	68.61±1.54
word uni- and bigrams, word lists	lemma	83.70±1.34	77.56±1.91	67.99±2.34
word unigrams	lemma	86.12±1.14	83.88±1.39	69.98±1.77
word uni- and bigrams	lemma	85.99±1.19	81.61±1.12	68.18±1.65
word lists	lemma	68.21±2.24	68.21±1.43	67.99±1.67
Majority vote classifier				67.99

As can be seen in Table 5.4, the Support Vector Machine (SVM) algorithm generally performs better than Logistic Regression (LR) and Naïve Bayes (NB). SVM achieves the classification accuracy between 68.21% (on discursive word lists) and 90.26% (on word unigrams), while LR follows with a performance between 68.21% (on discursive word lists) and 86.31% (word unigrams). Both SVM and LR outperform the majority vote classifier in all settings tested, whereby the setting with word lists as the only features outperforms the majority vote by only around 0.20%. In contrast, NB performs poorly and achieves an accuracy close to the baseline. When combined with word unigrams on the token (69.43%) or lemma level (69.98%), it achieves slightly better scores.

From the table it follows that the token-based features (word and character n-grams) have a more beneficial e□ect on the model performance than the lemma-based n-grams. Among the token-based features, word unigrams serve as the best features for SVM and LR. In the setting with token-based word unigrams, the highest accuracy is achieved, as the SVM performs with an accuracy of 90.26% ± 1.13%. SVM outperforms other algorithms in the setting with word uni- and bigrams, and character (2-4)-grams as well.

The SVM and the LR models perform best on token-based word unigrams with the accuracy scores of 90.26% ± 1.13% (SVM) and 86.31% ± 1.35% (LR). When the feature set is expanded to token-based uni- and bigrams, their performance decreases only slightly: by 0.63% (SMV) and by 1.64% (LR). This performance drop is low and possibly occurs due to overfitting of the models, as the size of the feature set of uni- and bigrams exceeds the size of the set with unigrams only. A similar phenomenon is observed when comparing the performance of SVM and LR on lemma-based word unigrams with the performance of these algorithms on lemma-based word uni- and bigrams.

The classifier performance does not seem to benefit from feature unions of word and character n-grams and word lists. We can expect that there is an overlap of features within these unions, so a feature selection method was applied, but the best performing setting with a feature union achieves 83.70% with SVM, which is lower than the best performing setting by 6.56%. Therefore, the experimental results suggest that a word

unigram representation of non-lemmatized text provides enough differentiating features to distinguish between female and male Twitter users for 90.26% of users in our dataset.

5.1.4 Most Informative Features

In the previous sections, the gender of Twitter users was predicted with rule-based and statistical models. In the experiments with statistical models, we compared the performance of various features, text forms and algorithms to find the best setting for distinguishing between female and male users. In this section, we focus on the features that serve as most informative (see Section 4.2.2.2) to the model with the highest classification accuracy score (90.26% $\pm$ 1.13%): the Support Vector Machine (SVM) algorithm used on word unigrams as features from non-lemmatized text, after the Twitter-specific elements were removed from text.

Each feature (in this case, each word unigram) is represented as a weight coefficient, whereby the features more distinctive for the male class have a negative value and the ones for the female class have a positive value. The top-ranked male feature has a coefficient of -2.56, and the 1,000th male feature has a coefficient of -2.75. The feature most informative of the female class has a coefficient of 3.74, and the 1,000th female-related feature has a coefficient of 0.309. The informative value of a feature correlates with a larger absolute value of its coefficient. We extracted a ranked list of 1,000 most informative features per class. Among the topmost features for both categories, word unigrams related to feminine and masculine referential gender can be found.

The most prominent are the verb l-participles: they take up over 14% of the 1,000 features in the features relating to the male class and about 13% of features associated with the female class. On the top of the male list are the l-participles of common verbs, e.g., *videl* [saw], *mislil* [thought], *bil* [was], *imel* [had], *vedel* [knew], and *gledal* [watched]. Examples of the non-standard spelling of l-participles with the reduction of the letter "l" occur: *mislu* [thought], *vidu* [saw], *vedu* [knew], *mogu* [could], *naredu* [did], *meu* [had]. Aside from l-participles, masculine forms of adjectives signal the male class: *ponosen* [proud], *vesel* [happy], and *prepričan* [sure]. For the female class, the topmost l-participles are also common verbs, e.g., *mislila* [thought], *rekla* [said], *bila* [was], *gledala* [watched], *prebrala* [read], and *dobila* [got]. Interestingly, some of the top ranked adjectives in the female list (*vesela* [happy], *ponosna* [proud], *prepričana* [sure]) are the same as in the male list, except they take the feminine form.

Even though the list comprises word unigrams that are not presented in context, tendencies in dominating topics can be observed in both classes. In the male list of features, three prominent (politics, technology, and sports) and some minor ones stand out. There is a large number of words relating to politics, whereby the words denote (un)official political bodies (*koalicija* [coalition], *oblast* [power], *levica* [the left], *stranka* [party], *desnica* [the right]), political systems (*demokracija* [democracy], *socializem* [socialism], *kapitalizem* [capitalism]), affiliations (*levičar* [leftist], *komunist* [communist]), state mechanisms (*davki* [taxes], *volitvah* [election], *referenduma* [referendum]) and Slovene politicians (*janković*, *bratušek*, *janše*, *cerarjeva*).

There are also references to foreign politics (*evrope* [Europe], *siriji* [Syria]) and terrorism (*isis*). The second topic present among the features associated with the male class is that of technology and motoring. More specifically, the words refer to computers and the Internet (*server*, *spam*, *internet*, *aplikacija* [application], *spletne* [web]), brands (*apple*, *nokia*, *samsung*), types of mobile phones and other gadgets (*iphone*, *nexus*, *ipad*, *gopro*) and operating systems (*android*, *ios*, *windows*, *osx*). There is also a number of car and car brand mentions (*avto/avtomobila* [car], *honda*, *ferrari*, *bmw*). Two words relate to playing video games (*ps4*, *gta*). Two other topics associated with entertainment are evident from

the features: sports and drinking. The sports vocabulary can be related to footbal (*nk* [FC], *bayern*, *nogomet* [football], *juve*, *fifa*), other sports (*f1*, *mtb*, *biathlon*), general references to spectator sports (*prvenstvo* [championship], *finale* [finals], *sodnikov* [judge], *derbi* [derby]) and Slovene sportspeople (*katanec*, *zavec*). In the topic of drinking, beer (*pivo/pir* [beer], *beer*), bars (*bar*, *gostilna* [pub]), and the state of being drunk (*pijan* [drunk]) are mentioned. As another minor topic, several references to female partners and women occur (*punca/punco/puncam* [girl or girlfriend], *zeno/ženo* [wife]) and a single kinship term (*strici* [uncle]). An observation of the use of *strici* [uncles] discloses that in a number of tweets by male users, the word does not denote a family member, but is rather used as a reference to people with political power who control the political decision making, even though they have no o□cial political function. In the tweets by male users, this term often occurs in phrases *strici iz ozadja* [uncles from the background], or *rdeči strici* [red uncles] and is usually related to members of the former Communist Party of Slovenia.

The list of features most informative for the female class also displays words relating to political topics, such as political bodies (*vlada* [government], *parlamenta* [parliament], *sodišče* [court], *ministrstvo* [ministry]) and Slovene politicians (*jankovic*, *kangler*, *fistravec*, *janša*). A di□erentiating aspect within the topic of politics is the features that are mentions of the refugee crisis by female users (*begunci* [refugee], *rigonce*).

Interestingly, many feminine forms of political functions occur in the female list (*županjo* [mayor], *poslanka* [MP], *podpredsednica* [vice president], *kandidatke* [candidate]), which might suggest that women mention more female politicians or that men use more generic masculine forms to refer to the function, but this requires support in further analysis. Another prominent topic is associated with food and beverages (smoothie, *kava* [co□ee], *orehe* [walnuts], *jagode* [strawberries], *večerja* [dinner], *paradižnik* [tomato]) with several mentions of sweets (*čokolada/čokoladno* [chocolate], *sladkanje* [eating sweets], *torta* [cake]). While kinship terms and general references to other people appear rarely in the male-related features, they take up an important part in the female feature list, where we can find mentions of family members (*babi* [grandma], *dedek* [grandpa], *otroki/otroški* [children], *starše/staršev* [parents]), men, women, and partners (*moz/moski* [man], *zenske/ženski* [woman]). A small number of words refer to education (*učitelje* [teachers], *vrtcih* [kindergarten], *šole* [school]). The vocabulary on the female list indicates two other minor topics on clothing and fashion (*čevlji* [shoes], *fashion*, *uhan/uhani* [earrings], *modna* [fashion], *torbice* [bag]), and nature and holidays (*julijci* [the Julian Alps] and *morja/morje* [sea]).

While common and proper nouns and adjectives suggest topical variation that assists in distinguishing between female and male Twitter users, the list of features additionally displays contrasting stylistic traits that are not bound to content. Pronouns appear in both feature lists, but they di□er in person and number. The personal and possessive pronouns among male-related features take the second person singular (*teboj* [you], *tvoj* [your]) and first person plural (*nami* [us], *našem* [our]). This changes when we observe the female list, as the pronouns appear in first person singular (*jaz/jz/js/meni/mene/zame* [I/me], *moji/mojega* [my]), dual (*midve* [the two of us], *najine* [our]), and plural (*nam* [we], *naše* [our]). Furthermore, personal pronouns refer to the second and third person (*vaša* [your], *jim/njimi* [them], *zanj* [for him]) and some impersonal and reflexive pronouns also appear (*vsi* [everyone], *nekih* [some], *vsak* [every(one)], *nekdo* [someone], *zase/sabo* [-self]), *svojega* [own]).

An important part of the female-related stylistic features comprises interjections, which are often spelled with repeating characters (character flooding). The interjections denote laughter (hihihi, hahahaha), agreement or excitement (*jaaaa/ja/jaaaaa/jaa, yessss/yes/jeee/jeeee/jp/jaaa/yeppp* [yeees], *juhej* [yippee], hura), worry (*joj/ajoj/ojoj/ojej* [oh

no]), admiration (*wauuuu/wau/vau* [wow], *bravoooo*), disgust (*fuj* [ew]), tenderness (*ii/iiii/iii* [aww]), pleasure from eating (*njami, nomnom* [yummy]) and negation (*nee* [noo]). Several interjections in the female list indicate replies or dialogue in general: *hej* [hey], *adijo* [bye], *hvala/tenks/ hvalaaa/tenkju* [thanks], *xo/xoxo*. There are also a number of interjections with character flooding that can have multiple meanings: *oo/ooo/oooo/ooooo* [oh], *mmm/mmmm* [hm], *uuu* [oh], *aaa* [argh].

A small number of interjections also defines the male class, whereby these also denote laughter (*hahahah*, *heh*) and agreement (*ok*, *jep* [yup]), or function as filler words (*hja* [well], *ajga* [yo]). Intensifiers are another expressive female-related feature: *ful* [really], *itak* [of course], *najbolj/najbl* [the most]. The most marked vocabulary of the male features are negatively connotated words, namely, example of profane language which ranges from vulgar references to sex (*jebeš/jebiga/jebes* [fuck it], *fak* [fuck], *jebejo* [they fuck]) and body (*pizda* [cunt], *kurac/kurc/kurcu* [dick], *prdec* [fart], *ass/riti/rit* [ass], *drek/sranje* [shit]), as well as insults (*kreten* [idiot]).

Adverbs are a strongly represented part of speech among the male-related features. This is especially true for temporal adverbs that describe frequency (*sedaj* [now], *zopet/ponovno* [again], *enkrat* [once], *nikoli* [never], *zmeraj* [always], *hkrati* [at the same time]) or a fixed point in time (*letos* [this year], *nocoj* [tonight], *nazadnje* [last time], *nekoč* [once]). A smaller number of temporal adverbs appear in the female list as well, again some referring to frequency (*vedno/skoz/skos* [always], *vsakič* [every time]) and a fixed point in time (*jutri* [tomorrow], *zdaj/zdej* [now]). Subordinating conjunctions (*dokler* [until]), as well as coordinating conjunctions (*ter* [and], *toda/vendar* [but], *oziroma* [or], *kajti* [because]) define the male class, while only two such examples (*zato* [so], *preden* [before]) are found in the female list.

5.2 Genderlect Analysis Based on Discursive Features

The central premise of genderlect studies (Tannen, 1990) is that men and women use the same language, but variations in terms of style, content, and communication goal that are associated with one's gender also occur. These variations are named genderlects (see Section 1.1). In the previous section, the di□erences between female and male language were examined by automatically distinguishing between male and female authors of Slovene tweets based on 1) use of grammatical gender as conditions in classification rules, 2) use of word and character n-grams from the text as features for statistical models.

In this section, we focus on the variation in vocabulary that is less bound to topic and more to the writing style of a particular blogger. It has been suggested by Pennebaker (2013) that language style can reveal traits of people's "personality, social connections and psychological states", whereby style is not represented in content words (nouns, regular and action verbs, most modifiers), but rather by function words (pronouns, articles, adpositions, auxiliary verbs, negations, conjunctions, quantifiers, and common adverbs). Furthermore, Newman et al. (2008) show that the di□erences between female and male speaking and writing have certain patterns across genres: male speakers/authors use more longer words, nouns, prepositions, numbers, swear words and longer sentences. In turn, women use more personal pronouns, verbs, references to negative emotions, and hedge phrases.

This section aims to recreate the experiments of Pennebaker and his associates (Pennebaker, 2013; Tausczik & Pennebaker, 2010; Newman et al., 2008), who used pre-defined dictionaries and the LIWC system (see Section 2.2) to analyze the psychological profile of English-speaking authors. Our goal is to use our own word lists to validate the findings of related work on Slovene tweets by female and male users. This is done by comparing the use of words that indicate a writing style, such as profane language (swear words and

vulgarisms), expressions of emotions (emoticons, words with a positive or negative sentiment), markers of intensity (intensifiers and hedges), and others (for the construction and content of word lists, see Section 4.5.1).

