FACULTY OF INFORMATION STUDIES
IN NOVO MESTO

D O C T O R A L D I S S E R T A T I O N

JOŽE BUČAR

FACULTY OF INFORMATION STUDIES
IN NOVO MESTO

DOCTORAL DISSERTATION

SENTIMENT BASED CLASSIFICATION OF THE WEB TEXTS

Adviser: Prof. Dr. Janez Povh
Co-adviser: Asst. Prof. Dr. Martin Žnidaršič

Novo mesto, July 2017 Jože Bučar

IZJAVA O AVTORSTVU

Podpisani Jože Bučar, študent FIŠ Novo mesto, izjavljam:

- ☐ da sem doktorsko disertacijo pripravljal samostojno na podlagi virov, ki so navedeni v doktorski disertaciji,
- ☐ da dovoljujem objavo doktorske disertacije v polnem tekstu, v prostem dostopu, na spletni strani FIŠ oz. v elektronski knjižnici FIŠ,
- ☐ da je doktorska disertacija, ki sem jo oddal v elektronski obliki identična tiskani verziji,
- ☐ da je doktorska disertacija lektorirana.

V Novem mestu, dne _____ Podpis avtorja _____

To my caring and supportive wife Tadeja.

Thank you so much for everything.

ABSTRACT

It has always been a challenging task to predict events in the near or distant future. People are interested in forecasting weather, earthquakes, floods, predicting economic, political and social changes, as well as the development of technology, sales products and sports outcomes. On the web, an enormous quantity of data is generated daily. We are practically deluged by all kinds of data – scientific, medical, financial, historical, health care, demographic, business, and other. Usually, there are not enough human resources to examine this data. However, from this chaotic cluster of data we strive to obtain valuable information, which may significantly impact strategic decisions of both business and individuals in the future. Predicting future trends and events has become easier and more efficient especially with the collaboration among scientists from various fields.

Sentiment analysis of web texts is an interesting and relevant research topic in this field. The aim of research described in this dissertation was to create specific language resources for sentiment analysis in Slovene, evaluate performance of sentiment based classification techniques and monitor the dynamics of sentiment, especially for the purpose of improving and contributing to computational analysis of texts in Slovene.

Here, we introduce the construction of Slovene web-crawled news corpora and a lexicon for sentiment analysis in Slovene. Besides their availability, we describe the methodology and the tools that were required for their construction. The corpora contain more than 250,000 documents with political, business, economic and financial content from five Slovenian media resources on the web that were published between 1st of September 2007 and 31st of January 2016. They include sentiment annotation on three levels of granularity: sentence, paragraph and document level. More than 10,000 of them were manually annotated as positive, negative or neutral. A Slovene sentiment lexicon, which is based on the annotated documents, contains more than 25,000 words with sentiment ratings, and is the first of this kind for Slovene. In detail, we describe the construction of these language resources, the manual annotation process and its characteristics. All developed resources are publicly available under Creative Commons copyright license.

We used the annotated documents to assess the sentiment classification approaches. Experimental performance evaluation of sentiment based classification techniques gives encouraging results. When classifying documents, in terms of time consumption and performance, the Multinomial Naïve Bayes and the Support Vector Machines approaches outperform the other classifiers. Also, consideration of smaller text segments, such as sentences, improves the performance. Models achieve F1-score value of 97.85% within the two-class (positive and negative) and 77.76% within the three-class (positive, negative and neutral) document-level sentiment based classification.

The sentiment analysis methodology was successfully used in the real-world applications for estimating the proportions of positive, negative and neutral news in the selected web media, and for monitoring the dynamics of sentiment. When estimating the proportions of positive, negative and neutral news, approximately half of the retrieved news is neutral. In general, the proportion of negative news is twice as high as the proportion of positive news. The study of sentiment dynamics shows that sentiment is on average more explicit at the beginning of documents and loses sharpness towards the end.

KEYWORDS: news corpus, sentiment analysis, lexicon, corpus linguistics, machine-learning, document classification, monitoring sentiment dynamics

POVZETEK

Napovedovanje dogodkov v bližnji ali daljni prihodnosti je od nekdaj veljalo za zahtevno. Ljudje se zanimajo za napovedi vremena, bližajočih se naravnih katastrof, gospodarskih, političnih in socialnih sprememb, kot tudi za trende v razvoju tehnologij, prodajo izdelkov in napovedovanje športnih izidov. Na svetovnem spletu se vsak dan objavi ogromna količina podatkov. Praktično smo zasuti z različnimi vrstami podatkov, ki izhajajo iz področij znanosti, zdravstva, financ, poslovanja, demografije, zgodovine in drugih, pri čemer nam v postopkih obdelave podatkov običajno primanjkuje človeških virov. Kljub vsemu si prizadevamo pridobiti dragocene informacije iz tega kaotičnega skupka podatkov, z namenom, da bi lahko v prihodnje izboljšali strateške odločitve tako posameznikov kot podjetij. Napovedovanje trendov in dogodkov v prihodnosti je postalo lažje in bolj učinkovito, še zlasti s sodelovanjem med znanstveniki z različnih področij.

Analiza sentimenta spletnih besedil je zanimivo in relevantno raziskovalno področje. Cilj raziskav v sklopu te disertacije je izdelava posebnih jezikovnih virov za analizo sentimenta, ocena učinkovitosti klasifikacijskih metod in spremljanje dinamike sentimenta, z namenom, da pripomoremo k boljšemu računalniškem razumevanju besedil v slovenskem jeziku.

V okviru te raziskave so opisani postopki za izgradnjo (s sentimentom) označenih korpusov novic in leksikona za analizo sentimenta v slovenskem jeziku. Poleg dostopnosti do razvitih jezikovnih virov so opisani tudi metodologija in orodja, ki so bila za to potrebna. Korpusi vsebujejo več kot 250 tisoč spletnih besedil ter vsebujejo politična, gospodarska in finančna besedila, ki so bila objavljena med 1 septembrom 2007 in 31 januarjem 2016 s strani petih spletnih medijev v Sloveniji. Dokumenti so bili označeni na treh nivojih, tj. na ravni dokumenta, na ravni odstavkov in na ravni stavkov. Več kot deset tisoč dokumentov je bilo ročno označenih kot pozitivni, negativni in nevtralni. Leksikon je bil zgrajen na osnovi označenega korpusa besedil. Vsebuje več kot 25 tisoč besed z dodeljenim sentimentom. Je prvi leksikon za analizo sentimenta v slovenščini, ki temelji na ročnem označevanju slovenskih besedil. Podrobno so opisani postopki izgradnje jezikovnih virov,

ročnega označevanja ter njihove lastnosti. Vsi viri so javno dostopni pod licenco Creative Commons.

V nadaljevanju je predstavljena študija ocene učinkovitosti klasifikacijskih metod, ki daje spodbudne rezultate. Pri klasifikaciji dokumentov se Naivni (večrazsežnostni) Bayesov klasifikator in Metoda podpornih vektorjev izkažeta kot najbolj učinkoviti metodi z vidika časovne zahtevnosti in različnih mer točnosti. Prav tako segmentacija besedil na manjše dele, kot na primer stavke, pripomore k boljšim rezultatom klasifikacije. Pri klasifikaciji dokumentov v dva razreda (pozitiven in negativen) dosežemo F1-oceno 97,85%, pri klasifikaciji dokumentov v tri razrede (pozitiven, negativen in nevtralen) pa 77,76%.

Principe analize sentimenta smo uspešno uporabili tudi pri ocenjevanju deleža pozitivnih, negativnih in nevtralnih novic izbranih spletnih medijev ter pri spremljanju dinamike sentimenta. V okviru ocenjevanja pozitivnih, negativnih in nevtralnih novic je bilo ugotovljeno, da je približno polovica izmed vseh pridobljenih novic nevtralnih. V splošnem je delež negativnih novic dvakrat večji od deleža pozitivnih novic. Študija dinamike sentimenta je pokazala, da je v povprečju sentiment močneje izražen na začetku dokumentov in izgublja svojo izraženost proti koncu dokumentov.

KLJUČNE BESEDE: korpus novic, analiza sentimenta, leksikon, korpusna lingvistika, strojno učenje, klasifikacija dokumentov, spremljanje dinamike sentimenta

PREFACE

"We are what we repeatedly do.
Excellence, then, is not an act,
but a habit."

– ARISTOTLE

Sentiment analysis combines many scientific fields, some of which are already well established, such as natural language processing, text analysis and computational analysis. But some are just emerging, such as affect analysis, which addresses the use of linguistic, acoustic and video information. Many new areas increased their set of computational procedures in the last decade, and consolidated their position as an independent scientific area, mainly due to intensive interdisciplinary collaboration. Sentiment classification is one of the main tasks in the field of sentiment analysis. More and more scientific publications are focused on sentiment classification. In general, it includes two basic categories. The first is mainly engaged with language resources, such as natural language corpora and sentiment lexicons. Researchers usually use specific natural language processing techniques combined with language resources to improve the overall performance of the sentiment classification. The second category strives to implement machine-learning techniques to apply sentiment classification.

The goal of this dissertation is to present the construction and use of language resources, i.e. annotated web-crawled corpora and a lexicon for sentiment analysis in the Slovenian language, along with the description of related evaluation methods, technologies and applications. The content within dissertation is naturally divided into six parts. The first (Chapter 1) introduces the basic concepts, provides the motivation along with scientific contributions, and presents the structure and the framework of this dissertation. The second includes Chapter 2 and examines some historical backgrounds of the corpus linguistics and sentiment analysis. The third part, in Chapters 3-4, deals with the construction of language resources and their availability. The fourth, which includes Chapters 5-7, covers applications of our language resources, along with the empirical research and

evaluation. The fifth part (Chapter 8) deals with hypotheses and their testing. Finally, the sixth part (Chapter 9) concludes this dissertation. The results bring us one-step closer to a better (computational) understanding of texts, particularly in the Slovenian language.

There is one relevant assumption regarding the prerequisite knowledge of readers. A reader with at least some basic knowledge of algorithms and probability should have no problems with reading this dissertation.

Acknowledgements

I would like to gratefully and sincerely thank Dr. Janez Povh and Dr. Martin Žnidaršič for their assistance, diligence, patience, and most importantly, their guidance during my doctoral study at Faculty of Information Studies in Novo mesto. Their enthusiasm encouraged me to gain invaluable experience in various fields. In the past few years, I was fortunate that I worked with the members of the research group of Dr. Povh and other co-workers who gave me the opportunity to work on real-world research projects and to meet wonderful people from academia and industry.

Many outstanding experts from the Department of Knowledge Technologies at the Jožef Stefan Institute helped me with experience and advice. Moreover, I truly thank Dr. Tomaž Erjavec for his guidance and support, when constructing the Slovene sentiment lexicon JOB 1.0 and publishing the developed resources at the CLARIN website.

I sincerely thank the School of Computer Science, The University of Manchester, especially Dr. Goran Nenadić and his team for their assistance and guidance during my three months research in Manchester. My gratitude goes to Dr. Ingo Mierswa, Dr. Simon Fischer and their co-workers for giving me the opportunity to work in the fascinating R&D team at Rapid-I GmbH in Dortmund. I also thank the Department of Informatics, The University of Rijeka, especially Dr. Sanda Martinčić - Ipšić and her team for their assistance, guidance in research career, and for providing me with the foundation for becoming a data scientist.

I thank the members of my doctoral committee for their valuable comments, suggestions, hard work and their expertise. Not only they helped me grow into respected Teaching Assistant and researcher but also into self-confident independent thinker.

I would also like to thank Marjana Miškovič for her assistance within administrative matters, Jernej Gabrič for IT services and support, and Katja Bogovič for proofreading this dissertation.

The greatest gratitude goes to my parents, Jože and Silva, my sisters, Vesna and Janja with their families, my friends, Dejan and Miloš, and also to my wife's family, especially to Zdravko and Olga. They have helped me in so many ways. Finally, and most importantly, I would like to express my special thanks to my wife Tadeja. She has taken care of almost everything at home, with a special care for our son, and put up with me during my research. Words cannot describe her support, selflessness, patience and unwavering love in the past six years, ever since we have been together. I dedicate this research to her.

Veliki Kamen, July 2017 Jože Bučar

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABC ..... American Broadcasting Company
ACL ..... Association for Computational Linguistics
AFINN ..... Affective Lexicon by Finn Årup Nielsen
ANEW ..... Affective Norms for English Words
BBC ..... British Broadcasting Corporation
BOW ..... Bag-Of-Words
CBS ..... Columbia Broadcasting System
CC BY-SA ..... Creative Commons copyright license Attribution-ShareAlike
COCA ..... Corpus of Contemporary American English
COHA ..... Corpus of Historical American English
CRISP-DM ..... CRoss-Industry Standard Process for Data Mining
CV ..... Cross-Validation
DT ..... Decision Table
GB ..... Gigabyte
GI ..... General Inquirer
GOS ..... Corpus of Spoken Slovene
HTML ..... HyperText Markup Language
ID ..... Identifier
IE ..... Information Extraction
IR ..... Information Retrieval
JOS ..... Linguistic Annotation of Slovene
KNN ..... k-Nearest Neighbour
LOB ..... Lancaster-Oslo/Bergen
MB ..... Megabyte

MPQA ..... Multi-Perspective Question Answering
MSD ..... Morphosyntactic Descriptions
NBC ..... National Broadcasting Company
NCR ..... National Cash Register
NBM ..... Multinomial Naïve Bayes
NLP ..... Natural Language Processing
PC ..... Personal Computer
POS ..... Part-Of-Speech
RAM ..... Random-Access Memory
RF ..... Random Forest
RSS ..... Rich Site Summary
SIGWAC ..... Special Interest Group on Web as Corpus
SE ..... Standard Error
SLR ..... Simple Logistic Regression
SMS ..... Short Message Service
SPSS ..... Statistical Package for the Social Sciences
SQL ..... Structured Query Language
SSKJ ..... Dictionary of the Slovenian Standard Language
SVM ..... Support Vector Machines
SVM-lin ..... Support Vector Machines with Linear Kernel
SVM-poly ..... Sequential Minimal Optimization with Polynomial Kernel
TF ..... Term Frequency
TF-IDF ..... Term Frequency - Inverse Document Frequency
URL ..... Uniform Resource Locator
UTF ..... Unicode Transformation Format
VP ..... Voted Perceptron
WaCky ..... The Web-As-Corpus Kool Yinitiative
WEKA ..... Waikato Environment for Knowledge Analysis

1 INTRODUCTION

The growing interest in efficient analysis of informal, subjective and opinionated web texts has led to a remarkable development in the field of sentiment analysis. Since 2010, there has been a rapid and steady growth in the number of scientific studies on this subject. Many papers report on the perception of emotion (i.e. sentiment) in text messages (Alm, Roth & Sproat, 2005; Thelwall, Buckley, Paltoglou, Cai & Kappas, 2012), for example, forecasting the outcomes of elections, based on comments found on Twitter and other social media resources (Tumasjan, Sprenger, Sandner & Welpe, 2010; Burnap, Gibson, Sloan, Southern & Williams, 2016), predicting future events (Asur & Huberman, 2010; Bothos, Apostolou & Mentzas, 2010), as well as addressing issues of global security, such as the global war on terror, etc. (Cheong & Lee, 2011; Wang, Gerber & Brown, 2012; Burnap et al., 2014).

Data scientists strive to improve the computational understanding of the world's languages. Therefore, it is not surprising that the availability and the use of language resources for the purposes of computational linguistics have increased significantly in recent years. Most of the language resources are in English; however, there is an increasing interest in other languages.

The topic of this dissertation is the analysis of sentiment of the web texts. We describe a procedure of building annotated web-crawled news corpora, a collection of various news corpora written in Slovene, present a construction of lexicon for analysing sentiment in the Slovenian language, and provide applicationand evaluation of sentiment analysis on developed language resources. These resources were built by web crawling in several attempts between 2013 and 2016. They contain sentiment annotations of political, business, economic and financial news, which was published between 1st of September 2007 and 31st of January 2016 from five Slovenian web media. The resources are freely available under the terms and conditions specified in Section 4.3.

Within this chapter, we first present basic concepts, applicationand objectives of sentiment analysis. Next, we explain the motivation for this dissertation, present the goals, research hypotheses, its scientific contributions and methodology. Finally, we provide the organizational structure of the dissertation.

1.1 Introducing Basic Concepts

In this section, we present basic concepts that are related to this dissertation. We first explain the importance of the web as a source for research activities of computational linguistics and text analytics. Also, we present the purpose and objectives of sentiment analysis and classification of documents.

1.1.1 The Web

In recent decades, the use of language technologies and resources has changed significantly, especially with the emergence of the web. The web experienced a great success all around the world with tremendous media support. It became new media and source for advertising and providing information. The growing interest has emerged rush to integrate new processes, features that can contribute to more efficient work.

Today, the web is a growing universe of websites and a huge repository of structured and unstructured data. With its varied and freely accessible data, it is a remarkable source of data for language and data scientists.

Although the structured type of data is easier for computer processing, there is a tendency to find a way to generate intelligence from documents containing unstructured information on the web. When dealing with textual data in particular, which is a conventional example of unstructured data, language scientists are more frequently turning to the web as a source. This is because it is so huge, it contains facts, emotions and opinions, which we can extract, or simply because it is free and constantly available (Kilgarriff & Grefenstette, 2003). The most common tool of mass data acquisition is a web crawler (internet bot or spider), which systematically visits the web and retrieves relevant information. It can extract multiple types of data, such as text, tables, images, links, videos and more. The amount of web content like customer feedback, competitor information,

client emails, tweets, press releases, legal filings, product and engineering documents, etc., rapidly grow. In addition, humankind is still hungry of knowledge derived from retrieved information.

Relevant information about a company, its structure, employees, activities, products and services can occur anywhere on the web. Although they can be either true or false, it has significant impact on public opinion and its response. However, more and more business, sale, finance, and other companies are aware of people's opinion. An increasing number of blogs, web sites, newsgroups, forums, chat rooms, etc., has allowed people to express and aggregate their feelings about products, services, events more intensively. It has made it possible to extract the opinions regardless of whether we are looking for opinions on the candidates related to upcoming elections or opinions about the holiday destination we tend to visit. As the phrase goes, *"The customer is king"*, in the eyes of the company, it is crucial to understand people's needs, feelings and satisfaction.

1.1.2 Corpus

Data scientists and linguists cannot actually work from observing a large amount of language use situated within its context in the world. So, instead, they simply use texts, and regard the textual content as a surrogate language in a real world context. A body of text is called a *corpus* - *corpus* is simply Latin for *"body"*, and when you have several such collections of texts, you have *corpora* (Manning & Schütze, 1999).

Oxford dictionaries (2017) define *corpus* as:

- *"A collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject."*
- *"A collection of written or spoken material in machine-readable form, assembled for the purpose of linguistic research."*

Corpora are used for a variety of purposes (Erjavec, 2010), such as construction of lexicons and other language resources, construction of grammar and other descriptions of linguistic structure, development of tools for translation, development of tools for learning languages, study of linguistic behaviour and language technologies.

There are many organizations that distribute text corpora for linguistic purposes,

such as Linguistic Data Consortium[1], European Language Resources Association[2], International Computer Archive of Modern and Medieval English[3], Oxford Text Archive[4], etc. Most of them are not freely available; in fact, they often charge a lot of money for their distribution. However, many are freely available on the web. Please see Chapter 2, where corpora are discussed in more detail.

1.1.3 Sentiment Analysis and Sentiment Classification

Subjectivity and Objectivity

Textual information can be categorized into facts and opinions. Fact is an objective expression about entities, events and their properties, items of information, or state of affairs existing, observed, or known to have happened, and which is confirmed or validated to such an extent that is considered reality.

We often use terms opinion and sentiment. On one hand, there are some differences among authors in interpretation regarding sentiment. In fact, in practice many researchers avoid definition of the *"sentiment"*. Boiy and co-authors (2007) for example, compare sentiments with emotions, judgments and ideas, which are prompted or colored by emotions. Liu (2010) defines emotions as subjective feelings and thoughts. Many researchers from various fields have been studying emotions, and jet, they cannot agree with a set of basic emotions. Peng and co-authors (2010) noted that people express six primary emotions, i.e., love, joy, surprise, sadness, fear and anger, which hold different intensities, and can be further subdivided into secondary and tertiary emotions. On the other hand, different authors define opinions in a similar way. Liu (2010), for example, treats opinions as expressions with subjective annotation that is based on personal interpretation, views, assumptions, emotions and judgements, or feelings toward entities, events and their properties. In everyday discussions, we often use subjective reviews. Also, when developing breaking research or just purchasing decision about a certain product, we often look for people that had experiences in a field of our research or product we are interested in. It

[1] `https://www.ldc.upenn.edu/`
[2] `http://www.elda.org/en/`
[3] `http://clu.uni.no/icame/`
[4] `https://ota.ox.ac.uk/`

is completely natural that we are looking for other's opinions.

An objective sentence expresses some factual information about the world, while a subjective sentence expresses some personal feelings or beliefs. Objective opinion implies a regular or comparative opinion, which usually expresses a desirable or undesirable fact, e.g., *"The more expensive Dell UltraSharp U3014 30" has better screen resolution than Dell UltraSharp U2412M 24"". "Cockta tastes better than Coca-Cola"* is a subjective statement that gives a regular or comparative opinion.

Semantics

Content of subjective opinions is highly context-sensitive, and its expressions often differ from person to person. Hence, we have to distinguish between subjective statements and false statements since subjective does not mean not true. Let me give an example with sentence *"Andy loves candy!"* Regardless of whether given sentence is true or not, in any case reflects Andy's feelings towards candy, which is that he enjoys candies.

Semantics is actually related to syntax. In most languages, the syntax is how you say or write something, where semantics is the meaning behind what you said or wrote. Let me use the previous example *"Andy loves candy!"* The syntax is represented with all the letters, words and punctuation in sentence, where semantics is actually the true meaning behind these words. At this point let us change this sentence with *"Andy ♡ candy!"* As you can see we have changed the syntax, however, notice that semantics of the sentence stays the same. Semantic orientation, which usually captures positive or negative evaluative factor, is a measure of subjectivity and opinion in texts.

Text Mining, Sentiment Analysis and Sentiment Classification

Text mining is an interesting area of data analysis that deals with a range of technologies for analysing and processing semi-structured and unstructured text data.

At the beginning of text mining activities scientists were dealing mainly with diverse forms of information retrieval and information summarization, like abstracts and grouping of documents (Lancaster, 1968; Salton & McGill, 1986). Later researchers focused on information extraction (Ready & Wintz, 1973; Rau, Jacobs & Zernik, 1989).

By tagging each document, we can extract the content and structure from a corpus of documents. Information extraction consists of an ordered series of steps designed to extract terms, attributes of the terms, facts, and events (Devore, Feldman & Sanger, 2009). Typical text mining tasks include classification and categorization of texts (documents), topic detection, sentiment analysis, summarization (summary) of texts, and the study of relationships between entities in the texts.

Miner and colleagues categorized text mining into seven sub-disciplines, based on the answers to the preceding questions (Miner et al., 2012):

- Search and information retrieval (IR): Storage and retrieval of text documents, including search engines and keyword search,
- Document clustering: Grouping and categorizing terms, snippets, paragraphs, or documents, using data mining clustering methods,
- Document classification: Grouping and categorizing snippets, paragraphs, or documents, using data mining classification methods, based on models trained on labelled examples,
- Web mining: Data and text mining on the Internet, with a specific focus on the scale and interconnectedness of the web,
- Information extraction (IE): Identification and extraction of relevant facts and relationships from unstructured text; the process of making structured data from unstructured and semi-structured text,
- Natural language processing (NLP): Low-level language processing and understanding tasks, e.g., tagging parts of speech (POS); often used synonymously with computational linguistics,
- Concept extraction: Grouping of words and phrases into semantically similar groups.

Listed practice areas overlap considerably, since most text mining approaches can be considered in multiple practice areas. A Venn diagram in Figure 1.1 visualizes this overlap between practice areas. It illustrates the intersection between the text mining areas (ovals) and specific text mining tasks (labels within ovals).

The popularity of social media, such as social networks and others, has escalated interest in sentiment analysis (Wright, 2009). Sentiment analysis is also known under other names, such as opinion mining, subjectivity analysis and appraisal extraction, which gen-

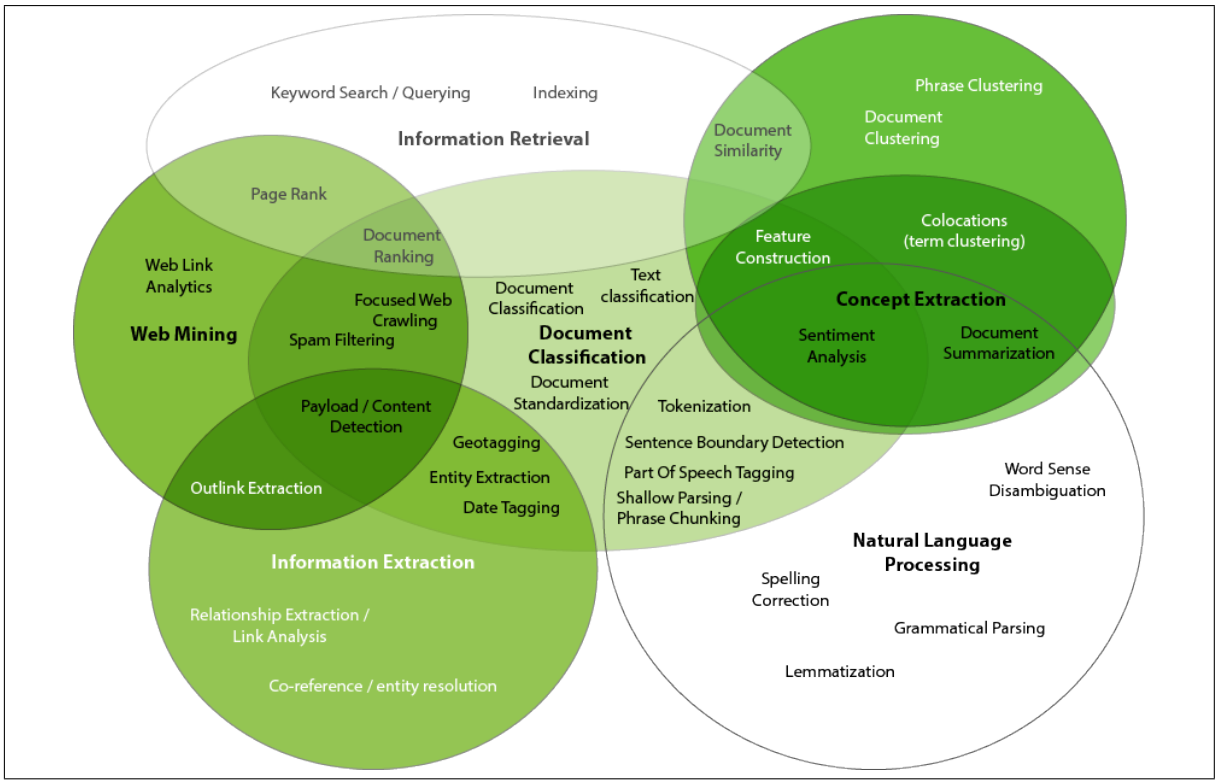

Figure 1.1: Text mining areas and tasks (Miner et al., 2012)

erally deals with subjective elements as sentiment units (words, phrases, sentences, or whole document). It is a challenging natural language processing and text mining problem, which brings together scientists from different fields like computational linguistics, data mining, computer science, machine-learning, graph theory, neural networks, sociology and psychology.

The purpose of sentiment analysis is to identify, extract and determine sentiment of source material through the expression and contextual polarity of the source. Opinion can be reflected through judgement or evaluation, emotional state of the subject (source), or state of emotional communication, by which they would like to impact on people's opinion or decision. Thus, it is a way, where people try to determine person's state of mind on specific subject they are talking about. Information as such can be extracted on-line from news articles, blogs, chats, forums, tweets, reviews, comments, and other web texts. Objectives of sentiment analysis are usually presented in three stages: sentiment detection and identification, polarity classification, and discovery of the opinion's holder and target (Pang & Lee, 2008; Mejova, 2009; Liu, 2010).

Sentiment classification is one of the most prominent techniques used in text mining. It is a supervised learning problem with usually three sentiment categories: positive, neg-

ative and neutral opinion or sentiment (Pang, Lee & Vaithyanathan, 2002; Melville, Gryc & Lawrence, 2009; Taboada, Brooke, Tofiloski, Voll & Stede, 2011). When classifying texts, we assign a known set of labels to unlabelled texts, using a model of text learned from texts with known labels. Advanced sentiment classification seeks even more complex emotional states, such as affection, pleasure, pride, regret, jealousy, agitation, etc.

Issues and Open Questions

Sentiment analysis faces many difficulties. One of the main reasons for the lack of study on opinions is the fact that before the emergence of the web there was not much opinionated text available. Some big concerns represent defining opinions and subjectivity, detection of negation, sarcasm, humour, opinion citations, quotations, speculations, and problems related with emotion and content perception. People commenting on the Internet often pay no attention to grammar. They change, duplicate, and omit letters, overreact, use figurative meaning, slang phrases, superlative words, abbreviations, uppercase letters, exclamation marks, etc.

It often occurs that in one situation a word contributes to positive sentiment, while in another situation it may be considered as negative. For example, let us take under consideration the word *"long"*. It makes a huge difference if a customer applies word *"long"* in conjunction with, for example, smart-phone's battery life expectancy, which would be a positive opinion; or if a customer relates it with smart-phone's start-up time, which, however, would give negative opinion. Understanding this problem and distinction between them means awareness that system trained to detect sentiment in one problem domain may not perform very well on another. Most of the related work has been done on product and movie reviews (Pang, Lee & Vaithyanathan, 2002; Taboada, Brooke, Tofiloski, Voll & Stede, 2011), where it is easy to identify the topic of the text. It is useful to pay attention to which characteristics of this product or service the writer is talking about: is it perhaps the smart-phone battery life or its start-up time that concerns consumers the most?

Another issue is negation. Sentence *"the lunch at your restaurant was tasty"*, however, has completely different meaning from *"the lunch at your restaurant wasn't tasty."*

One of the challenges in sentiment detection is how to detect and identify holder or

target. An opinion without its target being identified is of limited use. Understanding the significance of opinion holder and target contributes to the improvement of the sentiment analysis algorithms. For example, although the sentence *"although the service is not that great, I still love this restaurant"* clearly has a positive notation, however, we cannot say that this sentence produces only positive sentiment. Upper sentence has positive sentiment about the restaurant, but negative about its service.

Another problem is also quotation. Unlike usual topical analysis, sentiment statement authorship represents an integral part in the series of problems. If we follow daily reports or debates in parliament, we can discern that both news and political arguments are full of quotations and opinion citations. Some structures of sentences can be so complex that there can be some difficulties in sentiment detection. An article that includes daily news with political discussions, for example, would include not only quotations from the debaters, but also the pundits commenting on the debate, and perhaps even the author's stance on the issues.

Different people express their opinions differently. Even more, some people are contradictory in their statements, because they vary while expressing their own opinions. It is completely normal that people express both positive and negative opinions in their reviews or discussions. In most cases, this does not pose major obstacles to computational analysis, because it is quite successful by analysing sentences one at a time. Many traditional text-processing algorithms are not efficient enough, when small differences between two pieces of text occur. However, the more informal the media, the more likely people are to combine different opinions in the same sentence (Liu, 2010). Sentence, for example, *"Service was terrible but the food was excellent"* does not represent any problem to human perception; however it is a hard nut to crack for a computer. People often encounter the problem, how to express opinions, thoughts, or update their status in social network media like Twitter with limited number of signs. The consequence is that even other people sometimes find it difficult to understand what someone thought based on a short commentary or status line because there is simply lack of context. Meaning of commentary *"Food was as good as the last time"*, for example, is not clear enough because customer expressed the opinion based on previous experience, which is probably unknown to the vast majority of other readers.

One day, while checking forums related with textual sentiment analysis and natural

language processing, I encountered commentary, which I find amusing and entertaining at the same time. A linguistics professor was lecturing to her class one day. *"In English,"* she said, *"A double negative forms a positive. In some languages, though, such as Russian, a double negative is still a negative. However, there is no language wherein a double positive can form a negative."* A voice from the back of the room piped up, *"Yeah ... right"* (MARCUSQ, 2017).

1.2 Motivation

According to the World Wide Web Technology Surveys (2017), an estimation shows that the number of web users increased by more than eight times between 2000 and 2016, which represents more than 46% of the world's population.

The content on the web is written in many different languages. Xu (2000) estimated that English is used by 71% of all web pages, followed by Japanese (6.8%), German (5.1%), French (1.8%), Chinese (1.5%), Spanish (1.1%), Italian (0.9%) and Swedish (0.7%). In January 2017 it was recorded that 52.3% of all the web pages, whose content language is known, are in English, followed by Russian (6.4%), Japanese (5.7%), German (5.4%), Spanish (5.0%), French (4.0%), Portuguese (2.6%), Italian (2.3%), Chinese (2.0%) and Polish (1.7%). In 2016 we have observed a clear trend of reducing the proportion of pages on the web being written in English (-1.6%), German (-0.4%) and Polish (-0.2%), while the proportion of web pages increased for for Japanese (+0.7%), Persian (+0.4%), Russian, Spanish, Italian and Korean (all +0.2%).

Only 2 million people speak Slovene, which puts it on the 36th place among the most common languages on the web. The proportion of web pages written in Slovene increased from 0.081% to 0.091% in 2016 (World Wide Web Technology Surveys, 2017).

In the recent decade, sentiment analysis of web texts has experienced increased attention. Huge amount of textual information available on the web emerged a need to find and obtain relevant information for strategically supported decisions. Although the field of sentiment analysis is relatively young, both industry and academia understand advantages of sentiment extraction from web texts. As shown in Figure 1.2, the number of scientific publications that mention sentiment analysis, opinion mining and subjectivity analysis significantly increased in the past 10 years, which also reflects the interest in the

research area.

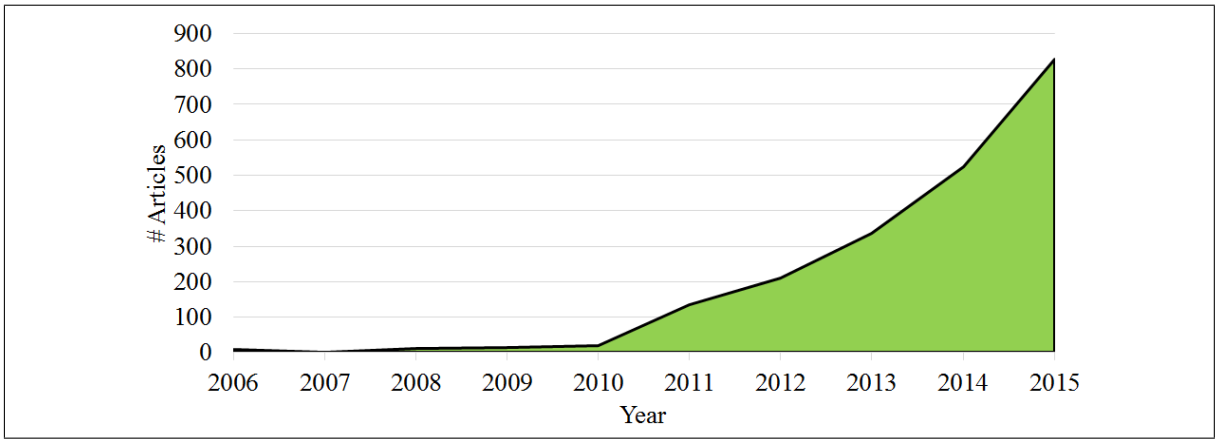

Figure 1.2: Scientific publications of *"sentiment analysis"*, *"opinion mining"* and *"subjectivity analysis"* (Web of Science, 2016)

Liu (2010), for example, introduced the problem of sentiment analysis by following review segment on iPhone: *"(1) I bought an iPhone few days ago. (2) It was such a nice phone. (3) The touch screen was really cool. (4) The voice was clear too. (5) Although the battery life was not long, that is ok for me. (6) However, my mother was mad with me as I did not tell her before I bought it. (7) She also thought the phone was too expensive, and wanted me to return it to the shop. . . ."*

A quick overview reveals that customer made this review after purchasing iPhone, as seen in sentence (1), which is actually neutral statement. We can detect positive sentiment or opinion in sentences (2, 3 and 4), while sentences (5, 6 and 7) include negative connotation. Furthermore, we can notice that there are different targets on which customer expresses the opinions. Sentence (2), for example, considers iPhone as a whole, while the following sentences (3, 4 and 5) reflect user's opinion on *"touch screen"*, *"voice quality"* and *"battery life"*. The opinion in sentence (6) actually refers to user and not iPhone. The last sentence (7) applies to iPhone again, more precisely on its price. It is also not so difficult to distinguish between two sources or holders of opinions in this text. Sentences (2, 3, 4 and 5) are related to the author of this iPhone review, while next two sentences (6 and 7) represent author's mother opinion.

In the 80's an article by communications professors Stone and Grusin found out that the average amount of negative news on ABC, CBS and NBC was 46.8% (Stone & Grusin, 1984). Since then the proportion of negative news has increased in most media. Negative news is often cheap and easy to produce; moreover, it makes profit to the media. They

draw more attention than similar positive news, which is reflected in longer viewing times (Ho, Chen & Sim, 2013; Trussler & Soroka, 2014; Vinkers, Tijdink & Otte, 2015; Kätsyri, Kinnunen, Kusumoto, Oittinen & Ravaja, 2016). Some media are obligated to regulate the proportion of positive and negative news. The Romanian senate, for instance, passed a law in 2008 requiring the media to provide their audiences with fifty percent positive news. The law has encountered a disapproval of the Romania's National Council for Audiovisual Broadcasting, which stated that news should reflect reality – whether positive or negative – independent of any laws (International Journalists' Network, 2008).

In 2012 a study was carried out, where over the period of 35 months - from December 2008 to October 2011 - a random selection of 2,386 headlines and short abstracts from 8 different Slovenian media were gathered using RSS aggregator (Kovačič, 2012). Evaluators were assessing RSS based on a question: *"How did this news make you feel?"*. To improve reliability of their research each headline and short summary (up to 250 characters) was evaluated twice by two independent native evaluators. Author also claims that evaluators' judgements were not influenced by any of the authors or support team. The results showed that in all tested media there is a strong negative bias (see Table 1.1).

Table 1.1: Media tone in different Slovenian media (Kovačič, 2012)

	Evaluation tone (count and % within media)			
Media	Positive	Neutral	Negative	Total
24ur	80 (25.7%)	79 (25.4%)	152 (48.9%)	311 (100%)
Finance	52 (17.2%)	63 (20.8%)	188 (62.0%)	303 (100%)
Rtvslo	42 (15.0%)	110 (39.3%)	128 (45.7%)	280 (100%)
Žurnal24	31 (10.3%)	114 (37.9%)	156 (51.8%)	301 (100%)

1.3 Goals, Hypotheses and Scientific Contributions

1.3.1 Goals

The goals of this research are grouped into 4 groups, containing one or more sub-goals:

1. Provide a comprehensive literature review.
    - In-depth overview of existing sentiment analysis approaches;

- Overview of significant corpora and lexicons for sentiment analysis;
- Overview of relevant studies and their performances.

2. Development of the language resources in the Slovenian language.
   - Build web crawlers and retrieve the news (texts) from digital archive of five Slovenian web media;
   - Clean the data;
   - Create a web application for manual annotation, annotate the data with several annotators at three levels of granularity;
   - Calculate the degree of agreement between annotators;
   - Develop an annotated news corpora and a lexicon for sentiment analysis in Slovene;
   - Provide access to publicly available tools and language resources with the terms and conditions of their use and distribution.

3. Propose a selection of the appropriate machine-learning classifiers and data pre-processing settings of the (web) texts in the Slovenian language.
   - Select the most appropriate classifiers for sentiment analysis of the (web) texts in the Slovenian language;
   - Select the most appropriate pre-processing settings for sentiment analysis of the (web) texts in the Slovenian language;
   - Find the impact of documents granularity on document classification;
   - Evaluate performance of the most appropriate classifiers and pre-processing settings for sentiment analysis of the (web) texts in the Slovenian language.

4. Provide real-life applications of the developed language resources.
   - Estimate the proportions of positive, neutral and negative news in five Slovenian web media;
   - Monitor the sentiment dynamics of the web media from different perspectives (within documents, over time, topic-sentiment and within authors of documents).

1.3.2 Hypotheses

The aim of this research is to create specific language resources for sentiment analysis, evaluate performance of sentiment based classification techniques, estimate the proportion of positive, negative and neutral news in the media, and monitor the dynamics of sentiment especially for the purpose of improving and contributing to computational analysis of texts in the Slovenian language. This research is based on four hypotheses concerning topical issues in the field of sentiment analysis. The following hypotheses have been investigated:

1. **Hypothesis 1 (H_1):** Appropriate selection of supervised machine-learning classifier and pre-processing settings can improve the classification performance.

2. **Hypothesis 2 (H_2):** Granulation of a document to smaller segments, such as sentences, can improve the classification performance.

3. **Hypothesis 3 (H_3):** The developed sentiment analysis tools, resources and methodology are applicable in real-life applications.

4. **Hypothesis 4 (H_4):** In the retrieved news with political, business, economic and financial content from five Slovenian web media, the proportion of negative news is greater than the proportion of positive news for all web media.

1.3.3 Scientific Contributions

The main contributions presented in this dissertation are:

1. Introduction and detailed description of the procedure used for building an annotated, web-crawled, news corpus, a collection of various news corpora written in Slovene. Three levels of sentiment granularity were used: sentence, paragraph and document level. Provision of all the newly developed language resources freely under the terms and conditions specified in Section 4.3.
2. Assessment of 9 machine-learning approaches to sentiment classification of Slovene

texts into two and three classes. In particular, Multinomial Naïve Bayes (NBM) reached F1-score above 97% for a two-class sentiment and more than 77% for a three-class sentiment classification when we used balanced training datasets and document sentiment scores based on sentence granularity.