5.2.1 Methodology and Experimental Setting

In this section, we compare the realization of various language styles in tweets produced by female and male authors by computing the Mann-Whitney U-test to test if the differences in the use of each style are statistically significant. Next, we compute Pearson's correlation coefficient and its squared value to measure the effect size (see Section 4.5).

For the experiments in target word usage based on style and discourse, we used the Slovene tweet corpus. More specifically, we used the lemmatized version of the corpus presented in Section 5.1.1, whereby the Twitter-specific communication elements were left in the corpus and not replaced with a constant string. All the testing of statistical significance is computed with an alpha level of 0.05.

5.2.2 Results of the Statistical Analysis

This section provides the results of the statistical analysis of how the words that mark a particular writing style are used by female and male Twitter users. We tested 1) whether the difference in use is statistically significant (using the Mann-Whitney *U*-test), 2) the correlation between word lists and gender (using Pearson's correlation coefficient), and 3) how much variance in use of these words can be explained by the gender of the user (using squared Pearson's correlation coefficient). The results are presented in Table 5.5. The word lists in the table are ordered by the ascending *p*-value obtained from the Mann-Whitney *U*-test.

Observing the statistics of the Mann-Whitney *U*-test, the word lists produce varying results in terms of statistical significance. The following eleven word lists display a statistically significant difference between female and male users ($p < 0.05$): *Emoticons and emojis*, *Emoticons*, *Emojis*, *Negative words*, *Function words*, *Profanity*, *Intensifiers*, *Positive words*, *New words*, *Emotional words*, and *Hedge words*. The shares of words from the rest (*Social words*, *Negation*, *Non-standard words*, *Modal verbs*, *Cognition verbs*, *Janes glossary*, and *Communication verbs*) are not used statistically significantcly different by female and male users.

As we can also see from Table 5.5, the sign of Pearson's correlation coefficient signals a correlation to the female (negative coefficient) or male class (positive coefficient). The list with the highest correlation coefficient is the *Emoticons and emojis* list that signals a low correlation of 0.1132 to the female class, whereby 1.28% of variance in the use of emoticons and emojis between female and male users is explained by the gender variable. Other lists indicate a statistically significant correlation to the female class (*Emoticons*, *Emojis*, *Function words*, *Positive words*, and *New words*), however their correlation coefficient is low and the user gender accounts for less than 1% of the variation in the use of these words. Three lists (*Negation*, *Cognition verbs*, and *Janes glossary*) display a correlation to the female class, but their p-value is greater than 0.05.

The list signaling the strongest correlation to the male class is the *Negative words* list; however, this correlation is low and only 1.02% of variance in the use of words with negative sentiment can be explained by the user gender. Similarly, other male-correlated word lists with statistically significant coefficients achieve a low correlation and effect size: *Profanity*, *Intensifiers*, and *Emotional words*, whereby the variance explained equals to 0.12% or less. In contrast, the correlation to the male class does not display a statistically significant

difference for the following lists: *Hedge words*, *Social words*, *Non-standard words*, *Modal verbs*, and *Communication verbs*.

Table 5.5: Comparison of word use in female and male tweets. The results include test statistics and *p*-value of the Mann-Whitney *U*-test and the point-biserial correlation coefficient (r_{pb}) and *p*-value and r_{pb}^2 as an effect size measure. A positive r_{pb} value indicates a correlation with male users, while a negative r_{pb} signals a correlation with female users.

Word list	**Mann-Whitney**		***rpb***		
	U	*p*-value	*rpb*	*p*-value	r_{pb}^2
Emoticons + emojis	3516810.5	4.4096e-37	-0.1132	3.8954e-19	0.0128
Emoticons	3656300.0	6.1996e-28	-0.0921	3.6326e-13	0.0084
Emojis	3844544.5	1.2312e-20	-0.0710	2.1804e-08	0.0050
Negative words	3612336.5	1.3074e-18	0.1010	1.5493e-15	0.0102
Function words	3891706.0	3.7570e-06	-0.0547	1.6449e-05	0.0030
Profanity	3966116.5	0.0001	0.0342	0.0070	0.0012
Intensifiers	3966116.5	0.0005	0.0296	0.0197	0.0009
Positive words	3973565.0	0.0006	-0.0434	0.0006	0.0019
New words	3979768.5	0.0007	-0.0303	0.0169	0.0009
Emotional words	3992074.0	0.0016	0.0260	0.0404	0.0007
Hedge words	4034112.5	0.0061	0.0113	0.3739	0.0001
Social words	4089628.5	0.0701	0.0072	0.5712	5.184e-05
Negation	4104683.5	0.1014	-0.0189	0.1374	0.0004
Non-standard words	4107031.5	0.1072	0.0236	0.0628	0.0006
Modal verbs	4155457.0	0.1129	0.0193	0.1294	0.0004
Cognition verbs	4185123.5	0.3505	-0.0183	0.1495	0.0003
Janes glossary	4174355.0	0.4274	-0.0027	0.8286	7.290e-06
Communication verbs	4185747.0	0.4560	0.0073	0.5653	5.329e-05

5.2.3 Visualization of Statistically Significant Differences

This section presents the visualization of how various writing styles are used by female and male Twitter users given the occurrence of target words from word lists that represent a particular type of discourse. In Section 5.2.1, these word lists were used in the gender prediction model as features, while Section 5.2.3 evaluated the differences in the use of target words statistically by applying the Mann-Whitney *U*-test and Pearson's correlation coefficient. Here, the stylistic word lists are interpreted given the visualization in the parallel coordinates plots, where each coordinate (vertical line) represents one word list. The input for plot construction is the share of target words per word total. The plot is used to observe the patterns between and within male and female Twitter users for each coordinate separately and as a whole, as each user is represented by one line and the classes differ in line color.

The visualization in Figure 5.1 includes the word lists that have displayed a statistically significant difference ($p < 0.05$) with the Mann-Whitney *U*-test given the gender of the author (see Table 5.5): *Emoticons*, *Emojis*, *Negative words*, *Function words*, *Profanity*, *Intensifiers*, *Positive words*, *New words*, *Emotional words*, *Hedge words*, and *Emoticons and emojis*. Because the frequency of words from the *Function words* list is greater compared to other lists, it, thus, determines the plot scale, the shares of *Function words* are divided by 5 (this is signaled by the asterisk in the plot). In the plot, the male class is represented with beige, and the female with turquoise lines.

Interesting remarks can be made by observing Figure 5.1. As can be seen in the plot, the lists *Intensifiers* and *Hedge words* display little variation between classes and for both lists, the shares generally score lower than 1%. In the *Intensifiers* coordinate, one male and one female user stand out slightly, which is also true for two male users in the *Hedge words* coordinate.

Some variation between the classes is exemplified in the *Function words*, *Positive words*, and *Emotional words* coordinates. Both classes, female and male, display a range between near 0% and 5.5% for *Function words*, whereby the more frequent users of function words are male and one female user. The pattern is similar for *Positive words*, where the shares of both classes vary between 0% and 4.0% with several male and female users lying near the upper level. The list of *Emotional words* is comprised of words from the *Positive words* and *Negative words* list, and thus reaches high share scores from near 0% to over 4.5% with most of the scores over 2.8% contributed by male users and by three female users.

Additionally, five word lists display interesting within-class use differences with minor or extreme outliers. In the *Negative words* coordinate, we can observe that both classes use between 0% and 1.7% of words with a negative sentiment per word total. The tweets of two male users contain slightly more negative words, while two male outliers use 2.6% and 3.3% of negative words per word total.

Male outliers are present in the coordinates of *Emojis* and *Emoticons*. In the tweets by a majority of female and male users, the emojis are used in 0% to 1.8% per word total. A group of seven female users reaches the share between 1.1% and 1.7%, while an extreme male outlier uses the most emojis in the tweet corpus: about 3.6% of all words he used are emoji symbols. Emoticons display a similar pattern: most users from both classes have the share of *Emoticons* below 1%. A small number of female and male users reach up to 1.5%, but the upper limit is determined by a male user with about 2.5% of emoticons per word total. Interestingly, the account with the most emoticons is not the account with the most emojis. The coordinate *Emoji_emoti* presents the scores for the list of emoticons and emojis. As can be seen from the plot, female users represent the more frequent users of these symbols, but two male users stand out as the most frequent.

Generally, vulgar and swear words do not occur frequently among the users in our corpus, as most have a score of 0.5% or below on the *Profanity* coordinate. Two male and one female author use about 0.7% of profane language per word total, while the user with the greatest share of such vocabulary is a male user reaching 1.3%. The only coordinate with an extreme female outlier is the *New words* list, where a single female user has about 5.1% of novel words per word total. Interestingly, another outlier in this coordinate is male (3.3%), followed by two female users (2.5%) and a group of male users and a single female user (between 0.5% and 1.6%), while the majority of other users from both classes use less than 0.5% of new vocabulary per word total.

Although the presented visualization points out some stylistic variation between female and male Twitter users, there is no clear or uniform variation between the groups. The statistically significant differences detected for eleven word lists also display a low effect size (see Table 5.5). While Figure 5.1 uncovers no strong variation between classes, it indeed shows interesting differences within classes. Some male users deviate from other male users by including more vocabulary with a negative sentiment, profane language, new words, emoticons and emojis than the majority of male users. Some female users also add more emoji symbols to their tweets and thus deviate from the main body of female users in Figure 5.1, while three female users stand out due to their use of new words with one of them marking the largest deviation of a single user in the corpus.

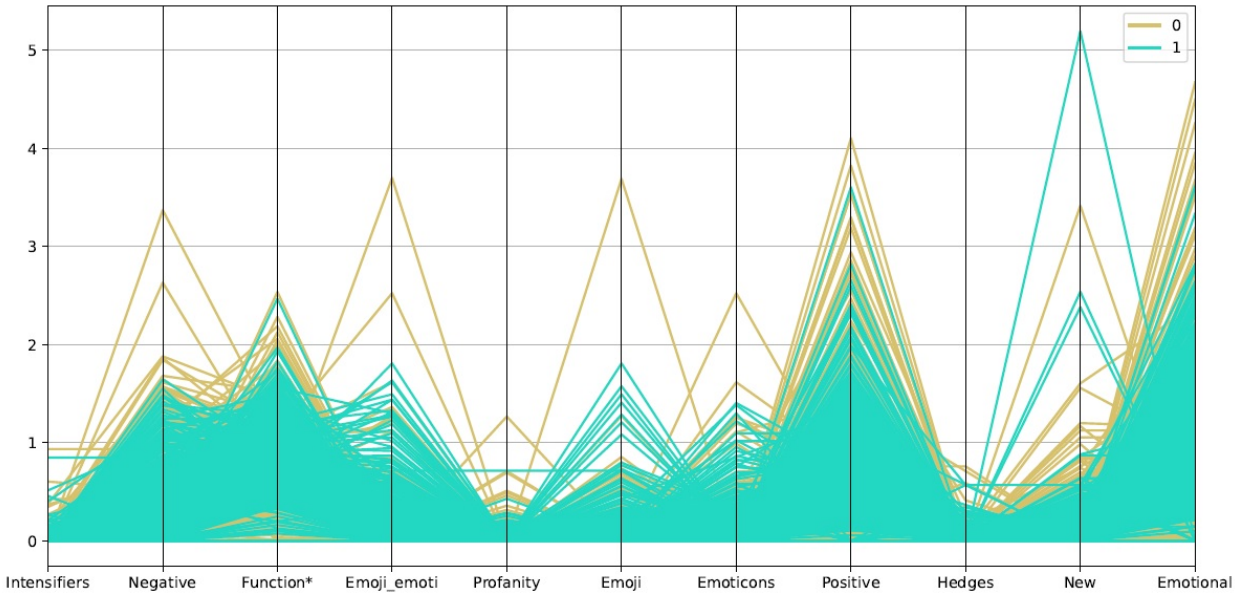

Figure 5.1: Parallel coordinates of the word lists that displayed a statistically significant difference (with Mann-Whitney *U*-test) between the occurrence in female and male tweets. From left to right: Intensifiers, Negative words, Function words*, Emoticons and emojis, Profanity, Emoji, Emoticons, Positive words, Hedges, New words, Emotional words. Beige (0) represents the male class, and turquoise (1) represents the female class. The "*" symbol signals that the scores were divided by 5.

Chapter 6

Analysis of Blog Entries

In the previous chapter, it was shown that various patterns in language carry information about the gender of the text author. More specifically, we observed that grammatical gender as well as general and more specific vocabulary can be used for predicting the gender of Twitter users. In this chapter, the same hypothesis is tested on a dataset of blog entries in Slovene, which is labeled manually with the author's gender. In Section 6.1, we explore the relationship between topical variation and the gender of bloggers by using document clustering to construct topic ontologies. In Section 6.2, we present the rule-based model as a baseline for gender prediction relying on referential gender, and describe the construction and performance of statistical models for gender prediction, where several experiments with features and algorithms are presented. The study presented in Section 6.3 is concerned less with the content and more with the writing style of female and male bloggers. We apply statistical testing to determine if the differences in stylistic choices between female and male bloggers are statistically significant.

6.1 Topic Ontologies of Slovene Blog Entries

This section focuses on the topics that are frequently covered by Slovene bloggers and attempts to detect the topical differences that occur given the gender of the author. This comparison is performed with the use of the OntoGen ontology editor (see Section 4.4). OntoGen generates hierarchical topic ontologies by employing the *k*-means clustering algorithm. This results in the identification of subtopics for each topic, enables the user to be involved in the process of ontology construction, and provides a visualization of the constructed ontologies. For each gendered group (female and male bloggers), a topic ontology is constructed, and its visualization allows for a direct comparison of topics covered by women and men in their blog entries. Furthermore, we construct a common topic ontology including entries by both blogger groups to compare the popularity of each topic among women and men. The results of topical studies are partially described in a published paper (Škrjanec & Pollak, 2016).