3. Analysis of the within-the-document sentiment dynamics of manually annotated documents and indication that the sentiment is on average sharper at the beginning of documents and leans to neutrality towards the end of documents.
4. Applications of the developed tools and resources, useful in real-life applications.
5. Analysis of sentiment expressed in Slovenian web media. Approximately half of the obtained news texts is neutral with negative news texts twice as much as positive.

1.4 Methodology

The research follows standardized procedure in data mining CRoss-Industry Standard Process for Data Mining - CRISP-DM (Chapman et al., 2000), which is described in Subsection 1.4. The methodology is based on qualitative and quantitative approaches to the identified research problem.

Within our research, we will select the data source in the Slovenian language that has not yet been classified and create an applicationto classify such data. News will be retrieved without comments from the digital archive of different Slovenian web media with political, business, economic and financial content.

Developed web crawlers will find and obtain objects within web pages and identify hierarchical relations between them to improve filtering process and retrieval, by selecting and eliminating certain kinds of objects, such as images. A module for automatic identification will capture all web texts with specified content from selected web media. HTML code retriever will enable automatic recognition of HTML code. For this purpose, we will develop customized text parsers for the default set of web media to enhance the acquisition of the content. Title and content will be extracted from a web page for further processing, while metadata writer will acquire parameters, such as URL address, date, author and keywords.

Furthermore, we will edit, clean and pre-process retrieved texts. Next, we will construct a stratified random sample from a population of retrieved news, approximately

2,000 news per web medium. Various native speaker annotators on different levels of granularity will manually annotate the sample . The annotators will be guided to specify the sentiment; detailed instructions will be given to ask annotators about their feelings after reading the news. The annotators will be independent; the degree of agreement among annotators will be calculated. This way, we will obtain manually annotated news corpora in Slovene.

Annotated corpora will then be ready to test and evaluate techniques for data classification, which is described in Subsection 1.4. Most commonly applied methods and algorithms will be compared with each other. When classifying textual information, it is necessary to evaluate results. Discussed hypotheses will be substantively and statistically analysed as described in Section 1.3.

1.4.1 CRISP-DM

The CRISP-DM (CRoss Industry Standard Process for Data Mining) was conceived in 1996 from collaboration of five companies including car manufacturer Daimler-Benz, insurance company OHRA, hardware and software manufacturer NCR Corp., data warehouse company Teradata, and statistical software maker SPSS, Inc. It describes commonly used procedures that help data miners to solve problems (Chapman et al., 2000).

CRISP-DM breaks the data mining process into following six phases: business (organizational) understanding, data understanding, data preparation, modelling, evaluation and deployment. The sequence of the phases is not fixed; moreover, data mining experts must often dive skilfully between the phases. In Figure 1.3 we present a cyclic process flow for data mining, based on the CRISP-DM. The feedback loop indicates that findings and lessons learned at any phase in the process can trigger a backward movement for corrections and refinements, and the completion of a process may lead to new and more focused discovery processes.

The initial phase requires determination of the purpose of the study from a business perspective. This is a crucial step in the data mining process, since we need to understand the business, processes and requirements to plan solutions for achieving the objectives. To understand the business perspective to the last detail we must often collaborate with the domain experts in order to develop an in-depth appreciation of the underlying system,

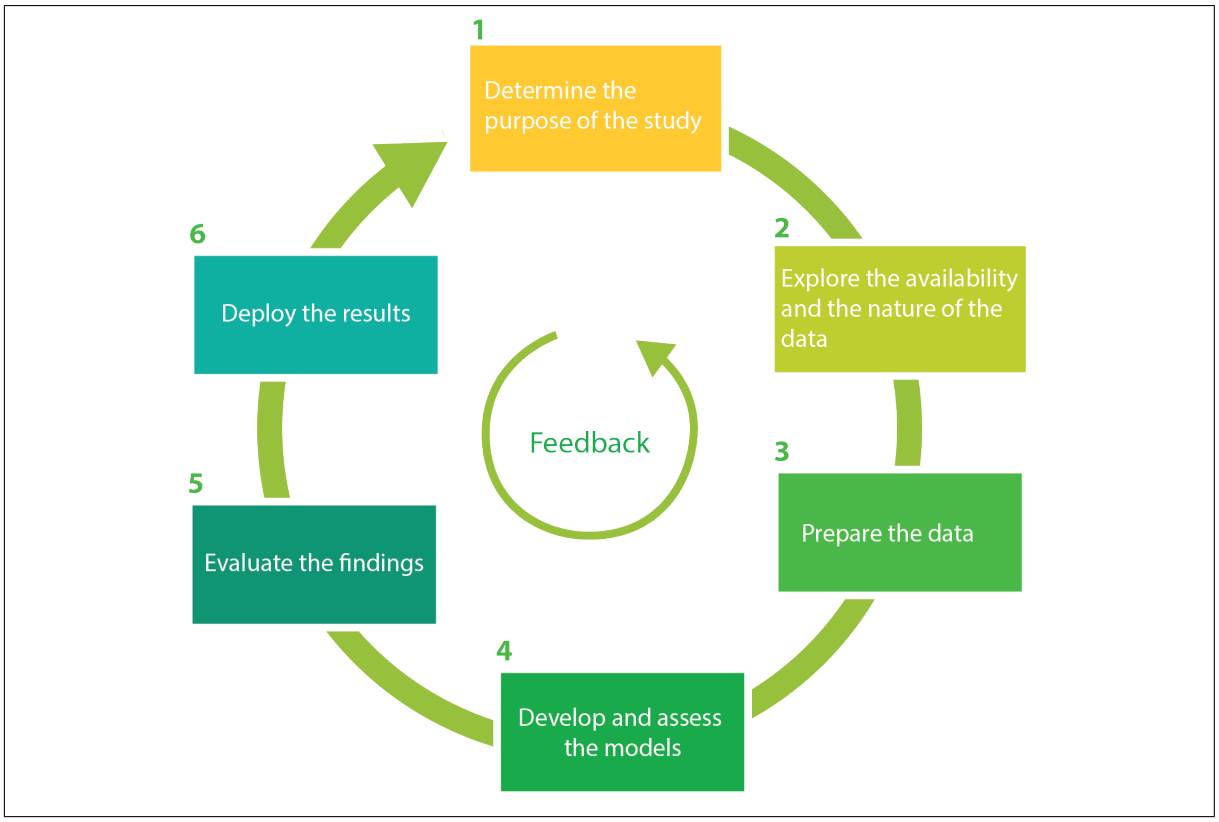

Figure 1.3: A CRISP-DM process flow for text mining (Chapman et al., 2000)

its structure, system constraints and the available sources.

Once we successfully determined business objectives, data mining goals and produced a project plan, we need to focus on how can we access and obtain the data, and whether the obtained data properly explains the nature of our problem. For that reason, we need to identify the data resources, determine their accessibility and usage, obtain and explore the data, as well as choose the appropriate data in relation to the scope of the study. Data miners spend some time to get familiar with the data. They also make some preliminary experiments to verify quality of the data, and to discover interesting content to form hypotheses or to find hidden information and patterns. In order to obtain high quality of data, we often need to collect large quantities of data from different sources.

The goal of the data preparation phase is to establish the final data set (from the raw data) as an input into the modelling phase. This phase includes selection, cleaning, construction, integration and formatting the data. In general, the sequence of the data preparation tasks is not rigorously determined, and we often need to repeat some stages within this phase.

The modelling phase covers all activities regarding selection and applicationof mod-

elling techniques, including adjustments and calibrations of their parameters, testing designs, building and assessing models. Similar to the previous phase, some tasks are likely to be performed multiple times, in order to find the optimal parameter settings of our models. Whenever we are dealing with techniques that require a specific form of the data, we often need to return to the data preparation phase.

Once we have built models, for which we consider sufficient in terms of accuracy and from a data analysis perspective, we need to evaluate the models. Therefore, we review all previous steps if the activities were properly executed, and if the constructed models achieve the business objectives. If we find our models as incomplete, e.g., they do not materialize our business problem in a proper way, or one of the key objectives was omitted or not sufficiently considered, then we need to repeat and correct these issues to validate. This step is meant to ensure that the models developed and verified are actually addressing the business objectives that they were built for, before moving to the deployment phase.

The final phase of the CRISP-DM comes into sight when the models and the modelling process successfully pass the assessment process. Depending on the requirements, the deployment phase can be as simple as generating a report that explains the findings of the study, or it can be as complex as integrating new knowledge into business intelligence system. This phase includes deployment plan, monitoring and maintenance plan, final report and project review. The models can be used repetitively, however, with new data the models need to be refined, since their accuracy and significance can decrease over time. The updates of business intelligence system can be performed manually or (preferably) automatically as new and relevant data is available. The construction of such advanced systems is a challenging task, it requires large investments (especially time any money), however, the results should fulfil the expectations of the data analyst and the customer.

1.4.2 Machine-learning Approach for Sentiment Analysis

Several approaches can be applied to sentiment analysis, i.e. lexicon-based, linguistic and machine-learning approach (Thelwall, Buckley, Paltoglou, Cai & Kappas, 2012).

For determining the sentiment in text, lexicon-based methods (Ding, Liu & Yu, 2008;

Taboada, Brooke, Tofiloski, Voll & Stede, 2011) use sentiment lexicons and dictionaries (see Section 2.2). More sophisticated approaches try to identify text features that could potentially be subjective in some contexts and then use contextual information to decide whether they are subjective in each new context (Wiebe, Wilson, Bruce, Bell & Martin, 2004). In practice, it turns out that these methods are much faster when compared to the linguistic and machine-learning approach, but are usually unable to adopt changes which may occur in data streams (Smailović, 2014).

The linguistic approach (Pang & Lee, 2008; Thet, Na, Khoo & Shakthikumar, 2009) determines sentiment polarity by analysing the grammatical structure of the text. This approach is time consuming, which might pose a great disadvantage in its applicationwithin sentiment analysis.

The machine-learning approach (Freitag, 1998; Pang, Lee & Vaithyanathan, 2002; Sebastiani, 2002; Kotsiantis, 2007; Boiy & Moens, 2009; Witten, Frank & Hall, 2013) is the most frequently used approach in the field of sentiment analysis. Machine learning techniques for sentiment classification gain interest because of their easier capability to model many features and in doing so, capturing context (Polanyi & Zaenen, 2006), their adaptability to changing input, and the possibility to measure degrees of uncertainty by which classifications are made (Boiy & Moens, 2009).

Sentiment classification of text can be performed on various levels: document, sentence or word level. Document-level classification aims to find a sentiment polarity for the whole document, whereas sentence-level or word-level classification can express a sentiment polarity for each sentence and even for each word. Most of studies focus on a document-level classification (Collomb, Costea, Joyeux, Hasan & Brunie, 2014).

Due to the limitations of lexicon-based and linguistic methods mentioned in the second and in the third paragraph in this Subsection, we discovered that supervised machine-learning approach is the most suitable among the three listed approaches for sentiment analysis of the Slovenian news texts, and that is why we used it within our study. Two recurring machine-learning issues are feature selection and selection of classification algorithm, which are described hereinafter.

Feature Selection

Feature selection is the process of selecting a subset of relevant features that we would like to use for building our model. In general, the process is used for three reasons (James, Witten, Hastie & Tibshirani, 2013):

- To simplify models, so that they are easier to interpret,
- To reduce computational time for training,
- To enhance generalization by reducing overfitting.

The central premise when using a feature selection technique is that the data contains many features that are either redundant or irrelevant, and can thus be removed without incurring much loss of information (Bermingham et al., 2015). Redundant or irrelevant features are two distinct notions, since one relevant feature may be redundant in the presence of another relevant feature with which it is strongly correlated (Guyon & Elisseeff, 2003).

In our experiments we used implementations within the WEKA[5] machine-learning toolkit, version 3.6.11 (Witten, Frank & Hall, 2013), which converts string attributes into a set of features representing word occurrence information from the text contained in the strings. The set of features is determined by:

- Tokenization

  Texts can be tokenized in different ways. It is a process breaking a stream of text into smaller segments, such as words, phrases, symbols, or other relevant segments called tokens. Whitespace characters (space, line break, punctuations) separate tokens. Within our study we used N-gram tokenizer, which finds unigrams (N = 1), bigrams (N = 2), trigrams (N = 3), etc., depending on the frequency and proximity of words.

- Transformation of upper case letters to lower case

  Words that contain capital letters can be transformed into words with lower case letters. This is one way to reduce the number of attributes, which enables faster

[5] Implementations in WEKA 3.6.11:
filter: *StringToWordVector*
filter: *AttributeSelection*

processing, analysis and classification. For example, the word *text* and *Text* have the same attribute *text* after transformation.

- Removal of stop words

  In computational linguistics, stop words are words which are filtered out before or after processing of text (Leskovec, Rajaraman & Ullman, 2014). In general, they refer to the most common words in a given language. Using the list of words for the Slovenian language enables us to study its impact on the quality and analysis of sentiment classification of texts. To access this list of words see the fourth paragraph in Section 4.3.

- Lemmatization and stemming

  Lemmatization is the process of determining the word's lemma, a single term that is set to a group of inflected forms of a word by its intended meaning. Stemming is the process of determining the word's stem, which is a shortened term with a common root. Stems are not necessarily correct linguistic terms, e.g., *have* → *hav*. Usually a part of the word, such as suffix, is removed. Unlike stems, lemmas always represent correct linguistic expressions, e.g., *running* → *run* or *puppy* → *dog*. In our study, we applied only algorithms for lemmatization.

- Term frequency (TF) and Term frequency - inverse document frequency (TF-IDF)

  The weight of a word that occurs in a document can be proportional to the TF or TF-IDF. TF simply represents the number of times a word occurs in a document. TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general (Leskovec, Rajaraman & Ullman, 2014).

- Minimum number of word frequency

  Word frequency enables to list the most frequently occurring words in texts. By increasing the minimum number of occurrences of words, we can reduce the number of features. In practice, it is not uncommon that texts, even after cleaning and pre-processing, are still burdened with spelling errors. To eliminate such words and words that occur very rarely in texts, we set the minimum number or word frequency to more than 1. Selection of a minimum number or word frequency

should be used with caution, since selection of larger numbers can lead to reduction of relevant features, which could significantly affect the results of the study. In our study, we set the minimum number or word frequency to 2 in all cases.

- Words to keep

  Here we specify the number of words per class to attempt to keep. In our case, we were dealing with items classified into three classes (positive, negative and neutral). Similar to the selection of the minimum number or word frequency, it should be used with caution, since selection of smaller numbers can lead to reduction of relevant features, which could significantly affect the results of the study.

- Normalization of the document length

  In our study, we consider texts (news) that differ significantly in length (number of characters). Therefore, it is necessary to take into account the length of the document besides word frequency. Therefore, we normalized word frequencies of a document by using average and actual document lengths.

- Feature subset selection method

  Within this stage, we evaluate the worth of a feature by computing the value of different feature subset selection methods with respect to the class. We tested three most popular ranker search methods, i.e., Chi-squared, Gain Ratio, and Information gain (Yang & Pedersen, 1997; Hall & Smith, 1998; Joachims, 1998; Forman, 2003; Zheng, Wu & Srihari, 2004; Karegowda, Manjunath & Jayaram, 2010).

  Information gain feature selection method evaluates the worth of an attribute by measuring the information gain with respect to the class (it measures information obtained (in bits) for class prediction of an arbitrary text document by evaluating the presence or absence of a feature in that text document). Information Gain is calculated by the feature's contribution on decreasing overall entropy (Yang & Pedersen, 1997) as:

$$Information\ gain(A) = Info(D) - Info_A(D) \tag{1.1}$$

  Where the entropy is calculated as $Info(D) = -\sum_{i=1}^{m}(P_i)log_2(P_i)$ and the amount of information in bits as $Info_A(D) = -\sum_{j=1}^{v}\frac{|D_j|}{|D|}\cdot Info(D_j)$. We use the following notation: m - the number of classes ($m = 2$ for binary classification), P_i - probability that a random instance in partition D belong to class c_i estimated as $|c_i, D|/|D|$ (i.e.

proportion of instances of each class), where $|c_i, D|$ stands for number of instances that belong to class c_i in partition D. A $\log_2$ function encodes information in bits. If we classify the instance in partition D on some feature attribute $A\{a_1, \ldots, a_v\}$, partition D will split into v partitions set $\{D_1, D_2, \ldots, D_v\}$. $|D_j|/|D|$ represents the weight of the j^{th} partition and $Info(D_j)$ is the entropy of partition D_j. The features are then ranked per the highest information gain score.

Gain Ratio feature selection method evaluates the worth of an attribute by measuring the gain ratio with respect to the class (it enhances Information Gain as it offers a normalized score of a feature's contribution to an optimal information gain based classification decision). Gain Ratio is used as one of disparity measures and the high gain ratio for selected feature implies that the feature will be useful for classification. It is calculated as (Yang & Pedersen, 1997):

$$Gain\,ratio(A) = \frac{Information\,gain(A)}{SplitInfo(A)} \tag{1.2}$$

Where the split information value corresponds to the potential information obtained by partitioning the training data set D into v partitions, resulting to v outcomes on attribute A, and is calculated as $SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \cdot log_2 \frac{|D_j|}{|D|}$.

Chi-squared feature selection method evaluates the worth of a feature by computing the value of the chi-squared statistic with respect to the class (it measures the association between the word feature t and its associated class c_i). Chi-squared statistics represents divergence from the distribution expected based on the assumption that the feature occurrence is independent ot the class value (Yang & Pedersen, 1997), and is calculated as:

$$Chi - squared\,statistics(t, c_i) = \frac{N \cdot (AD - BE)^2}{(A+E) \cdot (B+D) \cdot (A+B) \cdot (E+D)} \tag{1.3}$$

Where we use the following notation: A - the frequency when t and c_i co-occur, B - the frequency when t occurs without c_i, D - the frequency when neither c_i nor t occurs, E - the number representing events when c_i occurs without t, D - number of documents in the corpus.

Classification Algorithms

Machine-learning methods apply different learning algorithms to determine sentiment. The most popular are supervised methods, which require manually annotated data sets. Here, the data set is usually divided into a training set, on the basis of which we build a model that a classifier can learn, and testing set, which is used for evaluation of results. We applied a supervised machine-learning approach in our study.

In order to select the most suitable machine-learning algorithm for sentiment based classification of web texts, we tested several potentially applicable algorithms, such as k-Nearest Neighbour (KNN) (Cover & Hart, 1967; Aha, Kibler & Albert, 1991), Multinomial Naïve Bayes (NBM) (Lewis, 1998; McCallum & Nigam, 1998; Rennie, Shih, Teevan & Karger, 2003; Kibriya, Frank, Pfahringer & Holmes, 2004; Bermejo, Gámez & Puerta, 2011), Support Vector Machines (SVM-poly - Sequential Minimal Optimization with polynomial kernel and SVM-lin - Support Vector Machines with linear kernel (Boser, Guyon & Vapnik, 1992; Cortes & Vapnik, 1995; Hastie & Tibshirani, 1998; Hearst, Dumais, Osuna, Platt & Scholkopf, 1998; Platt, 1999; Keerthi, Shevade, Bhattacharyya & Murthy, 2001; Meyer, Leisch & Hornik, 2003; Steinwart & Christmann, 2008; Chang & Lin, 2011)), Random Forest (RF) (Breiman, 2001; Liaw & Wiener, 2002), C4.5 (Quinlan, 1993, 2014), Decision Table (DT) (Kohavi, 1995; Wang, Yu & Yang, 2002), Simple Logistic Regression (SLR) (Landwehr, Hall & Frank, 2005; Sumner, Frank & Hall, 2005; Hosmer, Lemeshow & Sturdivant, 2013) and Voted Perceptron (VP)(Freund & Schapire, 1999; Collins & Duffy, 2002). Based on the empirical assessment (see Section 5.1) we prefer the NBM and the SVM, which we describe hereinafter.

Naïve Bayes classifiers (Lewis, 1998; McCallum & Nigam, 1998; Rennie, Shih, Teevan & Karger, 2003; Kibriya, Frank, Pfahringer & Holmes, 2004; Bermejo, Gámez & Puerta, 2011) are widely used in machine-learning due to their efficiency and ability to combine evidence from a large number of features (Mitchell, 1997). Their origins date back to 1950s, however they remain popular especially in text mining, sentiment analysis and sentiment classification domains. Unlike some other classifiers, Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of features in a learning problem, and they are not computationally expensive. They represent a group of simple probabilistic classifiers, which are derived from the Bayes' rule (see Equation 1.4)

with a naïve assumption that the features are not dependent between each other (Lewis, 1998).

$$P(c_k|x_1, ..., x_n) = P(c_k|x) = \frac{P(x|c_k)P(c_k)}{P(x)} \tag{1.4}$$

In the Equations 1.4, 1.5 and 1.6, we use the following notation: x - predictor (data), which is represented with features $x_1, ..., x_n$, c_k - class (k possible outcomes or classes), $P(c_k|x)$ - the posterior probability of class given predictor, $P(c_k)$ - the prior probability of class, $P(x|c_k)$ - the likelihood which is the probability of predictor given class, $P(x)$ - the prior probability of predictor.

Multinomial Naïve Bayes classifier is typically used for document classification, where events represent the occurrence of a word in a single document.

$$P(x|c_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i {P_{ki}}^{x_i} \tag{1.5}$$

The likelihood of observing histogram x is given in Equation 1.5. With a Multinomial Naïve Bayes classifier, feature vectors represent the frequencies with which certain events have been generated by a multinomial (P_1, ..., P_n), where P_i is the probability that event i occurs (or k such multinomials in the multiclass case). So, P_{ki} is the probability that given event i occurs for class k. A feature vector $x = (x_1, ..., x_n)$ is then a histogram, with x_i counting the number of times event i was observed in a particular instance.

In Equation 1.6 we present Multinomial Naïve Bayes classifier, which becomes a linear classifier if we express it in log-space (Rennie, Shih, Teevan & Karger, 2003), where $b = \log P(c_k)$ and $w_{ki} = \log P_{ki}$.

$$\log P(c_K|x) \propto \log\left(P(c_k)\prod_{i=1}^{n} {P_{ki}}^{x_i}\right) = \log P(c_K) + \sum_{i=1}^{n} x_i \cdot \log P_{ki} = b + w_k^\top x \tag{1.6}$$

Rennie and colleagues (2003) discuss some issues with the multinomial assumption in the context of document classification and possible ways to alleviate those problems. With the appropriate pre-processing and the use of TF–IDF weights instead of TF and document length normalization, the Naïve Bayes classifier is competitive in this domain with more advanced methods including the SVM.

SVM classifiers are probably the most widely used classifiers in machine-learning. They achieve relatively robust pattern recognition performance using well established concepts in optimization theory (Bottou & Lin, 2007). Vapnik and Chervonenkis intro-

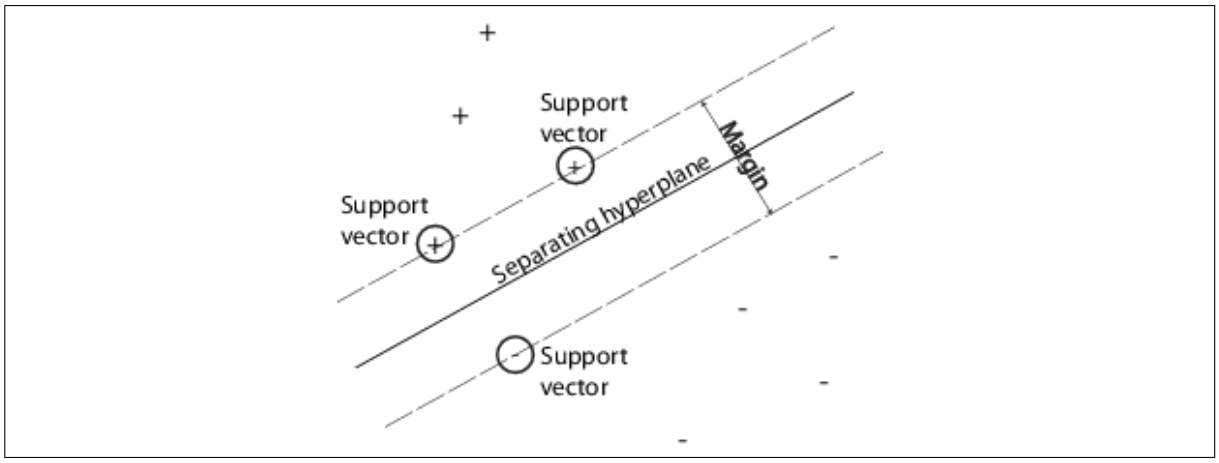

Figure 1.4: Linear SVM (MathWorks, 2017)

duced the original SVM algorithm in 1963. A linear SVM (see Figure 1.4) is presented with training data that was annotated as either positive or negative. These instances are expressed as points in the high dimensional space, which are separated by a hyperplane. There are several hyperplanes that could divide the training instances of different classes, however, the purpose of the SVM is to select the one that seperates those instances with the largest possible margin (Guyon, Weston, Barnhill & Vapnik, 2002).

Figure 1.4 presents the optimal hyperplane that separates positive and negative training instances with the maximal margin. The position of the optimal hyperplane is entirely determined by the sample s (called the *support vectors*) that are closest to the hyperplane. The larger margins imply lower classification error of potential new instances. The new instances are classified with the linear SVM decision function (see Equation 1.7). Depending on the side of hyperplane they appear, they are classified as positive, if the value $D(x)$ is positive, or negative, if the value $D(x)$ is negative (Guyon, Weston, Barnhill & Vapnik, 2002). In the Equation 1.7, x presents the feature vector with components x_i. The notation x_i denotes the i^{th} vector in a data set $\{(x_i, y_i)\}_{i=1}^{n}$, where y_i is the label associated with x_i. The objects x_i are called patterns or examples of some set X, b is the hyperplane bias, and w is the weight vector, which is obtained on the basis of support vectors (Ben-Hur & Weston, 2010).

$$D(x) = w^{\top} \cdot x + b \tag{1.7}$$

Boser, Guyon and Vapnik suggested a way to create nonlinear classifiers by applying the *kernel trick* to maximum-margin hyperplanes in 1992 (Boser, Guyon & Vapnik, 1992).

The classifiers are running on a similar principle, only that every dot product is replaced by a kernel function (not necessarily a linear one), such as polynomial, (Gaussian) radial, hyperbolic tangent, etc., allowing classifiers to fit the maximum margin hyperplane in a transformed feature space. In general, SVM classifiers are flexible and robust (not so sensitive to the number of features), however, their models are difficult or impossible to interpret.

We validate models by using cross-validation (CV) technique or by conventional validation, e.g., partitioning the data set depending on the percentage of documents included in training/testing set. Extensive experiments on various data sets and by different machine-learning algorithms have shown that it is recommended to use 10-fold CV method (Witten, Frank & Hall, 2013). In this case, each round of CV includes partitioning a sample of data set into ten complementary subsets, wherein the learning is performed on the training set (nine subsets), and validation analysis on the testing set (the remaining one). Each round of CV uses different partitions, and ten rounds reduce variability significantly. The validation of results is obtained by averaging the scores over the rounds.

Evaluation of Classification Algorithms

When classifying documents, it is necessary to evaluate results. Literature related to text and data mining most commonly defines five standard measures: accuracy, error, precision, recall, and F1-score to evaluate an algorithm's effectiveness on predicted category (Mitchell, 1997; Manning & Schütze, 1999; Sebastiani, 2002; Bishop, 2007; Qi & Davison, 2009; Liu, 2011; Miner et al., 2013).

Figure 1.5 presents the number and proportion of correctly and incorrectly classified

```
Correctly Classified Instances          60               90.9091%
Incorrectly Classified Instances         6                9.0909%

positive negative neutral    <-- classified as
   10        0        1     |  positive
    2       20        0     |  negative
    0        3       30     |  neutral
```

Figure 1.5: Confusion matrix

instances and confusion matrix. In this case we were classifying instances in three classes (*positive*, *negative* and *neutral*). The number of correctly classified instances is equal to 60 (Accuracy: 90.9091%), while the number of incorrectly classified documents is equal to 6 (Error: 9.0909%). The number of correctly classified documents is defined as the sum of the diagonal elements of the confusion matrix. The number of incorrectly classified documents is calculated as the difference between the total number of instances and the number of correctly classified instances.

Accuracy is determined as the fraction of correctly classified, and is calculated as the quotient of the number of correctly classified instances and the number of all instances (see Equation 1.8). Error is calculated as *1 - Accuracy*. Equation 1.8 presents accuracy, where C stands for confusion matrix and c_{ij} for individual items of confusion matrix C. The notation i denotes rows of confusion matrix C, j columns of confusion matrix C and n number of rows (columns) of confusion matrix C ($max\, i = max\, j = n$).

$$Accuracy = \frac{\sum_{i=j=1}^{n} c_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}} \tag{1.8}$$

Equations 1.9, 1.10 and 1.11 present the calculation of precision, recall and F1-score for class a. Precision of class a presents the proportion of all the instances classified as class a which are correctly classified as class a, while recall of class a presents the proportion of all the instances of class a that are correctly classified as class a. The F1-score, also known as F-score or F-measure, is a harmonic mean of precision and recall.

$$Precision(a) = \frac{TP(a)}{TP(a) + FP(a)} \tag{1.9}$$

$$Recall(a) = TPR(a) = \frac{TP(a)}{P(a)} \tag{1.10}$$

$$F1 - score(a) = 2 \cdot \frac{Precision(a) \cdot Recall(a)}{Precision(a) + Recall(a)} \tag{1.11}$$

For example given in Figure 1.5, the performance measures are calculated as:

- $Accuracy = \frac{60}{60+6} = 90.9091\%$,

- $Error = 1 - 90.9091\% = 9.0909\%$,
- $Precision = \frac{1}{3} \cdot (\frac{10}{12} + \frac{20}{23} + \frac{30}{31}) = 89.0213\%$,
- $Recall = \frac{1}{3} \cdot (\frac{10}{11} + \frac{20}{22} + \frac{30}{33}) = 90.9091\%$,
- $F1\text{-}score = 2 \cdot \frac{89.0213\% \cdot 90.9091\%}{89.0213\% + 90.9091\%} = 89.9553\%$.

1.5 Organisation of the Dissertation

In this dissertation, we first present some historical backgrounds of the corpus linguistics, and discuss the web as a corpus for linguistic research. Next, we describe the presence of languages on the web, and present the status of the Slovenian language on the web. Chapter 3 gives an overview of corpora construction, followed by the annotation process, exploration of the corpora. The lexicon construction and its exploration are described in Chapter 4, which ends with information on obtaining the data and tools used in the corpora construction. It also presents the format of language resources and the terms of use. In Chapters 5-7, we provide applications of the developed language resources. In detail, we evaluate the performance of sentiment based classification techniques, estimate the proportions of positive, negative and neutral news within the observed web media, and monitor the dynamics of sentiment. Chapter 8 deals with hypothesis testing. Finally, Chapter 9 concludes this dissertation.

2 RELATED WORK

Empirical and statistical analyses of language resources were successfully applied in the 1950s (Church & Mercer, 1993), and the impact has been remarkable in many fields, ranging from psychology to electronics.

One of the major milestones in modern corpus linguistics was Computational Analysis of Present-Day American English written by Kučera and Francis (1967). They analysed the Brown Corpus of current American English that contains 500 documents with approximately 1 million words (Kučera & Francis, 1967). This corpus motivated other researchers to build new similarly structured corpora, such as the Lancaster-Oslo/Bergen (LOB) Corpus (1960s British English), Kolhapur (late 1970s Indian English), Wellington (late 1980s and early 1990s New Zealand English), Australian Corpus of English (1986s Australian English), the Freiburg-Brown Corpus (early 1990s American English), and the Freiburg-LOB Corpus (1990s British English). The first computerized speech corpus, which contained one million words, was built in 1971 (Sankoff & Sankoff, 1973). It inspired Shana Poplack (1989) to assemble, transcript and concordance a mega corpus of spoken French in the Canadian capital region.

In the 1980s, breakthroughs in technology made possible not only the establishment and storage of corpora larger than ever before, but also the evolution of new models that could exploit statistical methods. Eventually, these corpora were introduced to computational linguistics in 1989 at the Computational Linguistics and Research in Humanities Panel at the 27th Annual Meeting of the Association for Computational Linguistics (ACL). Within their session, they reviewed current international guidelines for the encoding, interchange of machine-readable textual documents for research, and explored the mutual relevance of corpus-based language analysis and language-based corpus analysis. According to Kilgarriff and Grefenstette (2003), in those days corpora were large, messy, ugly objects, clearly lacking in theoretical integrity in all sorts of ways. Four years after the corpora were introduced to computational linguistics, Church and Mercer (1993) pub-

lished a special issue of the journal Computational Linguistics on the use of large corpora. In the mid-1990s, Marcus and his team (1993) built The Penn Treebank, a large, annotated corpus of English, which led to a breakthrough in the accuracy of natural language parsers for unrestricted text.

Corpus linguistics has led to the development of a number of research methods. For example, Wallis and Nelson (2001) proposed the so-called 3A perspective: annotation, abstraction and analysis. The annotation denotes the utilization of a scheme to texts, and may include part-of-speech (POS) tagging, parsing, structural (phrase or dependency) mark-up, syntactic bracketing, etc. Concordances, for example, position a word within its context, and thereby make it much easier to study how it is used in a language, both syntactically and semantically. Abstraction is understood to mean the mapping of terms to generate a theoretical model in order to perform a linguist-directed search, while analysis refers to statistical probing, manipulating and generalising from the dataset, in which they can incorporate statistical evaluations and optimisation of rule bases or knowledge-discovery methods.

2.1 The Web as Corpus

The use of the web as a corpus for linguistic research was proposed several times (Hofland, 2000; Rundell, 2000; Banko & Brill, 2001; Ghani, Jones & Mladenić; 2001; Fletcher, 2001; De Schryver, 2002; Volk, 2002; Renouf, 2003; Resnik & Smith, 2003; Robb, 2003). The advances in this field motivated Kilgarriff and Grefenstette (2003) to publish their work on the web as corpus. In addition, Kilgarriff was one of the founding members of the Special Interest Group on Web as Corpus (SIGWAC[6]) of the ACL, and a member of The Web-As-Corpus Kool Yinitiative (WaCky[7]). Since then there have been several studies on using the web for linguistic research, mostly based on queries within search engines or crawling –for more details, see the following papers and the references therein (Baroni & Bernardini, 2004; Rayson, Walkerdine, Fletcher & Kilgarriff, 2006; Hundt, Nesselhauf & Biewer, 2007; Ekbal & Bandyopadhyay, 2008; Taulé, Martí & Recasens, 2008; Baroni,

[6] SIGWAC is the leading professional organisation for web as corpus researchers: `https://www.sigwac.org.uk/`

[7] WaCky initiative aims to build a community, services and tools for retrieving, storing and annotating linguistic data from the web: `http://wacky.sslmit.unibo.it/`

Bernardini, Ferraresi & Zanchetta, 2009; Wu, Witten & Franken, 2010; Fletcher, 2012; Lyding et al., 2014; Schäfer, Barbaresi & Bildhauer, 2014; Biber, Egbert & Davies, 2015; Pecina et al., 2015; Asheghi, Sharoff & Markert, 2016).

A lot of effort is put into the process of acquiring and maintaining corpora. Entire departments and institutions are focused on building large and high-quality corpora (Spousta, 2006). The Google N-Gram corpus is the largest English corpus, with 155 billion words (Davies, 2016). Davies also constructed the largest structured corpus of historical English, i.e., the Corpus of Historical American English (COHA), which contains more than 400 million words (1810s-2000s), and the largest corpus of contemporary English, i.e., the Corpus of Contemporary American English (COCA), which contains 450 million words (1990s-present).

The European corpus initiative[8] is an example of a multilingual corpus, which contains over 98 million different words in English, French, German, Spanish, Dutch, Swedish, Dutch, Turkish, Bulgarian, Latin and Greek.

2.2 Corpora in Slovene

The two largest corpora of the Slovenian language comprise 1.2 billion words (tokens), i.e. the slWac[9] (Ljubešić & Erjavec, 2011; Erjavec, Ljubešić & Logar, 2015) and the Gigafida[10] (Berginc et al., 2012; Berginc & Ljubešić, 2013). The current version of slWac (v2.0) is a web-crawled corpus gathered mostly from the *.si* domain. It includes tokenization, POS tags as morphosyntactic descriptions (MSDs) and lemmatization with the ToTaLe[11] tool. The Gigafida was obtained from selected texts written in the Slovenian language of different genres and styles, mainly from newspapers, magazines and the web. It includes texts that were published between 1990 and 2011. In fact, the Gigafida consists of the FidaPLUS[12] corpus (Arhar, Gorjanc & Krek, 2007), the previous version of the Slovene reference corpus, along with the new material, which was obtained afterwards.

The 100 million word corpus KRES[13] (Berginc et al., 2012) is another Slovenian

[8] http://www.elsnet.org/eci.html
[9] http://nlp.ffzg.hr/resources/corpora/slwac/
[10] http://eng.slovenscina.eu/korpusi/gigafida/
[11] http://nl.ijs.si/analyse/
[12] http://www.fidaplus.net
[13] http://eng.slovenscina.eu/korpusi/kres/

corpus. The KRES was gathered from the Gigafida corpus and is a balanced corpus, especially by text types or genres.

Additionally, the one million word corpus JOS[14] (Erjavec, Fišer, Krek & Ledinek, 2010), a derived corpus from the FidaPLUS corpus, consists of lemmas and morphosyntactic descriptions.

The one million word corpus GOS[15] (Verdonik, Kosem, Vitez, Krek & Stabej, 2013; Žgank, Vitez & Verdonik, 2014) represents a spoken corpus in the Slovenian language. It includes the transcripts of approximately 120 hours of speech of different genres and styles: radio and TV shows, school lessons and lectures, private conversations between friends or within the family, work meetings, consultations, conversations in buying and selling situations, etc. However, there are two versions of spelling within this corpus (standardized and pronunciation-based).

Finally, there is the Janes corpus (Fišer, Smailović, Erjavec, Mozetič & Grčar, 2016). It contains 167 million words of Slovene user-generated content from five different platforms (tweets, forums, blogs and comments on news articles and on Wikipedia).

2.3 Lexicons for Sentiment Analysis

Lexicons have been widely used for sentiment and subjectivity analysis, as they represent a simple, yet effective way to build rule-based opinion classifiers (Perez-Rosas, Banea & Mihalcea, 2012). The lexicon-based approaches are usually domain-dependent since the subjectivity of most polarity words is very ambiguous.

One of the first-known, human-annotated lexicons for effect and opinion mining is the General Inquirer (GI) lexicon (Stone, Dunphy & Smith, 1966), which is still managed by the Harvard University. The GI lexicon contains 11,788 English words in alphabetical order (1,915 labelled as positive and 2,291 as negative, the rest as objective).

WordNet is lexical database for the English language (Miller, 1995; Fellbaum, 1998), which is maintained by the Cognitive Science Laboratory at Princeton University. The lexicon groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms (synsets). Each of 117,000 synsets is interlinked by means of conceptual-semantic and lexical rela-

[14] http://nl.ijs.si/jos/index-en.html
[15] http://eng.slovenscina.eu/korpusi/gos/

tions. WordNet was expanded to SentiWordNet (Esuli & Sebastiani, 2006; Baccianella, Esuli & Sebastiani, 2010), which supports opinion mining applications obtained by tagging all the WordNet 3.0 synsets by adding polarity and objectivity labels for each term. Several lexicons are based on WordNet's synsets, such as SentiWordNet[16], WordNet-Affect[17], Micro-WNOp[18], etc. SentiWordNet supports opinion mining applications obtained by tagging all the WordNet 3.0 synsets by adding polarity and objectivity labels for each term. WordNet-Affect and Micro-WNOp are based on WordNet synsets. The first is an extension of WordNet-Domains, which includes 1,903 synsets marked with the emotions within the affective labels. Its main part contains 539 nouns, 517 adjectives and 238 verbs. The second consists of 1,105 WordNet synsets manually annotated by a group of five human annotators. Similar to GI lexicon, each synset is assigned a score within the three categories (positive, negative and objective).

Liu et al. (2005) built a sentiment lexicon based on online customer reviews of products. The sentiment value of the text in English is estimated by adding the sentiment value in each word.

Wiebe and Riloff (2005) distributed OpinionFinder, a large lexicon of clues tagged with prior polarity (positive, negative and neutral), which has grown into the Multi-perspective Question Answering[19] (MPQA) subjectivity lexicon. The focus of their work was on disambiguating the contextual polarity of English words with positive or negative prior polarity, and the performance evaluation of features using several different machine-learning algorithms.

AFINN[20] (Nielsen, 2011) is another commonly used sentiment lexicon containing a list of English words rated for valence with an integer between -5 (negative) and +5 (positive). A new version ANEW[21] is based on multiple independent labels per item and provides, besides valence, also the arousal and dominance for each word.

Slovene WordNet (sloWNet) offers a lexical database in Slovene that organizes nouns, verbs, adjectives and adverbs in conceptual hierarchies, thereby linking semantically and lexically related concepts (Erjavec & Fišer, 2006). Two lexicons for Slovene were derived

[16] http://sentiwordnet.isti.cnr.it/
[17] http://wndomains.fbk.eu/wnaffect.html
[18] http://www-3.unipv.it/wnop/
[19] http://mpqa.cs.pitt.edu/
[20] http://neuro.imm.dtu.dk/wiki/AFINN
[21] http://neuro.imm.dtu.dk/wiki/A_new_ANEW:_evaluation_of_a_word_list_for_sentiment_analysis_in_microblogs

from translation of words from English. Martinc (2013) built the AFINN list of words, which is mainly intended for microblogging. Kadunc and Robnik-Šikonja (2016) developed an opinion lexicon with 90,620 positive and negative terms. The research activities on computational linguistics for the Slovenian language correspond to the number of people who speak Slovene. Bearing this in mind, we wish to support the diversity and richness of freely available language resources as well as the development of the Slovenian language in the future.

2.4 Sentiment Analysis

In the first studies that were dealing with sentiment analysis, the scientists were mainly focused on financial news (Hatzivassiloglou & McKeown, 1997), movie (Pang, Lee & Vaithyanathan, 2002) and product reviews (Turney, 2002).

Turney (2002) considered sentiment extraction from the whole document as basic information, which is commonly known as the document-level sentiment classification. He treated sentiment classification as detection of positive or negative sentiment. Others like to expand the basic task to classify a document's polarity or strength of opinion on a multi-way scale, for example: reviews measured in scales range from -5 to 5, where -5 refers to very negative, 0 to neutral, and 5 to very positive; movie reviews rated on either a 3- or a 4-star scale (Pang & Lee, 2008), or restaurant reviews that rated various aspects of the given restaurant on a 5-star scale (Snyder & Barzilay, 2007). Every concept is scored by its relation to associated sentiment words.

Durant and Smith (2006) investigated existing technologies and their utility for sentiment classification using dataset of political web logs over a two-year period. Godbole and co-authors (2007) worked on sources like newspapers and blog posts at the level of words, while others worked on documents, such as twitter posts (Smailović, Grčar, Lavrač & Žnidaršič, 2013; Ceron, Curini & Iacus, 2015).

Furthermore, Colbaugh and Glass (2010) estimated sentiment orientation of social media content. They illustrated its potential for security informatics through a case study involving the estimation of Indonesian public sentiment regarding the July 2009 Jakarta hotel bombings. Thelwall, Buckley and Paltoglou (2012) were investigating approaches to detect sentiment strength from six diverse social web datasets (MySpace,

Twitter, YouTube, Digg, Runners World and BBC Forums). Tourani and co-authors (2014) monitored sentiment, and evaluated the usage of automatic sentiment analysis to identify distress or happiness in a development team based on user and developer mailing lists.

Recently, Nakov et al. (2016) presented the development and evaluation of tasks on sentiment analysis in Twitter and other social media texts at the International workshop on semantic evaluation. Between 2013 and 2015, they created a large contextual and message-level polarity corpus consisting of tweets, SMS messages, LiveJournal messages, and a special test set of sarcastic tweets.

In general, we can divide studies related to the sentiment analysis into four main domains: business/financial (e.g. Hatzivassiloglou & McKeown, 1997; Das & Chen, 2001), film/movie-review (e.g. Pang, Lee & Vaithyanathan, 2002; Taboada, Brooke, Tofiloski, Voll & Stede, 2011), product-review (e.g. Turney, 2002; Kushal, Lawrence & Pennock, 2003; Glavaš, Korenčić & Šnajder, 2013), and political (e.g. Durant & Smith, 2006; Ceron, Curini & Iacus, 2015; Schatten, Seva & Đurić, 2015). The main source of such studies are the social media (e.g. Colbaugh & Glass, 2010; Nakov, 2016). In spite of the important role news play in our lives, the news genre has received much less attention within the sentiment analysis community. However, some studies explored sentiment analysis to investigate news articles (Balahur et al., 2013). Related to our effort, Reis et al. (2015) investigated the sentiment of the business news produced by four major global media corporations and the dynamics of sentiment in news over time.

Although subjectivity in news articles has traditionally tended to be implicit, the news stories still have their own biases. The growing trend to foster interactivity and more heavily report communication of internet users within the body of news articles is likely to make expression of subjectivity in news articles even more explicit (Abdul-Mageed & Diab, 2011).

Data scientists use several tools and methods to determine and classify the sentiment of digital text automatically. Sentiment classification has been studied by numerous researchers subsequently and might be the most widely studied problem in the field of sentiment analysis (see a survey in Paliouras, Papatheodorou, Karkaletsis & Spyropoulos, 2002). Most techniques apply supervised learning where a bag of individual words (unigrams) is the most commonly used documents representation. An early reference to

bag-of-words (BOW) in a linguistic context can be found in Zellig Harris's (1954) article on distributional structure. Use of various features and transformations was thoroughly assessed by researchers, such as TF and TF-IDF weighting schemes, POS tags, opinion words and phrases, negations, syntactic dependency (Liu, 2011).

It is difficult to compare performance of different studies between each other due to variety of resources and collections of documents that were used for training and testing. Godbole's team, who worked with newspaper news and blog articles texts at the level of words, achieved accuracy of 89%, when classifying texts into positive and negative, while others, who work with texts at the document level, in general achieve the accuracy around 90% for the two-class sentiment classification and around 65% for the three-class sentiment classification. Some studies have been done in domain of long documents such as product reviews, and in some tricky domains like political commentaries. Although a relatively high accuracy in document polarity labelling has been achieved, it is still a challenge to extract sentiment orientation, combined with emotion's intensity, its holder, and target.

A brief overview of related studies in the field of sentiment analysis and their performances is presented in Table 2.1.

Table 2.1: Sentiment Analysis: Studies and Performances

Study	Data source / Domain	# Categories	Performance
Hatzivassiloglou & McKeown (1997)	Wall Street Journal corpus / Financial	2 (pos & neg)	Acc: 92%
Das & Chen (2001)	Stock exchange messages / Financial	3 (pos, neg & amb)	Acc: 62%
Pang et al. (2002)	IMBb / Film	2 (pos & neg)	Acc: 83%
Turney (2002)	Epinions / Car, bank, movie, travel	2 (recomm & not recomm)	Acc: 66-84%
Kushal et al. (2003)	C|net, Amazon / Product-review	2 (pos & neg)	Acc: 89%
Kim & Hovy (2004)	DUC 2001 corpus	2 (pos & neg)	Acc: 81%
Devitt & Khurshid (2007)	Ryanair and Aer Lingus news / Financial	2 (pos & neg)	F: 47%
Godbole et al. (2007)	Newspaper news, blogs	2 (pos & neg)	Acc: 89%
Strapparava & Mihalcea (2007)	Newspaper news (NY Times, CNN, BBC, Google)	2 (pos & neg)	Acc: >90%, F: <31%
Ferguson et al. (2009)	Blogs / Financial	2, 3 (pos, neg & neu)	(2)Acc: 79%, (3)Acc: 63%
Melville et al. (2009)	Blogs, IMDb / Financial, political, film	2 (pos & neg)	Acc: 64-91%
Wilson et al. (2009)	MPQA corpus	2 (neu & other)	Acc: 83%
Colbaugh & Glass (2010)	Blogs, forums, IMBb / Film	2 (pos & neg)	Acc: 79%
Agarwal et al. (2011)	Twitter	2, 3 (pos, neg & neu)	(2)Acc, F: 75%, (3)Acc, F: 60%
Kouloumpis et al. (2011)	Twitter (HASH, EMOT) / Product-review	3 (pos, neg & neu)	Acc: 75%, F: 68%
Taboada et al. (2011)	Epinions, IMDb / Product, film	2 (recomm not recomm)	Acc: 76-82%
Mohammad et al. (2013)	Twitter	2 (pos & neg)	F: 69-89%
Smailović et al. (2013)	Twitter, SMS / Financial	3 (pos, neg & neu)	Acc: >85%
Balahur et al. (2013)	Newspaper quotes (EMM)	2 (pos & neg)	Acc: 82%
Hu et al. (2013)	Twitter / Political	2 (pos & neg)	Acc: 80%
Montejo-Ráez et al. (2014)	Twitter	2 (pos & neg)	Acc: 62.85% Acc: 62.85%
Habernal et al. (2015)	Facebook, IMDb / Product, film	3 (pos, neg & neu)	F: 75-79%
Vo & Zhang (2015)	Twitter	3 (pos, neg & neu)	Acc: 71%, F: 70%
Nakov (2016)	Twitter, SMS, LiveJournal	3 (pos, neg & neu)	F: 90%

Categories: pos - positive, neg - negative, neu - neutral, amb - ambiguous, recomm - recommended, not recomm - not recommended
Performances: Acc - accuracy; F - F1-score

2.5 Sentiment Annotation

The success of most studies depends on the quality of their knowledge bases; either lexicons containing the sentiment polarity of words, or quality annotation data for statistical training. Although most of manual intervention to assemble lexicons has been performed by bootstrapping (Wiebe & Riloff, 2005), it is difficult to bypass the manual annotation process which is often expensive and time-consuming (Hsueh, Melville & Sindhwani, 2009). Sentiment annotation is complex as it spans lexical, syntactic and semantic levels. Joshi et al. (2014) proposed a new metric called Sentiment Annotation Complexity (SAC), which uses four categories of linguistic features: lexical, syntactic, semantic and sentiment-related in order to capture the sub-processes of annotation.

The process of building the annotated corpus is often cyclical, with changes and adjustments to the annotation level and tasks since the data is further examined. Pustejovsky and Stubbs (2012) refer to the annotation process as the MATTER cycle that includes model, annotation, training, testing, evaluation, and revision. There are several systems, crowdsourcing platforms and tools[22] for retrieval, annotation and analysis of this sort of data.

Crowdsourcing is a popular approach of obtaining manual annotations (Von Ahn & Dabbish, 2004; Snow, O'Connor, Jurafsky & Ng, 2008). Hsueh, Melville and Sindhwani (2009) analysed the data from both expert and non-expert annotators recruited from the web services by exploring three selection criteria, i.e. noise level, sentiment ambiguity, and lexical uncertainty, in order to identify untrustworthy annotators, and select suitable items for the predictive modelling. As a result, they confirmed the utility of these criteria on improving quality of the data. However, most of the crowdsourcing problems can be avoided in annotation processes with small and purposely trained groups (Mozetič, Grčar & Smailović, 2016), like in our study (see Section 3.2).

[22] Apache OpenNLP, GATE, Lydia, MAE, MALLET, MPQA (OpinionFinder 2), Orange, Phyton (NLTK), QDA Miner Lite, R (tm), RapidMiner, TAMS Analyzer, WEKA, etc.

3 CONSTRUCTING ANNOTATED NEWS CORPORA IN SLOVENE

Within this chapter, we describe the procedure of the corpora construction, a collection of large (>3 million words within 10,427 documents) manually annotated news corpora, the annotation process, and provide relevant summary statistics about the corpora.

3.1 Corpora Construction

Manually sentiment annotated Slovenian news corpus SentiNews 1.0[23], which is introduced in this dissertation, is constructed on the basis of all Slovenian news texts with political, business, economic and financial content published between 1st of September 2007 and 31st of December 2013 retrieved from five widely read Slovenian web media resources (24ur[24], Dnevnik[25], Finance[26], Rtvslo[27], Žurnal24[28]). These five web media resources were chosen as they are very popular and at the same time have a well-organized digital news archive, which facilitates the acquisition of web texts for the selected period. The selection did not consider 6 web media (Delo, Radio1, Reporter, Siol, Slovenske novice and Večer), which were among the 200 most visited websites in Slovenia on 1st of January 2014, according to the Alexa[29] website. Thus, approximately 185 thousand of relevant texts were not dealt with.

Text for the corpora was obtained by crawling each of the selected web media resources[30]. Every piece of news was put in a separate textual file, containing the official URL of the web medium, the URL of the news, the date of publishing the news, the

[23] http://hdl.handle.net/11356/1110
[24] http://www.24ur.com/arhiv/novice/gospodarstvo/
[25] https://www.dnevnik.si/posel/novice/
[26] http://www.finance.si/danes/
[27] http://www.rtvslo.si/gospodarstvo/arhiv/
[28] http://www.zurnal24.si/archive/slovenija/
[29] http://www.alexa.com/topsites/countries/SI/
[30] http://hdl.handle.net/11356/1105

URL main:
www.24ur.com
URL:
http://www.24ur.com/novice/gospodarstvo/cenejse-gorivo.html
Date:
18.12.2007
Author:
M.K./Š.Z.
Keywords:
bencin, dizel, pocenitev, naftni, derivati
Title:
Cenejše gorivo
Summary:
Liter dizelskega goriva po novem stane 1,045 evra. Liter najbolj prodajanega 95-oktanskega bencina pa 1,033 evra.
Content:
Cene naftnih derivatov v Sloveniji so se spremenile. Cena za liter najbolj prodajanega 95-oktanskega bencina se je znižala za 0,015 evra na 1,033 evra.
Liter 98-oktanskega bencina se je pocenil za 0,010 evra na 1,054 evra. Dizelsko gorivo je cenejše za 0,041 evra in je zanj tako po novem potrebno odšteti 1,045 evra na liter. Od torka je cenejše tudi kurilno olje, in sicer za 0,030 evra, tako da je treba za liter odšteti 0,704 evra.

Figure 3.1: Sample format of raw files built by web crawlers written in R

author, the keywords, the title, the summary and the content of the news. The files are organized in such a way that the hashtag character indicates a new attribute, with each attribute stored in a new line, as presented in Figure 3.1. Similarly, within the attribute *Content*, each paragraph is stored separately on a new line. The dates are stored in the *dd.mm.yyyy* format.

First, we obtained 217,532 documents published between 1st of September 2007 and 31st of December 2013 (Finance - 110,841, Dnevnik - 47,684, Žurnal24 - 39,886, Rtvslo - 10,450 and 24ur - 8,671), which were the basis for a sample that we manually annotated (see Section 3.2). We subsequently obtained news that was published between 1st of January 2014 and 31st of January 2016 to estimate the proportion of positive, negative and neutral news (for details see Chapter 6).

Eventually, we stored the data in a MySQL database, and developed a web application[31] for retrieval, storage, annotation and sentiment allocation for the Slovene web texts. Initially, the web application was made for the company, which was involved in the early stages of our project. The company was interested in monitoring, collecting data

[31] `http://dejan.amadej.si/test/`

and information on the web, in aggregation and integration of retrieved data with MySQL databases, as well as in data analytic tools, which could improve company's future editor policy and its strategic business decisions. Since the systems and tools described in Section 2.5 did not fully satisfy the needs of the company, we developed a web application on our own. During the development of the web application, we encountered some challenges such as implementing our solutions in Google Custom Search Engine and retrieving the textual contents from HTML code, which we managed to overcome successfully.

3.2 Annotation Process

The data-retrieval process was followed by cleaning and pre-processing the data. We removed grammatical and spelling errors by using a spell-checker within text editor (Notepad++). Once we had set up an online environment for the annotation process, six native-speaker annotators were trained in two phases.

In the first phase, they obtained the basic guidelines for annotation and learned how to use the web application. Together with a referee, they annotated 10 news on three levels, i.e. document, paragraph and sentence level, and discussed about individual instances. The process of sentiment annotation consists of two sub-processes: comprehension, where the annotator understands the content, and sentiment judgment, where the annotator identifies the sentiment. Using the five-level Likert (1932) scale (1 - very negative, 2 - negative, 3 - neutral, 4 - positive and 5 - very positive) the annotators were told to specify evoked sentiment using the following instructions: *"Please specify the sentiment from the perspective of an average Slovenian web user. How did you feel after reading this news?"*

In the second phase, along with a referee, each of them annotated 50 news items individually. We analysed the agreement among the annotators, which indicated some issues with compound-complex sentences and the influence of their personal values, beliefs and attitudes. We discussed the instances with lower agreement and resolved the issues, which resulted in additional annotation guidelines. In the case of compound-complex sentences with more than one sentiment expressed, such as journalist's quotes and comments of politician's statements, we agreed on assigning the one, which prevails, or neutral sentiment in all other cases. Also, we agreed that the context should be taken into account,

but always in accordance with the instructions that were given.

Finally, the annotators manually annotated a stratified random sample of 10,427 documents independently, i.e. approximately 2,000 documents per web medium on the three levels of granularity (Žurnal24 - 2,212, Rtvslo - 2,163, 24ur - 2,103, Dnevnik - 2,048 and Finance 2,000). Again, they used the five-level Likert scale to annotate documents on the three levels of granularity, and followed the instructions they were given in the first and second phase. However, each annotator did not manually annotate all the items in the sample . Almost 9% of the news in the sample was annotated by all the annotators, and slightly more than 70% by at least two of them. The sentiment of an instance is defined as the average of the sentiment scores given by the different annotators. An instance was labelled as:

- Negative, if the average of given scores was less than or equal to 2.4,
- Neutral, if the average of given scores was between 2.4 and 3.6,
- Positive, if the average of given scores was greater than or equal to 3.6.

The annotators were paid to provide this annotation service. It took us nearly one year to manually annotate the sample . To evaluate the process of annotation, we explored correlation coefficients using various measures of inter-annotator agreement at three levels of granularity, as shown in Table 3.1 and Table 3.2.

Table 3.1: Values of Cronbach's alpha (α_C), Krippendorff's alpha (α_K), Fleiss' kappa (κ) and Kendall's coefficient of concordance (W) between the annotators, as well as the minima (min), maxima (max) and averages (avg) for the Pearson (r_P) and Spearman (r_S) correlation coefficients at the document, paragraph and sentence level of granularity

	Document level			Paragraph level			Sentence level		
α_C	0.903			0.862			0.856		
α_K	0.691			0.530			0.514		
κ	0.491			0.468			0.454		
W	0.679			0.593			0.586		
	min	max	avg	min	max	avg	min	max	avg
r_P	0.538	0.740	0.628	0.368	0.610	0.514	0.369	0.607	0.501
r_S	0.533	0.744	0.623	0.374	0.609	0.511	0.374	0.612	0.501

We also calculated correlation coefficients between average scores of documents based on average scores of:

- Documents and average scores of paragraphs (r_P: 0.900, r_S: 0.850),

- Documents and average scores of sentences (r_P: 0.896, r_S: 0.839),
- Paragraphs and average scores of sentences (r_P: 0.981, r_S: 0.979).

Table 3.2: Values of Pearson (r_P) and Spearman (r_S) correlation coefficients between 6 annotators (Ann#1 - Ann#6) at the three levels of granularity

	Document level					
	Ann#1	Ann#2	Ann#3	Ann#4	Ann#5	Ann#6
Ann#1	1 (1)	0.72 (0.71)	0.58 (0.59)	0.65 (0.63)	0.58 (0.58)	0.70 (0.69)
Ann#2	0.72 (0.71)	1 (1)	0.57 (0.57)	0.62 (0.60)	0.60 (0.60)	0.69 (0.68)
Ann#3	0.58 (0.59)	0.57 (0.57)	1 (1)	0.55 (0.54)	0.54 (0.53)	0.62 (0.62)
Ann#4	0.65 (0.63)	0.62 (0.60)	0.55 (0.54)	1 (1)	0.61 (0.60)	0.74 (0.74)
Ann#5	0.58 (0.58)	0.60 (0.60)	0.54 (0.53)	0.61 (0.60)	1 (1)	0.66 (0.66)
Ann#6	0.70 (0.69)	0.69 (0.68)	0.62 (0.62)	0.74 (0.74)	0.66 (0.66)	1 (1)
	Paragraph level					
	Ann#1	Ann#2	Ann#3	Ann#4	Ann#5	Ann#6
Ann#1	1 (1)	0.54 (0.52)	0.43 (0.44)	0.59 (0.58)	0.58 (0.56)	0.61 (0.61)
Ann#2	0.54 (0.52)	1 (1)	0.37 (0.37)	0.51 (0.51)	0.51 (0.51)	0.56 (0.56)
Ann#3	0.43 (0.44)	0.37 (0.37)	1 (1)	0.41 (0.40)	0.41 (0.43)	0.45 (0.45)
Ann#4	0.59 (0.57)	0.51 (0.51)	0.41 (0.40)	1 (1)	0.56 (0.54)	0.61 (0.61)
Ann#5	0.58 (0.56)	0.51 (0.51)	0.41 (0.43)	0.56 (0.54)	1 (1)	0.59 (0.61)
Ann6	0.61 (0.61)	0.56 (0.56)	0.45 (0.45)	0.61 (0.61)	0.58 (0.59)	1 (1)
	Sentence level					
	Ann#1	Ann#2	Ann#3	Ann#4	Ann#5	Ann#6
Ann#1	1 (1)	0.51 (0.50)	0.43 (0.43)	0.56 (0.56)	0.56 (0.55)	0.61 (0.61)
Ann#2	0.51 (0.50)	1 (1)	0.37 (0.37)	0.48 (0.47)	0.48 (0.47)	0.54 (0.54)
Ann#3	0.43 (0.43)	0.37 (0.37)	1 (1)	0.40 (0.40)	0.41 (0.42)	0.46 (0.46)
Ann#4	0.56 (0.56)	0.48 (0.47)	0.40 (0.40)	1 (1)	0.54 (0.53)	0.59 (0.59)
Ann#5	0.56 (0.55)	0.48 (0.47)	0.41 (0.42)	0.54 (0.53)	1 (1)	0.58 (0.59)
Ann#6	0.61 (0.61)	0.54 (0.54)	0.46 (0.46)	0.59 (0.59)	0.58 (0.59)	1 (1)

The first four internal consistency estimates of reliability for the scores, shown in Table 3.1, normally range between 0 and 1. The values closer to 1 indicate more agreement, when compared to the values closer to 0. The Cronbach's alpha values indicate a very good internal consistency at all levels of granularity. Normally, we refer a value greater than 0.8 to a good internal consistency, and above 0.9 to an excellent one (George & Mallery, 2001). The value of Krippendorff's alpha (Krippendorff, 2004) at the document level of granularity implies a fair reliability test, whereas its values at the paragraph level and sentence level are lower. The Fleiss' kappa values illustrate a moderate agreement among the annotators at all levels of granularity. In general, a value between 0.41 and 0.60

implies a moderate agreement, above 0.61 to a substantial agreement, and above 0.81 to an almost perfect agreement (Landis & Koch, 1977). The Kendall's values indicate a fair level of agreement between the annotators at all levels of granularity. Correspondingly, the Pearson and Spearman values range from -1 to 1, where 1 refers to the total positive correlation, 0 to no correlation, and -1 to the total negative correlation. The coefficients show moderate positive agreement among the annotators, but their values are decreasing when applied to the paragraph and the sentence level. Usually, the values above 0.3 refer to weak correlations, above 0.5 to a moderate, and above 0.7 to a strong correlation (Rumsey & Unger, 2015).

In addition, we observed the correlation between the annotators, and found that one of the annotators slightly differs from the rest, which results in overall lower correlations. Despite the clear instructions, the contents of the texts can sometimes be ambiguous, which makes the annotation more difficult. Our results support the claim by O'Hare et al. (2009), that it can be more difficult to accurately annotate sentences (or even phrases). In general, the sentiment scores by different annotators are more consistent at the document level than at the paragraph and sentence level. As an illustration, there were no major issues within the annotation process at the document level. The biggest inconsistency between the annotators was when they scored the same document with scores 3, 5, 2, 3 and 2. Within the paragraph and sentence level, for example, it occurred that two annotators scored the same paragraph (or sentence) with 1 and 5. Nevertheless, this did not occur in many cases (paragraph level: 27 out of 89,999, sentence level: 52 out of 168,899). For that reason, the anomalies, discussed in this paragraph, were not excluded from our annotated data sets as they preserve the information as it was obtained.

3.3 Exploring the Corpora

The results of our data retrieval and subsequent annotation efforts are three manually annotated news corpora (SentiNews 1.0) for three levels of granularity with 10,427 annotated documents. The news corpora include different components: document, paragraph or sentence identifier (*nid*, *pid*, *sid*), official URL of the web medium and URL of the news (*main_url*, *url*), *title*, *keywords*, body (*content*) of the news, *date*, reporter's or agency's name (*author*), manual annotation scores of 6 annotators (*Ann1* - *Ann6*), average and

Table 3.3: Attributes, descriptions and data types within the annotated news corpora

Attribute	Description	Data type
nid	News ID	Integer (1 to 12,540)
pid	Paragraph ID	Integer (1 to 94)
sid	Sentence ID	Integer (1 to 150)
main_url	Official URL of the web medium	String (5 web media)
url	URL of the news	String
title	Title of the news	String
keywords	Keywords of the news	String
content	Content of the news	String
date	Date of publishing the news	String (yyyy-mm-dd)
author	Author of the news	String
Ann1 - Ann6	Manual annotations from 6 annotators	Integer (1 to 5)
avg_sentiment	Average of scores (Ann1 - Ann6)	Float (1 to 5)
sd_sentiment	Standard deviation of scores (Ann1 - Ann6)	Float (0 to 2.828)
sentiment	Sentiment allocation according to *avg_sentiment* score (*positive* for scores $\geq$ 3.6, *negative* for scores $\leq$ 2.4 and *neutral* for scores between 2.4 and 3.6)	String (positive, negative, neutral)

standard deviation of the annotation scores (*avg_sentiment*, *sd_sentiment*), as well as the *sentiment* allocation (positive, negative or neutral). These components of the news corpora are denoted as attributes and are specified in Table 3.3.

The first row of each corpus includes the names of the attributes. Each instance that corresponds to the corpus is located on a new line. Also, the identifiers of instances are compatible within these corpora, e.g., the document ID (*nid*) in the sentence-based corpus corresponds to the document ID (*nid*) in the document/paragraph-based corpus.

We labelled instances within different levels of granularity on the basis of averaging the sentiment annotations:

- Document level (10,427 instances): 1,667 (15.97%) as positive, 3,337 (32.00%) as negative and 5,425 (52.03%) as neutral,
- Paragraph level (89,999 instances): 14,636 (16.26%) as positive, 23,721 (26.36%) as negative and 51,642 (57.38%),
- Sentence level (168,899 instances): 27,491 (16.28%) as positive, 45,170 (26.74%) as negative and 96,238 (56.98%) as neutral.

More than a half of the annotated news is labelled as neutral. Also, the proportion of

Table 3.4: Corpora statistical information

	Category			
Unit name	pos	neg	neu	Total
Documents	1,665	3,337	5,425	10,427
Paragraphs	14,636	23,721	51,642	89,999
Sentences	27,491	45,170	96,238	168,899
Words	497,686	1,068,547	1,695,094	3,261,327
Unique words	73,793	107,637	145,889	214,705
Avg word length (chars)	5.71	5.66	5.70	5.69
Avg sentence length (words)	20.25	20.81	20.67	20.65
Nouns	170,516	360,013	585,621	1,116,150
Verbs	79,824	190,537	295,088	565,449
Adjectives	62,101	124,186	200,348	386,635
Adverbs	24,676	52,285	80,992	157,953
Pronouns	24,249	54,070	83,614	161,933
Numeral Forms	21,558	39,344	62,315	123,217
Prepositions	59,140	126,017	198,738	383,895
Conjunctions	37,258	84,074	131,554	252,886
Particles	13,280	35,103	51,738	100,121
Interjections	29	56	31	116
Abbreviations	2,813	1,602	2,813	7,228
Residuals	2,242	1,260	2,242	5,744

news labelled as negative is approximately twice as large as the proportion of news labelled as positive. A more detailed overview of the corpora is given in Tables 3.4 and 3.5.

Table 3.4 presents the relevant statistical information about the manually annotated news corpora. These corpora contain 214,705 unique words, which were not lemmatized. In Table 3.5 we report the proportion of instances (in %) that we labelled as positive (*pos*), negative (*neg*) and neutral (*neu*) within three levels of granularity for five web media resources. Considering the annotated news, Finance publishes the most positive

Table 3.5: Proportion of instances (in %) labelled as positive, negative and neutral within each level of granularity in the observed web media

Web	Document level			Paragraph level			Sentence level		
medium	pos	neg	neu	pos	neg	neu	pos	neg	neu
24ur	16.26	39.99	43.75	10.72	24.86	64.42	10.67	25.12	64.21
Dnevnik	14.60	32.32	53.08	13.99	25.90	60.11	14.25	25.87	59.88
Finance	22.05	22.60	55.35	25.88	24.74	49.38	26.59	25.85	47.56
Rtvslo	13.76	37.06	49.18	13.38	27.69	58.93	12.96	26.98	60.07
Žurnal24	13.52	27.89	58.59	18.58	27.89	53.53	19.29	29.67	51.04

news per web medium (22.05%), while 24ur publishes the highest proportion of negative news (39.99%). When compared to the other media, Finance has the most balanced proportion of positive and negative news within the corpora.

4 A LEXICON FOR SENTIMENT ANALYSIS IN SLOVENE

In this chapter, we describe the construction and characteristics of a new lexicon that supports sentiment analysis in Slovene.

Sentiment analysis approaches that employ supervised learning usually depend on sentiment labels. These labels are in most cases created manually for a large amount of training items, which is costly and time consuming. Alternatively, one can use unsupervised and semi-supervised learning approaches, which usually rely on the use of lexicons of e.g., positive and negative terms.

A major advantage of inducing a lexicon directly from data is capturing domain specific effects. The lexicon-based techniques are also useful in systems for real-time analysis, such as monitoring of public sentiment toward presidential candidates during election campaigns for example. While there are several lexicons for sentiment analysis for English, those for Slovene are scarce and, to the best of our knowledge, none of them are built directly from a collection of manually annotated texts in Slovene.

4.1 Lexicon Construction

The Slovene sentiment lexicon JOB 1.0[32], a lexicon for sentiment analysis in Slovene, is constructed on the basis of the List of Slovenian headwords 1.1[33]. JOB 1.0 contains a list of 25,524 headwords from the list, extended with sentiment ratings based on the AFINN model with an integer between -5 (very negative) and +5 (very positive). The ratings are derived from the lemmatized version of Manually sentiment annotated Slovenian (sentence-based) news corpus SentiNews 1.0, described in Section 3.2.

Table 4.1, which provides the first insight into the lexicon, contains the attributes, descriptions and data types. The original sentence-level annotations are based on the

[32] http://hdl.handle.net/11356/1112
[33] http://hdl.handle.net/11356/1038

Table 4.1: Attributes, descriptions and data types within JOB 1.0

Attribute	Description	Data type
Word	Headword from the List of Slovenian headwords 1.1	String
AFINN	Rounded avg_AFINN score	Integer (-5 to +5)
freq	Headword frequency (total number of occurrences in the annotated sentence-based news corpus)	Integer (0 to 260,931)
avg_AFINN	Average of AFINN values in Manually sentiment annotated Slovenian (sentence-based) news corpus SentiNews 1.0 deducted by the average sentiment of the corpus	Float (-4.61 to +5.39)
sd_AFINN	Standard deviation of AFINN values in Manually sentiment annotated Slovenian (sentence-based) news corpus SentiNews 1.0	Float (0 to 7.071)

five-level Lickert scale (1 - very negative and 5 - very positive).

The structure of JOB 1.0 is shown in Figure 4.1. For every manually annotated sentence in the corpus, we made a linear transformation of the *avg_sentiment*, i.e. the average of the sentence-based scores (Ann1 - Ann6), from the Lickert model to the AFINN model to obtain the corresponding AFINN values. Score 1 within the Lickert model was transformed to -5 within AFINN, score 2.4 to -1.5 (negative sentiment), score 3.6 to 1.5 (positive sentiment), and the score 5 retained its value. For every headword in the list, we counted the number of occurrences, and calculated the average (*avg_AFINN*) and standard deviation (*sd_AFINN*) of the AFINN values of all the sentences where this headword appeared, using the annotated sentence-based news corpus.

Word	AFINN	freq	avg_AFINN	sd_AFINN
a	0	3415	-0.466	1.787
aa	-1	57	-1.116	1.367
ab	-1	6	-1.277	2.189
aba	-1	5	-0.610	1.046
abančen	0	3	-0.443	1.443
abc	0	28	0.121	1.564
abe	-1	1	-0.860	0.000
abeceda	0	1	0.390	0.000
abeceden	2	3	1.640	2.500
abecednik	2	1	1.640	0.000

Figure 4.1: Sample format of JOB 1.0. The first line of the lexicon contains the names of the attributes and every headword is stored in a new line along with the associated attributes. JOB 1.0 is alphabetically ordered and tab-separated.

Due to different proportions of documents, labelled as positive, negative and neutral, the most common words, such as *biti (to be)*, *v (in)*, *in (and)*, had a slightly negative *avg_AFINN* values. For that reason, the *avg_AFINN* was deducted by -0.390, which is the average sentiment of the corpus. Finally, we obtained the AFINN score for every headword in the list by rounding the *avg_AFINN* score.

4.2 Exploring the Lexicon

We assigned the *AFINN* score to 25,524 words within the List of Slovenian headwords 1.1. The main reason for not assigning the *AFINN* score to all the words from the list are the words such as *aaahah*, *aaajs*, etc., which are not included in our corpora. In most cases such words are not included in the Dictionary of the Slovenian Standard Language[34] or are unusual words for our corpora with political, economic and financial content, such as domain-dependent words.

Words that contain a negative *AFINN* score provoke a negative sentiment. In a similar way the words engage a tendency for a positive sentiment if their *AFINN* score is positive. Thus, we assigned a positive *AFINN* score to 7,976 words and a negative to 7,898 words.

In particular, we present some sentiment-bearing words, together with the assigned *AFINN* scores and their frequencies. In addition to JOB 1.0, we provide the English translations and the MSDs, as shown in Table 4.2. MSDs define the POS categories for Slovene and for each category its attributes and their values. They also define the mapping from these values into a position-based compact string encoding, the morphosyntactic descriptions (MSDs), and list all valid MSDs for Slovene. So, for example, the MSD *Ncmsan*, given in Table 4.2, stands for *Noun*, *Type = common*, *Gender = masculine*, *Number = singular*, *Case = accusative*, *Animate = no*.

At this point, a careful eye might notice that the words assigned with the *AFINN* score equal to -5 and +5 are not revealed. To clarify, we set the filter criteria in such a way as to generate only the words that express an intense positive or negative sentiment, and appear at least 10 times in the corpus. There were no words with an *AFINN* score equal to -5 and +5 that occur so often, so consequently they are not shown in Table 4.2.

[34] `http://www.fran.si/130/sskj-slovar-slovenskega-knjiznega-jezika/`

Table 4.2: Most common terms that express a sentiment in the annotated sentence-level news corpus, their translations, MSDs, AFINN scores and frequencies

Word	Word (Eng)	MSD	AFINN	freq
dota	dowry	Ncfsn	4	11
doživetje	experience	Ncnsn	4	21
kopel	bath	Ncfsn	4	13
ponareditev	forgery	Ncfsn	-3	11
pretepsti	(to) beat	Vmen	-3	10
splav	abortion, raft	Ncmsan	-3	16
trčiti	(to) collide	Vmen	-3	21
ubiti	(to) kill	Vmen	-3	34
zagoreti	(to) burn	Vmen	-3	20

Limiting our observations to the two most frequent words in Table 4.2, e.g. *ubiti (to kill)* and *doživetje (experience)*, we used histograms to illustrate the distribution of their frequencies across 11 groups within the AFINN model, as shown in Figure 4.2.

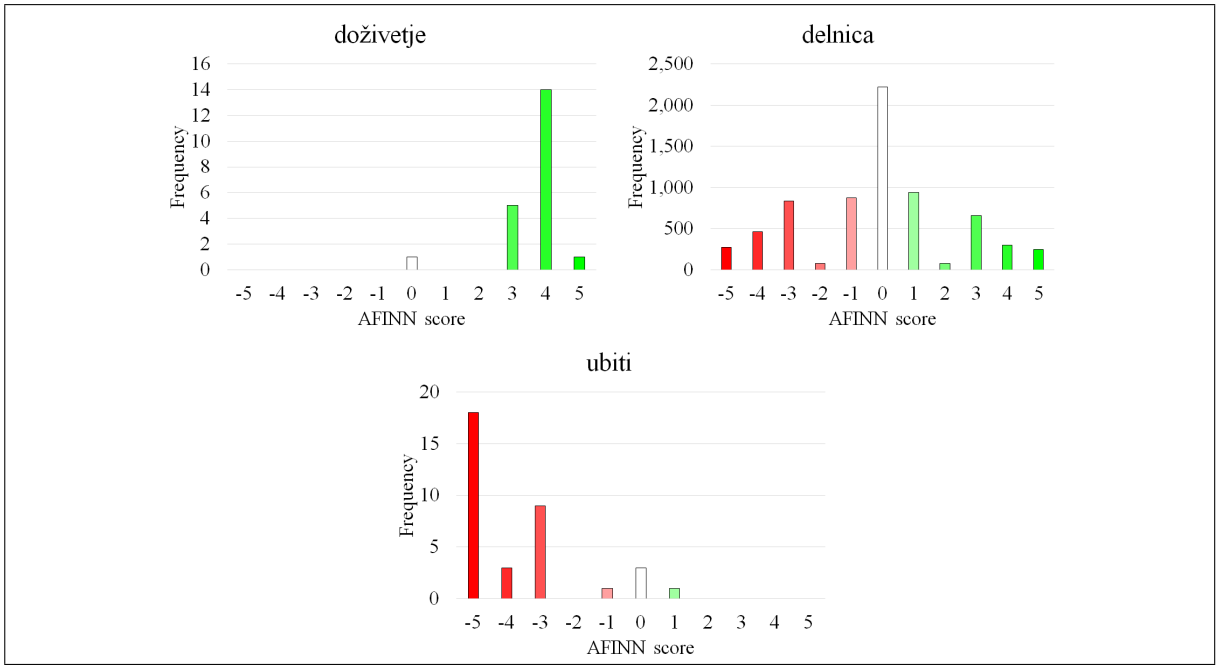

Figure 4.2: Words *doživetje* (left), *delnica* (middle), *ubiti* (right), and their frequencies across 11 groups within the AFINN model. The *AFINN* scores are coloured with more intense tones of green and red colour, to emphasize the sentiment polarity.

Their histograms are both asymmetrical, and the distribution of their frequencies is concentrated on either the positive or the negative side. The word *doživetje (experience)* has its peak at the *AFINN* score of +4. In contrast, the word *ubiti (to kill)* has its peak at the *AFINN* score of -5. Unsurprisingly, it appears that they both *doživetje* and *ubiti*

arouse a strong sentiment, each in its own way. The deduction of the average sentiment within the word *ubiti* decreased the negativity of the *AFINN* score from -4 to -3. The word *ubiti* generally provokes very negative sentiment, and therefore we would expect it to be -5. However, a more detailed inspection reveals its presence in sentences labelled as positive within our corpus, such as *"to kill two birds with one stone"*. In addition, we present a histogram of the word *delnica (capital stock)*. In contrast to the previous two words, this histogram is symmetrical. It reaches its peak at the *AFINN* score 0 (the frequency is equal to 7,024), and provokes neither only positive nor only negative sentiment.

4.3 Availability of the Developed Resources

Access to the resources is provided through the national technology infrastructure for language resources and tools CLARIN[35] and the GitHub[36] website.

Our resources include information about original texts, metadata and the annotation process. They are available in textual format using UTF-8 encoding and tab-separation. The resources were not lemmatised, MSD-tagged or linguistically processed in any way. All our resources are available under Creative Commons copyright license Attribution-ShareAlike 4.0 International[37] (CC BY-SA 4.0 or newer version).

[35] https://www.clarin.si/repository/xmlui/browse?value=Bu%C4%8Dar,%20Jo%C5%BEe&type=author

[36] https://github.com/19Joey85/Sentiment-annotated-news-corpus-and-sentiment-lexicon-in-Slovene

[37] https://creativecommons.org/licenses/by-sa/4.0/

5 PERFORMANCE EVALUATION OF SENTIMENT-BASED CLASSIFICATION TECHNIQUES

Sentiment classification might be the most widely studied problem in the field of sentiment analysis. Most techniques apply supervised learning, where a bag of words is the most commonly used representation. In this Chapter, we empirically evaluate the best approaches for two-class (positive and negative) and three-class (positive, negative and neutral) document-based sentiment classification of the Slovenian news texts.

5.1 Selection of Classifiers and Settings for Sentiment Classification

In our preliminary experiments (Bučar, Povh & Žnidaršič, 2016), we studied the performance of different classifiers within the two-class document-level sentiment classification.

5.1.1 Selection of Classifiers and Settings Choice

Initially, nine classifiers were evaluated on the document-based corpus, such as KNN, NBM, SVM (SVM-poly and SVM-lin), RF, C4.5, DT, SLR and VP (see second paragraph in *Classification Algorithms* in Subsection 1.4). In our experiments we used their implementations with default settings[38] from the WEKA machine-learning toolkit, version

[38] Implementations of classifiers in WEKA 3.6.11:
k-Nearest Neighbour (KNN): IBk -K 9 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.EuclideanDistance -R first-last""
Multinomial Naïve Bayes (NBM): NaiveBayesMultinomial
Support Vector Machine (SVM): SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"
Random Forest (RF): RandomForest -I 10 -K 0 -S 1
C4.5: J48 -C 0.25 -M 2
Decision Table (DT): DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"
Simple Logistic Regression (SLR): SimpleLogistic -I 0 -M 500 -H 50 -W 0.0
Voted Perceptron (VP): VotedPerceptron -I 1 -E 1.0 -S 1 -M 10000

3.6.11 (Witten, Frank & Hall, 2013). We used 10-fold CV technique within the two-class (positive and negative) and the three-class (positive, negative and neutral) document-level sentiment classification for all classifiers, using the BOW approach with a condition that a given term has to appear at least twice in the entire corpus (Manually sentiment annotated Slovenian (document-based) news corpus SentiNews 1.0). Within the two-class classification, we used 5,002 documents (1,665 documents labelled as positive and 3,337 as negative); and within the three-class classification we used 10,427 documents (1,665 documents labelled as positive, 3,337 as negative and 5,425 as neutral). The models included the following pre-processing techniques: term frequency - inverse document frequency (TF-IDF) weighting scheme, transformation of upper case letters to lower case, removal of stop words, combination of unigrams, bigrams, and trigrams.