6.1.1 Experimental Setting

This section presents the experimental setting of the topic ontology construction using the OntoGen tool (see Section 4.4) and the Janes blog corpus (see Section 3.1). Prior to ontology construction a subcorpus of blog entries was prepared. The lemmatized form of blog entries in Slovene that contain at least 100 words after stopword and punctuation removal are included into the subcorpus. The subcorpus comprises 3,771 entries by private female bloggers and 9,039 entries by private male bloggers. The subcorpus was parsed so

that each entry was written in a single line. Because OntoGen cannot process diacritics and other special characters, these were replaced by recognizable sequences (*č – cx*, *š – sx*, *ž – zx*, *đ – dzx*, *ć – cx*).

In the OntoGen tool, each document is represented in the BOW-format (see Section 4.1). The parameters for the BOW construction were set to word uni- and bigrams with the minimum document frequency of 10 for the gender-specific ontology, and 20 for the common ontology. The features were weighted using the TF-IDF scheme. This means each document was represented as a feature vector of TF-IDF weights of word uni- and bigrams that appear in at least 10 or 20 entries in the gender-specific and the common ontology, respectively. Once the corpus was imported into OntoGen, we experimented with several values of the k-parameter to build a number of k topic or subtopics (see Section 4.4 for the tool description). The tool made k topic suggestions, whereby each topic was represented by a set of keywords that helped us to put a label on it (e.g., the keywords *otrok* [child], *mama* [mother], *starš* [parent], *oče* [father], *družina* [family], *sin* [son] represent the *Family and parenthood* topic). Following this procedure, we imported three subcorpora into OntoGen: entries by female bloggers; entries by male bloggers; and entries by both groups. In the following sections, we present the topic ontology of each blogger group.

6.1.2 Topic Ontologies of Blog Entries by Female and Male Authors

First, the dataset with female- or male-only blog entries was imported into OntoGen. The topic ontology consisting of entries by female bloggers (Figure 6.1) comprises three main topics (*Health and environment*, *Current a airs*, and *Personal*) and 10 subtopics in total. The topic ontology by male bloggers (Figure 6.2) comprises six central topics (*Culture and entertainment*, *Current a airs*, *Personal*, *Roman Catholic Church*, *Environment and health*, and *Miscellaneous*) with 13 (sub)subtopics.

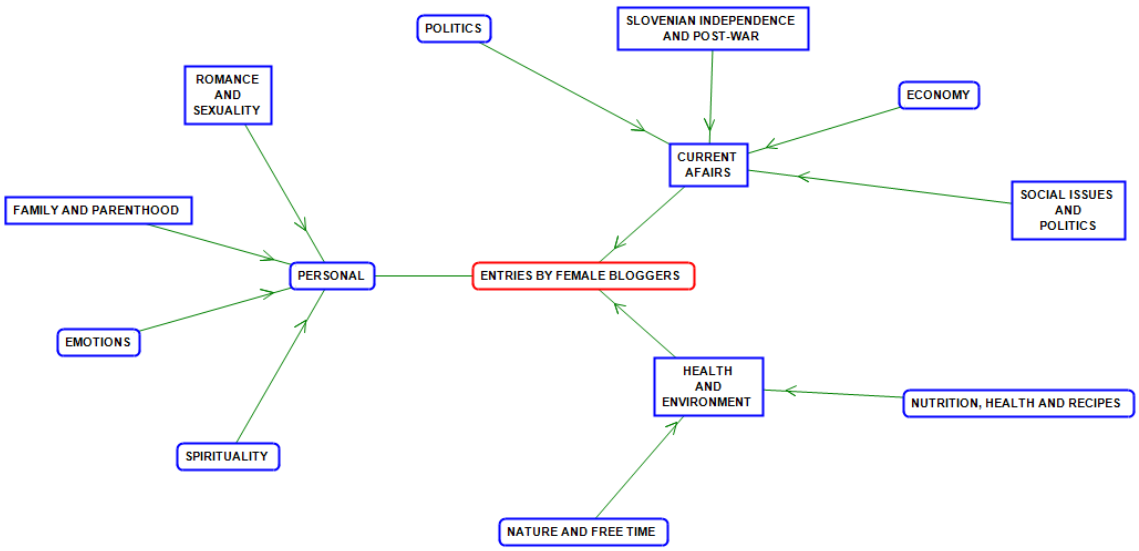

Figure 6.1: Topic ontology of entries by female bloggers.

A topical comparison of blog entries by female and male bloggers (Figures 6.1 and 6.2) shows some interesting similarities and di□erences. Both groups tend to write about the environment, nutrition, family and parenthood, sexuality, and politics, in particular within the subtopic on Slovene politics and the (post)independence era (Slovene Independence

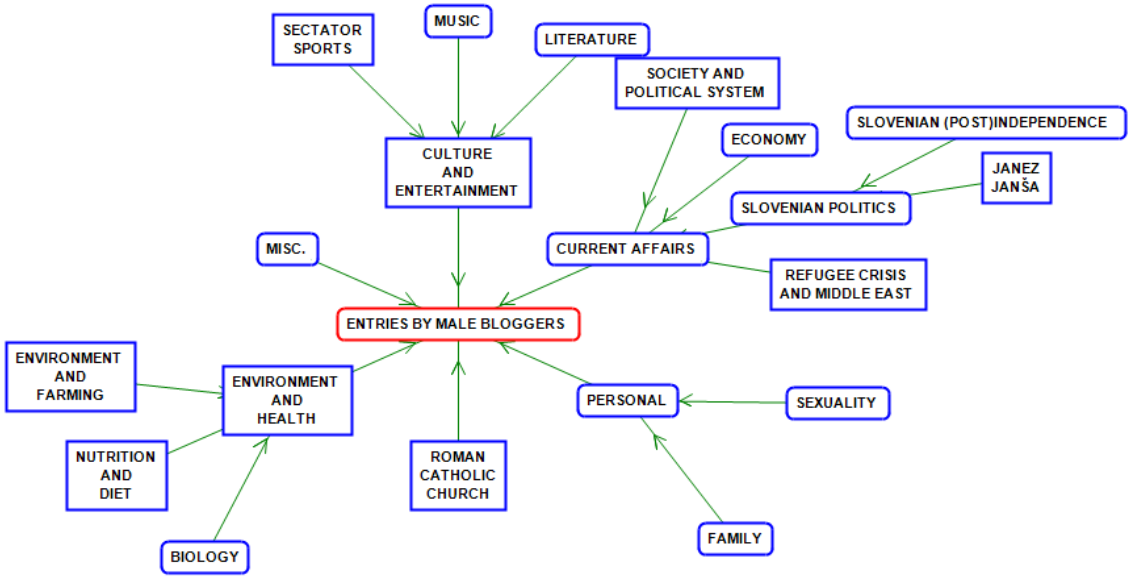

Figure 6.2: Topic ontology of entries by male bloggers.

War, post-war killings, and the role of former communists today). Another common topic is economy, mostly in connection to Slovene and EU economic issues. One of the more prominent topics among male bloggers concerns the Slovene politician Janez Janša and his trial for corruption. An evident topic regarding current a□airs is also the refugee crisis in the male ontology. In contrast to female bloggers, male authors contributed a significant number of entries on biology, spectator sports, music and literature. They also discuss the role of the Roman Catholic Church. In turn, female bloggers write more about spirituality in connection to various religious beliefs and nature. Emotions surface as an independent topic in the female ontology only. Female bloggers pay special attention to social politics and rights.

6.1.3 The Common Topic Ontology of Slovene Blog Entries

In addition to the female- and male-only ontologies, we constructed a topic ontology from the entire blog dataset (including both groups of users). The final form of this common topic ontology is split into 5 main topics (*Personal*, *Hobbies*, *Environment and health*, *Entertainment and culture*, *Politics, economy and society*) and 16 subtopics.

Table 6.1 presents the list of subtopics in the first column. The second column presents the number of entries by female bloggers that were clustered to this topic, as well as the relative frequency of these entries given the total count of female entries. The third column presents these absolute and relative frequencies for male bloggers.

The most popular topic among both groups of bloggers is *Free time and holidays*, while other topics display a divergence between the groups. The most frequent topics written about by female bloggers include *Family and parenthood*, *Romance and sexuality*, *Spirituality*, and *Food and recipes*, while the rest of the topics do not stand out. The blog entries by male authors strongly concern past and present political a□airs: *Slovene Independence and post-war*, *Economy*, and *Janez Janša*.

Table 6.1: Number of blog entries and shares of blog entries contributed to the ontology subtopic given the category total.

Subtopic	Entries by females (%)	Entries by males (%)
Nature and environment	29 (0.77)	60 (0.66)
Recipes and nutrition	276 (7.31)	541 (5.99)
Web and social media	71 (1.88)	183 (2.02)
Free time and holidays	1,199 (31.80)	1,924 (21.29)
Spirituality	348 (9.22)	248 (2.74)
Family and parenthood	697 (18.48)	362 (4.00)
Romance and sexuality	431 (11.43)	329 (3.64)
Slovene Independence and post-war	85 (2.25)	1,058 (11.70)
Current a□airs in politics	85 (2.25)	852 (9.43)
Economy	104 (2.76)	1,091 (12.07)
Janez Janša	106 (2.81)	950 (10.51)
Political system	129 (3.42)	503 (5.56)
Roman Catholic Church	65 (1.72)	303 (3.35)
Music	35 (0.93)	119 (1.32)
Literature	90 (2.39)	176 (1.95)
Spectator sports	21 (0.56)	340 (3.76)
Total	3,771	9,039

6.1.4 Discussion

The results of our comparison of topics covered by Slovene female and male bloggers relate to the findings of previous studies performed on English blogs. Argamon et al. (2007) also found that male authors blog more frequently about politics, business, the Internet, and religion. Concerning religion, Argamon et al. (2007) do not specify what this topic involves. Our analysis shows that female bloggers write about spirituality (9.22% of all entries by female authors), while blog entries by male authors more often deal with the Roman Catholic Church and its role in society (3.35% of entries by males). Argamon et al. (2007) also report that female bloggers in their corpus write about conversation, the domestic environment, fun, romance and swearing. While some of these categories did not stand out in the topic ontologies of Slovene blog entries, romance and the domestic environment display a strong tendency with female bloggers in our dataset as well. A comparable study was conducted by Schmid (2003) who observed the spoken part of the British National Corpus only to find an over-representation of female speakers in the topics concerning clothing, basic colors, home, food and drink, body and health, and people. Again, there is an overlap with our results in the topic of home, as well as food and drink. Schmid (2003) also finds that the domains of work, computing, sports, and public a□airs were shown to be more typical of the male subcorpus. This partially fits the behavior of the male bloggers in our corpus as well, as they display a stronger preference for the topics of sports and public a□airs.

To avoid over-generalization on gendered topics, it is important to take into account the distribution of blog entries among bloggers. Some topics are heavily dominated by a small number of bloggers or blog entries (*Nature and environment*, *Music*), but this is not visible in the ontology. When using quantitative methods to explore gender and language use, there seems to be a tendency towards favoring di□erences, while similarities are pushed into the background, what Baker (2014) calls the “di□erence mindset”. The findings of studies such as this one may suggest and show mostly the di□erences. However,

the language and topics of a single gendered group are not homogeneous, which is what Baker (2014) discovered when he contrasted single-gender parts of the spoken BNC among each other using Manhattan Distance for a list of keywords. He found that some pairs of same-gender parts vary more than pairs of mixed-gender combinations, thus making it di□cult to make broad statements without pinpointing the exceptions to the general rule.

6.2 Models for Automated Gender Prediction

In the previous section, we examined gender-related linguistic variation in blog entries by comparing the content of blog entries posted by female and male bloggers. In this section, we investigate whether the linguistic variation between female and male bloggers can provide predictive information to build a well-performing classification model for gender prediction. For this, we take into consideration the linguistic variation in terms of content, style, and referential gender, and build two types of prediction models. In Section 6.2.1, we describe the preparation of the Janes blog corpus for the classification task. Section 6.2.2 presents the rule-based model that relies on the use of grammatical gender (see Section 4.3). In Section 6.2.3, we provide the description of the experiments with statistical models, several algorithms, and features. In Section 6.2.4, we analyze the features that contributed most to the performance of the most successful statistical model.

6.2.1 Experimental Setting

The Janes blog corpus (Fišer, Erjavec, & Ljubešić, 2016) that is described in Section 3.1.2 was used for building predictive models. A subset of blog posts was collected for the task of gender prediction. Each blogger (author of blog posts) was annotated based on the account type (private or corporate) and gender (female, male or undefined). Since the goal of our experiments is the binary prediction of male or female author gender, only private male and female accounts were included in the experiments. Furthermore, only bloggers with a minimum of 10 blog posts in the Slovene language were included into the subcorpus. The final subcorpus contains 28,697 blog posts with more male (64.84%) than female users (35.16%); see Table 6.2. The entries of each individual blogger were concatenated into a single document, thus employing the profile-based approach to gender prediction (see Stamatatos (2009)).

Table 6.2: Blog corpus statistics: female and male private users.

	Users	Blog entries	Tokens
Female	157	9,056	3,939,315
Male	275	20,105	8,362,668
Female ($\geq$ 10 entries)	96	8,874	3,214,734
Male ($\geq$ 10 entries)	177	19,823	6,968,164

6.2.2 Rule-Based Model

The blog corpus described in Section 6.2.1 was processed by the classification rules in its tokenized and non-lemmatized version. The rule-based model observes the use of verb l-participles in self-referencing phrases and calculates an indicator for each author. See Section 4.3 for the description of the rule-based model. The rules classify the bloggers as *male*, *female*, or *undefined*. The results of the rule-based classification are presented in Table 6.3. As it can be seen from the table, the use of non-standard spellings *sm* and *nism*

does not improve or decrease the model performance. The highest classification accuracy (85.71%) is achieved when the indicator minimum is set to 3, outperforming the majority vote baseline by almost 21%.

Table 6.3: Results of gender identification by classification rules on blog entries.