Information about labelled instances within different levels of granularity show that there is a class imbalance in the three classes (for details see the third paragraph and Tables 3.4 and 3.5 in Section 3.3). The class imbalance problem typically occurs when there are many more instances of some classes than others. The standard classifiers tend to be overwhelmed by the large classes and ignore the small ones (Chawla, Japkowicz & Kotcz, 2004). We repeated the initial experiments with the same set of pre-processing options and classifiers, but with a balanced data set of 1,000 documents per class. For balancing the classes, we first used 2,000 documents, 1,000 documents labelled as positive and 1,000 as negative, and performed classification into the two sentiment categories within the balanced data set. Similarly, we classified 3,000 documents, i.e. 1,000 documents labelled as positive, 1,000 as negative and 1,000 as neutral, into three sentiment categories.

5.1.2 Results and Findings

In terms of time consumption and performance there are two classifiers, the NBM and the SVM, that perform best.

If we focus on two-class document-level classification with imbalanced classes, the best accuracy 87.68% is achieved by the SVM (F1-score: 90.98%) and the NBM (accuracy: 82.99%, F1-score: 86.98%) (see Table 5.1). The NBM has been shown as a very fast method, while the SVM computationally expensive. The RF reaches the accuracy of 74.61% (F1-score: 83.34%), however, it computes the results in a reasonable time only

Table 5.1: Performance evaluation (in %) within the two-class and three-class document-level sentiment classification for the KNN, NBM, SVM-poly and SVM-lin by using 10-fold CV with an imbalanced data set of documents (1,665 positive, 3,337 negative and 5,425 neutral)

Document-level based on the average scores of documents
(based on annotations at the document-level granularity)

Two-class

	KNN	NBM	SVM-poly	SVM-lin	RF
Accuracy	60.81 ± 9.73	**82.99 ± 1.73**	**87.21 ± 1.14***	**87.68 ± 0.85***	74.61 ± 1.34
Precision	67.46 ± 1.09	88.78 ± 1.10	88.52 ± 0.75	88.93 ± 0.55	74.12 ± 0.89
Recall	80.92 ± 30.06	85.29 ± 2.65	92.87 ± 1.35	93.14 ± 1.31	95.17 ± 1.18
F1-score	70.09 ± 17.91	**86.98 ± 1.47**	**90.64 ± 0.87***	**90.98 ± 0.67***	83.34 ± 0.87
Time [s]	30.91 ± 0.83	0.08 ± 0.04	271.20 ± 11.44	111.14 ± 12.74	22.64 ± 0.67

	C4.5	DT	SLR	VP
Accuracy	70.45 ± 2.02	x	x	x
Precision	78.82 ± 2.72	x	x	x
Recall	76.45 ± 3.72	x	x	x
F1-score	77.53 ± 1.65	x	x	x
Time [s]	9,105.75 ± 408.33	>24h	>24h	>24h

Three-class

	KNN	NBM	SVM-poly	SVM-lin	others
Accuracy	52.06 ± 0.27	**60.11 ± 1.16**	**60.02 ± 1.40**	**61.24 ± 1.26***	x
Precision	44.75 ± 44.85	59.46 ± 2.40	55.94 ± 2.06	58.20 ± 2.18	x
Recall	0.42 ± 0.49	56.91 ± 1.92	56.40 ± 2.75	56.16 ± 2.29	x
F1-score	0.83 ± 0.97	**58.11 ± 1.23**	**56.13 ± 1.83**	**57.14 ± 1.70**	x
Time [s]	123.07 ± 5.12	0.16 ± 0.04	4,662.58 ± 408.54	1,556.62 ± 57.01	>24h

- Hardware and software: 12 x (Intel® Core™ i7, 2.20GHz, 4 cores, 8GB RAM), Weka 3.6.11
- Pre-processing settings: TF-IDF, transformation to lower case, removal of stop words, combination of unigrams, bigrams, and trigrams, 1,000 features
- DT, SLR, VP take >24h for training and testing one model out of 10 (10-fold CV) within two-class and three-class document-level sentiment classification

* statistically significant (paired t-test) at the 0.05 significance level

within document level of granularity, whereas for paragraph level and sentence level it takes more than 24 hours with a standard PC for training and testing one model out of 10 within the 10-fold CV technique. The C4.5 achieves the accuracy of 70.45% (F1-score: 77.53%). This classifier turns out as a very time consuming since it takes more than two hours and a half for one fold to return results for the simplest experiment, i.e., classification into two classes at document level of granularity. This also applies to the DT, SLR, and VP since it takes more than 24 hours for the same task.

The best performance (accuracy) is achieved within two-class document-level sentiment classification. When classifying texts into two classes all selected performance measures (accuracy, precision, recall and F1-score) drop notably when segmenting texts

Table 5.2: Performance evaluation (in %) within the two-class and three-class paragraph-level sentiment classification for the KNN, NBM, SVM-poly and SVM-lin by using 10-fold CV with an imbalanced data set of paragraphs (14,636 positive, 23,721 negative and 51,642 neutral)

Paragraph-level based on the average scores of paragraphs
(based on annotations at the paragraph-level granularity)

Two-class

	KNN	NBM	SVM-poly	SVM-lin
Accuracy	62.59 ± 0.18	81.29 ± 0.64*	78.47 ± 0.64	77.66 ± 0.52
Precision	62.36 ± 0.10	85.00 ± 0.46	81.53 ± 0.60	80.75 ± 0.43
Recall	99.70 ± 0.12	84.68 ± 0.82	84.30 ± 0.81	83.87 ± 0.72
F1-score	76.72 ± 0.10	84.84 ± 0.55*	82.89 ± 0.53	82.28 ± 0.44
Time [s]	185.62 ± 6.13	0.10 ± 0.02	4,473 ± 217.60	6,246.86 ± 934.68

Three-class

	KNN	NBM	SVM-poly	SVM-lin
Accuracy	58.72 ± 0.11	60.47 ± 0.63	61.31 ± 0.62*	60.68 ± 0.50
Precision	69.97 ± 4.38	53.20 ± 1.00	51.54 ± 1.08	51.05 ± 0.77
Recall	3.27 ± 0.44	55.82 ± 1.48	48.49 ± 1.03	48.24 ± 0.92
F1-score	6.25 ± 0.81	54.48 ± 1.19*	49.96 ± 0.91	49.60 ± 0.76
Time [s]	1,031.29 ± 9.55	0.19 ± 0.03	31,377.97 ± 1,609.84	43,710.10 ± 6,916.54

* Hardware and software: 12 x (Intel® Core™ i7, 2.20GHz, 4 cores, 8GB RAM), Weka 3.6.11
* Pre-processing settings: TF-IDF, transformation to lower case, removal of stop words, combination of unigrams, bigrams, and trigrams, 1,000 features
- RF, C.4.5, DT, SLR, VP take >24h for training and testing one model out of 10 (10-fold CV) within two-class and three-class paragraph-level sentiment classification

* statistically significant (paired t-test) at the 0.05 significance level

from document level to sentence level (document level (accuracy, NBM: 82.99%, SVM: 87.68%), paragraph level (accuracy, NBM: 81.29%, SVM: 78.47%) and sentence level (accuracy, NBM: 78.93%, SVM: 78.19%); for more details see Tables 5.1 and 5.2 and 5.3). Also, the time needed for training and testing increases markedly, mainly depending on the number of documents. When comparing performance of two-class and three-class sentiment classification, the accuracy drops significantly for approximately 20-25% for all three levels of granulation.

Balancing the classes with 1,000 documents per class improves performance significantly. When classifying documents into two classes the NBM performs the best (F1-score: 93.09% ± 1.39%, Accuracy: 92.80% ± 1.46%), likewise, the SVM conducts similar results (F1-score: 92.14% ± 1.90%, Accuracy: 92.25% ± 1.81%). The NBM has proven to be a very fast method. On average, it took only one hundredth of a second with a 2.20 GHz Intel Core i7 processor and 8 GB RAM for training and testing one model out of 10 in the CV process. On the other hand, the SVM took almost 2 seconds for the same task. The

Table 5.3: Performance evaluation (in %) within the two-class and three-class sentence-level sentiment classification for the KNN, NBM, SVM-poly and SVM-lin by using 10-fold CV with an imbalanced data set of sentences (27,491 positive, 45,170 negative and 96,238 neutral)

Sentence-level based on the average scores of sentences
(based on annotations at the sentence-level granularity)

Two-class

	KNN	NBM	SVM-poly	SVM-lin
Accuracy	62.65 ± 0.14	78.93 ± 0.31*	78.19 ± 0.39	74.14 ± 0.57
Precision	62.53 ± 0.08	83.88 ± 0.29	81.20 ± 0.50	79.22 ± 0.60
Recall	99.62 ± 0.10	81.83 ± 0.52	84.48 ± 0.44	79.17 ± 0.80
F1-score	76.83 ± 0.07	82.84 ± 0.28	82.81 ± 0.28	79.198 ± 0.48
Time [s]	437.96 ± 1.78	0.12 ± 0.02	10,156.07 ± 352.97	50,983.46 ± 6,439.65

Three-class

	KNN	NBM	SVM-poly	SVM-lin
Accuracy	57.80 ± 0.13	59.14 ± 0.31	62.46 ± 0.21*	59.22 ± 0.30
Precision	55.59 ± 3.10	52.53 ± 0.59	54.77 ± 0.48	53.43 ± 0.58
Recall	3.60 ± 0.28	54.22 ± 0.61	46.46 ± 0.30	43.54 ± 0.55
F1-score	6.75 ± 0.51	53.36 ± 0.49*	50.27 ± 0.29	48.08 ± 0.49
Time [s]	2,181.02 ± 18.59	0.32 ± 0.02	55,405.68 ± 1,941.40	277,860.26 ± 35,421.56

* Hardware and software: 12 x (Intel® Core™ i7, 2.20GHz, 4 cores, 8GB RAM), Weka 3.6.11
* Pre-processing settings: TF-IDF, transformation to lower case, removal of stop words, combination of unigrams, bigrams, and trigrams, 1,000 features
- RF, C.4.5, DT, SLR, VP take >24h for training and testing one model out of 10 (10-fold CV) within two-class and three-class sentence-level sentiment classification

* statistically significant (paired t-test) at the 0.05 significance level

DT, SLR, and VP turned out to be computationally expensive since more than 24 hours elapsed for the training and testing one model. On the other hand, the KNN, RF and C4.5 performed significantly worse by using paired t-test at the 0.05 significance level. For these two reasons, we omitted some calculations that relate to these classifiers.

When classifying documents with a balanced data set into three classes, once again, the NBM performed the best (F1-score: 69.73% ± 3.39%, Accuracy: 70.07% ± 2.62%), followed by the SVM (F1-score: 65.33% ± 3.60%, Accuracy: 65.80% ± 2.59%). The other classifiers either achieved poor results, they did not exceed 50% (a random choice) within the F1-score and accuracy, or it took them more than 24 hours to train and test one model with 10-fold CV technique. Comparisons between the applied classifiers were performed using paired t-test at the significance level of 5%, whereby the NBM and the SVM outperformed the other classifiers. Therefore, we use these classifiers hereinafter.

Table 5.4: Performance evaluation (in %) within the two-class and three-class document-level sentiment classification for the KNN, the NBM, the SVM-poly and the SVM-lin by using 10-fold CV with a balanced data set of documents (1,000 positive, 1,000 negative and 1,000 neutral)

Document-level baesd on the average scores of documents
(based on annotations at the document-level granularity)

Two-class

	KNN	NBM	SVM-poly	SVM-lin	RF
Accuracy	53.40 ± 3.23	**92.80 ± 1.46**	**92.25 ± 1.81**	**87.15 ± 1.67**	81.50 ± 1.29
Precision	52.08 ± 2.26	91.98 ± 2.31	93.39 ± 2.33	86.45 ± 2.03	79.67 ± 2.16
Recall	92.60 ± 7.09	94.30 ± 2.16	91.00 ± 3.23	88.20 ± 3.43	84.70 ± 1.70
F1-score	66.50 ± 1.63	**93.09 ± 1.39**	**92.14 ± 1.90**	**87.26 ± 1.79**	82.08 ± 1.09
Time [s]	0.45 ± 0.06	0.01 ± 0.01	1.73 ± 0.19	1.35 ± 0.30	2.39 ± 0.09
	C4.5	DT	SLR	VP	
Accuracy	71.75 ± 3.17	x	x	x	
Precision	70.02 ± 3.33	x	x	x	
Recall	76.30 ± 3.40	x	x	x	
F1-score	72.99 ± 2.83	x	x	x	
Time [s]	43.30 ± 1.61	>24h	>24h	>24h	

Three-class

	KNN	NBM	SVM-poly	SVM-lin	RF
Accuracy	35.63 ± 1.49	**70.07 ± 2.62***	**65.80 ± 2.59**	**63.63 ± 3.14**	53.00 ± 2.45
Precision	34.63 ± 0.97	71.71 ± 3.07	62.86 ± 3.49	62.81 ± 2.99	50.70 ± 4.15
Recall	93.80 ± 2.94	68.00 ± 4.81	68.10 ± 4.58	65.20 ± 5.29	55.90 ± 5.26
F1-score	50.57 ± 1.26	**69.73 ± 3.39***	**65.33 ± 3.60**	**63.89 ± 3.35**	53.13 ± 4.44
Time [s]	1.94 ± 0.11	0.02 ± 0.01	14.72 ± 0.60	14.99 ± 1.88	2.95 ± 0.14
	C4.5	DT	SLR	VP	
Accuracy	49.50 ± 2.27	x	x	x	
Precision	48.36 ± 3.44	x	x	x	
Recall	52.20 ± 3.49	x	x	x	
F1-score	50.08 ± 2.27	x	x	x	
Time [s]	133.56 ± 4.54	>24h	>24h	>24h	

* Hardware and software: 12 x (Intel® Core™ i7, 2.20GHz, 4 cores, 8GB RAM), Weka 3.6.11
* Pre-processing settings: TF-IDF, transformation to lower case, removal of stop words, combination of unigrams, bigrams, and trigrams, 1,000 features
* DT, SLR, VP: takes >24h for training and testing one model out of 10 (10-fold CV) within three-class document-level sentiment classification

* statistically significant (paired t-test) at the 0.05 significance level

5.2 Feature Selection: Feature Vector Size and Its Impact on Performance

Here, we investigate the impact of feature vector size on performance (accuracy and F1-score) and computational complexity (time to train and test one model out of 10 (10-fold CV technique)).

5.2.1 Selection of Classifiers and Settings Choice

We performed series of tests where we tested three feature subset selection methods, i.e., Chi-squared, Gain Ratio and Information gain (Hall & Smith, 1998), combining various sizes of feature vector (from 100 to 30,000 features). In our experiments we used the implementations of the feature selection methods with default settings[39] from the WEKA machine-learning toolkit, version 3.6.11 (Witten, Frank & Hall, 2013).

In similar settings as introduced in Section 5.1, i.e., by using TF-IDF weighting scheme, upper case to lower case transformation, removal of stop words, and combination of unigrams, bigrams, and trigrams, we tested three algorithms, the NBM, the SVM-poly and the SVM-lin, for two-class and three-class sentiment classification tasks. Considering the granularity of the texts, we separately performed document-based sentiment classification on the average scores of documents, paragraphs and sentences by using 10-fold CV technique with an imbalanced data set of documents (1,665 positive, 3,337 negative and 5,425 neutral).

5.2.2 Results and Findings

Applying effective and efficient feature selection can enhance the performance of sentiment analysis in terms of accuracy, F1-score and time to train classifier (Sharma & Dey, 2012), like in our study. Based on our experiments, we find an interesting trend. Generally, we achieve the best performance (in terms of accuracy and F1-score) if we select between 2,000

[39] Implementations of feature selection methods in WEKA 3.6.11:
Chi-squared: ChiSquaredAttributeEval
Gain Ratio: GainRatioAttributeEval
Information gain: InfoGainAttributeEval

Table 5.5: Performance evaluation (in %) within the two-class and the three-class document-level sentiment classification for the NBM and the SVM by using 10-fold CV with an imbalanced data set of documents (1,665 positive, 3,337 negative and 5,425 neutral)

Document-level based on the average scores of documents (based on annotations at the document-level granularity)				
	Two-class		Three-class	
	NBM	SVM-poly	NBM	SVM-poly
Accuracy	91.10 ± 1.78*	89.92 ± 1.20	63.40 ± 1.22	64.59 ± 1.32*
F1-score	93.18 ± 1.39	92.49 ± 0.95	65.00 ± 1.70*	61.84 ± 1.49
Document-level based on the average scores of paragraphs (based on annotations at the paragraph-level granularity)				
	Two-class		Three-class	
	NBM	SVM-poly	NBM	SVM-poly
Accuracy	93.98 ± 1.34*	91.64 ± 1.01	65.64 ± 2.29	70.81 ± 1.36*
F1-score	95.33 ± 1.04*	93.64 ± 0.78	59.93 ± 2.38*	55.76 ± 1.95
Document-level based on the average scores of sentences (based on annotations at the sentence-level granularity)				
	Two-class		Three-class	
	NBM	SVM-poly	NBM	SVM-poly
Accuracy	94.53 ± 1.30*	92.45 ± 1.33	65.65 ± 1.23	72.72 ± 1.45*
F1-score	95.84 ± 1.00*	94.38 ± 0.99	60.13 ± 2.27*	54.15 ± 2.48
* statistically significant (paired t-test) at the 0.05 significance level				

and 5,000 features (irrespective to feature selection method and selection of classifier). In general, the performance rapidly improves to feature vector size equal to 5,000 and then starts decreasing. Our tests show that there is no significant difference between results conducted with different feature selection methods (see Figures 5.1, 5.2, 5.3, 5.4, 5.5 and 5.6). However, we used the Gain Ratio method, since it performs the best in most cases.

Again, our tests show that the NBM achieves better results than the SVM. However, the paired t-test with the significance level 0.05 shows that the NBM is not always significantly better compared to the SVM (see Table 5.5). Also, the tests show that SVM-poly in general achieves better performance when compared to SVM-lin. Consequently, we listed only relevant values for the SVM-poly in Table 5.5.

As expected, increasing the number of features reflects in increasing time that is needed for training and testing. We can notice that the SVM often outperforms the NBM when the feature size vector is relatively small. The NBM proves to be an extremely fast and effective method, while the SVM becomes more and more time-consuming method when increasing the size of feature vector.

Below we present the best F1-score (and accuracy) within two-class document-level classification depending on the Gain Ratio feature selection method and by applying three levels of granularity:

- Document-level based on the average scores of documents: **NBM: 93.18% ± 1.39% (91.10% ± 1.78%)**, SVM-poly: 92.49% ± 0.95% (89.92% ± 1.20%), SVM-lin: 91.13% ± 0.81% (87.99% ± 1.13%) (NBM and SVM-poly: 4,000 features, SVM-lin: 10,000 features),
- Document-level based on the average scores of paragraphs: NBM: 95.33% ± 1.04% (93.98% ± 1.34%), SVM-poly: 93.64% ± 0.78% (91.64% ± 1.01%), SVM-lin: 93.11% ± 0.82% (90.90% ± 1.04%) (NBM and SVM-poly: 4,000 features, SVM-lin: 10,000 features),
- Document-level based on the average scores of sentences: **NBM: 95.84% ± 1.00% (94.53% ± 1.30%)**, SVM-poly: 94.38% ± 0.99% (92.45% ± 1.33%), SVM-lin: 94.29% ± 1.30% (92.27% ± 1.79%) (NBM and SVM-poly: 4,000 features, SVM-lin: 10,000 features).

Furthermore, we present the best F1-score (and accuracy) within three-class document-level classification depending on the Gain Ratio feature subset selection method and by applying three levels of granularity:

- Document-level based on the average scores of documents: **NBM: 65.00% ± 1.70% (63.40% ± 1.22%)**, SVM-poly: 61.84% ± 1.49% (64.59% ± 1.32%), SVM-lin: 58.00% ± 1.23% (61.19% ± 1.35%) (NBM and SVM-poly: 3,000 features, SVM-lin: 2,000 features),
- Document-level based on the average scores of paragraphs: NBM: 59.93% ± 2.38% (65.64% ± 2.29%), SVM-poly: 55.76% ± 1.95% (70.81% ± 1.36%), SVM-lin: 52.43% ± 2.16% (67.14% ± 1.28%) (NBM and SVM-poly: 3,000 features, SVM-lin: 2,000 features),
- Document-level based on the average scores of sentences: **NBM: 60.13% ± 2.27% (65.65% ± 1.23%)**, SVM-poly: 54.15% ± 2.48% (72.72% ± 1.45%), SVM-lin: 50.82% ± 1.54% (68.71% ± 1.08%) (NBM and SVM-lin: 2,000 features, SVM-poly: 3,000 features).

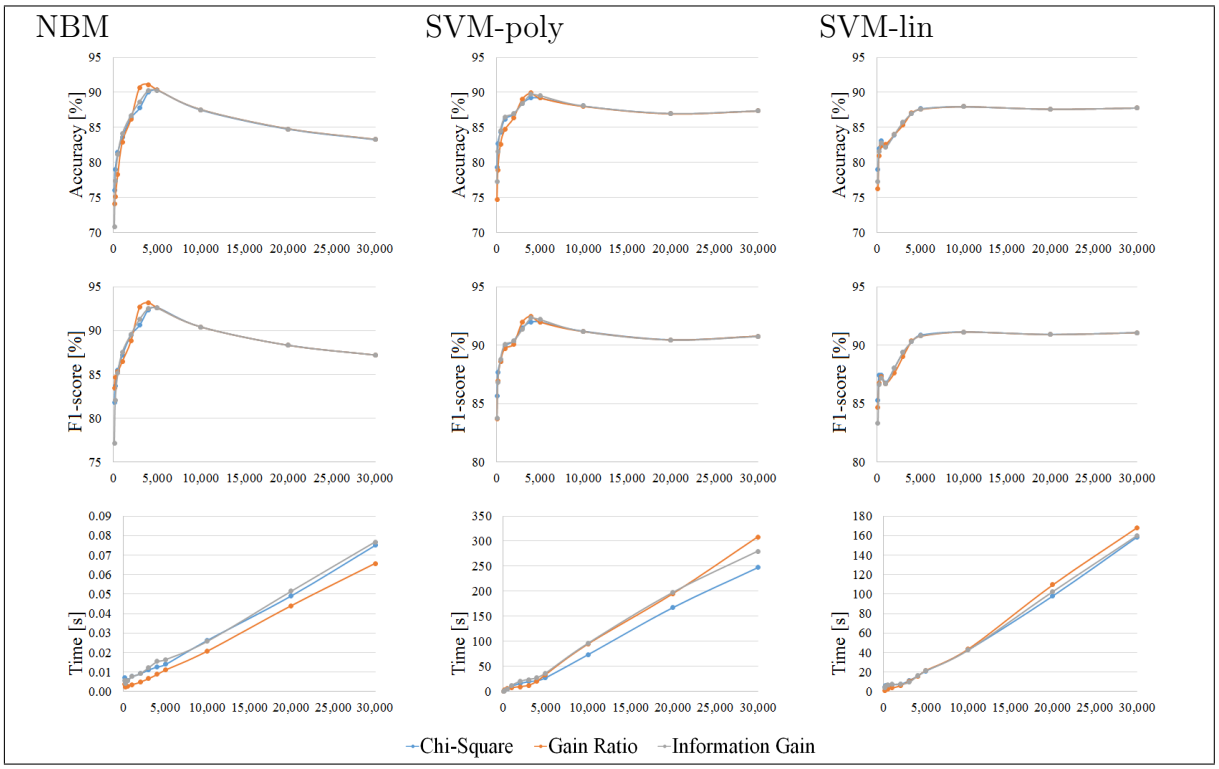

Figure 5.1: Performance evaluation (in %) according to the feature selection methods and feature vector size within the two-class document-level sentiment classification for the NBM, the SVM-poly and the SVM-lin by using 10-fold CV with an imbalanced data set of documents

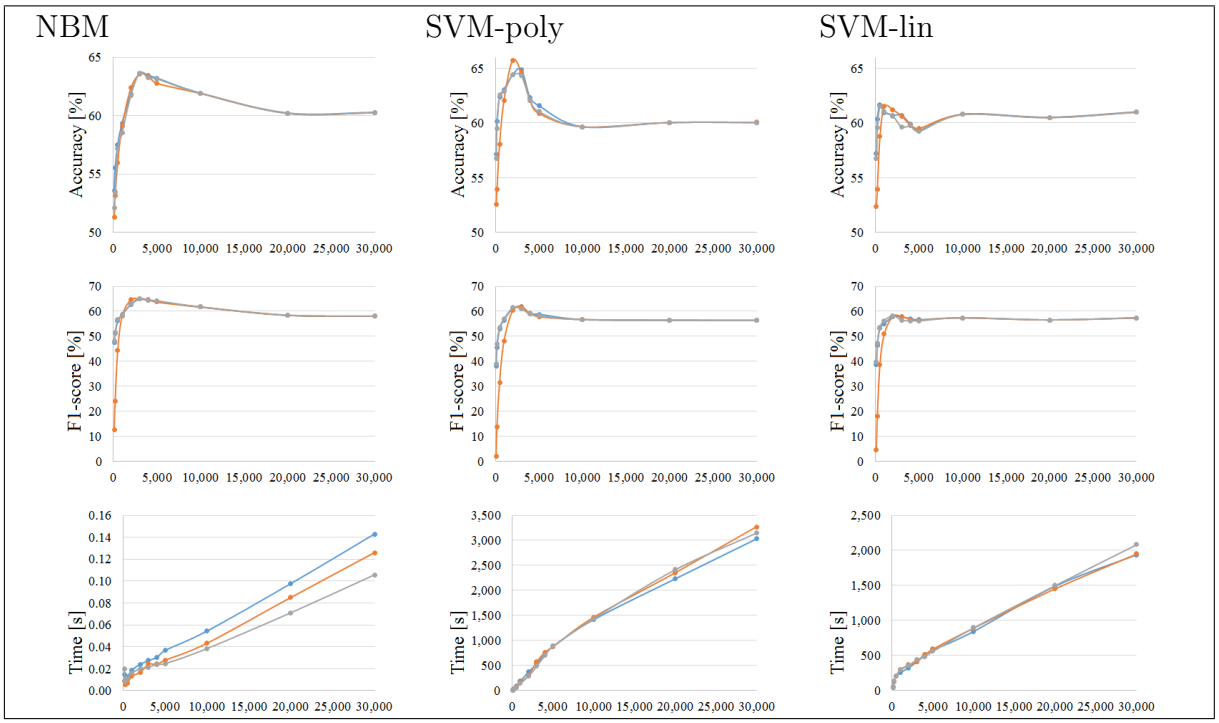

Figure 5.2: Performance evaluation (in %) according to the feature selection methods and feature vector size within the three-class document-level sentiment classification for the NBM, the SVM-poly and the SVM-lin by using 10-fold CV with an imbalanced data set of documents

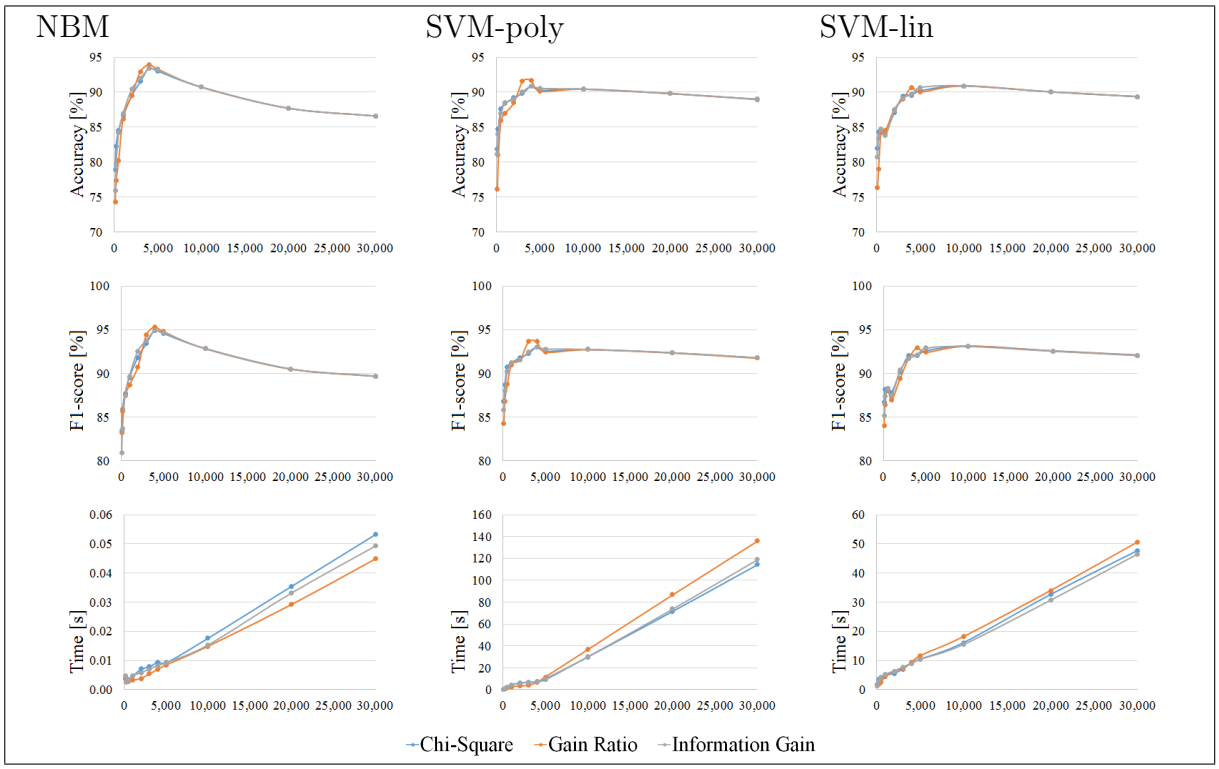

Figure 5.3: Performance evaluation (in %) according to the feature selection methods and feature vector size within the two-class document-level sentiment classification based on average scores of paragraphs for the NBM, the SVM-poly and the SVM-lin by using 10-fold CV with an imbalanced data set of documents

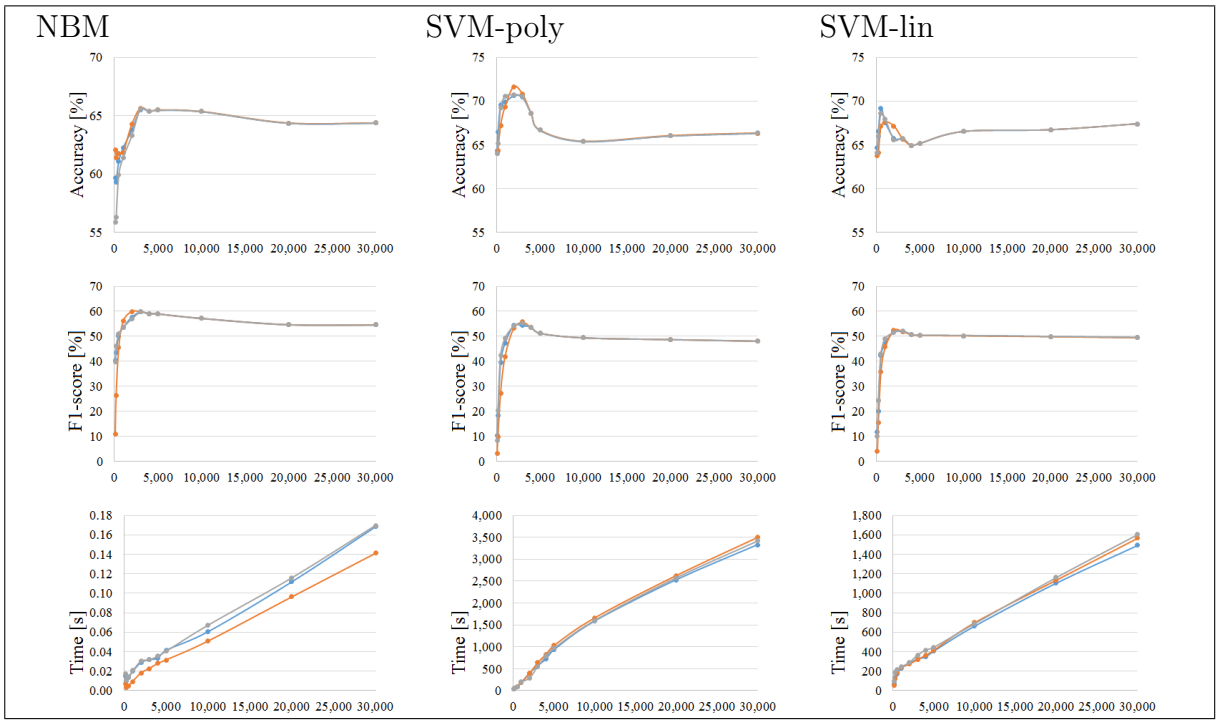

Figure 5.4: Performance evaluation (in %) according to the feature selection methods and feature vector size within the three-class document-level sentiment classification based on average scores of paragraphs for the NBM, the SVM-poly and the SVM-lin by using 10-fold CV with an imbalanced data set of documents

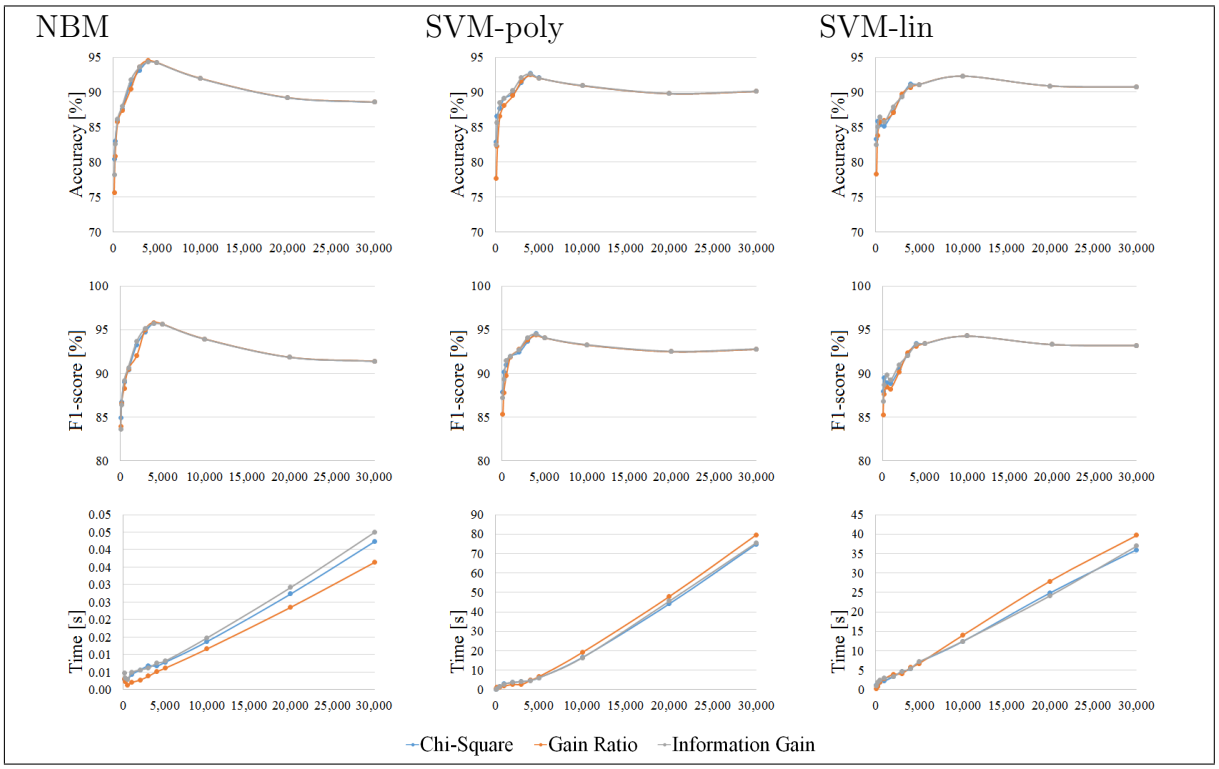

Figure 5.5: Performance evaluation (in %) according to the feature selection methods and feature vector size within the two-class document-level sentiment classification based on average scores of sentences for the NBM, the SVM-poly and the SVM-lin by using 10-fold CV with an imbalanced data set of documents

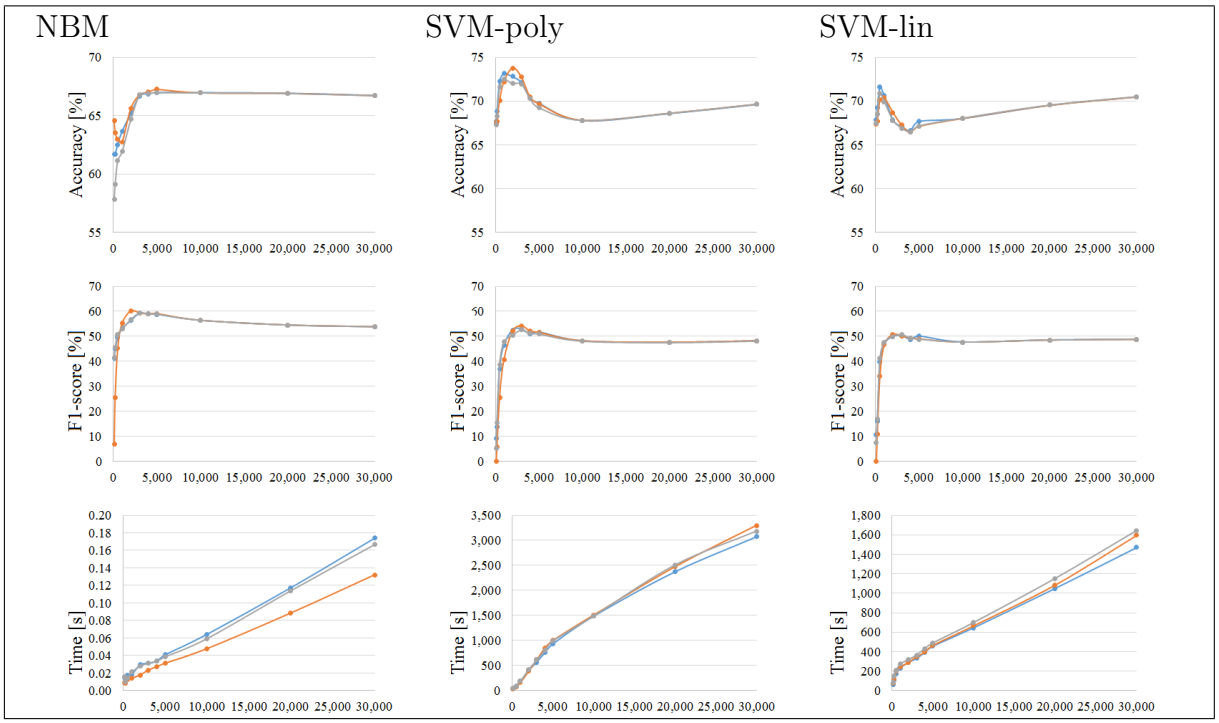

Figure 5.6: Performance evaluation (in %) according to the feature selection methods and feature vector size within the three-class document-level sentiment classification based on average scores of sentences for the NBM, the SVM-poly and the SVM-lin by using 10-fold CV with an imbalanced data set of documents

5.3 Performance Evaluation of Sentiment-based Classification Techniques

Sentiment classification might be the most widely studied problem in the field of sentiment analysis. Most techniques apply supervised learning, where a bag of words is the most commonly used representation. In this Section, we empirically evaluate the approaches for two-class (positive and negative) and three-class (positive, negative and neutral) document-based sentiment classification of the Slovenian news texts.

5.3.1 Selection of Classifiers and Settings Choice

As in the previous experiments in Sections 5.1 and 5.2, the results show that the NBM and the SVM significantly outperform the other classifiers in terms of classification performance and computational time consumption. Therefore, we focused on these two approaches to assess sentiment classification performance on the newly developed data resources. We tested a large set of pre-processing options, i.e. the term frequency (TF) or term frequency - inverse document frequency (TF-IDF) weighting scheme, the transformation of upper-case letters to lower-case, the removal of stop words, lemmatization, and different combinations of unigrams, bigrams and trigrams. We performed experiments both with balanced and imbalanced data sets of documents, and classified the documents in two ways to assess the impact of data granularity: (I) based on the average scores of documents and (II) based on the average scores of sentences. We omitted sentiment classification based on the average scores of paragraphs, since the results indicate that granulation of documents to paragraph level does not perform the best results (see Table 5.5).

5.3.2 Results and Findings

The experiments turned out to be computationally demanding, as it took us one month to build and evaluate the performance of 2,400 different predictive models on twelve desktop computers simultaneously. We present the best results (considering the pre-processing options) in terms of accuracy and F1-score within the two-class and the three-class document-based sentiment classification for the NBM and the SVM used on the imbalanced

Table 5.6: Performance evaluation (in %) within the two-class and the three-class document-level sentiment classification for the NBM and the SVM by using 5 times 10-fold CV with an imbalanced data set of documents (1,665 positive, 3,337 negative and 5,425 neutral)

Document-level based on the average scores of documents (based on annotations at the document-level granularity)				
	Two-class		Three-class	
	NBM	SVM	NBM	SVM
Accuracy	91.07 ± 0.96*	90.68 ± 1.18	64.32 ± 1.21	66.50 ± 1.41*
F1-score	93.19 ± 0.74	93.06 ± 0.88	65.97 ± 1.70*	63.42 ± 1.96
Document-level based on the average scores of sentences (based on annotations at the sentence-level granularity)				
	Two-class		Three-class	
	NBM	SVM	NBM	SVM
Accuracy	95.21 ± 0.98*	93.10 ± 1.18	66.46 ± 1.49	73.10 ± 1.23*
F1-score	96.38 ± 0.76*	94.86 ± 0.85	61.20 ± 2.21*	55.35 ± 2.31
* statistically significant (paired t-test) at the 0.05 significance level				

Table 5.7: Performance evaluation (in %) within the two-class and the three-class document-level sentiment classification for the NBM and the SVM by using 5 times 10-fold CV with a balanced data set of documents (1,000 positive, 1,000 negative and 1,000 neutral)

Document-level based on the average scores of documents (based on annotations at the document-level granularity)				
	Two-class		Three-class	
	NBM	SVM	NBM	SVM
Accuracy	92.89 ± 1.65	92.55 ± 1.64	73.09 ± 2.28*	67.94 ± 2.57
F1-score	93.12 ± 1.65*	92.48 ± 1.69	72.77 ± 3.44*	67.71 ± 3.18
Document-level based on the average scores of sentences (based on annotations at the sentence-level granularity)				
	Two-class		Three-class	
	NBM	SVM	NBM	SVM
Accuracy	**97.83 ± 0.98***	**96.27 ± 1.34**	**79.85 ± 1.93***	**76.20 ± 2.29**
F1-score	**97.85 ± 0.97***	**96.28 ± 1.34**	**77.76 ± 3.13***	**74.61 ± 3.16**
* statistically significant (paired t-test) at the 0.05 significance level				

data set of documents (see Table 5.6) and on the balanced data set of documents (see Table 5.7). We applied the paired t-tests between the NBM and the SVM at the significance level of 5%, in order to evaluate the classifiers within the two-class and three-class document-based sentiment classification. Relative to each classifier, the * sign means that the measure of this classifier is significantly better when compared to the other.

Within the balanced two-class document-based sentiment classification the NBM performed the best (F1-score: 97.85%, Accuracy: 97.83%), when a label was set by averaging scores of sentences (by using Gain Ratio feature selection method with 3,000

features, TF-IDF, without transforming upper case letters to lower case, by removal of stop words, using combination of unigrams, bigrams and trigrams, without lemmatization). The SVM achieved the best F1-score with 96.28% (Accuracy: 96.27%) with the same pre-processing options as the NBM, but with transformation of upper case letters to lower case and combination of unigrams and bigrams. To investigate results in more detail, see Tables 5.6 and 5.7, which summarizes the results presented in Tables 8.1, 8.2, 8.3 and 8.4.

Similarly, we achieve the best results within three-class document-based sentiment classification with a balanced data set (3,000 documents, 1,000 documents per class). Again, the NBM performed better than the SVM. The NBM carried out the best F1-score with 77.76% (Accuracy: 79.85%) by using Gain Ratio feature selection method with 3,000 features, TF-IDF, transforming upper case letters to lower case, removal of stop words, using combination of unigrams and bigrams, and without lemmatization. The SVM achieved the best F1-score with 74.61% (Accuracy: 76.20%) with the same pre-processing options as the NBM, but with combination of unigrams, bigrams and trigrams, and without removal of stop words.

The classifiers perform better on the balanced data, particularly in the three-class scenario. The results indicate that the use of sentence-level granularity is a better option, if available, as this approach in most cases yields better results than the document-level one. Overall, the NBM classifier mostly outperforms the SVM (statistically significant at the 0.05 significance level but it is not so significant from a practical point of view). The SVM classifier outperforms the NBM classifier in accuracy on the imbalanced three-class data. However, accuracy is a less appropriate measure in imbalanced settings. In terms of pre-processing options, an option shared by all the best solutions is the one of not performing lemmatization. All but one or two of such options also use transformation to lower case and stop word removal. Impact of the other options seems to be mixed.

6 ESTIMATING THE PROPORTIONS OF POSITIVE, NEGATIVE AND NEUTRAL NEWS

In order to estimate the proportions of positive, negative and neutral news up to 2016, we obtained all the Slovenian news texts that were published between 1st of September 2007 and 31st of January 2016 from the selected web media resources, i.e. 256,567 documents (Finance - 132,986, Dnevnik - 52,417, Žurnal24 - 47,735, Rtvslo - 13,420 and 24ur - 10,009).

In this experiment, our goal was to estimate the sentiment of 246,140 remaining documents that were not labelled manually. We applied the NBM predictive model[40], which was proven as the best within the three-class document-based sentiment classification in terms of F1-score and time complexity (F1-score: 77.76% ± 3.13%, see Table 5.7) to estimate the proportions of positive, negative and neutral news within the specified web media resources, as shown in Table 6.1.

6.1 Results and Conclusions

We estimated the proportion of positive, negative and neutral news within the specified web media, as shown in Table 6.1.

When comparing the outcomes with the results in Table 3.5, we notice many similarities. For example, the estimation shows that Finance (with 37%) publishes the most positive news, while 24ur and Rtvslo publish the biggest proportion of negative news per web medium (24ur: 42%, Rtvslo: 37%). In general, we estimate that all web media produce much more negative than positive news, with the exception of Finance. Reis et al. (2015) investigated sentiment of the business news produced by four major global media

[40] Pre-processing options: Gain Ratio feature selection method with 3,000 features, TF-IDF, transforming upper case letters to lower case, removal of stop words, using combination of unigrams and bigrams, and without lemmatization.

Table 6.1: Estimated proportions (in %) of positive, negative and neutral news with political, business, economic and financial content published between 1st of September 2007 and 31st of January 2016 from five Slovenian web media resources (n = 256,567) with corresponding values from Table 3.5 inside the brackets

	Positive	Negative	Neutral	Number of documents
24ur	20.20 (16.26)	42.07 (39.99)	37.73 (43.75)	10,009
Dnevnik	20.05 (14.60)	33.37 (32.32)	46.57 (53.08)	52,417
Finance	36.95 (22.05)	19.82 (22.60)	43.23 (55.35)	132,986
Rtvslo	14.23 (13.76)	36.51 (37.06)	49.27 (49.18)	13,420
Žurnal24	15.12 (13.52)	33.90 (27.89)	50.98 (58.59)	47,735

corporations – The New York Times, BBC, Reuters and Dailymail. Their study showed that these media produce between 40-60% negative news.

There are many similarities in comparison with results in Table 3.4. For example, the estimation shows that Finance (with 37%) publishes the most positive news, while 24ur and Rtvslo publish the largest proportion of negative news per web medium (24ur: 42%, Rtvslo: 37%). In general, all web media produce much more negative than positive news, with the exception of Finance.

We created Table 6.2 to compare results with the Kovačič' (2012) study (see Table 1.1). We can see similar results with most web media, with the exception of Finance. There could be many reasons for the difference, such as the difference in the number of retrieved news and in the content of news (in our case, we analysed the entire content of news containing political, business, economic and financial news, where Kovačič analysed RSS and was not limited only on political, business, economic and financial news).

As a result, we obtained another machine annotated news corpus. Automatically

Table 6.2: Media tone of political, business, economic and financial news from five Slovenian web media that were published between October 2008 and December 2011

	Evaluation tone (count and % within media)			
Media	Positive	Neutral	Negative	Total
24ur	759 (19.6%)	1,469 (37.9%)	1,645 (42.5%)	3,873 (100%)
Dnevnik	4,631 (23.7%)	8,548 (43.8%)	6,327 (32.44%)	19,506 (100%)
Finance	12,083 (36.2%)	14,294 (42.9%)	6,975 (20.9%)	33,352 (100%)
Rtvslo	499 (14.3%)	1,597 (45.6%)	1,405 (40.1%)	3,501 (100%)
Žurnal24	2,289 (16.9%)	6,846 (50.5%)	4,415 (32.58%)	13,550 (100%)

sentiment annotated Slovenian news corpus AutoSentiNews 1.0[41] is a large corpus with >92 million words in 256,567 documents. The structure of the corpus is very similar to the manually annotated corpus SentiNews 1.0, which is presented in Table 3.3. The news corpus includes the following attributes: *nid*, *main_url*, *url*, *title*, *keywords*, *content*, *date*, *author* and *sentiment* (see Table 6.3). Unlike in Section 3.2, the label (positive, negative and neutral) is estimated with machine-learning techniques.

Table 6.3: Attributes, descriptions and data types within the automatically annotated Slovenian news corpus AutoSentiNews 1.0

Attribute	Description	Data type
nid	News ID	Integer (1 to 256,567)
main_url	Official URL of the web medium	String (5 web media)
url	URL of the news	String
title	Title of the news	String
keywords	Keywords of the news	String
content	Content of the news	String
date	Date of publishing the news	String (yyyy-mm-dd)
author	Author of the news	String
sentiment	Estimated sentiment	String (positive, negative, neutral)

[41] `http://hdl.handle.net/11356/1109`

7 MONITORING THE DYNAMICS OF SENTIMENT

More and more people express their opinions through web media. Stakeholders use communication channels to monitor public opinion, so they can react or even revert the public opinion.

Data scientists monitor the dynamics of sentiment to find characteristic patterns, for tracking sudden changes or trends, and to predict the sentiment dynamics, which tend to be associated with economic, financial, political, or other events and issues. A vast number of experiments on using large-scale Twitter data have been produced recently (Bermingham & Smeaton, 2009; Ceron, Curini & Iacus, 2015; Burnap, Gibson, Sloan, Southern & Williams, 2016). Some experiments on large-scale Twitter data show that models can achieve accuracy above 85% on directional sentiment prediction (Nguyen, Wu, Chan, Peng & Zhang, 2012).

In this chapter, we present how we monitored the dynamics of sentiment in our labelled corpora.

7.1 Monitoring the Dynamics of Sentiment Within Documents

We investigated the dynamics of sentiment within Manually sentiment annotated Slovenian (sentence-based) news corpus SentiNews 1.0. Our goal was to find any patterns regarding the labelled documents and the web media.

First, we normalized the length of the documents. Second, we defined a sentiment score for every 10% of the document length based on the linear interpolation between two averaged sentence-based sentiment scores that were closest to our measurement. Third, we averaged all the interpolated sentiment scores for every 10% of the length of a document for each web medium.

7.1.1 Results and Conclusions

We present the dynamics of the average sentiment and the associated standard deviation through documents, which were labelled as positive, negative and neutral, as shown in Figure 7.1. The horizontal axis of the graphs shows the document length from 0% to 100% (for every 10%), while the average sentiment and standard deviation, which follow the five-level Lickert scale, appear on the vertical axis. The dynamics of the average sentiment of the documents is presented with a coloured line, where the green line refers to the documents that are manually labelled as positive, the red to documents that were labelled as negative, and the grey to documents that were labelled as neutral.

An interesting trend can be observed. The documents that were labelled as positive, in general hold very positive sentiment at the beginning of a document, but steadily lose the intensity of the positive sentiment with the length of the document. A similar trend can be observed within the documents that were labelled as negative. They also carry a very strong negative sentiment at the beginning of the document, while the sentiment intensity also weakens steadily towards the end of a document. However, the average sentiment within the documents that were labelled as neutral is levelled out. There is a similar trend within each individual web medium (see Figures 7.2, 7.3 and 7.4). The observation about the dynamics of sentiment inside documents is potentially very important as it indicates the varied influence of different sections of the document to the overall sentiment. This insight suggests that by using only the starting parts of the news we might be able to detect sentiment more efficiently and effectively.

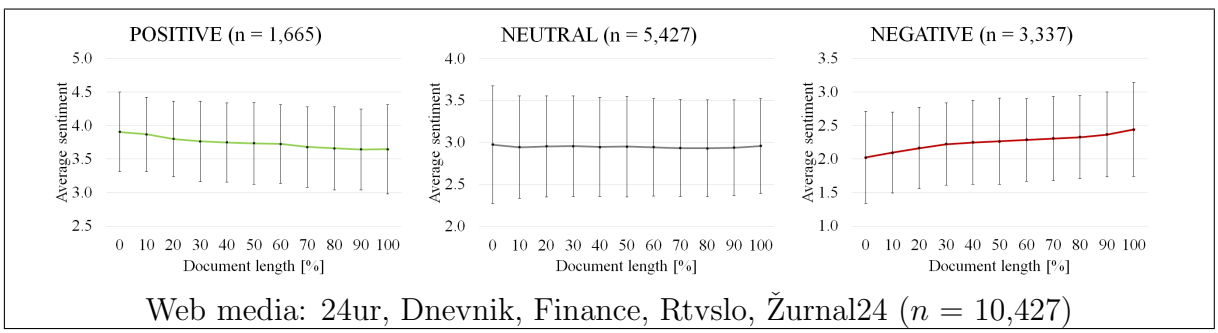

Figure 7.1: Dynamics of an average sentiment and standard deviation over the length (in %) of documents that were manually labelled as positive (left), neutral (middle) and negative (right) within the web media

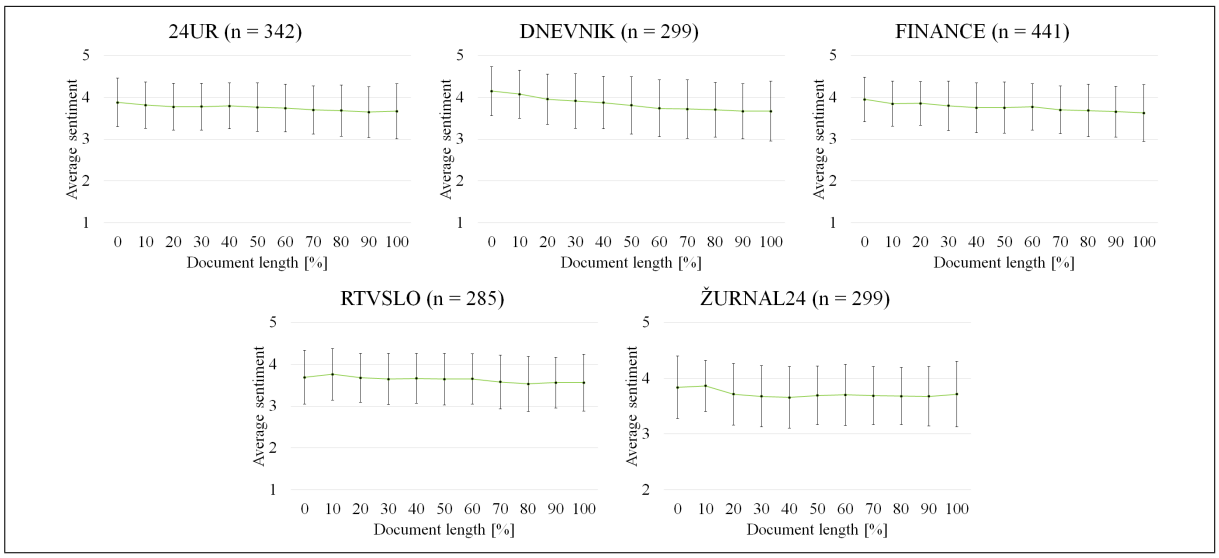

Figure 7.2: Dynamics of an average sentiment and standard deviation over the length (in %) of documents, which were manually labelled as positive in the web media

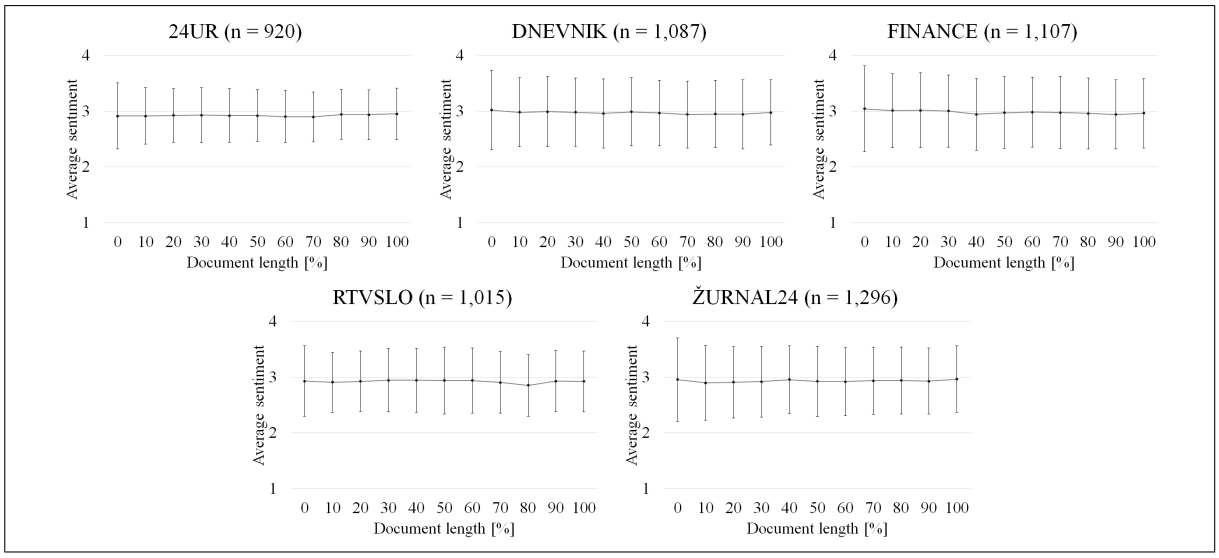

Figure 7.3: Dynamics of an average sentiment and standard deviation over the length (in %) of documents, which were manually labelled as neutral in the web media

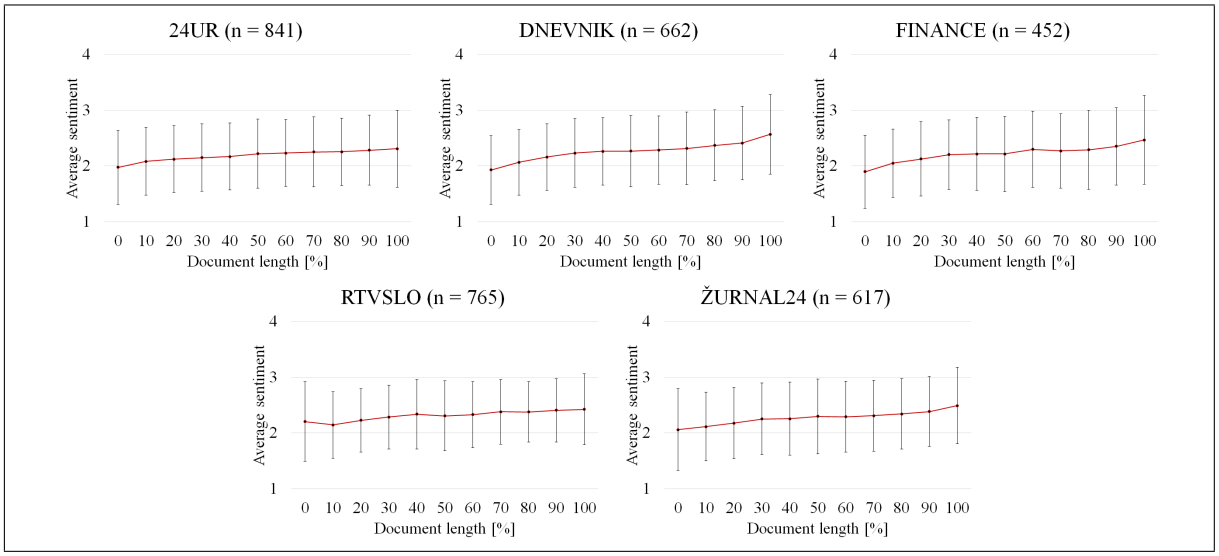

Figure 7.4: Dynamics of an average sentiment and standard deviation over the length (in %) of documents, which were manually labelled as negative in the web media

7.2 Monitoring the Dynamics of Sentiment Over Time

In this experiment, we were primarily interested in how the estimated sentiment proportions of positive, negative and neutral news changes over time within the individual web medium.

This study was derived from Automatically sentiment annotated Slovenian news corpus AutoSentiNews 1.0, as described in the last paragraph in Section 6.1.

7.2.1 Results and Conclusions

The results are presented in Figure 7.5, 7.6 and 7.7. The horizontal axis in the graphs shows different time periods between 1st of September 2007 and 31st of January 2016, while the vertical axes present the estimated sentiment proportions (in %) and total number of documents per time periods in the news and within the media.

The estimated sentiment proportions in the news alternate the most within 24ur, while in the others they seem to vary less. Also, Žurnal24's output suffers a dramatic fall in June 2014 (see Figure 7.5). To explain, in the middle of May 2014, the former owner of

Žurnal24 decided to close the company, and thereby to terminate the website. The new owners enabled further publiction of the news using this web medium. In general, most news were published in the spring, and the least in the summer, especially in August. Thus, the largest amount of news was published by Finance in March (12,488), and the least by 24ur in August (736). Finance produced the most positive news between 1st of September 2007 and 31st of January 2016, while 24ur produced the most negative news. Additional observations can be made on the time-associated data. Unsurprisingly, the number of news items that are published at weekends is much lower than on other days in a week, but they are obviously more positive (see Figure 7.7). The only exception is Finance, which publishes the most negative news at Saturdays. It may be that some negative financial news and events are deliberately made public late on Fridays after working hours of stock exchanges.

In addition, we used a popular tool (Mallet) to further explore our findings from the previous paragraph, and obtained some potentially interesting preliminary results. Our focus was to detect topics for Finance and all other media separately within weekends (see yellow dots in Figure 7.7). For topic retrieval, we defined two parameters: the number of expected topics (3) and the number of words per topic (10). The topics were extracted from the titles of AutoSentiNews 1.0, however, we removed the stop words for Slovene. Our results show that the news, which were published on Friday and Sunday within

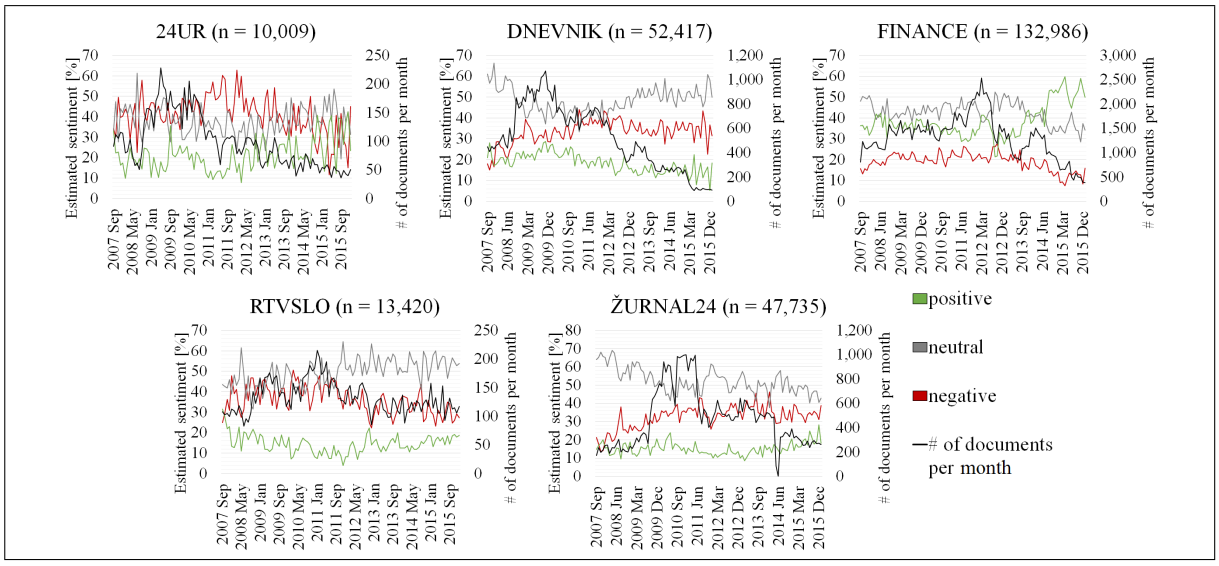

Figure 7.5: Estimated sentiment proportion (in %) in the news over time within 24ur (top left), Dnevnik (top middle), Finance (top right), Rtvslo (bottom left), Žurnal24 (bottom right)

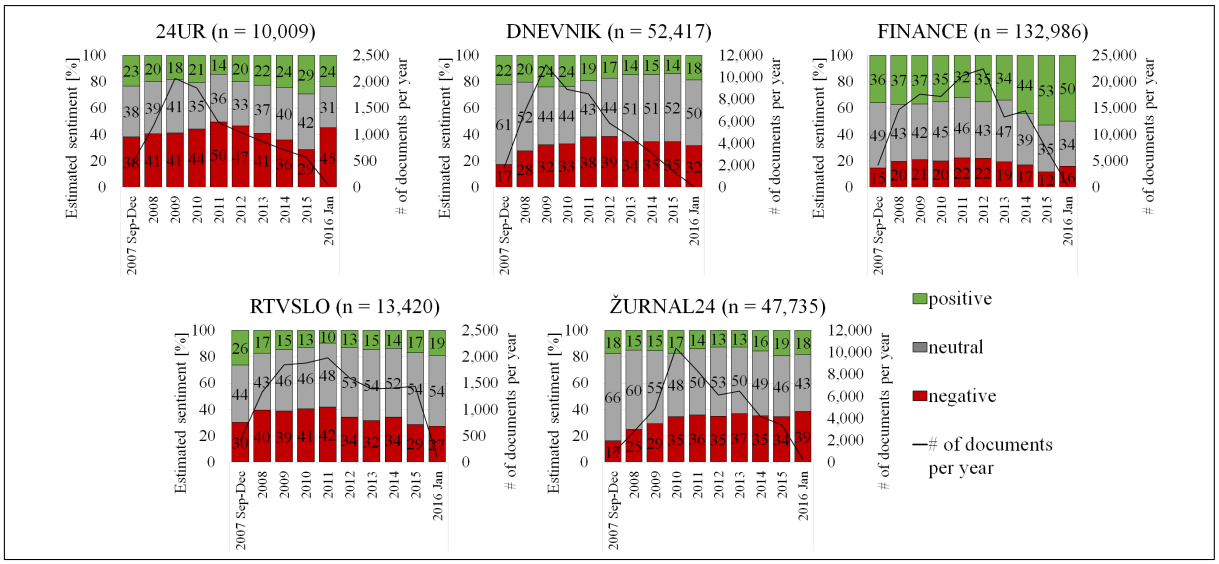

Figure 7.6: Estimated sentiment proportion (in %) in the news over years within 24ur (top left), Dnevnik (top middle), Finance (top right), Rtvslo (bottom left), Žurnal24 (bottom right)

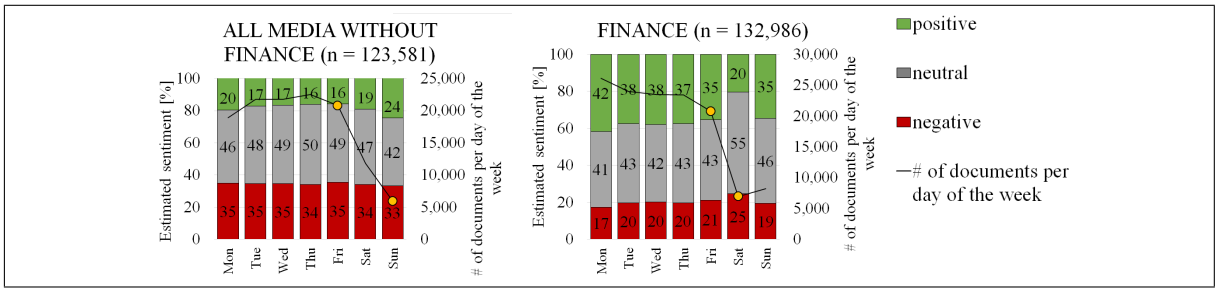

Figure 7.7: Estimated sentiment proportion (in %) in the news over days within all media without Finance (left), and in the news published in Finance (right)

all media without Finance include topics with subjects, organizations and institutions related to stock exchanges and banks. Moreover, we noticed a positive trend in the news published on Sunday within all media without Finance that contain a topic, which indicates a connection between the economic crisis and hope (topic words: government, Slovenia, companies, help, crisis, etc.). Similarly, the news that were published on Friday and Sunday within Finance contain topics related to financial reports, institutions, stock exchanges, banks and cash flows. The biggest difference is in the news published on Saturday, which include two topics. The first focuses on the foreign economy (topic words: USA, government, Obama, against, growth, forecast, etc.), and the second on the domestic economy (topic words: Slovenia, Pahor, Janša, Janković, Türk, banks, rush, sales, etc.).

7.3 Monitoring the Dynamics of Topic-sentiment

By monitoring the dynamics of topic-sentiment, such as people, places, companies, events, etc., we aim to explore their sentiment reputation in the web media over time. With a domain knowledge of social sciences, we can then relate the observed trends with actual events in the past, predict their sentiment reputation and possibly even events in the future.

This study was also derived from Automatically sentiment annotated Slovenian news corpus AutoSentiNews 1.0 (see the last paragraph in Section 6.1).

7.3.1 Results and Conclusions

We present the dynamics of the estimated sentiment proportion in the news with political, business, economic and financial content of the current president of the Republic of Slovenia, Borut Pahor, between 1st of September 2007 and 31st of January 2016, as shown in Figure 7.8. The horizontal axis of the graph shows the months, the period between 1st of September 2007 and 31st of January 2016, while the vertical axis indicates the estimated sentiment proportion (in %) in the news. The president's name appeared 11,648 times within the annotated corpus (AutoSentiNews 1.0) with 256,567 documents.

In 2004, Pahor was elected as member of the European Parliament. Following the victory at the parliamentary election of the Social Democrats in September 2008, Pahor was appointed as Prime Minister in November 2008. This phenomenon can be observed in

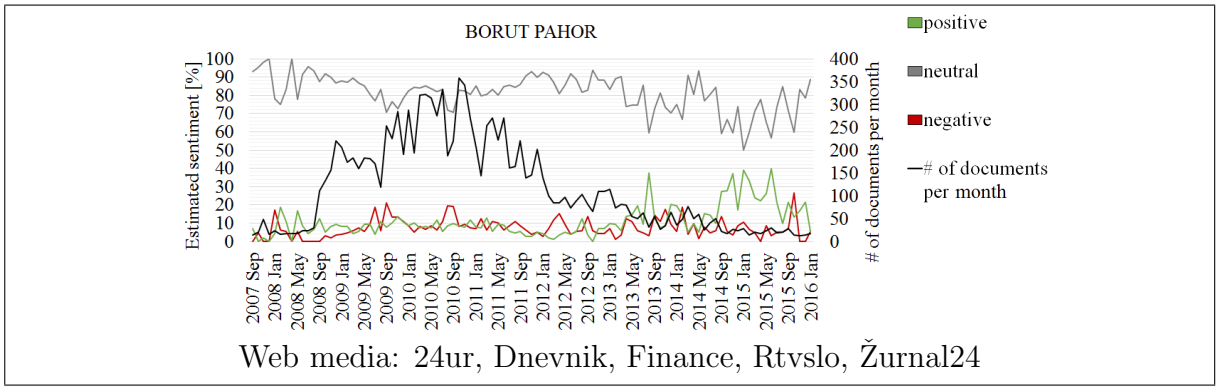

Figure 7.8: Estimated sentiment proportion (in %) in the news over time of the current Slovenian president in the web media

Figure 7.8, since his name appears considerably more often than before. In October 2010, Pahor met with Croatian Prime Minister Jadranka Kosor on the arbitration agreement between Slovenia and Croatia, which was ratified in April 2010. In September 2011, his government lost a confidence vote in the middle of an economic crisis and political tensions. He was replaced by Janez Janša in February 2012. During this period, when he was Prime Minister of Slovenia, he was mentioned much more often in the news. Pahor has been the president of Slovenia since December 2012, after convincing victory at the presidential election in the second round.

If we focus on the news that was estimated as positive, we can notice four peaks. The first can be observed in August 2013, when the president hosted a number of world leaders and businesspeople at the main economic and business conference in Slovenia. The second was in November 2014, which was the result of strengthening economic relations with Germany and China; he also actively participated in some charitable campaigns, and hosted an event at the 25th anniversary of children's rights. In January 2015 he hosted the president of Qatar, and attended several ceremonies and charity events. The last peak can be observed in June 2015, when his commitments to open markets for foreign capital and to an investor-friendly investment climate lead to events, which connected foreign businesspeople to representatives of leading companies in Slovenia.

However, when dealing with the news texts with political, business, economic and financial content that were estimated as negative, we can see that the dynamics of the estimated sentiment proportion is rather small. It is estimated that the largest proportion of negative news, where the current Slovenian president was mentioned, was published in October 2015. A more detailed view shows that the news texts were mainly dealing with issues in the migration crisis, and an affair involving a former member of the president's cabinet.

Overall, the studies have shown that developed tools for monitoring the sentiment dynamics are potentially helpful to find patterns and relations associated with specific people, places, companies, events, etc., but only to a certain extent. With this in mind, there is a legitimate challenge in the future to predict trends and events accurately.

7.4 Monitoring the Dynamics of Sentiment of Authors

At last, we investigate the dynamics of sentiment of authors of the news, in order to find characteristic patterns of their writing. A study of their writing styles could for instance show whether they tend to write more positive or negative news, and how their styles evolve through time. Once again, this study was also derived from Automatically sentiment annotated Slovenian news corpus AutoSentiNews 1.0 (see the last paragraph in Section 6.1).

7.4.1 Results and Conclusions

As an illustration, we explore the writing style of one Dnevnik journalist in Figure 7.9. He is one of the few authors who was regularly writing for the same web medium between 1st of September 2007 and 31st of January 2016, and published the largest number of news (3,146) within the web medium. The horizontal axis, within the left graph, shows the months, the period between 1st of September 2007 and 31st of January 2016, while the estimated sentiment proportion (in %) in the news is denoted on the vertical axis.

If we focus on the right graph in Figure 7.9, the horizontal axis shows the document length from 0% to 100% (for every 10%) while the estimated average sentiment and standard deviation, which follow the five-level Lickert scale, appear on the vertical axis. We used a heavy coloured line in order to present its dynamics. Since the estimated average sentiment is almost completely levelled out within the document length, we could not find

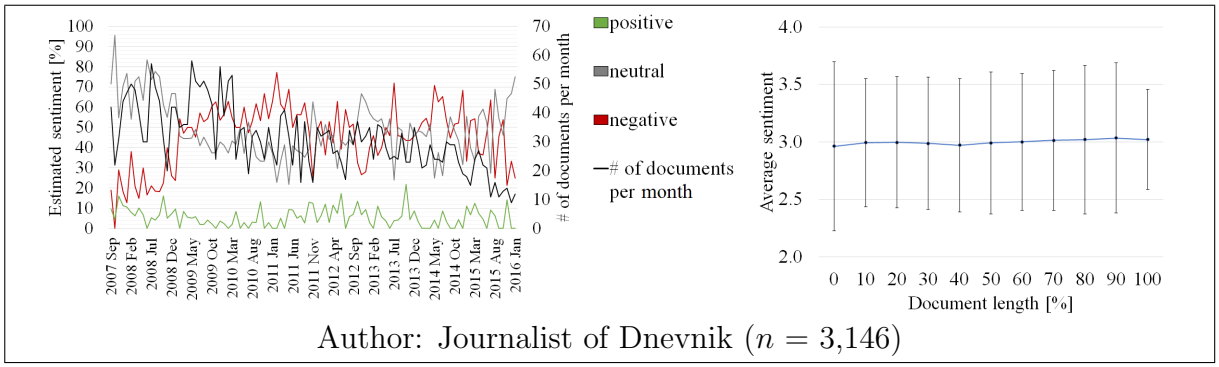

Figure 7.9: Estimated sentiment proportion (in %) in the news over time (left), as well as the estimated average sentiment and standard deviation over the length (in %) of the documents (right), which were written by the Dnevnik journalist

any specific writing style (within all of 3,146 documents) for the journalist of Dnevnik. In general, by averaging, points of interest in individual documents are lost.

Clearly, the author tends to write more negative news rather than positive. Moreover, the estimation shows that the journalist was at first publishing mainly the neutral news, and then the proportion of neutral news declined substantially, until it dropped off in February 2011. After a series of huge falls and jumps (from April 2011 to October 2015), the estimated proportion of neutral news ended with a rapid growth.

However, over the years of writing for the same web medium, the estimated proportion of negative news increased significantly, and reached its peak in February 2011. Then it declined in March and suffered a spectacular fall in November within the same year, but then made a significant recovery in May 2012. It dropped out again in November within the same year. After rising sharply during July 2013, it suffered another dramatic fall in August, and slightly recovered in March 2014. This was followed by a series of massive falls and jumps (from May 2014 to September 2015), and it ended with a considerable downturn. However, it seems that the estimated proportion of positive news does not vary to such a degree as compared to the estimated proportions of either negative or neutral news.

8 TESTING HYPOTHESES

Within this chapter, we present the methodology used to address hypotheses and obtained results.

8.1 Testing Hypothesis 1

8.1.1 Methodology

There are many existing methods for sentiment based document classification, however, within the H_1 hypothesis, we evaluated the following classifiers: KNN, NBM, SVM (SVM-poly and SVM-lin), RF, C4.5, DT, SLR and VP for two-class and three class document-based sentiment classification of the Slovenian news texts. We were interested in which classifier performs best using CV technique. The performance of classification was based on classification performance (accuracy and F1-score) and computational time consumption, however, the NBM and the SVM outperformed other classifiers (see Section 5.1). For this reason, we tested hypothesis H_1 only for the NBM and the SVM-poly. Also, we were interested in which pre-processing setting achieves the best result (see Section 5.3). Comparisons between the applied classifiers was performed using paired t-test (Student, 1908) and Wilcoxon (1945) signed-ranks test at the significance level of 5%.

If we define null and alternative hypotheses, we have H_0 : *There is no difference in classification performance between the NBM classifier and the SVM-poly classifier.* H_1 : *There is a difference in classification performance between the NBM classifier and the SVM-poly classifier.* H_0 denotes the null hypothesis and H_1 the alternative hypothesis. We test the hypotheses with the commonly used 0.05 significance level.

According to t-distribution with $n - 1$ degrees of freedom and critical value of the rejection region $t_{n-1,\alpha} = 2.10$, we calculate Student's statistics T (Student, 1908) as:

$$T = \frac{\bar{d}}{SE(\bar{d})} \tag{8.1}$$

where n is the sample size. Based on the NBM classifier x, the SVM-poly classifier y and the same selection of pre-processing settings i on each pair, the difference d^i is calculated as $d^i = x^i - y^i$. $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$ is the standard error of the mean difference, $\bar{d}$ is the mean difference and s_d is the standard deviation of the differences.

According to standard normal distribution critical value of the rejection region $z_\alpha = 1.96$, we calculate Wilcoxon's statistics W (Wilcoxon, 1945) as:

$$W = \frac{T - \frac{n\cdot(n+1)}{4}}{\sqrt{\frac{n\cdot(n+1)\cdot 2n+1}{24}}} \tag{8.2}$$

where n is the sample size and $T = min\{R_+, R_-\}$. Sum of (positive) ranks R_+ is computed as $R_+ = \sum_{i=1}^{n} Rank(d^i)$, where $d^i>0$, and sum of (negative) ranks R_- is calculated as $R_- = \sum_{i=1}^{n} Rank(d^i)$, where $d^i<0$. Based on the NBM classifier x, the SVM-poly classifier y and the same selection of pre-processing settings i on each pair, the difference d^i is calculated as $d^i = x^i - y^i$.

8.1.2 Results H_1

Our initial experiments have shown that the NBM and the SVM outperform other classifiers (see Section 5.1). The whole Chapter 5 explores the hypothesis H_1 in iterative stages. It also describes the methodology, examines its complexity, and discovers how classification performance evolves through specific stages.

The experiments, that followed, have shown that the NBM achieves the best F1-score within the two-class (97.85%) and three-class (77.76%) document-based (based on average sentiment scores of sentences) sentiment classification, while the SVM achieves 96.28% (within the two-class) and 74.61% (within the three-class) document-based sentiment classification (see Section 5.3). Within the two-class document-level sentiment classification the NBM achieved the best F1-score (97.85%) by using TF-IDF, without transforming upper case letters to lower case, by removal of stop words, using combination of unigrams, bigrams and trigrams, without lemmatization. Within the three-class

document-level sentiment classification the NBM achieved the best F1-score (77.76%) by using TF-IDF, transforming upper case letters to lower case, by without removal of stop words, using combination of unigrams, bigrams and trigrams, without lemmatization. In general, the NBM and the SVM perform (significantly) better using TF-IDF, combination of unigrams, bigrams and trigrams, when the feature vector size is appropriate, and when lemmatization is not included in the pre-processing settings.

To illustrate the process of testing hypothesis H_1, in Table 8.1, we apply the Wilcoxon signed-ranks test for F1-score within balanced two-class document-level sentiment classification (n = 2,000) based on average scores of documents for the NBM and the SVM-poly using pre-processing settings with $ID = 2$ by applying 5 times 10-fold CV. At first, we calculate the differences d^i of F1-scores between the NBM and the SVM-poly, that is the differences between $n = 50$ different values in columns 7 and 9 (in Table 8.1). We compute their ranks, followed by sum of (positive) ranks $R_+ = 1,230$, $R_- = 45$ and $T = 45$. Since, the computed test statistic $|W = -5.72|$ (see Equation 8.2) is greater than $z_\alpha = 1.96$, we reject the null hypothesis. In Table 8.2, we apply the paired t-test and Wilcoxon signed-ranks test for accuracy and F1-score within balanced three-class document-level sentiment classification (n = 3,000) based on average scores of documents for the NBM and the SVM-poly using various pre-processing settings by applying 5 times 10-fold CV (see Table 8.2). Table 8.3 presents the paired t-test and Wilcoxon signed-ranks test for accuracy and F1-score within balanced two-class document-level sentiment classification (n = 2,000) based on average scores of sentences for the NBM and the SVM-poly using various pre-processing settings by applying 5 times 10-fold CV (see Table 8.3). In Table 8.4, we apply the paired t-test and Wilcoxon signed-ranks test for accuracy and F1-score within balanced three-class document-level sentiment classification (n = 3,000) based on average scores of sentences for the NBM and the SVM-poly using various pre-processing settings by applying 5 times 10-fold CV (see Table 8.4).

Overall, the NBM classifier mostly outperforms the SVM (statistically significant (paired t-test and Wilcoxon signed-ranks test) at the 0.05 significance level). The SVM classifier outperforms the NBM classifier in accuracy and F1-score, when using specific pre-processing settings (combination of TF-IDF and trigrams or combination of TF and unigrams, bigrams and trigrams) within balanced two-class document-level sentiment classification (see Table 8.1). In terms of pre-processing options, an option shared by

all the best solutions is the one of not performing lemmatization. All but one or two of such options also use transformation to lower case and stop word removal. Impact of the other options seems to be mixed.