Indicator minimum	Node words	Classification accuracy
3	sem, nisem, bom	**85.71%**
3	sem, nisem, bom, sm, nism	**85.71%**
5	sem, nisem, bom	80.95%
5	sem, nisem, bom, sm, nism	80.95%
Majority vote baseline		64.84%

For the best performing rule-based model, a confusion matrix was constructed in Table 6.4. As can be seen from the table, 79.17% of female bloggers and 89.27% of male bloggers were correctly classified by the rules. We first observe the bloggers who were assigned the opposite gender. The model classified 4 female users as male. The rules found a very small number (less than 5) of gender markings in the entries by two of these bloggers, both of whom published narratives by male authors in first person singular. Similar is true for two other female bloggers misclassified as male for whom over 40 gender markings were found with the majority of them being male, because they cite longer passages from books or write fiction in first person singular with a male narrator. Only two male bloggers were assigned the female class. A closer reading of their blog entries revealed that the feminine referential gender occurred in longer testimonies of women, whereby again the first person narration was used.

Table 6.4: Confusion matrix for the optimal setting of the rule-based model on blog entries.

Predicted	Actual	
	female	male
female	76	2
male	4	158
undefined	16	17
Total	96	177

The model made most of the erroneous choices for both classes when assigning the undefined class: 16.67% of female bloggers and 9.60% of male bloggers were labeled as *undefined*. Among these 16 female bloggers, 6 display a great number of gender markings that are distributed almost equally among the feminine and masculine referential gender, because their blog entries are either comprised of dialogues as direct speech with female and male participants, or, again, entire blog entries are narrated by a male speaker. The rest of the female bloggers that were labeled as *undefined* seem to use few first person verb phrases with an auxiliary verb and l-participle, as 3 or fewer markings were found in their entries. Among these cases, it rarely occurs that the two components of the verb phrase are separated by more than one word (e.g., *Malo sem po naključju sledila* [I followed a bit by chance]). The referential gender is expressed in adjectives as well (e.g., *a v to sem prepričana* [I'm sure of it]).

Only 17 male bloggers were classified as undefined. In the blog entries by 4 of them, an almost even number of masculine and feminine gender markings in l-participles was found mostly due to direct speech in dialogues. For the rest of the incorrectly classified male users, the rules detected less than 3 uses of referential gender in verb phrases. We

searched their blog entries and generally a low number of self-references was found in verb and adjective phrases, even though at least ten of their blog entries were included in the experiments.

6.2.3 Statistical Models

This section presents the experiments with statistical models for gender prediction. For this task, we tested several settings by using various features and three different algorithms (Support Vector Machine SVM, Logistic Regression, and Naïve Bayes). Additionally, we used both the non-lemmatized and the lemmatized version of the tweet corpus.

The following features were employed:

- Word n-grams: uni- and bigrams;
- Character n-grams: (2-4)-grams;
- External word lists: for each list a feature was constructed, whereby the share of the words from the list occurring in the author's document was inserted as the feature value using Equation 4.6;
- Feature unions: experiments included single feature types, as well as feature unions.

In the experiments that used word and/or character n-grams, the feature space was reduced according to the following criteria: n-grams that were used in the entries of at least 5 and at most 218 (80.0%) authors were included as features. Furthermore, the SelectFromModel feature selection was applied (see Section 4.1). We employed 10-fold cross validation in the training and testing process. Table 6.5 provides the classification accuracy and standard deviation scores from 10-fold cross validation for different combinations of an algorithm, features, and text form. The majority class is provided as a baseline.

Table 6.5: Classification accuracy ± standard deviation scores obtained from 10-fold cross validation in gender prediction experiments with statistical models using various features, text forms, and three algorithms: Support Vector Machine (SVM), Logistic Regression (LR), and Naïve Bayes (NB). The majority vote classifier is provided as a baseline.

Feature	Form	SVM(%)	LR(%)	NB(%)
word unigram	token	**86.85±6.00**	81.67±5.89	64.89±8.21
word uni- and bigram	token	85.29±8.34	79.81±6.56	64.84±6.29
character (2-4)-gram	token	80.58±8.17	78.76±9.77	64.83±6.25
word uni- and bigrams, word lists	lemma	81.06±6.30	74.79±6.92	64.87±9.37
word unigrams	lemma	83.90±7.82	80.57±9.03	65.26±10.89
word uni- and bigrams	lemma	82.42±7.12	78.02±7.81	64.18±1.65
word lists	lemma	72.89±9.50	73.20±9.02	67.76±9.35
Majority vote classifier				64.85%

Several observations can be drawn from Table 6.5. Among the three learning algorithms, Support Vector Machine (SVM) generally performs better than Logistic Regression (LR) and Naïve Bayes (NB). The SVM achieves the best overall result (86.85% ± 6.00%) when token-based word unigrams are used as features. The SVM and LR outperform the majority vote baseline in all the settings. In contrast, NB performs better than the baseline in two settings only: when lemma-based word unigrams are used as features, it outperforms the baseline by 0.31%, and it achieves its highest accuracy (67.76% ± 9.35%) when used with word lists are features; however, the standard deviation of NB scores is

large, so the overall performance of NB can be generally described as poor, as it is close to the baseline.

With regard to different features, token-based word unigrams give the best result for SVM and LR. The performance of these two algorithms drops only slightly (by 1.56% for SVM, and 1.86% for LR) when token-based word bigrams are used in addition to the word unigrams and this can possibly be explained as a case of overfitting due to the large size of the feature space. Character (2-4)-grams give lower accuracy than any other setting with word n-grams for the SVM, while LR performs well on character n-grams with regard to the rest of the scores of this algorithm. For both, the SVM and LR, token-based n-grams serve as better features than lemma-based ones.

The feature union of lemma-based word uni- and bigrams and word lists does not contribute to the performance of SMV and LR much, as both algorithms achieve better results when lemma-based word uni- and bigrams are used without the word lists. When only word lists are used as features, SVM and LR achieve their overall worst accuracy (72.89% $\pm$ 9.50% for SVM, and 73.20% $\pm$ 9.02%). Interestingly, this setting is beneficial for NB, as it achieves its best general score (67.76% $\pm$ 9.35) on word lists.

6.2.4 Most Informative Features

The previous section was concerned with predicting the gender of blog authors based on various feature types of lemmatized and non-lemmatized text using three algorithms: linear Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression. The experimental results showed that the best performing algorithm was SVM on word unigrams as features from non-lemmatized text. Employing feature selection, it achieved the accuracy of 86.85% $\pm$ 6.00%. This section discusses the features ranked as most informative by the best performing model for the gender prediction of blog authors. We extracted a list of 1,000 such features for each class by using the methods described in Section 4.2.2.2.

Among the topmost features for both categories, features related to the grammatical gender of verb l-participles represent a large part: in the male category, 15% of the top 1,000 features belong to verb l-participles in the masculine form (e.g., *šel* [went], *videl* [saw], *dal* [gave]). In the female category, feminine verb l-participles take up 13% of the top 1,000 features (e.g., *imela* [had], *šla* [went], *vedela* [knew], *dobila* [got]).

Aside from l-participles, other parts of speech occur on the female and male list. High on the list for both categories are various adverbs, which can be divided into temporal, spatial, quantifying, expressive, and modal adverbs. Temporal adverbs are more frequent in the female category, whereby they indicate frequency (e.g., *znova* [again], *včasih* [sometimes], *pogosto* [often], *nikdar* [never]). In the male category, temporal adverbs relate more to a particular point in time (e.g., *nocoj* [tonight], *sinoči* [last night], *včeraj* [yesterday]). An interesting variation between the most informative features for the female and male class occur in pronouns. In the female category, most personal and possessive pronouns refer to first person singular (e.g., *moja/mojega/mojih* [my], *mene/zame/menoj* [me]) and first person dual (e.g., *naju/nama* [us], *najin* [our]). In the male category, fewer personal and possessive pronouns are found. In contrast to the female category of the list, they refer to first person plural (*naši/naše* [our]) or third person singular (e.g., *njej* [she], *njegovega/njegov* [his]) or plural (*njihovi* [their]).

While the number of gendered l-participles begins to drop slightly from the 100 feature on, the number of common and proper nouns increases, displaying biases of each category towards particular topics, many of which overlap with the topics detected in Section 6.1. Already on the top of the list of the most informative features, the female category is strongly associated with terms that denote the family and its members (e.g., *otroci/otroka/otroke* [children], *mama/mami* [mother/mommy], *očka* [daddy], *starši* [par-

ents], *sestro* [sister], *družina* [family], *otroštvo* [childhood]). Furthermore, references to romance and sexuality appear on the list (e.g., *spolnost* [sexuality], *ljubček* [love/lover], *zaljubljenost* [being in love], *seks* [sex]). The vocabulary regarding emotions, feelings, and emotional states has a great number of examples referring to either something positive (*ljubezen/ljubezni* [love], *strasti* [passion], *nasmeh* [smile]) or especially negative (*otožnost* [melancholy], *jokala* [cried], *solze* [tears], *samota* [solitude], *zavist* [envy], *žalostna* [sad], *sram* [shame], *sramota* [disgrace]).

The third topic that can be inferred as differentiating between female and male bloggers is food, as several references appear in the female list, e.g., *cvetače* [cauliflower], *zelenjave* [vegetables], *maslo* [butter], *kokosovo* [coconut], *testo* [dough], *penino* [sparkling wine], *cimet* [cinammon]). Issues with regard to health surface as a minor, but distinctive topic (e.g., *zdravljenje* [treatment], *kemoterapija/kemoterapije* [chemotherapy], *rakom/rak* [cancer], *zboleti* [fall ill], *cepiv* [vaccination]).

While few examples from the female list form the topic concerning current and past political affairs (e.g. *demokratske* [democratic], *socialisti* [socialists], isis), this topic connects the vocabulary of the male list the strongest. In the male list, the vocabulary refers to the state and its bodies of power (*država* [state], *sodišča* [court], *volitvah* [election], *vlade* [government], *referendum* [referendum]), political functions (*predsednik* [president], *državljanov* [citizens]), the Church, and different political and economic systems (*demokracija* [democracy], *kapitalizem* [capitalism]). Additionally, the male list comprises a number of terms associated with Yugoslavia (*udba* [the secret police of the Socialist Federal Republic of Yugoslavia], *komunisti* [communist], *jla* [Yugoslav People's Army]), the Second Word War (*nob* [National Liberation Army], *belogardisti* [White Guards]). The words politics and political appear in several forms (*politično/politična/politične/politična* [political]). References to sports occur exclusively on the male list: *ligi* [league], *žogo* [ball], *prvenstvo* [championship], *tekmo/tekem* [game]). In addition to topical variation, a stylistic trait also determines the male category, as examples of vulgar language (*hudiča* [devil], *jebo/jebe* [fuck], *scat* [piss]) and pejorative references to a minority group (*cigan/ciganov* [gypsy]) are found on the list.

6.3 Genderlect Analysis Based on Discursive Features

In this section, we present the analysis of writing style as displayed by female and male bloggers. We compare how these bloggers use the target words grouped into word lists (see Section 4.5.1) and test whether the differences in discourse are statistically significant. For this, we apply the Mann-Whitney U-test. Moreover, we compute Pearson's correlation coefficient to determine which discourses display a positive (or negative) correlation with which gender. We measure the effect of gender on the stylistic variation between the two groups with the squared value of Pearson's coefficient. The results of the statistical testing are provided in Section 6.3.2. Section 6.3.3 visually presents the use of discourses that return a statistically significant difference on the Mann-Whitney test.

6.3.1 Methodology and Experimental Setting

The methodology of the statistical analysis of discourse is presented in detail in Section 4.5. We compare the realization of various language styles in blog entries by female and male authors by computing the Mann-Whitney *U*-test to test if the differences in the use of each style are statistically significant. Next, we compute Pearson's correlation coefficient to measure the effect size. All the testing of statistical significance is computed with an alpha level of 0.05.

For the experiments in target word usage based on style and discourse, we used the Slovene blog corpus presented in Section 3.2. More specifically, we used the corpus version with the lemmatized blog posts of a single blogger combined into one document (female: 96 bloggers; male: 177 bloggers).

6.3.2 Results of the Statistical Analysis

This section presents the results of the Mann-Whitney *U*-test, which was used to test the statistical significance of differences in use for words belonging to a particular type of discourse collected in words lists. The effect size is measured with the squared Pearson's correlation coefficient. Table 6.6 shows the results of statistical tests: the *U*-score and *p*-value of the Mann-Whitney test, and the r_{pb} correlation coefficient and *p*-value of Pearson's correlation computed as a point-biserial coefficient. The words lists in the table are ordered by the ascending *p*-value of the Mann-Whitney *U*-test.

As we can see from Table 6.6, 12 of 16 collections of words are used differently between female and male bloggers in a statistically significant way according to the Mann-Whitney test ($p < 0.05$): *Social words*, *Cognition verbs*, *Janes glossary*, *Positive words*, *Profanity*, *Emoticons*, *New words*, *Non-standard words*, *Function words*, *Negative words*, *Modal verbs*, and *Negation words*. Four (*Intensifiers*, *Emotional words*, *Communication verbs*, and *Hedge words*) do not display a statistically significant variation.

Observing the sign of Pearson's coefficient, we can see that the list of *Social words* moderately correlates with the entries by female bloggers, and gender accounts for 15.4% of variance in the use of words from the list of *Social words*. The following lists display a weak correlation to the female class: *Cognition verbs*, *Positive words*, *Emoticons*, *Function words*, *Modal verbs*, *Negation words*, *Intensifiers*, *Emotional words*, *Communication verbs*, and *Hedge words*.

Word lists with a positive Pearson correlation coefficient indicate a weak association with entries by male bloggers (*Janes glossary*, *Profanity*, *New words*, *Non-standard words*, and *Negative words*). The highest effect size among these male-related lists is achieved by the *Janes glossary* list, whereby gender accounts for 5.0% of the variance in the occurrence of words from this list. The rest of the male-related lists display a low r_{pb}^2.

6.3.3 Visualization of Statistically Significant Differences

This section presents the visualization of target word shares as they occur in blog entries by female and male authors. We use the parallel coordinates plots that allow for the detection of interesting patterns and the observation of individual authors.