The series of experiments described in this Section (see Tables 8.1, 8.2, 8.3 and 8.4) and in Chapter 5 confirm the hypothesis H_1, which states that appropriate selection of supervised machine learning classifier and pre-processing settings can improve the classification performance.

8.2 Testing Hypothesis 2

8.2.1 Methodology

The Slovenian news texts were manually annotated as positive, negative and neutral on three levels of granularity, e.g., document level, paragraph level and sentence level. We explored, whether granulation of a document to smaller segments can improve classification performance.

Both classifiers, the NBM and the SVM, performed significantly better, when the average sentiment scores of sentences were used to determine the sentiment of a document. The results given in Tables 8.5, 8.6, 8.7 and 8.8 indicate that the hypothesis H_2, whether the granulation of documents to smaller segments (e.g. sentences) improves the classification performance, can be confirmed.

The NBM and the SVM outperformed other classifiers in terms of classification performance (accuracy and F1-score) and computational time consumption. For this reason, we tested only the NBM classifier and the SVM-poly classifier based on annotations at the document-level and sentence-level granularity. In a similar way to the H_1 hypothesis, comparisons between the applied classifiers was performed using paired t-test and Wilcoxon signed-ranks test at the significance level of 5%.

If we define null and alternative hypotheses, we have H_0 : *There is no difference in classification performance (accuracy and F1-score) between the sentence-level granulation and document-level granulation.* H_2 : *There is a difference in classification performance (accuracy and F1-score) between the sentence-level granulation and document-level granulation.* H_0 denotes the null hypothesis and H_2 the alternative hypothesis.

Table 8.1: Paired t-test and Wilcoxon signed-ranks test for accuracy and F1-score within balanced two-class document-level sentiment classification (n = 2,000) based on average scores of documents for the NBM and the SVM-poly using various pre-processing settings by applying 5 times 10-fold CV

Document-level based on the average scores of documents
(based on annotations at the document-level granularity)
(# documents: 2,000, feature selection method: Gain Ratio, # features: 3,000)
(x sign - included in the experiment)

	Pre-processing settings				NBM		SVM-poly	
ID	TF & TF-IDF	lower case	stop words	N-grams (1, 2, 3)	Accuracy	F1-score	Accuracy	F1-score
1	TF			1	87.35 ± 2.77	87.52 ± 2.66	86.81 ± 2.24	86.83 ± 2.15
2	TF-IDF			1	86.52 ± 2.77	88.33 ± 2.67*$^{\triangle}$	85.69 ± 2.36	85.63 ± 2.37
3	TF			2	87.79 ± 2.74*$^{\triangle}$	87.76 ± 2.73*$^{\triangle}$	85.67 ± 2.57	85.35 ± 2.72
4	TF-IDF			2	84.66 ± 2.53$^{\triangle}$	87.00 ± 2.41$^{\triangle}$	85.81 ± 2.27	85.51 ± 2.46
5	TF			3	75.07 ± 3.24	76.93 ± 2.86*$^{\triangle}$	74.53 ± 2.47	71.81 ± 3.21
6	TF-IDF			3	58.70 ± 3.62	69.80 ± 3.60	75.15 ± 2.50*$^{\triangle}$	72.55 ± 3.27$^{\triangle}$
7	TF			1+2	89.52 ± 2.28	89.44 ± 2.29	90.35 ± 1.76	90.31 ± 1.81
8	TF-IDF			1+2	92.22 ± 2.10*$^{\triangle}$	92.42 ± 2.03*$^{\triangle}$	89.64 ± 1.96	89.53 ± 2.01
9	TF			1+2+3	89.60 ± 2.53	89.51 ± 2.55	91.50 ± 2.03*$^{\triangle}$	91.49 ± 2.04*$^{\triangle}$
10	TF-IDF			1+2+3	92.80 ± 1.92*$^{\triangle}$	93.03 ± 1.95*$^{\triangle}$	90.60 ± 1.85	90.53 ± 1.91
11	TF	x		1	87.29 ± 2.58	87.54 ± 2.49	87.01 ± 2.05	86.96 ± 2.13
12	TF-IDF	x		1	85.87 ± 2.63	88.80 ± 2.22$^{\triangle}$	87.82 ± 2.29$^{\triangle}$	87.71 ± 2.32
13	TF	x		2	88.14 ± 2.46*$^{\triangle}$	88.09 ± 2.47*$^{\triangle}$	85.55 ± 2.50	85.32 ± 2.64
14	TF-IDF	x		2	86.62 ± 2.24	88.34 ± 2.01*$^{\triangle}$	86.83 ± 2.25	86.51 ± 2.33
15	TF	x		3	77.59 ± 2.77*$^{\triangle}$	78.93 ± 2.64*$^{\triangle}$	74.83 ± 3.14	72.59 ± 3.90
16	TF-IDF	x		3	61.26 ± 3.18	72.03 ± 2.93	74.84 ± 2.95*$^{\triangle}$	72.54 ± 3.64
17	TF	x		1+2	89.50 ± 2.35	89.43 ± 2.40	90.28 ± 2.00	90.20 ± 2.07
18	TF-IDF	x		1+2	92.40 ± 1.75*$^{\triangle}$	92.63 ± 1.70*$^{\triangle}$	90.06 ± 2.11	89.93 ± 2.24
19	TF	x		1+2+3	90.41 ± 2.39	90.34 ± 2.43	91.90 ± 1.64$^{\triangle}$	91.89 ± 1.66$^{\triangle}$
20	TF-IDF	x		1+2+3	92.89 ± 1.65	93.12 ± 1.65$^{\triangle}$	92.55 ± 1.64	92.48 ± 1.69
21	TF		x	1	86.85 ± 2.82	87.16 ± 2.67	86.89 ± 2.17	86.89 ± 2.14
22	TF-IDF		x	1	86.44 ± 2.90	88.39 ± 2.75*$^{\triangle}$	85.83 ± 2.39	85.69 ± 2.42
23	TF		x	2	87.79 ± 2.74*$^{\triangle}$	87.76 ± 2.73*$^{\triangle}$	85.67 ± 2.57	85.35 ± 2.72
24	TF-IDF		x	2	84.66 ± 2.53	87.00 ± 2.41$^{\triangle}$	85.81 ± 2.27$^{\triangle}$	85.51 ± 2.46
25	TF		x	3	75.07 ± 3.24	76.93 ± 2.86*$^{\triangle}$	74.53 ± 2.47	71.81 ± 3.21
26	TF-IDF		x	3	58.70 ± 3.62	69.80 ± 3.60	75.15 ± 2.50*$^{\triangle}$	72.55 ± 3.27$^{\triangle}$
27	TF		x	1+2	89.50 ± 2.29	89.46 ± 2.28	89.80 ± 1.69	89.82 ± 1.71
28	TF-IDF		x	1+2	92.12 ± 2.19*$^{\triangle}$	92.29 ± 2.08*$^{\triangle}$	89.98 ± 2.17	89.85 ± 2.23
29	TF		x	1+2+3	89.96 ± 2.58	89.99 ± 2.55	91.24 ± 1.84$^{\triangle}$	91.27 ± 1.82$^{\triangle}$
30	TF-IDF		x	1+2+3	92.78 ± 1.86*$^{\triangle}$	93.00 ± 1.87*$^{\triangle}$	90.55 ± 1.84	90.43 ± 1.91
31	TF	x	x	1	87.66 ± 2.33	88.02 ± 2.19$^{\triangle}$	86.92 ± 2.04	86.88 ± 2.07
32	TF-IDF	x	x	1	85.49 ± 2.50	88.60 ± 2.14	87.91 ± 2.28*$^{\triangle}$	87.78 ± 2.34
33	TF	x	x	2	88.14 ± 2.46*$^{\triangle}$	88.09 ± 2.47*$^{\triangle}$	85.55 ± 2.50	85.32 ± 2.64
34	TF-IDF	x	x	2	86.62 ± 2.24	88.34 ± 2.01*$^{\triangle}$	86.83 ± 2.25	86.51 ± 2.33
35	TF	x	x	3	77.59 ± 2.77*$^{\triangle}$	78.93 ± 2.64*$^{\triangle}$	74.83 ± 3.14	72.59 ± 3.90
36	TF-IDF	x	x	3	61.26 ± 3.18	72.03 ± 2.93	74.84 ± 2.95*$^{\triangle}$	72.54 ± 3.64
37	TF	x	x	1+2	89.38 ± 2.32	89.37 ± 2.34	90.17 ± 1.67	90.15 ± 1.74
38	TF-IDF	x	x	1+2	92.46 ± 1.78*$^{\triangle}$	92.75 ± 1.69*$^{\triangle}$	90.18 ± 1.97	90.07 ± 2.05
39	TF	x	x	1+2+3	90.68 ± 2.41	90.74 ± 2.39	92.06 ± 1.81$^{\triangle}$	92.04 ± 1.78$^{\triangle}$
40	TF-IDF	x	x	1+2+3	92.87 ± 1.54	93.12 ± 1.54$^{\triangle}$	92.43 ± 1.76	92.34 ± 1.83

* statistically significant (paired t-test) at the 0.05 significance level
$^{\triangle}$ statistically significant (Wilcoxon signed-ranks test) at the 0.05 significance level

Table 8.2: Paired t-test and Wilcoxon signed-ranks test for accuracy and F1-score within balanced three-class document-level sentiment classification (n = 3,000) based on average scores of documents for the NBM and the SVM-poly using various pre-processing settings by applying 5 times 10-fold CV

Document-level based on the average scores of documents
(based on annotations at the document-level granularity)
(# documents: 3,000, feature selection method: Gain Ratio, # features: 3,000)
(x sign - included in the experiment)

	Pre-processing settings				NBM		SVM-poly	
ID	TF & TF-IDF	lower case	stop words	N-grams (1, 2, 3)	Accuracy	F1-score	Accuracy	F1-score
1	TF			1	65.62 ± 2.33$^{\triangle}$	63.73 ± 3.82	63.57 ± 2.57	63.30 ± 3.40
2	TF-IDF			1	67.92 ± 2.51*$^{\triangle}$	66.30 ± 4.07$^{\triangle}$	63.47 ± 2.69	62.69 ± 3.73
3	TF			2	68.05 ± 2.61*$^{\triangle}$	67.65 ± 2.86*$^{\triangle}$	63.07 ± 2.34	64.24 ± 2.83
4	TF-IDF			2	69.67 ± 2.26*$^{\triangle}$	69.94 ± 2.82*$^{\triangle}$	64.25 ± 2.54	64.99 ± 3.04
5	TF			3	57.92 ± 2.75	59.28 ± 3.58*$^{\triangle}$	57.59 ± 2.84	54.22 ± 4.88
6	TF-IDF			3	55.95 ± 2.68	56.51 ± 3.07$^{\triangle}$	57.69 ± 3.03$^{\triangle}$	54.52 ± 4.95
7	TF			1+2	67.96 ± 2.07$^{\triangle}$	65.70 ± 3.70	66.61 ± 2.93	65.74 ± 3.67
8	TF-IDF			1+2	72.64 ± 2.35*$^{\triangle}$	72.07 ± 3.49*$^{\triangle}$	67.05 ± 2.70	66.34 ± 3.66
9	TF			1+2+3	66.83 ± 2.12$^{\triangle}$	64.61 ± 3.75	64.69 ± 3.00	64.34 ± 3.66
10	TF-IDF			1+2+3	70.21 ± 2.36*$^{\triangle}$	68.77 ± 3.84*$^{\triangle}$	64.71 ± 3.16	64.67 ± 3.92
11	TF	x		1	64.97 ± 2.21	62.89 ± 3.80	64.92 ± 2.86	64.97 ± 3.52$^{\triangle}$
12	TF-IDF	x		1	67.40 ± 2.27*$^{\triangle}$	65.73 ± 3.24$^{\triangle}$	63.68 ± 2.62	63.65 ± 3.68
13	TF	x		2	67.78 ± 2.45*$^{\triangle}$	67.16 ± 3.50$^{\triangle}$	64.15 ± 2.35	64.40 ± 3.53
14	TF-IDF	x		2	70.86 ± 2.24*$^{\triangle}$	70.83 ± 3.05*$^{\triangle}$	64.07 ± 2.48	64.17 ± 3.63
15	TF	x		3	58.57 ± 2.65*$^{\triangle}$	58.80 ± 3.52*$^{\triangle}$	55.90 ± 2.80	53.98 ± 4.65
16	TF-IDF	x		3	57.03 ± 2.52	57.61 ± 3.53$^{\triangle}$	56.25 ± 2.74	54.11 ± 4.84
17	TF	x		1+2	67.25 ± 2.13	65.29 ± 3.45	67.43 ± 2.71	67.77 ± 3.50$^{\triangle}$
18	TF-IDF	x		1+2	73.09 ± 2.28*$^{\triangle}$	72.77 ± 3.44*$^{\triangle}$	67.94 ± 2.57	67.71 ± 3.18
19	TF	x		1+2+3	66.52 ± 2.18$^{\triangle}$	63.86 ± 3.77	65.25 ± 1.96	64.66 ± 2.75
20	TF-IDF	x		1+2+3	70.31 ± 2.02$^{\triangle}$*	69.13 ± 3.32*$^{\triangle}$	64.93 ± 2.49	64.70 ± 3.13
21	TF		x	1	65.34 ± 2.52*$^{\triangle}$	63.35 ± 3.80	62.25 ± 2.78	62.26 ± 3.54
22	TF-IDF		x	1	68.06 ± 2.80*$^{\triangle}$	66.99 ± 4.30*$^{\triangle}$	63.12 ± 2.26	62.40 ± 3.60
23	TF		x	2	68.05 ± 2.61*$^{\triangle}$	67.65 ± 2.86*$^{\triangle}$	63.07 ± 2.34	64.24 ± 2.83
24	TF-IDF		x	2	69.67 ± 2.26*$^{\triangle}$	69.94 ± 2.82*$^{\triangle}$	64.25 ± 2.54	64.99 ± 3.04
25	TF		x	3	57.92 ± 2.75	59.28 ± 3.58*$^{\triangle}$	57.59 ± 2.84	54.22 ± 4.88
26	TF-IDF		x	3	55.95 ± 2.68	56.51 ± 3.07$^{\triangle}$	57.69 ± 3.03$^{\triangle}$	54.52 ± 4.95
27	TF		x	1+2	68.34 ± 2.16	66.68 ± 3.74	67.36 ± 2.74	66.41 ± 3.60
28	TF-IDF		x	1+2	72.80 ± 2.34*$^{\triangle}$	71.83 ± 3.32*$^{\triangle}$	67.83 ± 2.65	67.18 ± 3.45
29	TF		x	1+2+3	67.15 ± 2.01$^{\triangle}$	64.92 ± 3.67	65.85 ± 3.28	65.79 ± 4.21
30	TF-IDF		x	1+2+3	70.61 ± 2.27*$^{\triangle}$	69.62 ± 3.47*$^{\triangle}$	65.11 ± 2.93	65.05 ± 3.92
31	TF	x	x	1	65.07 ± 2.17*	63.04 ± 3.78	64.71 ± 2.62	64.87 ± 3.68$^{\triangle}$
32	TF-IDF	x	x	1	67.26 ± 2.25*$^{\triangle}$	65.57 ± 3.20	64.27 ± 2.44	64.52 ± 3.50
33	TF	x	x	2	67.78 ± 2.45*$^{\triangle}$	67.16 ± 3.50$^{\triangle}$	64.15 ± 2.35	64.40 ± 3.53
34	TF-IDF	x	x	2	70.86 ± 2.24*$^{\triangle}$	70.83 ± 3.05*$^{\triangle}$	64.07 ± 2.48	64.17 ± 3.63
35	TF	x	x	3	58.57 ± 2.65*$^{\triangle}$	58.80 ± 3.52*$^{\triangle}$	55.90 ± 2.80	53.98 ± 4.65
36	TF-IDF	x	x	3	57.03 ± 2.52	57.61 ± 3.53$^{\triangle}$	56.25 ± 2.74	54.11 ± 4.84
37	TF	x	x	1+2	67.37 ± 2.14	65.50 ± 3.57	66.99 ± 2.51	67.24 ± 3.48$^{\triangle}$
38	TF-IDF	x	x	1+2	71.40 ± 2.00*$^{\triangle}$	70.93 ± 3.11*$^{\triangle}$	66.97 ± 2.64	67.42 ± 3.69
39	TF	x	x	1+2+3	66.82 ± 2.07*$^{\triangle}$	64.53 ± 3.79	64.78 ± 2.10	64.54 ± 3.22
40	TF-IDF	x	x	1+2+3	70.31 ± 2.19*$^{\triangle}$	69.58 ± 3.68*$^{\triangle}$	65.60 ± 2.81	65.03 ± 3.23

* statistically significant (paired t-test) at the 0.05 significance level
$^{\triangle}$ statistically significant (Wilcoxon signed-ranks test) at the 0.05 significance level

Table 8.3: Paired t-test and Wilcoxon signed-ranks test for accuracy and F1-score within balanced two-class document-level sentiment classification (n = 2,000) based on average scores of sentences for the NBM and the SVM-poly using various pre-processing settings by applying 5 times 10-fold CV

Document-level based on the average scores of sentences
(based on annotations at the sentence-level granularity)
(# documents: 2,000, feature selection method: Gain Ratio, # features: 3,000)
(x sign - included in the experiment)

	Pre-processing settings				NBM		SVM-poly	
ID	TF & TF-IDF	lower case	stop words	N-grams (1, 2, 3)	Accuracy	F1-score	Accuracy	F1-score
1	TF			1	92.89 ± 1.40*△	93.08 ± 1.32*△	91.35 ± 1.70	91.33 ± 1.72
2	TF-IDF			1	92.83 ± 1.63△	93.02 ± 1.55*△	91.12 ± 1.93	91.04 ± 1.97
3	TF			2	93.38 ± 1.66*△	93.42 ± 1.61*△	90.25 ± 1.75	90.02 ± 1.81
4	TF-IDF			2	93.32 ± 2.06*△	93.45 ± 1.99*△	89.79 ± 1.75	89.50 ± 1.83
5	TF			3	80.00 ± 2.90	81.36 ± 2.47*△	79.55 ± 2.91	77.13 ± 3.61
6	TF-IDF			3	79.74 ± 3.05	81.30 ± 2.55*△	79.21 ± 2.51	76.54 ± 3.33
7	TF			1+2	95.33 ± 1.14	95.36 ± 1.11	95.01 ± 1.38	94.98 ± 1.40
8	TF-IDF			1+2	96.32 ± 1.07*△	96.38 ± 1.02*△	94.14 ± 1.48	94.08 ± 1.52
9	TF			1+2+3	96.22 ± 1.04*△	96.22 ± 1.04△	95.42 ± 1.27	95.41 ± 1.28
10	TF-IDF			1+2+3	97.78 ± 0.97*△	97.80 ± 0.96*△	95.87 ± 1.29	95.82 ± 1.30
11	TF	X		1	92.82 ± 1.16*△	92.99 ± 1.11*△	91.23 ± 1.67	91.26 ± 1.67
12	TF-IDF	X		1	93.75 ± 1.26*△	93.90 ± 1.21*△	91.08 ± 2.04	91.08 ± 2.05
13	TF	X		2	93.24 ± 1.75*△	93.22 ± 1.77*△	89.81 ± 1.65	89.67 ± 1.76
14	TF-IDF	X		2	94.06 ± 1.63*△	94.08 ± 1.64*△	89.25 ± 2.09	89.04 ± 2.22
15	TF	X		3	81.06 ± 2.47*△	81.89 ± 2.32*△	77.97 ± 3.09	75.98 ± 3.57
16	TF-IDF	X		3	80.28 ± 2.54△	81.26 ± 2.50*△	78.28 ± 2.90	76.21 ± 3.43
17	TF	X		1+2	95.41 ± 1.15	95.46 ± 1.11	95.63 ± 1.34	95.63 ± 1.33
18	TF-IDF	X		1+2	96.48 ± 0.93*△	96.55 ± 0.89*△	94.93 ± 1.34	94.88 ± 1.38
19	TF	X		1+2+3	96.28 ± 1.02	96.28 ± 1.02	95.86 ± 1.32	95.84 ± 1.35
20	TF-IDF	X		1+2+3	97.41 ± 0.92*△	97.44 ± 0.91*△	96.13 ± 1.47	96.09 ± 1.49
21	TF		X	1	92.39 ± 1.44△	92.65 ± 1.34△	91.39 ± 1.80	91.36 ± 1.80
22	TF-IDF		X	1	93.47 ± 1.53*△	93.62 ± 1.48*△	90.66 ± 1.79	90.60 ± 1.85
23	TF		X	2	93.38 ± 1.66*△	93.42 ± 1.61*△	90.25 ± 1.75	90.02 ± 1.81
24	TF-IDF		X	2	93.32 ± 2.06*△	93.45 ± 1.99*△	89.79 ± 1.75	89.50 ± 1.83
25	TF		X	3	80.00 ± 2.90	81.36 ± 2.47*△	79.55 ± 2.91	77.13 ± 3.61
26	TF-IDF		X	3	79.74 ± 3.05	81.30 ± 2.55*△	79.21 ± 2.51	76.54 ± 3.33
27	TF		X	1+2	95.42 ± 1.12	95.46 ± 1.08	95.46 ± 1.30	95.43 ± 1.32
28	TF-IDF		X	1+2	96.42 ± 1.06*△	96.48 ± 1.01*△	94.56 ± 1.43	94.51 ± 1.47
29	TF		X	1+2+3	96.03 ± 1.07△	96.02 ± 1.07△	95.42 ± 1.29	95.44 ± 1.28
30	TF-IDF		X	1+2+3	97.83 ± 0.98*△	97.85 ± 0.97*△	95.73 ± 1.23	95.67 ± 1.24
31	TF	X	X	1	92.53 ± 1.21△	92.74 ± 1.15△	91.41 ± 1.91	91.46 ± 1.86
32	TF-IDF	X	X	1	93.96 ± 1.33*△	94.09 ± 1.28*△	91.56 ± 2.15	91.54 ± 2.14
33	TF	X	X	2	93.24 ± 1.75*△	93.22 ± 1.77*△	89.81 ± 1.65	89.67 ± 1.76
34	TF-IDF	X	X	2	94.06 ± 1.63*△	94.08 ± 1.64*△	89.25 ± 2.09	89.04 ± 2.22
35	TF	X	X	3	81.06 ± 2.47*△	81.89 ± 2.32*△	77.97 ± 3.09	75.98 ± 3.57
36	TF-IDF	X	X	3	80.28 ± 2.54△	81.26 ± 2.50*△	78.28 ± 2.90	76.21 ± 3.43
37	TF	X	X	1+2	95.41 ± 1.08	95.45 ± 1.06	96.27 ± 1.34△	96.28 ± 1.34△
38	TF-IDF	X	X	1+2	96.30 ± 0.99*△	96.37 ± 0.95*△	95.08 ± 1.30	95.04 ± 1.34
39	TF	X	X	1+2+3	96.26 ± 1.07	96.24 ± 1.08	95.98 ± 1.33	95.97 ± 1.34
40	TF-IDF	X	X	1+2+3	97.48 ± 0.93*△	97.51 ± 0.91*△	95.99 ± 1.34	95.94 ± 1.37

* statistically significant (paired t-test) at the 0.05 significance level
△ statistically significant (Wilcoxon signed-ranks test) at the 0.05 significance level

Table 8.4: Paired t-test and Wilcoxon signed-ranks test for accuracy and F1-score within balanced three-class document-level sentiment classification (n = 3,000) based on average scores of sentences for the NBM and the SVM-poly using various pre-processing settings by applying 5 times 10-fold CV

Document-level based on the average scores of sentences
(based on annotations at the sentence-level granularity)
(# documents: 3,000, feature selection method: Gain Ratio, # features: 3,000)
(x sign - included in the experiment)

	Pre-processing settings				NBM		SVM-poly	
ID	TF & TF-IDF	lower case	stop words	N-grams (1, 2, 3)	Accuracy	F1-score	Accuracy	F1-score
1	TF			1	72.12 ± 2.30*△	68.99 ± 3.49△	69.71 ± 2.25	67.43 ± 3.18
2	TF-IDF			1	74.19 ± 2.09*△	71.54 ± 3.36*△	70.36 ± 2.19	68.24 ± 3.20
3	TF			2	75.13 ± 2.46*△	72.59 ± 3.74*△	68.34 ± 2.59	66.63 ± 3.09
4	TF-IDF			2	75.79 ± 2.39*△	73.43 ± 3.47*△	68.71 ± 2.45	66.53 ± 3.25
5	TF			3	63.72 ± 2.27△	61.73 ± 3.30*△	62.05 ± 2.43	57.00 ± 3.64
6	TF-IDF			3	62.25 ± 2.30	60.27 ± 3.38*△	61.55 ± 2.39	56.33 ± 3.62
7	TF			1+2	75.30 ± 2.20	73.01 ± 3.51	75.32 ± 2.23	73.28 ± 3.30
8	TF-IDF			1+2	79.33 ± 1.92*△	77.11 ± 3.31*△	74.11 ± 1.93	71.73 ± 2.50
9	TF			1+2+3	76.36 ± 2.23	73.18 ± 3.49	76.24 ± 2.30	74.53 ± 3.03△
10	TF-IDF			1+2+3	79.59 ± 1.93*△	77.64 ± 2.96*△	76.13 ± 1.79	74.47 ± 2.49
11	TF	x		1	72.05 ± 2.30△	68.92 ± 3.29	70.31 ± 2.49	68.46 ± 3.18
12	TF-IDF	x		1	73.99 ± 2.58*△	71.10 ± 3.71△	71.33 ± 2.30	69.24 ± 3.06
13	TF	x		2	75.31 ± 2.60*△	72.58 ± 3.87*△	70.03 ± 2.82	67.81 ± 3.83
14	TF-IDF	x		2	77.91 ± 2.08*△	75.46 ± 3.34*△	70.29 ± 2.21	68.27 ± 3.23
15	TF	x		3	62.90 ± 2.63△	60.71 ± 3.52*△	60.86 ± 2.73	56.55 ± 3.42
16	TF-IDF	x		3	61.81 ± 2.57△	59.23 ± 3.55*△	60.84 ± 2.56	56.08 ± 3.15
17	TF	x		1+2	75.88 ± 2.23	73.27 ± 3.62	75.35 ± 2.12	74.08 ± 2.59
18	TF-IDF	x		1+2	79.72 ± 1.91*△	77.63 ± 3.20*△	74.69 ± 1.99	73.42 ± 2.87
19	TF	x		1+2+3	76.28 ± 2.08	73.12 ± 3.33	75.72 ± 2.09	74.12 ± 2.78
20	TF-IDF	x		1+2+3	79.07 ± 1.92*△	76.52 ± 3.18△	76.20 ± 2.29	74.61 ± 3.16
21	TF		x	1	72.06 ± 2.42*△	69.28 ± 3.60△	69.51 ± 2.54	66.77 ± 3.24
22	TF-IDF		x	1	74.46 ± 2.05*△	72.07 ± 3.14*△	69.97 ± 2.70	68.06 ± 3.78
23	TF		x	2	75.13 ± 2.46*△	72.59 ± 3.74*△	68.34 ± 2.59	66.63 ± 3.09
24	TF-IDF		x	2	75.79 ± 2.39*△	73.43 ± 3.47*△	68.71 ± 2.45	66.53 ± 3.25
25	TF		x	3	63.72 ± 2.27△	61.73 ± 3.30*△	62.05 ± 2.43	57.00 ± 3.64
26	TF-IDF		x	3	62.25 ± 2.30	60.27 ± 3.38*△	61.55 ± 2.39	56.33 ± 3.62
27	TF		x	1+2	75.54 ± 2.20	73.45 ± 3.48	74.89 ± 2.02	72.90 ± 2.72
28	TF-IDF		x	1+2	79.24 ± 1.91*△	77.05 ± 3.26*△	74.07 ± 1.91	71.93 ± 2.74
29	TF		x	1+2+3	76.45 ± 2.25△	73.35 ± 3.46	75.49 ± 2.11	74.16 ± 2.88
30	TF-IDF		x	1+2+3	79.38 ± 1.93*△	77.56 ± 2.97*△	75.81 ± 2.01	74.10 ± 2.75
31	TF	x	x	1	72.43 ± 2.41*△	70.29 ± 3.29△	69.75 ± 2.59	67.53 ± 3.61
32	TF-IDF	x	x	1	74.21 ± 2.57*△	71.42 ± 3.66△	71.49 ± 2.06	69.54 ± 2.84
33	TF	x	x	2	75.31 ± 2.60*△	72.58 ± 3.87*△	70.03 ± 2.82	67.81 ± 3.83
34	TF-IDF	x	x	2	77.91 ± 2.08*△	75.46 ± 3.34*△	70.29 ± 2.21	68.27 ± 3.23
35	TF	x	x	3	62.90 ± 2.63△	60.71 ± 3.52*△	60.86 ± 2.73	56.55 ± 3.42
36	TF-IDF	x	x	3	61.81 ± 2.57△	59.23 ± 3.55*△	60.84 ± 2.56	56.08 ± 3.15
37	TF	x	x	1+2	76.13 ± 2.00△	73.57 ± 3.39	75.26 ± 2.39	74.21 ± 3.17
38	TF-IDF	x	x	1+2	79.85 ± 1.93*△	77.76 ± 3.13*△	75.20 ± 2.31	74.10 ± 3.07
39	TF	x	x	1+2+3	76.72 ± 2.03	73.80 ± 3.22	75.75 ± 2.35	74.19 ± 2.90
40	TF-IDF	x	x	1+2+3	78.90 ± 1.85*△	76.37 ± 3.09△	75.89 ± 2.48	74.12 ± 3.33

* statistically significant (paired t-test) at the 0.05 significance level
△ statistically significant (Wilcoxon signed-ranks test) at the 0.05 significance level

8.2.2 Results H_2

Both classifiers, the NBM and the SVM, performed significantly better, when the average sentiment scores of sentences were used to determine the sentiment of a document.

To illustrate the process of testing hypothesis H_2, in Table 8.5, we apply the Wilcoxon signed-ranks test for F1-score within balanced two-class document-level sentiment classification (n = 2,000) based on average scores of sentences and documents for the NBM using pre-processing settings with ID = 1 by applying 5 times 10-fold CV. First, we calculate the differences d^i for the NBM of F1-scores based on average scores of sentences and documents, that is the differences between n = 50 different values in columns 7 and 9 (in Table 8.5). We compute their ranks, followed by sum of (positive and negative) ranks R_+ = 1,272, R_- = 3 and T = 3. Since, the computed test statistic $|W = -6.13|$ (see Equation 8.2) is greater than z_α = 1.96, we reject the null hypothesis. In Table 8.6, we apply the paired t-test and Wilcoxon signed-ranks test for accuracy and F1-score within balanced two-class document-level sentiment classification (n = 2,000) based on average scores of sentences and documents for the SVM-poly using various pre-processing settings by applying 5 times 10-fold CV. Table 8.7 presents the paired t-test and Wilcoxon signed-ranks test for accuracy and F1-score within balanced three-class document-level sentiment classification (n = 3,000) based on average scores of sentences and documents for the NBM using various pre-processing settings by applying 5 times 10-fold CV. In Table 8.8, we apply the paired t-test and Wilcoxon signed-ranks test for accuracy and F1-score within balanced three-class document-level sentiment classification (n = 3,000) based on average scores of sentences and documents for the SVM-poly using various pre-processing settings by applying 5 times 10-fold CV.

Overall (in all cases), both classifiers, the NBM classifier and the SVM classifier, perform significantly better (statistically significant (paired t-test and Wilcoxon signed-ranks test) at the 0.05 significance level), when the average sentiment scores of sentences is used to determine the sentiment of a document.

The series of experiments described in this Section (see Tables 8.5, 8.6, 8.7 and 8.8) and in Section 5.3 confirm the hypothesis H_2, which states that granulation of a document to smaller segments, such as sentences, can improve the classification performance.

Table 8.5: Paired t-test and Wilcoxon signed-ranks test for accuracy and F1-score within balanced two-class document-level sentiment classification (n = 2,000) based on average scores of sentences and documents for the NBM using various pre-processing settings by applying 5 times 10-fold CV

Document-level based on the average scores of sentences and documents
(based on annotations at the sentence-level and document-level granularity)
(# documents: 2,000, feature selection method: Gain Ratio, # features: 3,000)
(x sign - included in the experiment)

	Pre-processing settings				NBM (based on the average of sentences)		NBM (based on the average of documents)	
ID	TF & TF-IDF	lower case	stop words	N-grams (1, 2, 3)	Accuracy	F1-score	Accuracy	F1-score
1	TF			1	92.89 ± 1.40*△	93.08 ± 1.32*△	87.35 ± 2.77	87.52 ± 2.66
2	TF-IDF			1	92.83 ± 1.63*△	93.02 ± 1.55*△	86.52 ± 2.77	88.33 ± 2.67
3	TF			2	93.38 ± 1.66*△	93.42 ± 1.61*△	87.79 ± 2.74	87.76 ± 2.73
4	TF-IDF			2	93.32 ± 2.06*△	93.45 ± 1.99*△	84.66 ± 2.53	87.00 ± 2.41
5	TF			3	80.00 ± 2.90*△	81.36 ± 2.47*△	75.07 ± 3.24	76.93 ± 2.86
6	TF-IDF			3	79.74 ± 3.05*△	81.30 ± 2.55*△	58.70 ± 3.62	69.80 ± 3.60
7	TF			1+2	95.33 ± 1.14*△	95.36 ± 1.11*△	89.52 ± 2.28	89.44 ± 2.29
8	TF-IDF			1+2	96.32 ± 1.07*△	96.38 ± 1.02*△	92.22 ± 2.10	92.42 ± 2.03
9	TF			1+2+3	96.22 ± 1.04*△	96.22 ± 1.04*△	89.60 ± 2.53	89.51 ± 2.55
10	TF-IDF			1+2+3	97.78 ± 0.97*△	97.80 ± 0.96*△	92.80 ± 1.92	93.03 ± 1.95
11	TF	x		1	92.82 ± 1.16*△	92.99 ± 1.11*△	87.29 ± 2.58	87.54 ± 2.49
12	TF-IDF	x		1	93.75 ± 1.26*△	93.90 ± 1.21*△	85.87 ± 2.63	88.80 ± 2.22
13	TF	x		2	93.24 ± 1.75*△	93.22 ± 1.77*△	88.14 ± 2.46	88.09 ± 2.47
14	TF-IDF	x		2	94.06 ± 1.63*△	94.08 ± 1.64*△	86.62 ± 2.24	88.34 ± 2.01
15	TF	x		3	81.06 ± 2.47*△	81.89 ± 2.32*△	77.59 ± 2.77	78.93 ± 2.64
16	TF-IDF	x		3	80.28 ± 2.54*△	81.26 ± 2.50*△	61.26 ± 3.18	72.03 ± 2.93
17	TF	x		1+2	95.41 ± 1.15*△	95.46 ± 1.11*△	89.50 ± 2.35	89.43 ± 2.40
18	TF-IDF	x		1+2	96.48 ± 0.93*△	96.55 ± 0.89*△	92.40 ± 1.75	92.63 ± 1.70
19	TF	x		1+2+3	96.28 ± 1.02*△	96.28 ± 1.02*△	90.41 ± 2.39	90.34 ± 2.43
20	TF-IDF	x		1+2+3	97.41 ± 0.92*△	97.44 ± 0.91*△	92.89 ± 1.65	93.12 ± 1.65
21	TF		x	1	92.39 ± 1.44*△	92.65 ± 1.34*△	86.85 ± 2.82	87.16 ± 2.67
22	TF-IDF		x	1	93.47 ± 1.53*△	93.62 ± 1.48*△	86.44 ± 2.90	88.39 ± 2.75
23	TF		x	2	93.38 ± 1.66*△	93.42 ± 1.61*△	87.79 ± 2.74	87.76 ± 2.73
24	TF-IDF		x	2	93.32 ± 2.06*△	93.45 ± 1.99*△	84.66 ± 2.53	87.00 ± 2.41
25	TF		x	3	80.00 ± 2.90*△	81.36 ± 2.47*△	75.07 ± 3.24	76.93 ± 2.86
26	TF-IDF		x	3	79.74 ± 3.05*△	81.30 ± 2.55*△	58.70 ± 3.62	69.80 ± 3.60
27	TF		x	1+2	95.42 ± 1.12*△	95.46 ± 1.08*△	89.50 ± 2.29	89.46 ± 2.28
28	TF-IDF		x	1+2	96.42 ± 1.06*△	96.48 ± 1.01*△	92.12 ± 2.19	92.29 ± 2.08
29	TF		x	1+2+3	96.03 ± 1.07*△	96.02 ± 1.07*△	89.96 ± 2.58	89.99 ± 2.55
30	TF-IDF		x	1+2+3	97.83 ± 0.98*△	97.85 ± 0.97*△	92.78 ± 1.86	93.00 ± 1.87
31	TF	x	x	1	92.53 ± 1.21*△	92.74 ± 1.15*△	87.66 ± 2.33	88.02 ± 2.19
32	TF-IDF	x	x	1	93.96 ± 1.33*△	94.09 ± 1.28*△	85.49 ± 2.50	88.60 ± 2.14
33	TF	x	x	2	93.24 ± 1.75*△	93.22 ± 1.77*△	88.14 ± 2.46	88.09 ± 2.47
34	TF-IDF	x	x	2	94.06 ± 1.63*△	94.08 ± 1.64*△	86.62 ± 2.24	88.34 ± 2.01
35	TF	x	x	3	81.06 ± 2.47*△	81.89 ± 2.32*△	77.59 ± 2.77	78.93 ± 2.64
36	TF-IDF	x	x	3	80.28 ± 2.54*△	81.26 ± 2.50*△	61.26 ± 3.18	72.03 ± 2.93
37	TF	x	x	1+2	95.41 ± 1.08*△	95.45 ± 1.06*△	89.38 ± 2.32	89.37 ± 2.34
38	TF-IDF	x	x	1+2	96.30 ± 0.99*△	96.37 ± 0.95*△	92.46 ± 1.78	92.75 ± 1.69
39	TF	x	x	1+2+3	96.26 ± 1.07*△	96.24 ± 1.08*△	90.68 ± 2.41	90.74 ± 2.39
40	TF-IDF	x	x	1+2+3	97.48 ± 0.93*△	97.51 ± 0.91*△	92.87 ± 1.54	93.12 ± 1.54

* statistically significant (paired t-test) at the 0.05 significance level
△ statistically significant (Wilcoxon signed-ranks test) at the 0.05 significance level

Table 8.6: Paired t-test and Wilcoxon signed-ranks test for accuracy and F1-score within balanced two-class document-level sentiment classification (n = 2,000) based on average scores of sentences and documents for the SVM-poly using various pre-processing settings by applying 5 times 10-fold CV

Document-level based on the average scores of sentences and documents
(based on annotations at the sentence-level and document-level granularity)
(# documents: 2,000, feature selection method: Gain Ratio, # features: 3,000)
(x sign - included in the experiment)