The visualization plots are presented as parallel coordinates, whereby each target word list is represented by a vertical line. Each colored line represents one blogger (turquoise lines represent female bloggers; beige lines represent male bloggers). The shares of target words (see Equation 4.6) were imported as the input. The visualization (Figure 6.3) includes the lists where the difference between female and male bloggers was shown to be statistically significant ($p < 0.05$) with the Mann-Whitney *U*-test. The following twelve word lists were included: *Social words*, *Janes glossary*, *Cognition words*, *Positive words*, *Profanity*, *Emoticons*, *New words*, *Non-standard words*, *Function words*, *Negative words*, *Modal verbs*, and *Negation words*. Because the frequency of the words from *Positive words*, *Negative words*, and *Function words* is greater and thus determines the scale's upper limit, the calculated frequencies were divided by five, which is indicated by the "*" symbol in the names of variables. Bloggers from both groups use words from the *New words* list with similar frequency, while one female and one male blogger use them more frequently than others: in entries by both, these words take up 0.3%. The occurrence of verbs from

Table 6.6: Comparison of word use in female and male blog entries. The results include test statistics and *p*-value of the Mann-Whitney *U*-test and the point-biserial correlation coe□cient (r_{pb}) and *p*-value and r_{pb}^2 as an e□ect size measure. A positive r_{pb} value indicates a positive correlation with male users, while a negative r_{pb} signals a positive correlation with female users.

Word list	**Mann-Whitney test**		r_{pb}		
	U	*p*-value	r_{pb}	*p*-value	r_{pb}^2
Social words	4312.0	9.266e-12	-0.393	1.643e-11	0.154
Cognition verbs	5781.0	6.540e-06	-0.280	2.691e-06	0.078
Janes glossary	5972.0	2.535e-05	0.224	0.0002	0.050
Positive words	6227.0	0.0001	-0.203	0.0007	0.041
Profanity	6364.0	0.0003	0.149	0.014	0.022
Emoticons	6691.0	0.001	-0.231	0.0001	0.053
New words	6613.5	0.001	0.151	0.013	0.023
Non-standard words	6650.5	0.001	0.184	0.002	0.034
Function words	6678.0	0.002	-0.157	0.010	0.025
Negative words	6885.0	0.005	0.13	0.031	0.017
Modal verbs	7301.0	0.028	-0.163	0.007	0.027
Negation words	7321.0	0.030	-0.131	0.031	0.017
Intensifiers	8019.0	0.222	-0.056	0.359	0.003
Emotional words	8098.0	0.261	-0.023	0.708	5e-4
Communication verbs	8309.0	0.382	-0.001	0.982	1e-5
Hedge words	8310.0	0.383	-0.002	0.972	4e-6

Modal verbs varies within and between groups: in the entries by the majority of bloggers, modal verbs occur in 0.1% to 1.0% of words. On the bottom, an over-representation of male bloggers can be observed, whereas one female blogger uses modal verbs exceptionally often, as they take up 2% of her word total.

In the plot (Figure 6.3), the behavior of bloggers for each word list can be observed. Female and male bloggers seem to make fairly equal choices in terms of frequency for the use of the following styles: *New words*, *Negative words*, *Positive words*, *Modal verbs*, *Function words*, and *Negation words*. The *Negative words* coordinate indicates that the greatest and smallest share of words that carry a negative sentiment occur in the entries by male bloggers. A similar pattern can be observed for *Function words*, both the most and the least frequent users are female bloggers. The scores for *Negation words* vary between 0% and 3% for most bloggers regardless of the gender. The *Positive words* coordinate displays a uniform use by both groups of bloggers with shares varying between 0.4% and 1.3%.

Female bloggers use words from the following lists more frequently than male bloggers: *Social words*, *Emoticons*, and *Cognition verbs*. In the *Social words* coordinate, there is a female outlier with the words of social contact taking up almost 4% of all words in her blog entries, followed by two other female bloggers and a male blogger with the social words share of roughly 2.3%. With the rest of the authors from both groups, the shares fall between 0.5% and 1.9%, whereby the bottom-most shares occur in blog entries by male authors. A similar pattern is present in the *Emoticons* coordinate: while most of the blog entries regardless of their author's gender contain between 0.0% and 0.3%, four female bloggers stand out: in the entries by three of them, there are 0.5% to 0.6% emoticons, and the blogger with most emoticons is a female blogger with emoticons taking up 1.0% of the words she has written. The plot shows a more dynamic distribution for *Cognition verbs*,

as the shares of these verbs vary across both groups between 0.0% and 2.0%. Entries by a single female blogger stand out with a share of 2%, followed by two male bloggers with around 1.6% and 1.4% of cognition verbs in their entries.

A more excessive use of certain discourses by male bloggers can be seen in the coordinates of *Profanity*, *Non-standard words*, and *Janes glossary*. While most entries by female bloggers are comprised of 0.25% or less profane vulgar language, many male bloggers use more: up to 0.5%. Two female bloggers are an exception with 0.5% and entries by one male blogger set the limit at 1.1% of profane language. A similar phenomenon can be observed in the *Non-standard words* coordinate. Most female bloggers use few of the words from this list, whereas they seem frequent in entries by male bloggers, with one of them using non-standard vocabulary in 1.4% of words. Since the *Janes glossary* word list includes all of the words from *Non-standard words*, the behavior of female and male bloggers in the use of *Janes glossary* words reflects the one for *Non-standard words*. However, the list of *Janes glossary* contains more words, so the shares of both groups are larger, whereby the most frequent users of this vocabulary are male. Four of them stand out with the most frequent user achieving 1.13%.

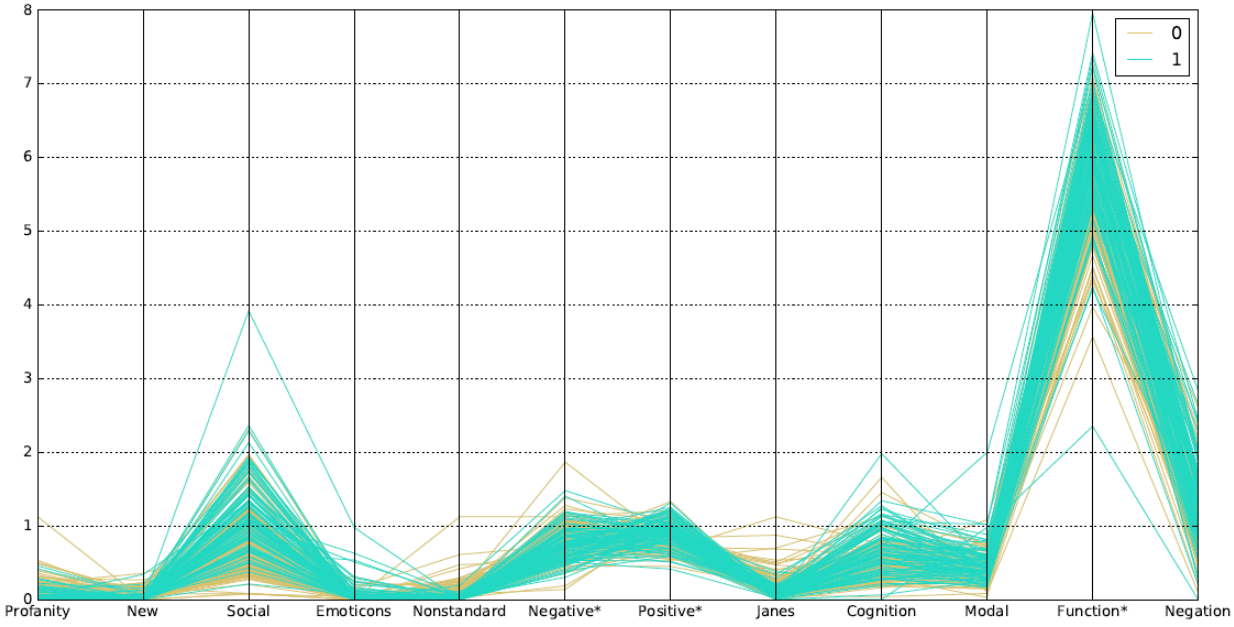

Figure 6.3: Parallel coordinates of word lists that displayed a statistically significant difference between the occurrence in female and male blog entries. From left to right: Profanity, New words, Social words, Emoticons, Non-standard words, Negative words*, Positive words*, Janes words, Cognition verbs, Modal verbs, Function words*, Negation words. Beige (0) represents the male class, and turquoise (1) represents the female class. The symbol "*" signals that the scores were divided by 5.

Chapter 7

Comparative Analysis and Cross-Genre Experiments on Twitter and Blog Corpora

In this chapter, we compare the results of the analyses from Chapters 5 and 6, which focus on gender-related linguistic variation in Slovene tweets and blog entries, respectively. Each of the chapters discusses a single genre and this chapter aims to determine which differences between the language of women and men occur across both genres and which differences are bound to one or the other. Firstly, we contrast the performance of classification models for gender prediction and the results of the statistical analyses of writing styles. Secondly, the chapter presents a set of cross-genre experiments which aim to explore the robustness of statistical models for gender prediction beyond a single genre. For this, we again use the Slovene tweet and blog corpora to train a predictive model on one genre and test it on the other.

7.1 Comparative Analysis

This section is concerned with the question of whether the linguistic variations between female and male Twitter users resemble the variations that occur between female and male bloggers. We compare the performance of rule-based and statistical models for gender prediction in tweets and blog entries. Furthermore, we investigate whether the statistical models of tweets and blog rely on the same features. Finally, we contrast the results of statistical analyses concerning the writing style, whereby we explore whether an overlap in the patterns of gender-related stylistic variation exists.

7.1.1 Performance of Gender Identification Models

In Sections 5.1.2 and 6.2.2, we applied a model based on classification rules to predict the gender of tweet and blog authors. The rules searched for self-references in verb phrases that comprise an auxiliary verb and an l-participle. The model classified each author as *male*, *female*, or *undefined* in case when there were not more than 70% of gender markings in favor of male or female gender. On the Twitter corpus with a majority class baseline of roughly 68% in favor of male users, the most beneficial setting of the rule based-model achieved the classification accuracy only slightly above the majority baseline with 68.56% (see Table 5.2). The reason for the erroneous classification of Twitter users lies in noisy documents (i.e. automatically generated tweets) and non-standard spellings of l-participles (e.g., *kloniro* instead of *kloniral*). The rules performed more successfully on blog entries,

as the highest classification accuracy achieved was 85.71% (see Table 6.3), thus improving the majority classifier by almost 21%, whereby the majority class is again male. When applied to blog entries, the rules made most mistakes in classifying female bloggers as male due to the use of the masculine referential gender in long passages by male narrators.

Interestingly, the use of non-standard spellings (*sm* and *nism*) of auxiliary verbs to find gender indicators did not change the performance of rules in blogs, whereas it increased the model accuracy in tweets by 1.6%. From this, it can be concluded that the referential gender in blog entries is used in a more predictable way, i.e. within a verb phrase, where the auxiliary verb and the l-participle both occur, lie close together, and are written with standard endings. The criteria for the rule-based gender prediction used in Fišer, Erjavec, and Ljubešić (2016) di□er from the ones presented in Sections 5.1.2 and 6.2.2, but we can nevertheless draw a comparison. The classification rules in Fišer, Erjavec, and Ljubešić (2016) correctly classified the gender of roughly 76% of Twitter users and 78% of bloggers. Their classification model outperforms ours in tweets, but not in blog entries, as our rule-based model works better on blog entries.

In Sections 5.1.3 and 6.2.3, we studied whether the linguistic variation between women and men can be used for building a statistical gender prediction model. A set of experiments was conducted employing three machine learning algorithms (Support Vector Machine, Naïve Bayes, and Logistic Regression) and various features (character and word n-grams, as well as style-related features based on the relative frequency of particular words). The models were tested in 10-fold cross validation. The most beneficial features for gender prediction in tweets and blogs are token-based word unigrams when learned with the linear SVM algorithm. On tweets, the statistical classifier achieved the accuracy and standard deviation of 90.26% $\pm$ 1.13% and thus exceeded the majority vote classifier by around 21% (see Table 5.4), as well as the rule-based classifier by almost 22% (see Table 5.2) and the rule-based classifier described in Fišer, Erjavec, and Ljubešić (2016) by 14%.

When applied to the blog corpus, the setting with token-based word unigrams and SVM as the learning algorithm yields a classification accuracy of 86.85% $\pm$ 6.00%. The model outperforms the majority vote baseline by 22% (see Table 6.5) and the rule-based classifier only by about 1% (see Table 6.3). The statistical model also exceeds the accuracy achieved by the rule classifier described in Fišer, Erjavec, and Ljubešić (2016) by roughly 7%.

7.1.2 Most Informative Features

In order to gain insight into the best performing statistical models for gender prediction, we conducted an analysis of the most informative features, i.e. features that were assigned the largest weights and that the model deemed best at di□erentiating between female and male authors. Sections 5.1.4 and 6.2.4 provide a detailed analysis of the top 1,000 features for each class in tweets and blog entries, respectively. Generally, the female and male lists display linguistic variation in terms of referential gender, topic, and style.

According to the gender prediction model on tweets, l-participles and adjectives with encoded masculine or feminine gender marking carry most information about the female and male class, as they are positioned on the top of the female and male feature list. Personal and possessive pronouns in general and especially in first person point to the female class. The female list also comprises a number of interjections that indicate interaction and emotions and are often combined with expressive spelling, i. e. character flooding. Various adverbs, conjunctions, and prepositions are associated with the male class in tweets, while less features of these parts of speech occur in the female list. A very di□erentiating male trait is also profane language. Although the feature lists indicate topical variations between female and male Twitter users, these appear to be less informative than gram-

matical and stylistic differences. Both groups of users tweet about politics. Additionally, female Twitter users tend to focus on the topics of family, food and beverage, while male Twitter users write more about technology and gadgets, sports, and drinking.

The analysis of the most informative features in predicting gender from blog entries shows similar tendencies. The topmost features are l-participles and adjectives: they take the feminine form in the female list, and the masculine form in the male list. Aside from markings of referential gender, the features in the female list display a strong bias towards the topics of family, romance and emotions, and food and beverage. The same topical tendencies of female bloggers are evident from the topic ontology of female blog entries (see Figure 6.1). The female-related features also include frequency adverbs and personal and possessive pronouns in first person singular.