	Pre-processing settings				SVM-poly (based on the average of sentences)		SVM-poly (based on the average of documents)	
ID	TF & TF-IDF	lower case	stop words	N-grams (1, 2, 3)	Accuracy	F1-score	Accuracy	F1-score
1	TF			1	91.35 ± 1.70*△	91.33 ± 1.72*△	86.81 ± 2.24	86.83 ± 2.15
2	TF-IDF			1	91.12 ± 1.93*△	91.04 ± 1.97*△	85.69 ± 2.36	85.63 ± 2.37
3	TF			2	90.25 ± 1.75*△	90.02 ± 1.81*△	85.67 ± 2.57	85.35 ± 2.72
4	TF-IDF			2	89.79 ± 1.75*△	89.50 ± 1.83*△	85.81 ± 2.27	85.51 ± 2.46
5	TF			3	79.55 ± 2.91*△	77.13 ± 3.61*△	74.53 ± 2.47	71.81 ± 3.21
6	TF-IDF			3	79.21 ± 2.51*△	76.54 ± 3.33*△	75.15 ± 2.50	72.55 ± 3.27
7	TF			1+2	95.01 ± 1.38*△	94.98 ± 1.40*△	90.35 ± 1.76	90.31 ± 1.81
8	TF-IDF			1+2	94.14 ± 1.48*△	94.08 ± 1.52*△	89.64 ± 1.96	89.53 ± 2.01
9	TF			1+2+3	95.42 ± 1.27*△	95.41 ± 1.28*△	91.50 ± 2.03	91.49 ± 2.04
10	TF-IDF			1+2+3	95.87 ± 1.29*△	95.82 ± 1.30*△	90.60 ± 1.85	90.53 ± 1.91
11	TF	X		1	91.23 ± 1.67*△	91.26 ± 1.67*△	87.01 ± 2.05	86.96 ± 2.13
12	TF-IDF	X		1	91.08 ± 2.04*△	91.08 ± 2.05*△	87.82 ± 2.29	87.71 ± 2.32
13	TF	X		2	89.81 ± 1.65*△	89.67 ± 1.76*△	85.55 ± 2.50	85.32 ± 2.64
14	TF-IDF	X		2	89.25 ± 2.09*△	89.04 ± 2.22*△	86.83 ± 2.25	86.51 ± 2.33
15	TF	X		3	77.97 ± 3.09*△	75.98 ± 3.57*△	74.83 ± 3.14	72.59 ± 3.90
16	TF-IDF	X		3	78.28 ± 2.90*△	76.21 ± 3.43*△	74.84 ± 2.95	72.54 ± 3.64
17	TF	X		1+2	95.63 ± 1.34*△	95.63 ± 1.33*△	90.28 ± 2.00	90.20 ± 2.07
18	TF-IDF	X		1+2	94.93 ± 1.34*△	94.88 ± 1.38*△	90.06 ± 2.11	89.93 ± 2.24
19	TF	X		1+2+3	95.86 ± 1.32*△	95.84 ± 1.35*△	91.90 ± 1.64	91.89 ± 1.66
20	TF-IDF	X		1+2+3	96.13 ± 1.47*△	96.09 ± 1.49*△	92.55 ± 1.64	92.48 ± 1.69
21	TF		X	1	91.39 ± 1.80*△	91.36 ± 1.80*△	86.89 ± 2.17	86.89 ± 2.14
22	TF-IDF		X	1	90.66 ± 1.79*△	90.60 ± 1.85*△	85.83 ± 2.39	85.69 ± 2.42
23	TF		X	2	90.25 ± 1.75*△	90.02 ± 1.81*△	85.67 ± 2.57	85.35 ± 2.72
24	TF-IDF		X	2	89.79 ± 1.75*△	89.50 ± 1.83*△	85.81 ± 2.27	85.51 ± 2.46
25	TF		X	3	79.55 ± 2.91*△	77.13 ± 3.61*△	74.53 ± 2.47	71.81 ± 3.21
26	TF-IDF		X	3	79.21 ± 2.51*△	76.54 ± 3.33*△	75.15 ± 2.50	72.55 ± 3.27
27	TF		X	1+2	95.46 ± 1.30*△	95.43 ± 1.32*△	89.80 ± 1.69	89.82 ± 1.71
28	TF-IDF		X	1+2	94.56 ± 1.43*△	94.51 ± 1.47*△	89.98 ± 2.17	89.85 ± 2.23
29	TF		X	1+2+3	95.42 ± 1.29*△	95.44 ± 1.28*△	91.24 ± 1.84	91.27 ± 1.82
30	TF-IDF		X	1+2+3	95.73 ± 1.23*△	95.67 ± 1.24*△	90.55 ± 1.84	90.43 ± 1.91
31	TF	X	X	1	91.41 ± 1.91*△	91.46 ± 1.86*△	86.92 ± 2.04	86.88 ± 2.07
32	TF-IDF	X	X	1	91.56 ± 2.15*△	91.54 ± 2.14*△	87.91 ± 2.28	87.78 ± 2.34
33	TF	X	X	2	89.81 ± 1.65*△	89.67 ± 1.76*△	85.55 ± 2.50	85.32 ± 2.64
34	TF-IDF	X	X	2	89.25 ± 2.09*△	89.04 ± 2.22*△	86.83 ± 2.25	86.51 ± 2.33
35	TF	X	X	3	77.97 ± 3.09*△	75.98 ± 3.57*△	74.83 ± 3.14	72.59 ± 3.90
36	TF-IDF	X	X	3	78.28 ± 2.90*△	76.21 ± 3.43*△	74.84 ± 2.95	72.54 ± 3.64
37	TF	X	X	1+2	96.27 ± 1.34*△	96.28 ± 1.34*△	90.17 ± 1.67	90.15 ± 1.74
38	TF-IDF	X	X	1+2	95.08 ± 1.30*△	95.04 ± 1.34*△	90.18 ± 1.97	90.07 ± 2.05
39	TF	X	X	1+2+3	95.98 ± 1.33*△	95.97 ± 1.34*△	92.06 ± 1.81	92.04 ± 1.78
40	TF-IDF	X	X	1+2+3	95.99 ± 1.34*△	95.94 ± 1.37*△	92.43 ± 1.76	92.34 ± 1.83

* statistically significant (paired t-test) at the 0.05 significance level
△ statistically significant (Wilcoxon signed-ranks test) at the 0.05 significance level

Table 8.7: Paired t-test and Wilcoxon signed-ranks test for accuracy and F1-score within balanced three-class document-level sentiment classification (n = 3,000) based on average scores of sentences and documents for the NBM using various pre-processing settings by applying 5 times 10-fold CV

Document-level based on the average scores of sentences and documents
(based on annotations at the sentence-level and document-level granularity)
(# documents: 3,000, feature selection method: Gain Ratio, # features: 3,000)
(x sign - included in the experiment)

	Pre-processing settings				NBM (based on the average of sentences)		NBM (based on the average of documents)	
ID	TF & TF-IDF	lower case	stop words	N-grams (1, 2, 3)	Accuracy	F1-score	Accuracy	F1-score
1	TF			1	72.12 ± 2.30*△	68.99 ± 3.49*△	65.62 ± 2.33	63.73 ± 3.82
2	TF-IDF			1	74.19 ± 2.09*△	71.54 ± 3.36*△	67.92 ± 2.51	66.30 ± 4.07
3	TF			2	75.13 ± 2.46*△	72.59 ± 3.74*△	68.05 ± 2.61	67.65 ± 2.86
4	TF-IDF			2	75.79 ± 2.39*△	73.43 ± 3.47*△	69.67 ± 2.26	69.94 ± 2.82
5	TF			3	63.72 ± 2.27*△	61.73 ± 3.30*△	57.92 ± 2.75	59.28 ± 3.58
6	TF-IDF			3	62.25 ± 2.30*△	60.27 ± 3.38*△	55.95 ± 2.68	56.51 ± 3.07
7	TF			1+2	75.30 ± 2.20*△	73.01 ± 3.51*△	67.96 ± 2.07	65.70 ± 3.70
8	TF-IDF			1+2	79.33 ± 1.92*△	77.11 ± 3.31*△	72.64 ± 2.35	72.07 ± 3.49
9	TF			1+2+3	76.36 ± 2.23*△	73.18 ± 3.49*△	66.83 ± 2.12	64.61 ± 3.75
10	TF-IDF			1+2+3	79.59 ± 1.93*△	77.64 ± 2.96*△	70.21 ± 2.36	68.77 ± 3.84
11	TF	x		1	72.05 ± 2.30*△	68.92 ± 3.29*△	64.97 ± 2.21	62.89 ± 3.80
12	TF-IDF	x		1	73.99 ± 2.58*△	71.10 ± 3.71*△	67.40 ± 2.27	65.73 ± 3.24
13	TF	x		2	75.31 ± 2.60*△	72.58 ± 3.87*△	67.78 ± 2.45	67.16 ± 3.50
14	TF-IDF	x		2	77.91 ± 2.08*△	75.46 ± 3.34*△	70.86 ± 2.24	70.83 ± 3.05
15	TF	x		3	62.90 ± 2.63*△	60.71 ± 3.52*△	58.57 ± 2.65	58.80 ± 3.52
16	TF-IDF	x		3	61.81 ± 2.57*△	59.23 ± 3.55*△	57.03 ± 2.52	57.61 ± 3.53
17	TF	x		1+2	75.88 ± 2.23*△	73.27 ± 3.62*△	67.25 ± 2.13	65.29 ± 3.45
18	TF-IDF	x		1+2	79.72 ± 1.91*△	77.63 ± 3.20*△	73.09 ± 2.28	72.77 ± 3.44
19	TF	x		1+2+3	76.28 ± 2.08*△	73.12 ± 3.33*△	66.52 ± 2.18	63.86 ± 3.77
20	TF-IDF	x		1+2+3	79.07 ± 1.92*△	76.52 ± 3.18*△	70.31 ± 2.02	69.13 ± 3.32
21	TF		x	1	72.06 ± 2.42*△	69.28 ± 3.60*△	65.34 ± 2.52	63.35 ± 3.80
22	TF-IDF		x	1	74.46 ± 2.05*△	72.07 ± 3.14*△	68.06 ± 2.80	66.99 ± 4.30
23	TF		x	2	75.13 ± 2.46*△	72.59 ± 3.74*△	68.05 ± 2.61	67.65 ± 2.86
24	TF-IDF		x	2	75.79 ± 2.39*△	73.43 ± 3.47*△	69.67 ± 2.26	69.94 ± 2.82
25	TF		x	3	63.72 ± 2.27*△	61.73 ± 3.30*△	57.92 ± 2.75	59.28 ± 3.58
26	TF-IDF		x	3	62.25 ± 2.30*△	60.27 ± 3.38*△	55.95 ± 2.68	56.51 ± 3.07
27	TF		x	1+2	75.54 ± 2.20*△	73.45 ± 3.48*△	68.34 ± 2.16	66.68 ± 3.74
28	TF-IDF		x	1+2	79.24 ± 1.91*△	77.05 ± 3.26*△	72.80 ± 2.34	71.83 ± 3.32
29	TF		x	1+2+3	76.45 ± 2.25*△	73.35 ± 3.46*△	67.15 ± 2.01	64.92 ± 3.67
30	TF-IDF		x	1+2+3	79.38 ± 1.93*△	77.56 ± 2.97*△	70.61 ± 2.27	69.62 ± 3.47
31	TF	x	x	1	72.43 ± 2.41*△	70.29 ± 3.29*△	65.07 ± 2.17	63.04 ± 3.78
32	TF-IDF	x	x	1	74.21 ± 2.57*△	71.42 ± 3.66*△	67.26 ± 2.25	65.57 ± 3.20
33	TF	x	x	2	75.31 ± 2.60*△	72.58 ± 3.87*△	67.78 ± 2.45	67.16 ± 3.50
34	TF-IDF	x	x	2	77.91 ± 2.08*△	75.46 ± 3.34*△	70.86 ± 2.24	70.83 ± 3.05
35	TF	x	x	3	62.90 ± 2.63*△	60.71 ± 3.52*△	58.57 ± 2.65	58.80 ± 3.52
36	TF-IDF	x	x	3	61.81 ± 2.57*△	59.23 ± 3.55*△	57.03 ± 2.52	57.61 ± 3.53
37	TF	x	x	1+2	76.13 ± 2.00*△	73.57 ± 3.39*△	67.37 ± 2.14	65.50 ± 3.57
38	TF-IDF	x	x	1+2	79.85 ± 1.93*△	77.76 ± 3.13*△	71.40 ± 2.00	70.93 ± 3.11
39	TF	x	x	1+2+3	76.72 ± 2.03*△	73.80 ± 3.22*△	66.82 ± 2.07	64.53 ± 3.79
40	TF-IDF	x	x	1+2+3	78.90 ± 1.85*△	76.37 ± 3.09*△	70.31 ± 2.19	69.58 ± 3.68

* statistically significant (paired t-test) at the 0.05 significance level
△ statistically significant (Wilcoxon signed-ranks test) at the 0.05 significance level

Table 8.8: Paired t-test and Wilcoxon signed-ranks test for accuracy and F1-score within balanced three-class document-level sentiment classification (n = 3,000) based on average scores of sentences and documents for the SVM-poly using various pre-processing settings by applying 5 times 10-fold CV

Document-level based on the average scores of sentences and documents
(based on annotations at the sentence-level and document-level granularity)
(# documents: 3,000, feature selection method: Gain Ratio, # features: 3,000)
(x sign - included in the experiment)

	Pre-processing settings				SVM-poly (based on the average of sentences)		SVM-poly (based on the average of documents)	
ID	TF & TF-IDF	lower case	stop words	N-grams (1, 2, 3)	Accuracy	F1-score	Accuracy	F1-score
1	TF			1	69.71 ± 2.25*△	67.43 ± 3.18*△	63.57 ± 2.57	63.30 ± 3.40
2	TF-IDF			1	70.36 ± 2.19*△	68.24 ± 3.20*△	63.47 ± 2.69	62.69 ± 3.73
3	TF			2	68.34 ± 2.59*△	66.63 ± 3.09*△	63.07 ± 2.34	64.24 ± 2.83
4	TF-IDF			2	68.71 ± 2.4*△5	66.53 ± 3.25*△	64.25 ± 2.54	64.99 ± 3.04
5	TF			3	62.05 ± 2.43*△	57.00 ± 3.64*△	57.59 ± 2.84	54.22 ± 4.88
6	TF-IDF			3	61.55 ± 2.39*△	56.33 ± 3.62*△	57.69 ± 3.03	54.52 ± 4.95
7	TF			1+2	75.32 ± 2.23*△	73.28 ± 3.30*△	66.61 ± 2.93	65.74 ± 3.67
8	TF-IDF			1+2	74.11 ± 1.93*△	71.73 ± 2.50*△	67.05 ± 2.70	66.34 ± 3.66
9	TF			1+2+3	76.24 ± 2.30*△	74.53 ± 3.03*△	64.69 ± 3.00	64.34 ± 3.66
10	TF-IDF			1+2+3	76.13 ± 1.79*△	74.47 ± 2.49*△	64.71 ± 3.16	64.67 ± 3.92
11	TF	X		1	70.31 ± 2.49*△	68.46 ± 3.18*△	64.92 ± 2.86	64.97 ± 3.52
12	TF-IDF	X		1	71.33 ± 2.30*△	69.24 ± 3.06*△	63.68 ± 2.62	63.65 ± 3.68
13	TF	X		2	70.03 ± 2.82*△	67.81 ± 3.83*△	64.15 ± 2.35	64.40 ± 3.53
14	TF-IDF	X		2	70.29 ± 2.21*△	68.27 ± 3.23*△	64.07 ± 2.48	64.17 ± 3.63
15	TF	X		3	60.86 ± 2.73*△	56.55 ± 3.42*△	55.90 ± 2.80	53.98 ± 4.65
16	TF-IDF	X		3	60.84 ± 2.56*△	56.08 ± 3.15*△	56.25 ± 2.74	54.11 ± 4.84
17	TF	X		1+2	75.35 ± 2.12*△	74.08 ± 2.59*△	67.43 ± 2.71	67.77 ± 3.50
18	TF-IDF	X		1+2	74.69 ± 1.99*△	73.42 ± 2.87*△	67.94 ± 2.57	67.71 ± 3.18
19	TF	X		1+2+3	75.72 ± 2.09*△	74.12 ± 2.78*△	65.25 ± 1.96	64.66 ± 2.75
20	TF-IDF	X		1+2+3	76.20 ± 2.29*△	74.61 ± 3.16*△	64.93 ± 2.49	64.70 ± 3.13
21	TF		X	1	69.51 ± 2.54*△	66.77 ± 3.24*△	62.25 ± 2.78	62.26 ± 3.54
22	TF-IDF		X	1	69.97 ± 2.70*△	68.06 ± 3.78*△	63.12 ± 2.26	62.40 ± 3.60
23	TF		X	2	68.34 ± 2.59*△	66.63 ± 3.09*△	63.07 ± 2.34	64.24 ± 2.83
24	TF-IDF		X	2	68.71 ± 2.45*△	66.53 ± 3.25*△	64.25 ± 2.54	64.99 ± 3.04
25	TF		X	3	62.05 ± 2.43*△	57.00 ± 3.64*△	57.59 ± 2.84	54.22 ± 4.88
26	TF-IDF		X	3	61.55 ± 2.39*△	56.33 ± 3.62*△	57.69 ± 3.03	54.52 ± 4.95
27	TF		X	1+2	74.89 ± 2.02*△	72.90 ± 2.72*△	67.36 ± 2.74	66.41 ± 3.60
28	TF-IDF		X	1+2	74.07 ± 1.91*△	71.93 ± 2.74*△	67.83 ± 2.65	67.18 ± 3.45
29	TF		X	1+2+3	75.49 ± 2.11*△	74.16 ± 2.88*△	65.85 ± 3.28	65.79 ± 4.21
30	TF-IDF		X	1+2+3	75.81 ± 2.01*△	74.10 ± 2.75*△	65.11 ± 2.93	65.05 ± 3.92
31	TF	X	X	1	69.75 ± 2.59*△	67.53 ± 3.61*△	64.71 ± 2.62	64.87 ± 3.68
32	TF-IDF	X	X	1	71.49 ± 2.06*△	69.54 ± 2.84*△	64.27 ± 2.44	64.52 ± 3.50
33	TF	X	X	2	70.03 ± 2.82*△	67.81 ± 3.83*△	64.15 ± 2.35	64.40 ± 3.53
34	TF-IDF	X	X	2	70.29 ± 2.21*△	68.27 ± 3.23*△	64.07 ± 2.48	64.17 ± 3.63
35	TF	X	X	3	60.86 ± 2.73*△	56.55 ± 3.42*△	55.90 ± 2.80	53.98 ± 4.65
36	TF-IDF	X	X	3	60.84 ± 2.56*△	56.08 ± 3.15*△	56.25 ± 2.74	54.11 ± 4.84
37	TF	X	X	1+2	75.26 ± 2.39*△	74.21 ± 3.17*△	66.99 ± 2.51	67.24 ± 3.48
38	TF-IDF	X	X	1+2	75.20 ± 2.31*△	74.10 ± 3.07*△	66.97 ± 2.64	67.42 ± 3.69
39	TF	X	X	1+2+3	75.75 ± 2.35*△	74.19 ± 2.90*△	64.78 ± 2.10	64.54 ± 3.22
40	TF-IDF	X	X	1+2+3	75.89 ± 2.48*△	74.12 ± 3.33*△	65.60 ± 2.81	65.03 ± 3.23

* statistically significant (paired t-test) at the 0.05 significance level
△ statistically significant (Wilcoxon signed-ranks test) at the 0.05 significance level

8.3 Testing Hypothesis 3

Within our study, we developed tools that enable mass data acquisition (web crawlers) and web application for retrieval, storage, annotation and sentiment allocation of web texts in the Slovenian language. Also, we developed several language resources in the Slovenian language, such as corpus of retrieved (raw) news containing 256,567 news, manually annotated news corpora at three levels of granularity including 10,427 news, automatically annotated news corpora with 256,567 news, and a lexicon for sentiment analysis. Developed tools and language resources are publicly available under the terms of use (see Section 4.3).

The developed sentiment analysis methodology was successfully used in real-world applications for estimating the proportions of positive, negative and neutral news in the selected web media and for monitoring the dynamics of sentiment. When estimating the proportion of positive, negative and neutral news, the experiments show that approximately half of the retrieved news is neutral. The proportion of negative news is estimated twice as high as the proportion of positive news. All estimations are based on documents that are dealing with political, business, economic and financial news that were published between 1st of September 2007 and 31st of January 2016 in five Slovenian web media (24ur, Dnevnik, Finance, Rtvslo and Žurnal24). Monitoring the dynamics of sentiment is another interesting area. Monitoring dynamics by time–series keywords and web media is useful for tracking sudden changes or trends, which tend to be associated with political and economic issues. Examining authors, for example, gives us an insight into characteristic patterns of writing. A study of their writing styles show whether they prone to write more positive or negative news articles and how their writing styles evolve through time.

The Section 1.2 and Chapters 5-7 indicate that the hypothesis H_3, which deals with the real-life applicability of the developed sentiment analysis tools, resources and methodology, can be confirmed.

8.4 Testing Hypothesis 4

8.4.1 Methodology

We were motivated to test this hypothesis, since literature indicates that the proportion of negative news has increased in most media (Stone & Grusin, 1984; International Journalists' Network, 2008, Kovačič, 2012; Ho, Chen & Sim, 2013; Trussler & Soroka, 2014; Vinkers, Tijdink & Otte, 2015; Kätsyri et al., 2016).

Regarding the H_4 we test the hypothesis about the proportion of negative news in relation to positive news (we excluded neutral news), which contain political, business, economic and financial content from five Slovenian web media.

If we define null and alternative hypotheses, we have $H_0 : \pi^i_{NEG} = \pi^i_{POS}$ and $H_4 : \pi^i_{NEG} > \pi^i_{POS}$, where H_0 means the null hypothesis, H_1 the alternative hypothesis, i the web medium (Rtvslo, 24ur, Dnevnik, Finance, Žurnal24), π^i_{NEG} is the population proportion for negative sentiment in web medium i, and π^i_{POS} is the population proportion for positive sentiment in web medium i. We test the hypothesis with the commonly used 0.05 significance level. According to standard normal distribution critical value of the rejection region z_α is specified and *Z-test* statistics calculated:

$$Z = \frac{p^i_{NEG} - p^i_{POS}}{SE^i} \tag{8.3}$$

where $SE^i = \sqrt{p^i \cdot (1 - p^i) \cdot (\frac{1}{n^i_{NEG}} + \frac{1}{n^i_{POS}})}$ is the standard error in web medium i, $p^i = \frac{p^i_{NEG} \cdot n^i_{NEG} + p^i_{POS} \cdot n^i_{POS}}{n^i_{NEG} + n^i_{POS}}$ is the pooled sample proportion in web medium i, p^i_{NEG} is the sample proportion for negative sentiment in web medium i, p^i_{POS} is the sample proportion for positive sentiment in web medium i, n^i_{NEG} is the number of news with negative sentiment in web medium i, and n^i_{POS} is the number of news with positive sentiment in web medium i. *Z-test* statistics is then compared to the critical value z_α. Then, we use tables of the z-distribution ($z_\alpha = 1.645$) to compare the value of *Z-test* statistics (Z), which gives the $p - value$ for the *Z-test*. P-value is based on the standard normal distribution, which has a mean of 0 and a standard deviation of 1. If the absolute value of *Z-test* statistics (Z) is greater than the value of z-distribution ($z_\alpha = 1.645$), then we can reject H_0 and accept H_4 with the selected significance level. We tested the hypothesis with series of *Z-test*

statistics (for five web media i).

8.4.2 Results H_4

To confirm the H_4 hypothesis, we estimated the proportion of positive and negative news, obtained all Slovenian news two times. In Table 8.9 we test the hypothesis on (document-based) news corpus SentiNews 1.0 (manually annotated Slovenian news published between 1st of September 2007 and 31st of December 2013, i.e., 10,427 documents (24ur - 2,103, Dnevnik - 2,048, Finance - 2,000, Rtvslo - 2,163 and Žurnal24 - 2,212)). In Table 8.10 we test the hypothesis on (document-based) news corpus AutoSentiNews 1.0 (automatically sentiment annotated Slovenian news published between 1st of September 2007 and 31st of January 2016, i.e., 256,567 documents (24ur - 10,009, Dnevnik - 52,417, Finance - 132,986, Rtvslo - 13,420 and Žurnal24 - 47,735)).

Table 8.9: Data and results when testing the H_4 hypothesis on (document-based) news corpus SentiNews 1.0 (manually sentiment annotated Slovenian news)

Web medium	n_{POZ}	n_{NEG}	n	p_{POZ}	p_{NEG}	p	SE	Z	$p-value$
24ur	342	841	1,183	0.289	0.711	0.589	0.032	13.368	0
Dnevnik	299	662	961	0.311	0.689	0.571	0.034	10.954	0
Finance	441	452	893	0.494	0.506	0.500	0.033	0.368	0.356
Rtvslo	284	765	1,049	0.271	0.729	0.605	0.034	13.500	0
Žurnal24	299	617	916	0.326	0.674	0.560	0.035	9.926	0

Table 8.10: Data and results when testing the H_4 hypothesis on (document-based) news corpus AutoSentiNews 1.0 (automatically sentiment annotated Slovenian news)

Web medium	n_{POZ}	n_{NEG}	n	p_{POZ}	p_{NEG}	p	SE	Z	$p-value$
24ur	2,022	4,211	6,233	0.324	0.676	0.562	0.013	26.160	0
Dnevnik	10,511	17,494	28,005	0.375	0.625	0.531	0.006	40.488	0
Finance	49,133	26,361	75,494	0.651	0.349	0.545	0.004	-79.348	1
Rtvslo	1,909	4,899	6,808	0.280	0.720	0.596	0.013	33.179	0
Žurnal24	7,216	16,183	23,399	0.308	0.692	0.573	0.007	54.739	0

The results of *Z-tests* show that the hypothesis H_4 can be confirmed with the 0.05 significance level for all web media except Finance. The proportion of negative news

is greater than the proportion of positive news (in all web media except Finance) in the retrieved news with political, business, economic and financial content from five Slovenian web media.

9 CONCLUSIONS AND FUTURE WORK

9.1 Conclusions

A rapid growth of information available on the web, such as customer feedback, competitor information, client emails, tweets, press releases, legal filings, product and engineering documents, etc., has increased the interest in the analysis of informal, subjective and opinionated web content. Especially business industry quickly realized the importance of extracting opinionated texts from the web, for the purpose of carrying quality control as well as marketing research for selling their products and services.

In the past decades, we have witnessed an exceptional development of language technologies, resources and applications. However, data scientists still encounter some challenges within their research. The most common are related to emotion and content perception, defining opinions and subjectivity properly, dealing with opinion citations, quotations, speculations and negation, detecting sarcasm and humour, as well as issues related to semantics and grammar.

In this dissertation, we describe the process of obtaining a collection of annotated web-crawled news corpora and a lexicon for sentiment analysis in Slovene, evaluate performance of sentiment based classification techniques and monitor the dynamics of sentiment in our labelled corpora.

We retrieved more than 250 thousand news with political, business, economic and financial content from five Slovenian web media (24ur, Dnevnik, Finance, Rtvslo, Žurnal24) between 1st of September 2007 and 31st of January 2016. Moreover, more than ten thousand of them we manually annotated as positive, negative and neutral on three levels of granularity (document level, paragraph level and sentence level). Six different measures, such as Cronbach's alpha, Krippendorff's alpha, Fleiss' kappa, Kendall's coefficient of concordance, as well as Pearson and Spearman correlation coefficients, were used to evaluate the process of annotation. In general, all measures indicate a good internal con-

sistency at all levels of granularity; however, their values are decreasing steadily when applying to paragraph level and sentence level of granularity. The language resources are freely available under the Creative Commons license license Attribution-ShareAlike 4.0 (see Section 4.3). Through our work, we support the open source community, in order to allow future researchers to contribute to (computational) linguistics community.

Next, we present performance evaluation of sentiment-based classification techniques of the obtained language resources. Firstly, we empirically evaluated the best approaches for two-class and three-class document-based sentiment classification of the Slovenian news texts, where different classifiers and pre-processing options were tested. We have shown that the Multinomial Naïve Bayes classifier achieves the best F1-score within the two-class (97.85%) and three-class (77.76%) document-based sentiment classification based on annotations (average scores of sentences) at the sentence-level granularity. Secondly, we applied the Multinomial Naïve Bayes classifier to estimate the sentiment of unlabelled documents, and eventually the proportions on positive, negative and neutral news in the web media. The results show that half of news is neutral, whereas the proportion of negative news is approximately twice as large as the proportion of positive news.

At last, we monitored the dynamics of sentiment in our labelled corpora. An important discovery was found when monitoring dynamics of sentiment within documents. We noticed that the sentiment is more emphasized at the beginning of news. This discovery might have a significant impact on the practice of sentiment analysis in the news.

9.2 Future Work

The language resources are relatively easy to build, therefore, we welcome other researchers, in particular representatives of other small language groups, to construct new language resources in a similar way. In the future, we plan to enrich, develop, and increase the size of the language resources, coupled with a wider range of media, and if possible, to proceed with a comprehensive manual annotation of the news.

In the past decades, we have witnessed an exceptional development of language technologies, resources and applications. However, data scientists still encounter many challenges within their research. The most common are related to emotion and content perception, defining opinions and subjectivity properly, dealing with opinion citations,

quotations, speculations and negation, detecting sarcasm and humour, as well as issues related to semantics and grammar. Here, we see a huge potential for further research.

Currently, we are focusing on modelling methods, where we can further combine topics with monitoring sentiment over time, which should lead to a novel result. In the future, we wish to use the language resources within market-oriented projects, which would encourage the demand for service, such as tracking up to date accurate information about products, services and events, as well as evaluating the results for potential clients. Also, we intend to further investigate different approaches and techniques, which can improve the achieved performances, especially within the three-class document-based sentiment classification.

10 REFERENCES

Abdul-Mageed, M. & Diab, M. T. (2011). Subjectivity and Sentiment Annotation of Modern Standard Arabic Newswire. *Proceedings of 5th Linguistic Annotation Workshop*, pp. 110–118. Association for Computational Linguistics.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O. & Passonneau, R. (2011). Sentiment Analysis of Twitter Data. *Proceedings of Workshop on Languages in Social Media*, pp. 30–38. Association for Computational Linguistics.

Aha, D. W., Kibler, D. & Albert, M. K. (1991). Instance-based Learning Algorithms. *Machine Learning 6*(1):37–66.

Alm, C. O., Roth, D. & Sproat, R. (2005). Emotions from Text: Machine Learning for Text-based Emotion Prediction. *Proceedings of conference on human language technology and empirical methods in natural language processing*, pp. 579–586. Association for Computational Linguistics.

Arhar, Š., Gorjanc, V. & Krek, S. (2007). FidaPLUS Corpus of Slovenian: The New Generation of Slovenian Reference Corpus: Its Design and Tools. *Proceedings of Corpus Linguistics Conference*.

Asheghi, N. R., Sharoff, S. & Markert, K. (2016). Crowdsourcing for Web Genre Annotation. *Language Resources and Evaluation* pp. 1–39.

Asur, S. & Huberman, B. A. (2010). Predicting the Future with Social Media. *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference*, volume *1*, pp. 492–499. IEEE.

Baccianella, S., Esuli, A. & Sebastiani, F. (2010). Sentiwordnet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of Language Resources and Evaluation Conference*, volume *10*, pp. 2200–2204.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van Der Goot, E., Halkia, M., ... Belyaeva, J. (2013). Sentiment Analysis in the News.

Banko, M. & Brill, E. (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. *Proceedings of 39th Annual Meeting on Association for Computational Linguistics*, pp. 26–33. Association for Computational Linguistics.

Baroni, M. & Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. *Proceedings of Language Resources and Evaluation Conference*, pp. 1313–1316.

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-crawled Corpora. *Language Resources and Evaluation 43*(3):209–226.

Ben-Hur, A. & Weston, J. (2010). A User's Guide to Support Vector Machines. *Data Mining Techniques for the Life Sciences* pp. 223–239.

Berginc, N. L. & Ljubešić, N. (2013). Gigafida in slWaC: Tematska Primerjava. *Slovenščina 2*(1):178–110.

Berginc, N. L., Grčar, M., Brakus, M., Erjavec, T., Holdt, Š. A., Krek, S., ... Kosem, I. (2012). *Korpusi Slovenskega Jezika Gigafida, KRES, ccGigafida in ccKRES: Gradnja, Vsebina, Uporaba*. Trojina, Zavod za Uporabno Slovenistiko.

Bermejo, P., Gámez, J. A. & Puerta, J. M. (2011). Improving the Performance of Naive Bayes Multinomial in E-mail Foldering by Introducing Distribution-based Balance of Datasets. *Expert Systems with Applications 38*(3):2072–2080.

Bermingham, A. & Smeaton, A. F. (2009). A Study of Inter-annotator Agreement for Opinion Retrieval. *Proceedings of 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 784–785. ACM.

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., ... Haley, C. S. (2015). Application of High-dimensional Feature Selection: Evaluation for Genomic Prediction in Man. *Scientific Reports 5*.

Biber, D., Egbert, J. & Davies, M. (2015). Exploring the Composition of Searchable Web: A Corpus-based Taxonomy of Web Registers. *Corpora 10*(1):11–45.

Bishop, C. (2007). *Pattern Recognition and Machine Learning.* Springer, New York.

Boiy, E. & Moens, M. F. (2009). A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts. *Information Retrieval 12*(5):526–558.

Boiy, E., Hens, P., Deschacht, K. & Moens, M. F. (2007). Automatic Sentiment Analysis in On-line Text. *Proceedings of International Conference on Electronic Publishing*, pp. 349–360.

Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. *Proceedings of 5th Annual Workshop on Computational Learning Theory*, pp. 144–152. ACM.

Bothos, E., Apostolou, D. & Mentzas, G. (2010). Using Social Media to Predict Future Events with Agent-based Markets. *IEEE Computer Society 25*(6):50–58.

Bottou, L. & Lin, C. J. (2007). Support Vector Machine Solvers. *Large Scale Kernel Machines* pp. 301–320.

Breiman, L. (2001). Random Forests. *Machine Learning 45*(1):5–32.

Burnap, P., Gibson, R., Sloan, L., Southern, R. & Williams, M. (2016). 140 Characters to Victory?: Using Twitter to Predict the UK 2015 General Election. *Electoral Studies 41*:230–233.

Burnap, P., Williams, M. L., Sloan, L., Rana, O., Housley, W., Edwards, A., ... Voss, A. (2014). Tweeting the Terror: Modelling the Social Media Reaction to the Woolwich Terrorist Attack. *Social Network Analysis and Mining 4*(1):206.

Bučar, J., Povh, J. & Žnidaršič, M. (2016). Sentiment Classification of Slovenian News Texts. *Proceedings of 9th International Conference on Computer Recognition Systems (CORES 2015)*, pp. 777–787. Springer.

Ceron, A., Curini, L. & Iacus, S. M. (2015). Using Sentiment Analysis to Monitor Electoral Campaigns Method Matters—evidence from the United States and Italy. *Social Science Computer Review 33*(1):3–20.

Chang, C. C. & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST) 2*(3):27.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., ... Wirth, R. (2000). *CRISP-DM - Step-by-step Data Mining Guide.*

Chawla, N. V., Japkowicz, N. & Kotcz, A. (2004). Special Issue on Learning from Imbalanced Data Sets. *ACM Sigkdd Explorations Newsletter 6*(1):1–6.

Cheong, M. & Lee, V. (2011). A Microblogging-based Aapproach to Terrorism Informatics: Exploration and Chronicling Civilian Sentiment and Response to Terrorism Events via Twitter. *Information Systems Frontiers 13*(1):45–59.

Church, K. W. & Mercer, R. L. (1993). Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics 19*(1):1–24.

Colbaugh, R. & Glass, K. (2010). Estimating Sentiment Orientation in Social Media for Intelligence Monitoring and Analysis. *Proceedings of International Conference on Intelligence and Security Informatics (ISI)*, pp. 135–137. IEEE.

Collins, M. & Duffy, N. (2002). New Ranking Algorithms for Parsing and Tagging: Kernels Over Discrete Structures, and the Voted Perceptron. *Proceedings of 40th Annual Meeting on Association for Computational Linguistics*, pp. 263–270. Association for Computational Linguistics.

Collomb, A., Costea, C., Joyeux, D., Hasan, O. & Brunie, L. (2014). *A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation.*

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

Cover, T. & Hart, P. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory 13*(1):21–27.

Das, S. & Chen, M. (2001). Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards. *Proceedings of Asia Pacific Finance Association Annual Conference (APFA)*, volume *35*, page 43. Bangkok, Thailand.

Davies, M. (2016). *Google's English Language Corpus, Drawn from Google Books.* Retrieved May 27, 2016, from `http://googlebooks.byu.edu/x.asp`.

De Schryver, G. M. (2002). Web for/as Corpus: A Perspective for the African Languages. *Nordic Journal of African Studies 11*(2):266–282.

Devitt, A. & Khurshid, A. (2007). Sentiment Polarity Identification in Financial News: A Cohesion-based Approach. *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics*.

Devore, J., Feldman, R. & Sanger, J. (2009). *The Text Mining Handbook*. JSTOR.

Ding, X., Liu, B. & Yu, P. S. (2008). A Holistic Lexicon-based Approach to Opinion Mining. *Proceedings of International Conference on Web Search and Data Mining*, pp. 231–240. ACM.

Durant, K. T. & Smith, M. D. (2006). Mining Sentiment Classification from Political Web Logs. *Proceedings of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Ekbal, A. & Bandyopadhyay, S. (2008). A Web-based Bengali News Corpus for Named Entity Recognition. *Language Resources and Evaluation 42*(2):173–182.

Erjavec, T. (2010). *Korpusno Jezikoslovje in Jezikovne Tehnologije*. Department of Knowledge Technologies Jožef Stefan Institute, Retrieved April 22, 2015, from `nl.ijs.si/et/teach/ung10-kj/ung10-kj-01.ppt`.

Erjavec, T. & Fišer, D. (2006). Building Slovene WordNet. *Proceedings of Language Resources and Evaluation Conference*, volume *6*, pp. 1678–1683.

Erjavec, T., Fišer, D., Krek, S. & Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. *Proceedings of Language Resources and Evaluation Conference*. Citeseer.

Erjavec, T., Ljubešić, N. & Logar, N. (2015). The slWaC Corpus of the Slovene Web. *Informatica 39*(1):35–42.

Esuli, A. & Sebastiani, F. (2006). Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of Language Resources and Evaluation Conference*, volume *6*, pp. 417–422. Citeseer.

Fellbaum, C. (1998). *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Ferguson, P., O'Hare, N., Davy, M., Bermingham, A., Sheridan, P., Gurrin, C., ... Smeaton, A. F. (2009). Exploring the Use of Paragraph-level Annotations for Sentiment Analysis of Financial Blogs.

Fišer, D., Smailović, J., Erjavec, T., Mozetič, I. & Grčar, M. (2016). Sentiment Annotation of Slovene User-generated Content. *Proceedings of Conference Language Technologies and Digital Humanities*.

Fletcher, W. H. (2001). Concordancing the Web with KWiCFinder. *Proceedings of 3rd North American Symposium on Corpus Linguistics and Language Teaching*, pp. 1–16. Citeseer.

Fletcher, W. H. (2012). Corpus Analysis of the World Wide Web. *The Encyclopedia of Applied Linguistics* .

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research 3*(Mar):1289–1305.

Freitag, D. (1998). Information Extraction from HTML: Application of a General Machine Learning Approach. *AAAI/IAAI*, pp. 517–523.

Freund, Y. & Schapire, R. E. (1999). Large Margin Classification Using the Perceptron Algorithm. *Machine Learning 37*(3):277–296.

George, D. & Mallery, P. (2001). SPSS for Windows Step by Step: A Simple Guide and Reference 10.0 Update. *Allyn and Bacon, Toronto* .

Ghani, R., Jones, R. & Mladenić, D. (2001). Mining the Web to Create Minority Language Corpora. *Proceedings of 10th International Conference on Information and Knowledge Management*, pp. 279–286. ACM.

Glavaš, G., Korenčić, D. & Šnajder, Jan (2013). Aspect-oriented Opinion Mining from User Reviews in Croatian. *ACL 2013* page 18.

Godbole, N., Srinivasaiah, M. & Skiena, S. (2007). Large-scale Sentiment Analysis for News and Blogs. volume *7*, page 21.

Guyon, I. & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research 3*(Mar):1157–1182.

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.

Habernal, I., Ptáček, T. & Steinberger, J. (2015). Supervised Sentiment Analysis in Czech Social Media. *Information Processing & Management 51*(4):532–546.

Hall, M. A. & Smith, L. A. (1998). Practical Feature Subset Selection for Machine Learning.

Harris, Z. (1954). The Structure of Language, Chapter Distributional Structure.

Hastie, T. & Tibshirani, R. (1998). Classification by Pairwise Coupling. *The Annals of Statistics 26*(2):451–471.

Hatzivassiloglou, V. & McKeown, K. R. (1997). Predicting the Semantic Orientation of Adjectives. *Proceedings of 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 174–181. Association for Computational Linguistics.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. (1998). Support Vector Machines. *IEEE Intelligent Systems and Their Applications 13*(4):18–28.

Ho, S. S., Chen, V. & Sim, C. C. (2013). The Spiral of Silence: Examining How Cultural Predispositions, News Attention, and Opinion Congruency Relate to Opinion Expression. *Asian Journal of Communication 23*(2):113–134.

Hofland, K. (2000). A Self-expanding Corpus Based on Newspapers on the Web. *Proceedings of Language Resources and Evaluation Conference*.

Hosmer, D. W., Lemeshow, S. & Sturdivant, R. (2013). *Applied Logistic Regression*, volume *398*. John Wiley & Sons.

Hsueh, P. Y., Melville, P. & Sindhwani, V. (2009). Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria. *Proceedings of NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pp. 27–35. Association for Computational Linguistics.

Hu, X., Tang, L., Tang, J. & Liu, H. (2013). Exploiting Social Relations for Sentiment Analysis in Microblogging. *Proceedings of 6th ACM International Conference on Web Search and Data Mining*, pp. 537–546. ACM.

Hundt, M., Nesselhauf, N. & Biewer, C. (2007). Corpus Linguistics and the Web. *Corpus Linguistics and the Web*, pp. 1–5. Brill.

International Journalists' Network (2008). *Negative News Harmful or Just Reality.* Retrieved June 15, 2014, from `http://ijnet.org/community/groups/10189/negative-news-harmful-or-just-reality`.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume *6*. Springer.

Joachims, T. (1998). *Text Categorization with Support Vector Machines: Learning with Many Relevant Features.* Springer.

Joshi, A., Mishra, A., Senthamilselvan, N. & Bhattacharyya, P. (2014). Measuring Sentiment Annotation Complexity of Text. *Proceedings of 54th Annual Meeting of Association for Aomputational*, pp. 36–41.

Kadunc, K. & Robnik-Šikonja, M. (2016). Analiza Mnenj s Pomočjo Strojnega Učenja in Slovenskega Leksikona Sentimenta pp. 83–89.

Karegowda, A. G., Manjunath, A. S. & Jayaram, M. A. (2010). Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection. *International Journal of Information Technology and Knowledge Management 2*(2):271–277.

Kätsyri, J., Kinnunen, T., Kusumoto, K., Oittinen, P. & Ravaja, N. (2016). Negativity Bias in Media Multitasking.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C. & Murthy, K. R. K. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation 13*(3):637–649.

Kibriya, A. M., Frank, E., Pfahringer, B. & Holmes, G. (2004). Multinomial Naive Bayes for Text Categorization Revisited. *Australasian Joint Conference on Artificial Intelligence*, pp. 488–499. Springer.

Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics 29*(3):333–347.

Kim, S. & Hovy, E. (2004). Determining the Sentiment of Opinions. *Proceedings of 20th International Conference on Computational Linguistics*, page 1367. Association for Computational Linguistics.

Kohavi, R. (1995). The Power of Decision Tables. *European Conference on Machine Learning*, pp. 174–189. Springer.

Kotsiantis, S. (2007). Supervised Machine Learning: A Review of Classification Techniques.

Kouloumpis, E., Wilson, T. & Moore, J. D. (2011). Twitter Sentiment Analysis: The Good the Bad and the Omg! *Proceeding of International AAAI Conference on Weblogs and Social Media 11*:538–541.