According to the male list, a range of conjunctions and adverbs (especially temporal adverbs expressing a point in time) point to male bloggers. However, the vocabulary on politics, economy, and sports represents the most indicative set of word unigrams after l-participles and adjectives in the masculine form. The popularity of these topics among male bloggers is confirmed in the topic ontology of blog entries by male authors (see Figure 6.2). Given the male list, the topic of sport seems to be differentiating between female and male bloggers, even though the topic itself is not a dominant one among male bloggers compared to politics, economy and current affairs (see Table 6.1). The male list contains also examples of profane language, but these are not as frequent as in the male list of Twitter users.

7.1.3 Statistical Analyses of Writing Style

In this section, we compare the results obtained in the statistical analyses of writing styles from Sections 5.2.2 and 6.3.2 performed on tweets and blog entries, respectively. For this, we measured the share of chosen words from pre-defined word lists (see Section 4.5). We applied the Mann-Whitney U-test on tweets and blog entries to contrast female and male authors according to their writing style. Furthermore, Pearson's correlation coefficient was used to determine which style relates to each of the genders. The squared value of Pearson's coefficient was used as the measure of effect size.

Tables 5.5 and 6.6 present the results of statistical tests on tweets and blog entries, respectively. From the tables, several observations about the genres can be made. To some extent, tweets and blogs display similar results. In both text genres, we obtain a statistically significant difference between female and male authors in terms of the use of the following styles: *Emoticons*, *Function words*, *Profanity*, *Positive words*, *Negative words*, and *New words*. Furthermore, the direction of the correlation between the style and gender is the same across both genres for the following lists: *Emoticons*, *Function words*, and *Positive words* correlate positively with the female bloggers and Twitter users, while *Profanity* and *Negative words* stand in positive correlation with male bloggers and Twitter users. The *New words* list deviates from this, as it correlates positively with female Twitter users (*rpb*=-0.0303, p =0.0169) and male bloggers (*rpb*=+0.151, p =0.013). It should be noted that the correlation is low for all of the above-mentioned lists.

The relative frequency scores of the target word use displaying a statistically significant difference were visually presented in parallel coordinates in Figure 5.1 and Figure 6.3 for tweets and blogs, respectively. The plots allowed for an observation of between- and within-class variation. As can be seen from both Figure 5.1 and Figure 6.3, there is no clear or uniform deviation between the female and male authors, as the lines representing female and male authors generally overlap. However, interesting within-gender variation is evident from the plots: in the tweet plot (Figure 5.1) outliers appear in the *Negative words*, *Profanity*, *Emoji*, *Emoticons*, and *New words* coordinates, whereas the blog plot (Figure

6.3) displays outlying authors in *Social words*, *Profanity*, *Emoticons*, *Non-standard words*, and *Negative words* coordinates. An interesting observation about the outliers is that their deviation from their gendered group does not position them closer to their opposite-gender counterparts, but rather pushes them away from both gendered groups.

Comparing the lists that occur in both plots (*Emoticons*, *Function words*, *Profanity*, *Positive words*, *Negative word*, and *New words*), we can see that generally, the vocabulary with a negative sentiment, and profane language appear more frequently in blog entries, rather than tweets. This is especially true for the words from the *Function words* list, as the mean of shares in tweets (4.62%) is substantially smaller than the mean in blogs entries (29.55%). On average, emoticons and new words are used in tweets and blog entries with roughly similar frequencies.

Interestingly, the only list among the total of 18 lists which displays a medium correlation is the *Social words* list that contains words related to social interactions and processes. Male and female bloggers use the words from this list differently in a way that is statistically significant, as female bloggers, as opposed to male ones, include more vocabulary associated with social interactions. This results in a medium correlation between female bloggers and the use of these target words (*rpb*=-0.393, $p < 0.05$).

7.1.4 Discussion

In this section, we draw on the findings of related works concerning gender-related linguistic variation in UGC genres. The results of our analysis of the language use of women and men in Slovene tweets and blog entries has shown that the linguistic differences between women and men can be employed to build a classification model for gender prediction. An analysis of the most informative features of statistical models for predicting author gender in tweets and blogs indicates that female and male authors use language differently in terms of grammatical gender, topic, and writing style. Many of our findings corroborate previous studies for other languages.

Our analysis of the most informative features from SVM classifiers has shown that female bloggers and Twitter users differ from their male counterparts by the use of personal pronouns, especially in a self-referencing context. The same observation has been made for English Facebook chats (Schwartz et al., 2013), English blog entries (Schler et al., 2006), and a corpus of various written and spoken sources of English (Newman et al., 2008), as well as in our preliminary study of Slovene tweets (Verhoeven et al., 2017).

Female Twitter users in our data opt for expressive spellings in the form of character flooding, which was found by Bamman et al. (2014) as well. According to our results, female Twitter and blog users display non-verbal and language-independent expressiveness with emoticons and emojis more often than male users, which has been supported in the studies by Ljubešić et al. (2017), Bamman et al. (2014), Schwartz et al. (2013), Newman et al. (2008). In contrast, Bamman et al. (2014), Schwartz et al. (2013), Newman et al. (2008) report that male authors use more swear words than female authors do, which has been confirmed in our analysis of the most informative features on the one hand, and by positive Pearson's correlation coefficient between male authors (both Twitter users and bloggers) and the relative frequency of profane words in the text on the other hand.

According to Newman et al. (2008) and Schwartz et al. (2013), female authors refer more often to social processes. The feature analysis of our models for gender prediction in tweets and blogs has indeed shown that references to family relate to female authors. However, the difference in the use of words referring to social interaction has been found statistically significant and correlated to female authors in blog entries, but this has not been the case with tweets as the difference between the genders in terms of this feature is statistically insignificant. The same is true for references to psychological or cognitive

processes, which are associated with female authors (Schwartz et al., 2013; Newman et al., 2008). The use of verbs from the *Cognition verbs* list is correlated with female bloggers (*rpb*=-0.280, p <0.001), but not with female Twitter users.

Some researchers claim a gender prediction model should be topic-independent to provide a reliable generalization over female and male language use (see Daelemans (2013)). Nevertheless, the tendencies of female authors to focus more on their personal lives and romance and the tendencies of male authors to write more than women about religion, politics, business, and the Internet have been found in English blogs and successfully applied to gender prediction (Schler et al., 2006). This is supported in our analysis of tweets and blogs, where we found that topical cues indeed make up an important part of the differentiating features especially for gender prediction in blog entries (see Section 7.1.2).

7.2 Cross-Genre Experiments for Automated Gender Prediction

This section presents the set of cross-genre experiments with gender prediction models that are trained on a corpus comprising one UGC genre and that are tested on a corpus of a different UGC genre. For this we use the Janes tweet and blog corpora described in Sections 3.1.1 and 3.1.2, respectively. We use the tweet and blog subcorpora described in Sections 5.1.1 and 6.2.1, respectively. In this section, we aim to explore the robustness of the best performing statistical gender prediction models that were trained on each genre and tested in a 10-fold cross validation setting. We obtain the highest accuracy of 90.26% on tweets (see Section 5.1.3) and 86.85% on blogs (see Section 6.2.3).

The comparative analysis of results from classification experiments on tweets and blogs has shown there is a certain overlap in the features accounted for carrying the most information about the female and male class (see Section 7.1.2). Most notably, these features include feminine and masculine referential gender in verb l-participles and adjectives. The topical tendency of female authors towards the subjects of family, food and beverage is observed in tweets as well as blog entries. A common feature that distinguishes male bloggers and Twitter users from their female counterparts is the use of conjunctions, adverbs, and profane language in terms of style, and politics, and sports in terms of content. We expect that these similarities will be proven useful in cross-genre experiments.

7.2.1 Experimental Setting

A classification model for gender prediction was built for each of the two genres. In the model construction, we applied the settings that proved to be most beneficial for the classification accuracy in the 10-fold cross validation experiments from Sections 5.1.3 and 6.2.3 for tweets and blogs, respectively. In the cross-genre setting, each model is trained on the entire corpus and tested on the unseen test set, which comprises documents from the other UGC genre.

For gender prediction in tweets, the SVM algorithm performed best on token word unigrams as features, while the corpus was preprocessed prior to model construction, whereby the Twitter-specific elements (hashtags, URLs, and user mentions) were removed from the tweets (see Table 5.4). Additionally, we built a classifier using the same setting, except the lemmatized preprocessed tweet text was used. In a 10-fold cross validation the models performed with the accuracy of 90.26% and 85.99% on tokens and lemmas, respectively.

For gender prediction in blog entries, we also experimented with the settings and found that the SVM algorithm trained on token word unigrams performs the best gender prediction in blog entries (see Table 6.5). In a 10-fold cross validation the models performed

with the accuracy of 86.85% and 82.42% on tokens and lemmas, respectively.

7.2.2 Results of Cross-Genre Experiments

The results of cross-genre gender prediction experiments are presented in Table 7.1. For each experiment, the table describes the text form and genre of the train and test set, the classification accuracy of the cross-genre models as well as the test set majority classifier as the baseline for comparison. The text form column applies to both the train and test set.

From the table, several observations can be made. The classification models trained on tweets and tested on blog entries outperform the test set majority class in the case of non-lemmatized and lemmatized text. The best performance of the cross-genre setting is achieved by the model trained on token-based word unigrams from tweets and tested on blog entries, as it obtains the accuracy of 87.91% and thus outperforms the test set baseline by 23.06%. The model trained on lemmatized tweets and tested on lemmatized blog entries outperforms the baseline by 11.71%.

Table 7.1: Results of gender prediction in cross-genre experiments using the word unigram features and the SVM learning algorithm.

Form	Train set	Test set	Test set majority class	Classification accuracy
token	tweets	blogs	64.85%	**87.91%**
lemma	tweets	blogs	64.85%	76.56%
token	blogs	tweets	67.99%	68.48%
lemma	blogs	tweets	67.99%	68.00%

In the settings with blog entries as the train set, the gender prediction model achieves poor classification accuracy. It barely outperforms the test set baseline in the case of token-based text, where the model accuracy is greater than the baseline by roughly 0.5%. When lemmatized text is used in the blog train set and tweet test set, the model classifies the instances into the majority class.

A surprising result of the cross-genre experiments is the performance of the model trained on tweets and tested on blogs in the case of non-lemmatized text, as it achieves 87.91% accuracy. In this setting, the blog authors are classified with a higher accuracy than using the rule-based model (which achieved 85.71% accuracy; see Table 6.3). Furthermore, this cross-genre model performs better than the best 10-fold cross validation model on blogs, as the single genre blog model achieves the accuracy of 86.85% (see Table 6.5). In comparison to the 10-fold cross validation setting of the gender prediction in tweets (see Table 5.4), the cross-genre model trained on tweets performs slightly worse, as the classification accuracy decreases by 2.35%.

7.2.3 Discussion

Aside from the cross-genre setting, the size of the train and test sets should also be taken into consideration when interpreting the results. The tweet corpus consists of 6,203 instances, while the blog corpus is considerably smaller as it comprises 273 instances. Furthermore, both datasets contain more male than female authors. The results from Table 7.1 can lead to several observations. Given that the tweet corpus is much larger than the blog corpus, the model trained on tweets may perform better when faced with a smaller test set. In contrast, training on a small dataset of blogs may result in overfitting to the

train set, so the model fails to outperform the baseline accuracy when applied to the much larger tweet corpus.

This drop of classification accuracy in cross-genre and cross-topic settings is observed in Sarawgi et al. (2011). The accuracy of their gender prediction models decreased when the train set comprised 280 gender-balanced blog entries, while the model was tested on a dataset of 200 gender-balanced scientific papers. The BOW-model performed best and achieved an accuracy of 61%, which is lower than the cross-validated train set by about 10%.

Similarly, the overall winners of the PAN 2016 author profiling task (Busger op Vollenbroek et al., 2016) report on the poorer gender classifier performance in cross-genre settings. The task of PAN AP 2016 (Rangel et al., 2016) was to predict age and gender in English, Spanish, and Dutch documents. All the datasets included in the task were balanced in gender. The above mentioned overall winners (Busger op Vollenbroek et al., 2016) trained an SVM classifier using a variety of features: n-grams (based on words, characters, and POS-tags), capitalization, punctuation, word and text length, vocabulary richness, topic-related words, and emoticons. For English and Spanish, the training data comprised tweets by 436 and 250 users for each language, respectively. In cross-validated train sets, their model achieved an accuracy of 70.67% for English and 70.85% for Spanish. The test set comprised blog entries; moreover, the size of the test sets was substantially smaller for both languages: 78 users for English, and 56 for Spanish. On the test set, their model performed with an accuracy of only 64.10% for English, while it displayed an improved score for Spanish with 71.43%. On the Dutch train set, which included tweets, the model performed well with an accuracy of 72.12%, while the score decreased on the test set, which comprised reviews, achieving only the baseline accuracy of 49.60%. The authors speculate that either blogs seem more similar to tweets in terms of how the author's gender surfaces, or their model performed worse for Dutch because the size of the train set was smaller than the test set.

Chapter 8

Conclusions, Further Work, and Lessons Learned

In this chapter, we briefly summarize the results of our research on gender-related linguistic variation in Slovene tweets and blog entries. We present the conclusions drawn from our study. Additionally, we suggest potential improvements and ideas for further work. At the end, we share the lessons learned during our study.

8.1 Conclusions

Variation is a part of the social nature of language. Recent advances in language technologies and natural language processing have enabled researchers to model this variation using automated methods on large collections of textual data. The studies of linguistic variation are more and more frequently applied to user-generated content (UGC) due to its accessibility, size, and production in real time. In this thesis, we discuss the linguistic variation that occurs among the speakers taking their gender into account. More specifically, we use machine learning techniques and statistical analysis to analyze how gender surfaces in the language use of Slovene female and male Twitter and blog users.

We investigate whether the gender-related linguistic variation can provide predictive information to di□erentiate between male and female UGC authors automatically. For this, we build two types of classification models for gender prediction. The first model type is based on manually written classification rules that predict the author's gender given the use of grammatical gender in self-referencing verb phrases. The model was applied to the Twitter and blog corpora. The second model type is a statistical classifier. We constructed several statistical classifiers by experimenting with various machine learning algorithms (Support Vector Machine, Logistic Regression, and Naïve Bayes) and features (word and character n-grams, and external word lists) to find the most suitable setting for gender prediction in tweets and blog entries. The highest accuracy of statistical models in tweets as well as blog entries was achieved by using the SVM algorithm and unigrams based on word tokens as features.

We compared the performance of the rule-based and statistical models and reached some interesting conclusions with regard to text genre. The rule-based classifier assigns the label *female* or *male* if the number of gender markings in l-participles exceeds a certain threshold. Otherwise, the authors are classified as *undefined*. The rule-base model performs well on blog entries achieving 85.71% accuracy, thus outperforming the majority baseline by almost 21%. The rule-based classifier is less successful in predicting the author's gender in tweets, as it outperforms the 67.99% majority baseline by only 0.57%. However, the best performing statistical models work well on blogs as well as tweets, achieving the

classification accuracy of 90.26% ± 1.13% on tweets, and 86.85% ± 6.00% on blog entries. It should be noted that in contrast to the rule-based model, statistical models are binary classifiers with the *female* and *male* labels as possible classes.

These most accurate statistical models were further analyzed with regard to the most informative features for the female and male authors. Among the top ranked features, we can find verb l-participles and adjectives in the masculine form for the male authors, and in the feminine form for the female authors. Apart from referential gender, features that imply a topical bias between female and male authors stand out as well. Male Twitter and blog users have in common the topic of politics and sports, as opposed to female Twitter and blog users who differ from their male counterparts by the topics of family and food. Pronouns in first person singular also distinguish female Twitter and blog users from their male counterparts. The topical bias divides female and male bloggers strongly, as the differentiating features of male bloggers largely comprise the vocabulary on political issues, while female bloggers diverge from male bloggers by the vocabulary on emotions, romance and sexuality, and health issues. Male bloggers are distinguished by profane language, but in a very minor sense. While topical variation carries predictive information for the author's gender in tweets as well, stylistic differences also stand out, as profane language is strongly associated with male users, while interjections characterize female users. Based on the performance of both model types and on the feature analysis, we can confirm Hypothesis 1 (see Section 1.3) which states textual content, referential gender, and writing style contribute to the prediction of the author's gender.

The analysis of features ranked as most informative by the statistical models has provided insight into the linguistic variation between women and men. The analysis has shown that several features appear to be stereotypical in terms of the predominant topics (e.g., men: politics and sports; women: family and emotions) and linguistic style (e.g., women: interjections; men: profane language). However, even if the statistical models predict the gender of the author with high accuracy, one should note that features encoding stereotypical linguistic behavior do not generalize over all the authors in our data. This calls for a more detailed error analysis of the predictive models and the results of our work can serve as the base for computational approaches to inter- and intra-gender linguistic variation. Thus, Hypothesis 2, which states that features from predictive models can contribute to the sociolinguistic understanding of gender differences in writing, can also be confirmed.

Aside from predictive modelling, we approach the gender-related linguistic variation in tweets and blog entries with statistical testing of the differences in writing style. We used lists of target words that denote particular traits of linguistic style (profane language, positive or negative sentiment vocabulary, non-standard and new vocabulary, verbs of cognitive processes or communication, vocabulary of social interaction, markers of intensity, hedging, or negation, emoticons, emojis, function words, and modal verbs) to compute the relative frequency of these target words contained in the documents of each user. The Mann-Whitney U-test was applied to test the statistical significance of the variation between female and male authors with regard to the relative frequency of these target words. For some word lists, the Mann-Whitney U-test showed that the difference between female and male authors is statistically significant (see 5.2.2 and 6.3.2 for tweets and blog entries, respectively). However, when we calculated Pearson's coefficient as the measure of effect size, we discovered that the effect of the authors' gender on the stylistic variation is small. The only word list that displayed a medium effect of gender on the variation is the *Social words* list in blog entries. The comparative analysis of statistical results in both tweets and blog entries has shown that 6 target word lists display a statistically significant difference in both genres, namely, *Negative words*, *Positive words*, *New words*, *Profanity*, *Function words*, and *Emoticons* (see Section 7.1.3). Hypothesis 3 states that statistically

significant differences between the writing style of female and male authors in tweets and blogs can be found. To conclude, Hypothesis 3 can be partly confirmed. The gender-related variation in tweets is significant in 10 word lists (see Table 5.5), while this is true for 12 word lists in blog entries (see Table 6.6). Among all the word lists, 6 display a statistically significant difference between women and men in both genres.

In the thesis, we explored the gender-related linguistic variation in Slovene tweets and blog entries separately. Section 7.1.2 provides a comparison of the results and points out the similarities and differences found in the variation between Twitter and blog users. In addition, we conducted cross-genre experiments to test the performance of statistical models for gender prediction when they are trained on one UGC genre and tested on the other. The experimental setting, results, and discussion thereof are presented in Sections 7.2.1, 7.2.2, and 7.2.3, respectively. The performance of the SVM classifier that was trained on tweets and tested on blogs performed well over the majority baseline; however, the classifier trained on blogs and tested on tweets achieved poor accuracy as it barely outperformed the majority baseline. Thus, we can neither confirm nor reject Hypothesis 4, which states that models for gender prediction trained on a single UGC genre do no support generalization that goes beyond that genre. It should be noted that the size of the blog corpus is much smaller than the size of the Twitter corpus, which possibly affects the poor performance of the blog-trained corpus. On the other hand, the tweet-trained model can potentially be used for annotating the authors' gender in other UGC genres included in the Janes corpus.

8.2 Further Work

In this section we present the further work we plan to undertake in terms of developing the methods for analyzing gender-related linguistic variation, as well as applying the classification models to data with no author gender annotation.

With regard to the rule-based classification models, we propose several ideas for their improvement. Their poor performance on tweets can partially be explained by the noise (automatically generated tweets with the masculine gender used as neutral). Thus, the model performance could benefit from a thorough cleaning in the corpus post-processing phase. Furthermore, citations in tweets and blogs could be detected given the use of quotation marks in tweets or given the reference to the author and source of longer citation in blog entries. Furthermore, the analysis of the most informative features has shown that adjectives in the feminine and masculine grammatical gender surface as informative for female and male authors, respectively. The classification rules could be modified in terms of including adjective endings in the gender indicator score.

Our statistical models for gender prediction performed well when evaluated with cross validation. We plan to explore the classifier performance with regard to the size of the training corpus by plotting a learning curve, especially given that the amount of manually annotated data can be substantially smaller than the amount of data available for our experiments. We plan to apply the gender prediction model built for the PAN shared (Martinc et al., 2017) task to Slovene UGC. We intend to perform further cross-genre experiments with the tweet-trained classifier and the Janes subcorpora that have no manual gender labels (forum posts, news comments, and Wikipedia talk pages), but have been annotated automatically by Fišer, Erjavec, and Ljubešić (2016), who used the rule-based implementation described in Section 3.3.

The extraction of the features deemed most important for classifying between female and male authors opens new opportunities for the analysis of gender and language use. For the models that performed best on each of the two UGC genres, the feature lists will be made publicly available. The lists can be used for further detailed analysis in context.

This especially applies to the vocabulary that is not topic-based, namely, pronouns and adverbs. For example, Pennebaker (2013) focuses largely on the use of various pronouns as employed by di□erent gender, age, and personality groups.

In the construction of topic ontologies from blog entries, each document was assigned to a topic. In the future, we plan to analyze the linguistic di□erences between female and male bloggers within the documents that belong to the same topic. In that way, we expect to gain more insight into variation within a specific topic as well as more focus on stylistic variation.

In order to conduct a study that employs the methods of computational sociolinguistics (Nguyen et al., 2016) or social psychology (Pennebaker, 2013), the used corpus should be rich in author and text metadata, thus ensuring a more complex analysis of linguistic variation on several levels. This is something that could be taken into consideration in further research of linguistic variation in Slovene UGC.

8.3 Lessons Learned

This section summarizes the lessons learned during the process of building, evaluating, and comparing models for automated gender identification in Slovene tweets and blog entries.

- **Selecting the approach to automated gender prediction in Slovene documents**
  The rule-based classification model has a the great advantage in that it does not require labeled data to learn from. The rule-based classifier also performs well in general, as it achieves around 99.10% precision for female as well as male Twitter users (see Table 5.2), 95% precision for female bloggers, and 98.75% for male bloggers (see Table 6.3). However, data that is prone to noise (as was shown in tweets) might make the gender prediction task more di□cult, as additional cleaning is required. If labeled training data is available, a statistical classifier could outperform the rule-based model and allow also for feature interpretation for a sociolinguistic analysis.

- **Features and algorithms**
  Word and character n-grams in the token and lemma form were experimented with as features. Stylistic features were also tested, as well as the union of several features. Generally the token-based word unigrams performed best. The experiments were carried out using three machine learning algorithms (Naïve Bayes, Support Vector Machine, and Logistic Regression), among which the Support Vector Machine performed best on blogs, as well as tweets.

- **Size of the train and test sets in cross-genre experiments**
  The gender prediction model trained on tweets was evaluated as very accurate on blog entries, whereas the opposite setting was shown to result in poor classification. However, the Twitter corpus exceeds the blog corpus in size. If the learning algorithm is provided with a training set with enough instances, it can be generalized for gender prediction beyond a single genre.

References

Álvarez-Carmona, M., López-Monroy, A., Montes-y-Gómez, M., Villaseñor-Pineda, L., & Escalante, H. (2015). Inaoe's participation at pan'15: Author profiling task. In L. Cappellato, N. Ferro, G. Jones, & E. San Juan (Eds.), *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, *12*(9).

Baker, P. (2014). *Using corpora to analyze gender*. London: Bloomsbury.

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

Biber, D. (2007). *Discourse on the move: Using corpus analysis to describe discourse structure*. John Benjamins Publishing.

Bramer, M. (2013). *Principles of data mining*. London: Springer.

Brank, J., Mladenić, D., & Grobelnik, M. (2010). Feature construction in text mining. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning*. Springer.

Busger op Vollenbroek, M., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., ... Nissim, M. (2016). GronUP: Groningen User Profiling—Notebook for PAN at CLEF 2016. In K. Balog, L. Cappellato, N. Ferro, & C. Macdonald (Eds.), *Clef 2016 evaluation labs and workshop – working notes papers*. Évora, Portugal: CEUR-WS.org.

Butler, J. (1990). *Gender trouble: Feminism and the subversion of identity*. New York: Routledge.

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

Čibej, J. (2016). Framework for an analysis of slovene regional language variants on twitter. In *Proceedings of the 4th conference on cmc and social media corpora for the humanities* (pp. 17–21). Ljubljana, Slovenia: Academic Publishing Division of the Faculty of Arts of the University of Ljubljana.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cox, D. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, *20*(2), 215–242.

Dadvar, M. & de Jong, F. (2012). Cyberbullying detection: A step toward a safer internet yard. In *Proceedings of the 21st international conference on world wide web* (pp. 121–126). WWW '12 Companion. Lyon, France: ACM.

Daelemans, W. (2013). Explanation in computational stylometry. In *Proceedings of the 14th international conference on computational linguistics and intelligent text processing - volume 2* (pp. 451–462). CICLing'13. Springer. Berlin, Heidelberg.

Dhillon, I., Kogan, J., & Nicholas, C. (2004). Feature selection and document clustering. In M. W. Berry (Ed.), *A comprehensive survey of text mining*. New York: Springer.

Dobrovoljc, K., Krek, S., & Rupnik, J. (2012). Skladenjski razčlenjevalnik za slovenščino. In T. Erjavec & J. Ž. Gros (Eds.), *Zbornik 15. mednarodne multikonference informacijska družba - is 2012, zvezek c* (pp. 42–47). Institut Jožef Stefan.

Fišer, D., Erjavec, T., & Ljubešić, N. (2016). Janes v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0*, *4* (2), 67–99.

Fišer, D., Erjavec, T., & Ljubešić, N. (2017). The compilation, processing and analysis of the janes corpus of slovene user-generated content. In *Corpus de communication médiée par les réseaux: Construction, structuration, analyse*. Paris, France: L'Harmattan.

Fišer, D., Smailović, J., Erjavec, T., Mozetič, I., & Grčar, M. (2016). Sentiment annotation of the Janes corpus of slovene user-generated content. In *Proceedings of the 10th language technologies and digital humanities conference* (pp. 65–70). Ljubljana, Slovenija.

Fortuna, B., Grobelnik, M., & Mladenić, D. (2005). Semi-automatic construction of topic ontologies. In M. Ackermann, B. Berendt, M. Grobelnik, A. Hotho, D. Mladenić, G. Semeraro, . . . M. van Someren (Eds.), *Semantics, web and mining, joint international workshop, ewmf 2005 and kdo 2005*. Porto, Portugal: Springer.

Fortuna, B., Grobelnik, M., & Mladenić, D. (2007). Ontogen: Semi-automatic ontology editor. In *Hci international* (pp. 309–318). Beijing, China.

Gantar, P., Škrjanec, I., Fišer, D., & Erjavec, T. (2016). Slovar tviterščine. In *Zbornik konference jezikovne tehnologije in digitalna humanistika* (pp. 71–76). Ljubljana, Slovenija.

Gergen, K. J. & Shotter, J. (1989). *Texts of identity*. London: Sage.

Goddard, A. & Patterson, L. M. (2000). *Language and gender*. London: Routledge.

Gorjanc, V. (2007). Kontekstualizacija oseb ženskega in moškega spola v slovenskih tiskanih medijih. In *43. seminar slovenskega jezika, literature in kulture. Stereotipi v slovenskem jeziku, literaturi in kulturi: Zbornik predavanj* (pp. 173–180). Ljubljana, Slovenija: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete Univerze v Ljubljani.

Gravetter, F. J. & Wallnau, L. B. (2013). *Statistics for the behavioral sciences*. Belmont: Wadsworth.

Grčar, M. & Krek, S. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In *Proceedings of the 8th language technologies conference* (Vol. 100, pp. 89–94). Ljubljana, Slovenia: IJS.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46* (1–3), 389–422.

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

Hu, M. & Liu, B. (2004). Mining opinion features in customer reviews. In *Proceedings of AAAI Conference on Artificial Intelligence* (pp. 755–760). San Jose, California.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London New York: Continuum.

Jurafsky, D. & Martin, J. H. (2009). *Speech and language processing, second edition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Juršič, M., Mozetič, I., Erjavec, T., & Lavrač, N. (2010). LemmaGen: Multilingual lemmatisation with induced ripple-down rules. *J. UCS*, *16* (9), 1190–1214.

Kadunc, K. & Robnik-Šikonja, M. (2016). Analiza mnenj s pomočjo strojnega učenja in slovenskega leksikona sentimenta. In *Zbornik konference jezikovne tehnologije in digitalna humanistika* (pp. 83–89). Ljubljana, Slovenija.

Kapočiūtė-Dzikienė, J., Šarkutė, L., & Utka, A. (2014). Automatic author profiling of Lithuanian parliamentary speeches: Exploring the influence of features and dataset sizes. In *Human Language Technologies – The Baltic Perspective, Proceedings of the Sixth International Conference Baltic HLT 2014*. Kaunas, Lithuania.

Kapočiūtė-Dzikienė, J., Utka, A., & Šarkutė, L. (2015). Authorship attribution and author profiling of Lithuanian literary texts. In *Proceedings of the 5th workshop on balto-slavic natural language processing*. Hissar, Bulgaria.

Kilgarri□, A., Baisa, V., Bušta, J., Jakubiček, M., Kovar, V., Michelfeit, J., ... Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1* (1), 7–36.

Kobayashi, M. & Aono, M. (2007). Vector space models for search and cluster mining. In M. W. Berry & M. Castellanos (Eds.), *Survey of text mining: Clustering, classification, and retrieval*. Springer.

Koletnik, A., Grm, A., & Gramc, M. (2015). *Vsi spoli so resnični: transspolnost, transseksualnost in cisspolna nenormativnost*. Ljubljana: Društvo informacijski center Legebitra.

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

Koppel, M., Schler, J., & Argamon, S. (2008). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, *60* (1), 9–26.

Krek, S., Gantar, P., Dobrovoljc, K., & Škrjanec, I. (2016). Označevanje udeleženskih vlog v učnem korpusu za slovenščino. In *Zbornik konference jezikovne tehnologije in digitalna humanistika* (pp. 106–110). Ljubljana, Slovenija: Znanstvena založba Filozofske fakultete v Ljubljani.

Kunst Gnamuš, O. (1995). Razmerje med spolom kot potezo reference in spolom kot slovnično kategorijo. *Jezik in slovstvo*, *40* (7), 255–262.

Lako□, R. T. (1975). *Language and woman's place*. New York: Harper & Row.

Litvinova, T., Seredin, P., & Litvinova, O. (2015). Using part-of-speech sequences frequencies in a text to predict author personality: A corpus study. *Indian Journal of Science and Technology*, *8* (9), 93–97.

Ljubešić, N. & Erjavec, T. (2016). Corpus vs. lexicon supervision in morphosyntactic tagging: The case of slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 1527–1531). Portorož, Slovenia: European Language Resources Association (ELRA).

Ljubešić, N., Erjavec, T., & Fišer, D. (2016). Corpus-based diacritic restoration for south slavic languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 3612–3616). Portorož, Slovenia: European Language Resources Association (ELRA).

Ljubešić, N. & Fišer, D. (2016). Private or corporate? predicting user types on twitter. In *Proceedings 2016 The 2nd Workshop on Noisy User-generated Text (W-NUT)* (pp. 38–46). Osaka, Japan.

Ljubešić, N., Fišer, D., & Erjavec, T. (2014). Tweetcat: A tool for building Twitter corpora of smaller languages. In *Proceedings of the 9th language resources and evaluation conference (lrec 2014)*. Reykjavik, Iceland: ELRA.

Ljubešić, N., Fišer, D., & Erjavec, T. (2017). Language-independent Gender Prediction on Twitter. In *Proceedings of NLP+CSS: Second Workshop on Natural Language Processing and Computational Social Science* (pp. 1–6). Vancouver, Canada: ACL.

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S., & Škrjanec, I. (2015). Predicting the level of text standardness in user-generated content. In *Proceedings of the 10th ranlp 2015 conference* (pp. 371–378). Hissar, Bulgaria.

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., & Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

López-Monroy, A., Montes-y-Gómez, M., Escalante, H., & Villaseñor-Pineda, L. (2014). Using Intra-Profile Information for Author Profiling—Notebook for PAN at CLEF 2014. In L. Cappellato, N. Ferro, M. Halvey, & W. Kraaij (Eds.), *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, She eld, UK*. CEUR-WS.org.

Lui, M. & Baldwin, T. (2012). Langid.py: An o□-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*. Jeju, Korea: ACL.

Macaulay, R. K. (2005). *Talk that counts: Age, gender, and social class di erences in discourse*. Oxford University Press.

Mandravickaitė, J. & Krilaviius, T. (2017). Stylometric analysis of parliamentary speeches: Gender dimension. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. Valencia, Spain.

Mann, H. B. & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18* (1), 50–60.

Martinc, M., Škrjanec, I., Zupan, K., & Pollak, S. (2017). PAN 2017: Author Profiling - Gender and Language Variety Prediction. In L. Cappellato, N. Ferro, L. Goeuriot, & T. Mandl (Eds.), *Working notes papers of the CLEF 2017 evaluation labs*. Dublin, Ireland: CLEF and CEUR-WS.org.

Meina, M., Brodzinska, K., Celmer, B., Czokow, M., Patera, M., Pezacki, J., & Wilk, M. (2013). Ensemble-based classification for author profiling using various features. In *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*. Valencia, Spain: CLEF.

Motschenbacher, H. (2010). *Language, Gender and Sexual Identity: : Poststructuralist Perspectives*. Amsterdam: John Benjamins.

Mukherjee, A. & Liu, B. (2010). Improving gender classification of blog authors. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 207–217). Association for Computational Linguistics.

Newman, M., Groom, C., Handelman, L., & Pennebaker, J. W. (2008). Gender di□erences in language use: An analysis of 14,000 text samples. *Discourse Processes*, *45* (3), 211–236.

Nguyen, D., Dogruoz, A. S., Rose, C. P., & de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics*, *42* (3), 537–593.

Nguyen, D., Trieschnigg, R., Dogruoz, A., Gravel, R., Theune, M., Meder, T., & de Jong, F. (2014). Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of the 25th International Conference on Computational Linguistics, COLING 2014* (pp. 1950–1961). Association for Computational Linguistics.

Osrajnik, E., Fišer, D., & Popič, D. (2015). Primerjava rabe ekspresivnih ločil v tvitih slovenskih uporabnikov in uporabnic. In *Zbornik konference Slovenščina na spletu in v novih medijih* (pp. 50–74). Ljubljana, Slovenia: Znanstvena založba Filozofske fakultete.

Pan, S. J. & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22* (10), 1345–1359.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennebaker, J. W. (2013). *The secret life of pronouns: What our words say about us*. Bloomsbury Publishing USA.

Prabhakaran, V., Reid, E. E., & Rambow, O. (2014). Gender and power: How gender and gender environment a□ect manifestations of power. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1965–1976). ACL.

Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015). Overview of the 3rd author profiling task at PAN 2015. In *CLEF 2015 Working Notes*. CEUR.

Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., ... Daelemans, W. (2014). Overview of the author profiling task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers*. She□eld, UK: CEUR-WS.org.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., & Inches, G. (2013). Overview of the author profiling task at pan 2013. In *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers* (pp. 23–26). Valencia, Spain: CLEF.

Rangel, F., Rosso, P., Potthast, M., & Stein, B. (2017). Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In L. Cappellato, N. Ferro, L. Goeuriot, & T. Mandl (Eds.), *Working notes papers of the clef 2017 evaluation labs*. CEUR Workshop Proceedings. CLEF and CEUR-WS.org.

Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., & Stein, B. (2016). Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations. In *Working Notes Papers of the CLEF 2016 Evaluation Labs, CEUR Workshop Proceedings*. CLEF and CEUR-WS.org.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on search and mining user-generated contents* (pp. 37–44). ACM.

Santosh, K., Bansal, R., Shekhar, M., & Varma, V. (2013). Author profiling: Predicting age and gender from blogs. In P. Forner, R. Navigli, & D. Tufis (Eds.), *Clef 2013 evaluation labs and workshop – working notes papers*. Valencia, Spain: CLEF.

Sarawgi, R., Gajulapalli, K., & Choi, Y. (2011). Gender attribution: Tracing stylometric evidence beyond topic and genre. In *Proceedings of the fifteenth conference on computational natural language learning* (pp. 78–86). Association for Computational Linguistics.

Sboev, A., Litvinova, T., Gudovskikh, D., Rybka, R., & Moloshnikov, I. (2016). Machine learning models of text categorization by author gender using topic-independent features. In *Proceedings of the 5th international young scientist conference on computational science*. Krakow, Poland: Procedia Computer Science.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. (2006). E□ects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (Vol. 6, pp. 199–205).

Schmid, H.-J. (2003). Do men and women really live in di□erent cultures? evidence from the bnc. In A. Wilson, P. Rayson, & T. McEnery (Eds.), *Corpus Linguistics by the Lune: A Festschrift for Geo rey Leech*. Bern, Switzerland: Peter Lang.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, *8*(9).

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM COMPUTING SURVEYS*, *34*(1), 1–47.

Senellart, P. & Blondel, V. (2007). Automatic discovery of similar words. In M. W. Berry & M. Castellanos (Eds.), *Survey of text mining: Clustering, classification, and retrieval*. New York: Springer.

Škrjanec, I. & Pollak, S. (2016). Topic ontologies of the Slovene blogosphere: A gender perspective. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities* (pp. 62–65). Ljubljana, Slovenia: Academic Publishing Division of the Faculty of Arts of the University of Ljubljana.

Smailović, J., Grčar, M., Lavrač, N., & Žnidarčič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, *285* (100), 181–203.

Speer, S. A. (2005). *Gender talk: Feminism, discourse and conversation analysis*. London: Routledge.

Spender, D. (1980). *Man made language*. London: Routledge Kegan Paul.

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

Tannen, D. (1990). *You just don't understand: Women and men in conversation*. New York: Ballantine Books.

Tausczik, Y. & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, *29* (1), 24–54.

Verhoeven, B., Daelemans, W., & Plank, B. (2016). TwiSty: a multilingual Twitter stylometry corpus for gender and personality profiling. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Verhoeven, B., Škrjanec, I., & Pollak, S. (2017). Gender profiling for Slovene Twitter communication: The influence of gender marking, content and style. In *Proceedings of the EACL workshop, The 6th Workshop on Balto-Slavic Natural Language Processing* (pp. 119–125). Valencia, Spain: The Association for Computational Linguistics.

Vogel, A. & Jurafsky, D. (2012). He said, she said: Gender in the acl anthology. In *Proceedings of the acl-2012 special workshop on rediscovering 50 years of discoveries* (pp. 33–41). ACL.

Weikert, M. & Motschenbacher, H. (2015). Structural gender trouble in Croatian. In M. Hellinger & H. Motschenbacher (Eds.), *Gender Across Languages. The Linguistic Representation of Women and Men. Volume IV* (pp. 49–95). John Benjamins.

Widdowson, H. G. (2004). *Text, context, pretext: Critical issues in discourse analysis*. Oxford: Blackwell.

Witten, I. H. & Frank, E. (2005). *Data mining: Practical machine learning. tools and techniques*. San Francisco: Morgan Kaufmann.

Yang, Y., Pan, S., Downey, D., & Zhang, K. (2014). Active learning with constrained topic model. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 30–33). ACL.

Zwitter Vitez, A. (2011). Povej mi karkoli in povem ti, kdo si: Ugotavljanje avtorstva besedil. In *Obdobja 30: Meddisciplinarnost v slovenistiki* (pp. 565–570). Ljubljana, Slovenija: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete.

Bibliography

Publications Related to the Thesis

Conference and Workshop Papers

Martinc, M., Škrjanec, I., Zupan, K., & Pollak, S. (2017). PAN 2017: Author Profiling - Gender and Language Variety Prediction. In L. Cappellato, N. Ferro, L. Goeuriot, & T. Mandl (Eds.), *Working notes papers of the CLEF 2017 evaluation labs*. Dublin, Ireland: CLEF and CEUR-WS.org.

Škrjanec, I. & Pollak, S. (2016). Topic ontologies of the Slovene blogosphere: A gender perspective. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities* (pp. 62–65). Ljubljana, Slovenia: Academic Publishing Division of the Faculty of Arts of the University of Ljubljana.

Verhoeven, B., Škrjanec, I., & Pollak, S. (2017). Gender profiling for Slovene Twitter communication: The influence of gender marking, content and style. In *Proceedings of the EACL workshop, The 6th Workshop on Balto-Slavic Natural Language Processing* (pp. 119–125). Valencia, Spain: The Association for Computational Linguistics.

Biography

Iza Škrjanec was born on 10 March 1993 in Ljubljana, Slovenia, where she also completed her primary and secondary education. In 2012, she started the study of Interlingual Communication at the Faculty of Arts of the University in Ljubljana. In 2015, she finished her BA studies after defending her thesis “Comparison of the language manuals Slovenski Pravopis and Chicago Manual of Style according to the proper name capitalization rules” under the supervision of Prof. Dr. David Limon.

In the same year, she enrolled to the MSc programme of Information and Communication Technologies at the Jožef Stefan International Postgraduate School in Ljubljana to study under the supervision of Prof. Dr. Nada Lavrač and working supervision of Dr. Senja Pollak.

In her studies, she primarily focuses on the aspect of gender in language, as she compares the language of men and women in social media. She is also working in the field of shallow semantic parsing and lexicography. She has presented her work at several Slovene and international conferences.