Kovačič, A. (2012). Using RSS to Aggregate News on the Internet – Case Study Media Tone. *Proceedings of 4th International Conference on Information Technologies and Information Society (ITIS 2012)*, pp. 1–6.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*. Sage.

Kushal, D., Lawrence, S. & Pennock, D. M. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. *Proceedings of 12th International Conference on World Wide Web*, pp. 519–528. ACM.

Kučera, H. & Francis, W. N. (1967). *Computational Analysis of Present-day American English*. Brown University Press.

Lancaster, F. W. (1968). *Information Retrieval Systems; Characteristics, Testing, and Evaluation*. Wiley.

Landis, R. J. & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics* pp. 159–174.

Landwehr, N., Hall, M. & Frank, E. (2005). Logistic Model Trees. *Machine Learning 59*(1-2):161–205.

Leskovec, J., Rajaraman, A. & Ullman, J. D. (2014). *Mining of Massive Datasets*. Cambridge University Press.

Lewis, D. D. (1998). Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. *European Conference on Machine Learning*, pp. 4–15. Springer.

Liaw, A. & Wiener, M. (2002). Classification and Regression by RandomForest. *R News 2*(3):18–22.

Likert, R. (1932). *A Technique for the Measurement of Attitudes*.

Liu, B. (2010). Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing 2*:627–666.

Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage data*. Springer, NY.

Liu, B., Hu, M. & Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. *Proceedings of 14th International Conference on World Wide Web*, pp. 342–351. ACM.

Ljubešić, N. & Erjavec, T. (2011). HrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. *Text, Speech and Dialogue - 14th International Conference (TSD 2011), Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, Lecture Notes in Computer Science, pp. 395–402. Springer.

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., ... Pirrelli, V. (2014). The PAISA Corpus of Italian Web Texts. *Proceedings of 9th Web as Corpus Workshop (WaC-9)*, pp. 36–43.

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, volume *999*. MIT Press.

Marcus, M. P., Marcinkiewicz, M. A. & Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics 19*(2):313–330.

MARKUSQ (2017). *Sentiment Analysis for Twitter in Python*. Retrieved April 3, 2013, from `http://stackoverflow.com/questions/573768/sentiment-analysis-for-twitter-in-python`.

Martinc, R. (2013). *Measuring Sentiment on Social Network Twitter: Designing a Tool and Evaluation*. University of Ljubljana, Faculty of Social Sciences.

MathWorks (2017). *Support Vector Machines for Binary Classification*. Retrieved April 23, 2017, from `https://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html?requestedDomain=www.mathworks.com`.

McCallum, A. & Nigam, K. (1998). A Comparison of Event Models for Naïve Bayes Text Classification. *AAAI-98 Workshop on Learning for Text Categorization*, volume *752*, pp. 41–48. Citeseer.

Mejova, Y. (2009). Sentiment Analysis: An Overview.

Melville, P., Gryc, W. & Lawrence, R. D. (2009). Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification. *Proceedings of 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1275–1284. ACM.

Meyer, D., Leisch, F. & Hornik, K. (2003). The Support Vector Machine Under Test. *Neurocomputing 55*(1):169–186.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM 38*(11):39–41.

Miner, G., Delen, D., Elder, E., Fast, A., Hill, T. & Nisbet, R. A. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press.

Mitchell, T. (1997). Machine Learning.

Mohammad, S. M., Kiritchenko, S. & Zhu, X. (2013). NRC-Canada: Building the State-of-the-art in Sentiment Analysis of Tweets.

Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T. & Ureña-López, L. A. (2014). Ranked Wordnet Graph for Sentiment Polarity Classification in Twitter. *Computer Speech & Language 28*(1):93–107.

Mozetič, I., Grčar, M. & Smailović, J. (2016). Multilingual Twitter Sentiment Classification: The Role of Human Annotators. *PloS One 11*(5):1–26.

Nakov, P., Rosenthal, S., Kiritchenko, S., Mohammad, S. M., Kozareva, Z., Ritter, A., ... Zhu, X. (2016). Developing a Successful SemEval Task in Sentiment Analysis of Twitter and Other Social Media Texts. *Language Resources and Evaluation 50*(1):35–65.

Nguyen, L. T., Wu, P., Chan, W., Peng, W. & Zhang, Y. (2012). Predicting Collective Sentiment Dynamics from Time-series Social Media. *Proceedings of 1st International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 6. ACM.

Nielsen, F. Å. (2011). A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. *Proceedings of 1st Workshop on Making Sense of Microposts: Big Things Come in Small Packagess*, pp. 93–98.

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., ... Smeaton, A. F. (2009). Topic-dependent Sentiment Analysis of Financial Blogs. *Proceedings of 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, pp. 9–16. ACM.

Oxford Dictionaries (2017). *Definition of Corpus in English.* Retrieved July 13, 2016, from `https://en.oxforddictionaries.com/definition/corpus`.

Paliouras, G., Papatheodorou, C., Karkaletsis, V. & Spyropoulos, C. D. (2002). Discovering User Communities on the Internet Using Unsupervised Machine Learning Techniques. *Interacting with Computers 14*(6):761–791.

Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval 2*(1-2):1–135.

Pang, B., Lee, L. & Vaithyanathan, S. (2002). Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. *Proceedings of ACL-02 Conference on Empirical Methods in Natural Language Processing*, volume *10*, pp. 79–86. Association for Computational Linguistics.

Pecina, P., Toral, A., Papavassiliou, V., Prokopidis, P., Tamchyna, A., Way, A., ... van Genabith, J. (2015). Domain Adaptation of Statistical Machine Translation with Domain-focused Web Crawling. *Language Resources and Evaluation 49*(1):147–193.

Peng, J., Zeng, D. D., Zhao, H. & Wang, F. (2010). Collaborative Filtering in Social Tagging Systems Based on Joint Item-tag Recommendations. *Proceedings of 19th ACM International Conference on Information and Knowledge Management*, pp. 809–818. ACM.

Perez-Rosas, V., Banea, C. & Mihalcea, R. (2012). Learning Sentiment Lexicons in Spanish. *Proceedings of Language Resources and Evaluation Conference*, volume *12*, page 73.

Platt, J. C. (1999). Fast Training of Support Vector Machines Using Sequential Minimal Optimization. *Advances in Kernel Methods* pp. 185–208.

Polanyi, L. & Zaenen, A. (2006). Contextual Valence Shifters. *Computing Attitude and Affect in Text: Theory and Applications*, pp. 1–10. Springer.

Poplack, S. (1989). The Care and Handling of a Megacorpus: The Ottawa-Hull French Project. *Language Change and Variation 4*.

Pustejovsky, J. & Stubbs, A. (2012). *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc.

Qi, X. & Davison, B. D. (2009). Web Page Classification: Features and Algorithms. *ACM Computing Surveys (CSUR) 41*(2):12.

Quinlan, J. R. (1993). C4. 5: Programming for Machine Learning. *Morgan Kauffmann* page 38.

Quinlan, J. R. (2014). *C4. 5: Programs for Machine Learning*. Elsevier.

Rau, L. F., Jacobs, P. S. & Zernik, U. (1989). Information Extraction and Text Summarization Using Linguistic Knowledge Acquisition. *Information Processing & Management 25*(4):419–428.

Rayson, P., Walkerdine, J., Fletcher, W. H. & Kilgarriff, A. (2006). Annotated Web as Corpus. *Proceedings of 2nd International Workshop on Web as Corpus*, pp. 27–33. Association for Computational Linguistics.

Ready, P. & Wintz, P. (1973). Information Extraction, SNR Improvement, and Data Compression in Multispectral Imagery. *IEEE Transactions on Communications 21*(10):1123–1131.

Reis, J., Benevenuto, F., Vaz de Melo, P., Prates, R., Kwak, H. & An, J. (2015). Breaking the News: First Impressions Matter on Online News.

Rennie, J. D., Shih, L., Teevan, J. & Karger, D. R. (2003). Tackling the Poor Assumptions of Naive Bayes Text Classifiers. *ICML*, volume *3*, pp. 616–623. Washington DC).

Renouf, A. (2003). WebCorp: Providing a Renewable Data Source for Corpus Linguists. *Language and Computers 48*(1):39–58.

Resnik, P. & Smith, N. A. (2003). The Web as a Parallel Corpus. *Computational Linguistics 29*(3):349–380.

Robb, T. (2003). Google as a Quick 'n Dirty Corpus Tool. *The Electronic Journal for English as a Second Language 7*(2).

Rumsey, D. J. & Unger, D. (2015). *U Can: Statistics for Dummies*. John Wiley.

Rundell, M. (2000). The Biggest Corpus of All. *Humanising Language Teaching 2*(3).

Salton, G. & McGill, M. J. (1986). Introduction to Modern Information Retrieval .

Sankoff, D. & Sankoff, G. (1973). Sample Survey Methods and Computer-assisted Analysis in the Study of Grammatical Variation. *Canadian Languages in Their Social Context* pp. 7–63.

Schäfer, R., Barbaresi, A. & Bildhauer, F. (2014). Focused Web Corpus Crawling. *Proceedings of 9th web as corpus workshop (WaC-9)*, pp. 9–15.

Schatten, M., Seva, J. & Đurić, B. O. (2015). An Introduction to Social Semantic Web Mining & Big Data Analytics for Political Attitudes and Mentalities Research. *European Quarterly of Political Attitudes and Mentalities 4*(1):40.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys (CSUR) 34*(1):1–47.

Sharma, A. & Dey, S. (2012). Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis. *IJCA Special Issue on Advanced Computing and Communication Technologies for HPC Applications 3*:15–20.

Smailović, J. (2014). *Sentiment Analysis in Streams of Microblogging Posts*. Ph.D. thesis, PhD Thesis, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia.

Smailović, J., Grčar, M., Lavrač, N. & Žnidaršič, M. (2013). Predictive Sentiment Analysis of Tweets: A Stock Market Application. *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, pp. 77–88. Springer.

Snow, R., O'Connor, B., Jurafsky, D. & Ng, A. Y. (2008). Cheap and Fast—But is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 254–263. Association for Computational Linguistics.

Snyder, B. & Barzilay, R. (2007). Multiple Aspect Ranking Using the Good Grief Algorithm. *HLT-NAACL*, pp. 300–307.

Spousta, M. (2006). Web as a Corpus. *WDS'06 Proceedings of Contributed Papers. Prague. Czech Republic: Matfyzpress* pp. 179–84.

Steinwart, I. & Christmann, A. (2008). *Support Vector Machines*. Springer Science & Business Media.

Stone, G. C. & Grusin, E. (1984). Network TV as the Bad News Bearer. *Journalism and Mass Communication Quarterly 61*(3):517–592.

Stone, P. J., Dunphy, D. C. & Smith, M. S. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Strapparava, C. & Mihalcea, R. (2007). Semeval-2007 task 14: Affective Text. *Proceedings of 4th International Workshop on Semantic Evaluations*, pp. 70–74. Association for Computational Linguistics.

Student (1908). The Probable Error of a Mean. *Biometrika* pp. 1–25.

Sumner, M., Frank, E. & Hall, M. (2005). Speeding Up Logistic Model Tree Induction. *European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 675–683. Springer.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. (2011). Lexicon-based Methods for Sentiment Analysis. *Computational Linguistics 37*(2):267–307.

Taulé, M., Martí, M. A. & Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. *Proceedings of Language Resources and Evaluation Conference.*

Thelwall, M., Buckley, K. & Paltoglou, G. (2012). Sentiment Strength Detection for the Social Web. *Journal of American Society for Information Science and Technology 63*(1):163–173.

Thet, T. T., Na, J. C., Khoo, C. S. G. & Shakthikumar, S. (2009). Sentiment Analysis of Movie Reviews on Discussion Boards Using a Linguistic Approach. *Proceedings of 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, pp. 81–84. ACM.

Tourani, P., Jiang, Y. & Adams, B. (2014). Monitoring Sentiment in Open Source Mailing Lists-exploratory Study on the Apache Ecosystem. *Proceedings of Conference of the Center for Advanced Studies on Collaborative Research (CASCON)*, pp. 74–95.

Trussler, M. & Soroka, S. (2014). Consumer Demand for Cynical and Negative News Frames. *The International Journal of Press/Politics 19*(3):360–379.

Tumasjan, A., Sprenger, T. O., Sandner, P. G. & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment. *Procedeedings of International Conference on Weblogs and Social Media 10*(1):178–185.

Turney, P. D. (2002). Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of 40th Annual Meeting on Association for Computational Linguistics*, pp. 417–424. Association for Computational Linguistics.

Verdonik, D., Kosem, I., Vitez, A. Z., Krek, S. & Stabej, M. (2013). Compilation, Transcription and Usage of a Reference Speech Corpus: The Case of the Slovene Corpus GOS. *Language Resources and Evaluation 47*(4):1031–1048.

Vinkers, C. H., Tijdink, J. K. & Otte, W. M. (2015). Use of Positive and Negative Words in Scientific PubMed Abstracts Between 1974 and 2014: Retrospective Analysis. *BMJ 351*:h6467.

Vo, D. & Zhang, Y. (2015). Target-dependent Twitter Sentiment Classification with Rich Automatic Features. *Proceedings of 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pp. 1347–1353.

Volk, M. (2002). Using the Web as Corpus for Linguistic Research. *Catcher of the Meaning. Pajusalu, R., Hennoste, T.(Eds.). Dept. of General Linguistics 3.*

Von Ahn, L. & Dabbish, L. (2004). Labeling Images with a Computer Game. *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, pp. 319–326. ACM.

Žgank, A., Vitez, A. Z. & Verdonik, D. (2014). The Slovene BNSI Broadcast News Database and Reference Speech Corpus GOS: Towards the Uniform Guidelines for Future Work. *Proceedings of Language Resources and Evaluation Conference.*

Wallis, S. & Nelson, G. (2001). Knowledge Discovery in Grammatically Analysed Corpora. *Data Mining and Knowledge Discovery 5*(4):305–335.

Wang, G. Y., Yu, H. & Yang, D. C. (2002). Decision Table Reduction Based on Conditional Information Entropy. *Chinese Journal of Computers 25*(7):759–766.

Wang, X., Gerber, M. S. & Brown, D. E. (2012). Automatic Crime Prediction Using Events Extracted from Twitter Posts. *SBP 12*:231–238.

Web of Science (2016). *Number of Scientific Publications that Mention Sentiment Analysis.* Retrieved February 23, 2016, from `webofknowledge.com`.

Wiebe, J. & Riloff, E. (2005). Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. *Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 486–497. Springer.

Wiebe, J., Wilson, T., Bruce, R., Bell, M. & Martin, M. (2004). Learning Subjective Language. *Computational Linguistics 30*(3):277–308.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin 1*(6):80–83.

Wilson, T., Wiebe, J. & Hoffmann, P. (2009). Recognizing Contextual Polarity: An Exploration of Features for Phrase-level Sentiment Analysis. *Computational Linguistics 35*(3):399–433.

Witten, I. H., Frank, E. & Hall, M. A. (2013). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

World Wide Web Technology Surveys (2017). *Usage of Content Languages for Websites*. Retrieved January 2, 2017, from `http://w3techs.com/technologies/overview/content_language/all`.

Wright, A. (2009). Mining the Web for Feelings, Not Facts. *New York Times 24*.

Wu, S., Witten, I. H. & Franken, M. (2010). Utilizing Lexical Data from a Web-derived Corpus to Expand Productive Collocation Knowledge. *ReCALL 22*(01):83–102.

Xu, J. L. (2000). Multilingual Search on the World Wide Web. *Proceedings of Hawaii International Conference on System Sciences (HICSS)*, volume *33*.

Yang, Y. & Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. *Icml*, volume 97, pp. 412–420.

Zheng, Z., Wu, X. & Srihari, R. (2004). Feature Selection for Text Categorization on Imbalanced Data. *ACM Sigkdd Explorations Newsletter 6*(1):80–89.

SUBJECT INDEX

INDEX OF AUTHORS

BIBLIOGRAPHY

Bučar, J. & Povh, J. (2013). A KNN Based Algorithm for Text Categorization. *Proceed--ings of the 12th International Symposium on Operational Research*, pp. 367-372.

Bučar, J. & Povh, J. (2013). Sentiment Analysis in Web Text: An Overview. *Proceed--ings of the 7th European Computing Conference (EEC'13)*, pp. 154-159.

Bučar, J., Zidar, D. & Mertik, M. (2013). Web Clipping Case Study : Digital Trace of SMEs: (Preliminary Work). *Proceedings of the 4th RapidMiner Community Meeting and Conference: (RCOMM 2013)*, pp. 219-224.

Bučar, J., Zidar, D. & Mertik, M. (2013). Case study web clipping : the preliminary work. *Proceedings of the 5th International Conference on Information Technologies and Information Society and Conference: (ITIS 2013)*, pp. 36-39.

Bučar, J. & Povh, J. (2014). Case Study: Web Clipping and Sentiment Analysis.*Proceed--ings of the 6th International Conference on Information Technologies and Info--rmation Society (ITIS 2014)*, pp. 75-80.

Bučar, J. (2015). A research voucher case study : web clipping. *Applications of ICT in social sciences*, pp. 89-98, PL Academic Research.

Bučar, J., Povh, J. & Dobrovoljc, A. (2015). Web Clipping and Sentiment Analysis of Slovenian news articles. *Social Sciences via Network Analysis and Computation*, pp. 11-20, Peter Lang AG.

Barišić, R., & Bučar, J. (2015). Machine learning in classification of news portal articles. *Social Sciences via Network Analysis and Computation*, pp. 21-27, Peter Lang AG.

Bučar, J., Povh, J. & Žnidaršič, M. (2016). Sentiment Classification of the Slovenian News Texts. *Proceedings of the 9th International Conference on Computer Recognition Systems (CORES 2015)*, pp. 777-787, Springer.

Bučar, J., Povh, J. & Žnidaršič, M. (2017). Annotated News Corpora and a Lexicon for Sentiment Analysis in Slovene. *Language Resources and Evaluation*, pp. 1-23. Springer. (Date submitted: 1 November 2016, Status: Under review)

BIOGRAPHY

Jože Bučar was a Young Researcher at the Laboratory of Data Technologies at the Faculty of Information Studies between December 2012 and May 2016. Within his research at the Faculty of Information Studies his primary research interests were applications of data mining as well as quantitative and qualitative analysis, algorithm development, machine learning and statistical analysis of complex networks. Jože Bučar studied geodesy at the Faculty of Civil and Geodetic Engineering at University of Ljubljana, Slovenia, where he received M.Sc. in 2011. He is currently working as a Senior Adviser at the Real Estate Mass Valuation System, Surveying and Mapping Authority of the Republic of Slovenia.

KLASIFIKACIJA SPLETNIH BESEDIL NA OSNOVI IZRAŽENOSTI SENTIMENTA

UVOD

Vedno večje zanimanje za učinkovito analizo neformalnih, subjektivnih in s sentimentom izraženih spletnih besedil je pripomoglo k izjemnemu razvoju na področju analize sentimenta. Število znanstvenih publikacij se je na temo analize sentimenta od leta 2010 hitro in enakomerno povečevalo (glej Sliko 1.2). Mnogo člankov je bilo objavljenih na temo zaznavanja čustev (razpoloženja) v besedilnih sporočilih (Alm, Roth & Sproat, 2005; Thelwall, Buckley, Paltoglou, Cai & Kappas, 2012), napovedovanja rezultatov volitev na podlagi sporočil na vse bolj priljubljeni mikroblogerski platformi Twitter in drugih družbenih omrežjih (Tumasjan, Sprenger, Sandner & Welpe, 2010; Burnap, Gibson, Sloan, Southern & Williams, 2016), napovedovanja dogodkov in izidov v prihodnosti (Asur & Huberman, 2010; Bothos, Apostolou & Mentzas, 2010), kot tudi na temo svetovnega boja proti terorizmu itd. (Cheong & Lee, 2011; Wang, Gerber & Brown, 2012; Burnap in drugi, 2014).

Znanstveniki (na področju lingvistike, podatkovnega rudarjenja in drugih področjih), ki se ukvarjajo z analizo besedil, si prizadevajo k boljšemu računalniškemu razumevanju svetovnih jezikov. Zato ni presenetljivo, da sta se razpoložljivost in uporaba jezikovnih virov v zadnjih letih znatno povečali ravno zaradi potreb računalniškega jezikoslovja. Večina jezikovnih virov je na voljo v angleškem jeziku, vendar pa se zanimanje po virih v drugih jezikih hitro povečuje.

Vsebina te disertacije se nanaša na analizo sentimenta in zaznavanje sentimenta v spletnih besedilih. V disertaciji je podrobno opisan postopek izdelave označenih korpusov, tj. zbirke raznih korpusov spletnih novic (besedil) v slovenskem jeziku. Predstavljen je postopek izdelave leksikona za analizo sentimenta v slovenščini, uporabnost in vrednotenje

rezultatov na razvitih jezikovnih virih. Viri so bili pridobljeni s pomočjo spletnih pajkov razvitih izključno za ta namen, in sicer v več poskusih med letoma 2013 in 2016. Vsebujejo označena besedila petih slovenskih spletnih medijev s politično, gospodarsko, poslovno in finančno vsebino, ki so bila objavljena med 1. septembrom 2007 in 31. januarjem 2016. Viri so prosto dostopni pod pogoji, določenimi v Sekciji 4.3.

V okviru slovenskega prevoda te disertacije bodo v nadaljevanju predstavljeni osnovni koncepti, uporaba in cilji analize sentimenta. Pojasnjena bo motivacija, raziskovalne hipoteze in metodologija. Nazadnje bodo predstavljeni še cilji, znanstveni doprinosi raziskave in zaključek.

OBSEŽEN POVZETEK

Predstavitev osnovnih konceptov

V zadnjih desetletjih se je uporaba jezikovnih virov in tehnologij bistveno spremenila, predvsem zaradi pojava svetovnega spleta. Splet, ki doživlja izreden uspeh tudi s podporo globalnih in lokalnih medijev, je prevladujoči medij za oglaševanje, deljenje in pridobivanje informacij. Naraščajoče zanimanje njegovih uporabnikov je botrovalo k njegovemu hitremu razvoju, vključevanju vedno novih procesov, zmogljivosti in funkcionalnosti. Danes je splet nepredstavljiva množica spletnih strani in hkrati tudi ogromno skladišče strukturiranih in nestrukturiranih (prosto dostopnih) podatkov.

Strukturiran tip podatkov na spletu je enostavnejši za računalniško obdelavo. Raziskovalci si prizadevajo poiskati nove vzorce in trende prav na podlagi dokumentov, ki vsebujejo nestrukturirane podatke. Ogromno količino podatkov najenostavneje pridobimo s pomočjo spletnih pajkov, ki sistematično iščejo vsebine po spletu in pridobivajo ustrezne informacije. Vedno več blogov, spletnih strani, novičarskih skupin, forumov, klepetalnic itd. ljudem omogoča, da lahko na hiter in enostaven način izražajo svojega mnenja in izkušenje o izdelkih, storitvah in dogodkih. Poleg tega je človeštvo še vedno lačno znanja, ki izhaja iz pridobljenih informacij.

Korpus

Beseda *korpus* izhaja iz latinske besede za telo (angl. *body*). V splošnem korpus predstavlja zbirko dokumentov v pisni, zvočni ali drugi računalniško berljivi obliki, in se uporablja za raziskave naravnih jezikov (npr. prepoznavanje govora in strojnega prevajanja v računalniški lingvistiki, razvoj jezikovnih orodij, gradnjo leksikonov in drugih jezikovnih virov, gradnjo slovnice in drugih jezikovnih struktur). Običajno so opremljeni z označbami, kot npr. z osnovnimi oblikami besed (lemami), skladenjskimi in oblikoskladenjskimi označbami in drugimi. Poznamo več vrst korpusov. Glede na velikost in namen jih delimo na referenčne, govorne, specializirane, vzorčne, vzporedne in druge.

Analiza sentimenta in klasifikacija

Besedila lahko po njihovi naravi delimo na dejstva in mnenja. Dejstva so objektivni izrazi o entitetah, dogodkih in njihovih lastnostih, podatkih ali stanju obstoječih, opazovanih ali splošno znanih entitetah, ki so se zgodile oziroma so potrjene do te mere, da jih štejemo za resnične. Mnenja so subjektivni izrazi, ki izražajo čustva, ocene ali občutke ljudi do subjektov, dogodkov in njihovih lastnosti. Objektivni stavek izraža dejanske, realne informacije o svetu, medtem ko subjektivni stavek izraža osebne občutke in prepričanja. Tako je stavek *"Dražji Dell UltraSharp U3014 30" ima boljšo ločljivost zaslona od Dell UltraSharp U2412M 24""* objektiven, medtem ko stavek *"Cockta je boljšega okusa kot Coca-Cola"* predstavlja subjektivno izjavo.

Rudarjenje besedil je ena izmed zanimivih jezikovnih tehnologij na področju analize podatkov. Obravnava vrsto tehnologij za analizo in obdelavo polstrukturiranih in nestrukturiranih besedil. Na začetku so se raziskovalci ukvarjali predvsem z različnimi oblikami pridobivanja in povzemanja informacij, kot na primer tvorjenje povzetkov in združevanje dokumentov v posamezne skupine (Lancaster, 1968; Ready & Wintz, 1973; Salton & McGill, 1986; Rau, Jacobs & Zernik, 1989). Običajne naloge rudarjenja besedil vključujejo klasifikacijo in kategorizacijo besedil, odkrivanje in zaznavanje vsebine (teme) besedil, analizo sentimenta, povzemanje informacij in preučevanje odnosov med entitetami v besedilih. Miner in drugi (2012) so razdelili področje rudarjenja besedil v sedem skupin, in sicer v iskanje in pridobivanje informacij, klasifikacijo dokumentov, katego-

rizacijo dokumentov, spletno rudarjenje, ekstrakcijo informacij, računalniško obdelavo naravnega jezika in ekstrakcijo koncepta.

Priljubljenost družabnih omrežij je povečala zanimanje za analizo sentimenta (Wright, 2009). Analiza sentimenta predstavlja zahtevne procese pri obdelavi naravnega jezika in izzivih rudarjenja besedil, ter hkrati združuje znanstvenike iz različnih področij, kot so računalniška lingvistika, podatkovno rudarjenje, računalništvo, strojno učenje, teorija grafov, nevronske mreže, sociologija in psihologija. Namen analize sentimenta je identifikacija, ekstrakcija in določitev sentimenta izvornega besedila. Subjektivnost izjave se lahko določi na podlagi presoje ali vrednotenja, zaznave čustvenega stanje vira izjave ali čustvenega stanja v procesu komuniciranja, s katerim vir želi vplivati na mnenje ali odločitev ljudi.

Pri klasifikaciji sentimenta v besedilih gre za nadzorovano učenje, kjer imamo običajno tri kategorije besedil s pozitivnim, negativnim in nevtralnim sentimentom (Pang, Lee & Vaithyanathan, 2002; Melville, Gryc & Lawrence, 2009; Taboada, Brooke, Tofiloski, Voll & Stede, 2011). Klasifikacija sentimenta je ena izmed najpogosteje uporabljenih tehnik, ki jih uporabljamo pri rudarjenju besedil. Cilj je, da še neoznačenim besedilom dodelimo znani nabor kategorij, pri čemer uporabimo model, ki vključuje lastnosti besedil z označbami. Peng in soavtorji (2010) so ugotovili, da ljudje izražamo šest primarnih čustev, in sicer ljubezen, veselje, presenečenost, žalost, strah in jezo, ki so lahko različnih intenzitet ter jih je mogoče nadalje razdeliti na sekundarna in terciarna čustva. Napredni modeli stremijo k zaznavi tudi bolj kompleksnih čustvenih stanj, kot so naklonjenost, zadovoljstvo, ponos, obžalovanje, ljubosumnost, vznemirjenost in drugi.

Analiza sentimenta se sooča s številnimi izzivi. Eden od glavnih razlogov za pomanjkanje raziskav o izraženosti sentimenta je dejstvo, da pred nastankom spleta ni bilo na voljo veliko besedil, ki bi izražala sentiment. Največji izzivi predstavljajo zaznavo subjektivnih besedil, označevanje besedil, določevanje sentimenta, odkrivanje negacije, sarkazma, humorja, citatov, špekulacij in drugih izzivov, povezanih z zaznavanjem čustev in vsebin. Ljudje, ki izražajo svoja mnenja na spletu, pogosto ne posvečajo veliko pozornosti slovnici. Zato ni neobičajno, da spreminjajo, podvajajo in izpuščajo črke, prekomerno reagirajo, uporabljajo izraze s prenesenimi pomeni, uporabljajo sleng, superlative, okrajšave, velike črke, ločila (npr. klicaje) itd.

V okviru naše študije bomo sledili CRISP-DM postopku (Chapman in drugi, 2000),

standardiziranemu postopku v podatkovnem rudarstvu. CRISP-DM postopek podatkovnega rudarjenja je sestavljen iz naslednjih šestih faz: organizacijskega razumevanja, razumevanja podatkov, priprave podatkov, modeliranja, vrednotenja in implementacije.

Naša raziskava bo temeljila na podatkih spletnih novic v slovenskem jeziku, ki še niso bili uporabljeni za namene analize sentimenta. Izdelali bomo orodja za njihovo pridobivanje ter spletno aplikacijo za ročno označevanje novic. S pomočjo spletnih pajkov bomo iz digitalnih arhivov petih slovenskih spletnih medijev pridobili vsebino in metapodatke novic (naslov, vsebino brez komentarjev, datum, avtorja, URL naslov in ključne besede) s politično, gospodarsko, poslovno in finančno vsebino. Pridobljene podatke bomo najprej očistili ter pripravili v obliko, ki bo primerna za označevanje in nadaljnjo analizo. Iz celotne populacije pridobljenih novic bomo izbrali naključen stratificiran vzorec s približno 2.000 dokumenti na spletni medij. Vzorec s približno 10.000 novicami bo ročno označen s strani več označevalcev na treh nivojih granulacije, in sicer na nivoju dokumentov, odstavkov in stavkov. Na podlagi krajšega usposabljanja in podrobnih navodil bodo označevalci neodvisno drug od drugega označili vzorec novic. Označevalci bodo na lestvici od 1 do 5 (1 - zelo negativno, 2 - negativno, 3 - nevtralno, 4 - pozitivno in 5 - zelo pozitivno) izrazili svoje občutke, ki jih bodo občutili po prebrani vsebini. Procesu označevanja bo sledil izračun stopnje ujemanja označevanja med različnimi označevalci. Na ta način bomo pridobili ročno označene korpuse v slovenskem jeziku. Označeni korpusi bodo nato služili za vrednotenje klasifikacijskih tehnik v procesu testiranja različnih tehnik nadzorovanega strojnega učenja. V okviru naše raziskave bomo uporabili najpogosteje uporabljene metode ter jih med seboj primerjali.

Obstaja več pristopov, ki jih je mogoče uporabiti za analizo sentimenta, pristopi, ki temeljijo na leksikonih, jezikoslovni pristopi in pristopi z uporabo metod nadzorovanega strojnega učenja (Thelwall, Buckley, Paltoglou, Cai & Kappas, 2012). Pristopi z uporabo metod strojnega učenja so najpogosteje uporabljeni pristopi na področju analize sentimenta (Freitag, 1998; Pang, Lee & Vaithyanathan, 2002; Sebastiani, 2002; Kotsiantis, 2007; Boiy & Moens, 2009; Witten, Frank & Hall, 2013). Zaradi nekaterih omejitev leksikalnih in jezikoslovnih pristopov smo ugotovili, da so za analizo sentimenta novic v slovenskem jeziku najbolj primerni ravno pristopi nadzorovanega strojnega učenja, zato jih bomo uporabili v naši študiji.

Klasifikacijo besedil lahko izvajamo na različnih nivojih, na nivoju dokumentov,

stavkov, besed itd. Klasifikacija na nivoju dokumentov stremi k učinkoviti določitvi sentimenta na ravni dokumentov, medtem ko klasifikacija na nivoju stavkov oziroma besed skuša učinkovito določiti sentiment na ravni stavkov oziroma besed. Večina študij se osredotoča predvsem na klasifikacijo sentimenta na nivoju dokumentov, pri čemer večji izziv predstavljata ustrezen izbor značilk in izbor klasifikacijskih metod. Pri postopku izbire značilk izbiramo podmnožico ustreznih značilk, ki bi jih radi uporabili pri gradnji našega modela. V splošnem se postopek uporablja iz treh razlogov: (i) poenostavitve modelov, tako da jih je lažje interpretirati, (ii) zmanjšanja časovne zahtevnosti v okviru strojnega učenja, (iii) izboljšanja generalizacije na račun nižje stopnje prilagajanja podatkom. V naših eksperimentih bomo uporabili implementacije rešitev znotraj orodja za strojno učenje WEKA (Witten, Frank & Hall, 2013). Z uporabo tega orodja lahko pretvorimo besedila v niz značilk, ki predstavljajo pogostost pojava značilnih besed v besedilih. Izbira značilk je odvisna od tokanizacije, lematizacije, krnjenja, preoblikovanja velikih črk v male, odstranitve seznama blokiranih besed, frekvence besed, določitve najmanjšega števila frekvence besed, normalizacije dolžine dokumentov, števila značilk, za katere želimo, da ustrezno opisujejo naš model, ter drugih.

Klasifikacijske metode, validacija in vrednotenje

Metode strojnega učenja uporabljajo različne učne algoritme za določanje sentimenta v besedilih. Znanstveniki pri svojem delu najpogosteje uporabljajo nadzorovane metode, pri katerih je potrebno podatke ročno označiti. Nabor podatkov je potrebno razdeliti na učno množico, na podlagi katere zgradimo model, in testno množico, ki nam služi za klasifikacijo besedil in vrednotenje rezultatov učinkovitosti klasifikacije. V okviru naše raziskave smo uporabili nadzorovane metode strojnega učenja.

Z namenom, da z vidika učinkovitosti in časovne zahtevnosti izberemo najprimernejši klasifikator za klasifikacijo spletnih besedil v slovenskem jeziku, smo testirali več najpogosteje uporabljenih metod, kot so: Metoda k najbližjih sosedov (KNN) (Cover & Hart, 1967; Aha, Kibler & Albert, 1991), Naivni (večrazsežnostni) Bayes (NBM) (Lewis, 1998; McCallum & Nigam, 1998; Rennie, Shih, Teevan & Karger, 2003; Kibriya, Frank, Pfahringer & Holmes, 2004; Bermejo, Gámez & Puerta, 2011), Metoda podpornih vektorjev (SVM-poly and SVM-lin) (Boser, Guyon & Vapnik, 1992; Cortes & Vapnik, 1995;

Hastie & Tibshirani, 1998; Hearst, Dumais, Osuna, Platt & Scholkopf, 1998; Platt, 1999; Keerthi, Shevade, Bhattacharyya & Murthy, 2001; Meyer, Leisch & Hornik, 2003; Steinwart & Christmann, 2008; Chang & Lin, 2011), Naključni gozdovi (RF) (Breiman, 2001; Liaw & Wiener, 2002), C4.5 (Quinlan, 1993, 2014), Odločitvene tabele (DT) (Kohavi, 1995; Wang, Yu & Yang, 2002), Enostavna logistična regresija (SLR) (Landwehr, Hall & Frank, 2005; Sumner, Frank & Hall, 2005; Hosmer, Lemeshow & Sturdivant, 2013), Perceptron - nevronske mreže (VP) (Freund & Schapire, 1999; Collins & Duffy, 2002). Na podlagi empiričnih testiranj (glej Sekciji 5.1 in 5.3) smo v nadaljevanju naše raziskave izbrali klasifikatorja NBM in SVM, ki sta dosegla statistično značilno najboljše rezultate izmed vseh testiranih metod.

V strojnem učenju se prečno preverjanje (CV) pogosto uporablja za ugotavljanje točnosti modelov, kar naj bi povečalo zaupanje v rezultate pridobljene z večkratnim ponovnim učenjem in testiranjem. To pomeni, da iz celotne populacije enot izberemo določen delež izmed vseh enot ter jih uporabimo za izdelavo modela, preostali del pa za testiranje. Obsežni poskusi na različnih podatkovnih nizih ter z uporabo različnih klasifikacijskih metod so pokazali, da je za izdelavo modela pri taki razdelitvi podatkov priporočljivo ta postopek ponoviti desetkrat, kar pomeni, da so vse enote devetkrat vključene v izdelavo modela, enkrat pa v validacijo modelov (Witten, Frank & Hall, 2013).

Pri klasifikaciji dokumentov pa je pomembno tudi vrednotenje zgrajenih napovednih modelov. Literatura, ki je povezana s podatkovnim rudarjenjem oziroma rudarjenjem besedil, najpogosteje omenja naslednjih pet standardnih mer za učinkovito vrednotenje rezultatov: točnost, napako, preciznost, priklic in F1-oceno (Mitchell, 1997; Manning & Schütze, 1999; Sebastiani, 2002; Bishop, 2007; Qi & Davison, 2009; Liu, 2011; Miner in drugi, 2013). Točnost predstavlja delež pravilno klasificiranih enot, in se izračuna kot kvocient števila pravilno klasificiranih enot s številom vseh enot (glej Enačbo 1.8). Napaka je določena kot *1 - Točnost*. Izračun preostalih mer (preciznosti, priklica in F1-ocene) za vrednotenje modelov je predstavljen z Enačbami 1.9, 1.10 in 1.11. Preciznost predstavlja razmerje med vsemi pozitivnimi primeri, ki smo jih pravilno napovedali, in številom vseh primerov, za katere smo napovedali, da pripadajo pozitivnemu razredu. Priklic je delež resničnih pozitivnih primerov klasifikatorja, in predstavlja razmerje med številom vseh pozitivnih primerov, ki jih pravilno napovemo, in številom vseh pozitivnih primerov. F1-ocena predstavlja harmonično sredino preciznosti in priklica.

Motivacija

Ocene spletnega portala World Wide Web Technology Surveys (2017) kažejo, da se je število spletnih uporabnikov med letoma 2000 in 2016 povečalo za več kot osemkrat, kar predstavlja več kot 46% svetovnega prebivalstva.

Vsebina na spletu je napisana v različnih jezikih. Xu (2000) je ocenil, da angleščino uporablja 71% vseh spletnih strani, sledijo ji japonščina (6,8%), nemščina (5,1%), francoščina (1,8%), kitajščina (1,5%), španščina (1,1%), italijanščina (0,9%) in švedščina (0,7%). Podatki iz meseca januarja 2017 kažejo, da je 52,3% spletnih strani napisanih v angleškem jeziku, sledijo ruščina (6,4%), japonščina (5,7%), nemščina (5,4%), španščina (5,0%), francoščina (4,0%), portugalščina (2,6%), italijanščina (2,3%), kitajščina (2,0%) in poljščina (1,7%). V letu 2016 smo opazili trend zmanjševanja deleža strani na spletu, napisanih v angleškem (-1,6%), nemškem (-0,4%) in poljskem jeziku (-0,2%), medtem ko se je delež spletnih strani povečal za japonski (+0,7%), perzijski (+0,4%), ruski, španski, italijanski in korejski (vsi +0,2%).

Slovenščino uporablja približno 2 milijona ljudi, kar jo uvršča na 36. mesto med najpogosteje uporabljenimi jeziki na spletu. Delež spletnih strani, napisanih v slovenščini, se je v koledarskem letu 2016 povečal z 0,081% na 0,091% (World Wide Web Technology Survey, 2017).

Analizi sentimenta je predvsem v zadnjem desetletju namenjena večja pozornost. Ogromna količina tekstovnih informacij, ki so na voljo na spletu, je spodbudila potrebo po iskanju in pridobivanju informacij za strateško podprte odločitve. Čeprav je področje analize sentimenta relativno mlado, sta tako industrija kot akademski svet prepoznala prednosti pridobivanja sentimenta iz spletnih besedil. V zadnjih desetih letih se je število znanstvenih publikacij, ki omenjajo analizo sentimenta znatno povečalo, kar hkrati odraža tudi zanimanje za to raziskovalno področje (glej Sliko 1.2).

Liu (2010) je predstavil problem analize sentimenta s komentarjem o iPhone telefonu: *"(1) Pet dni nazaj sem kupil iPhone. (2) Bil je tako lep telefon. (3) Zaslon na dotik je bil res dober. (4) Tudi zvok je bil brezhiben. (5) Čeprav življenjska doba baterije ni bila dolga, mi je to povsem zadostovalo. (6) Moja mati je bila jezna, ker ji nisem povedal, preden sem ga kupil. (7) Prepričana je bila, da je telefon predrag, in je hotela, da ga vrnem nazaj v trgovino ..."*

Po pregledu zgornjega komentarja lahko rečemo, da je uporabnik objavil svoje mnenje po nakupu iPhone telefona, kot je razvidno iz stavka (1), kar je dejansko nevtralna izjava. V stavkih (2, 3 in 4) lahko zaznamo njegovo zadovoljstvo z izdelkom (pozitivni sentiment), medtem ko stavki (5, 6 in 7) vključujejo negativno konotacijo. Poleg tega lahko opazimo, da uporabnik v drugem stavku izraža mnenje o telefonu kot celoti, medtem ko v naslednjih treh stavkih (3, 4 in 5) izrazi svoje mnenje o zaslonu na dotik, kakovosti zvoka in življenjski dobi baterije. Šesti stavek se dejansko nanaša na uporabnika in ne na telefon. Zadnji stavek (7) se ponovno nanaša na telefon, natančneje na njegovo (previsoko) ceno. Opazimo lahko tudi, da stavki (2, 3, 4 in 5) predstavljajo mnenje uporabnika, medtem ko zadnja dva stavka (6 in 7) predstavljata mnenje uporabnikove matere.

Stone in Grusin sta v 80. letih objavila članek, v katerem sta ugotovila, da je delež negativnih novic objavljenih v medijih ABC, CBS in NBC znašal 46,8% (Stone & Grusin, 1984). Od takrat se je delež negativnih vsebin v večini medijev samo povečeval. Negativne novice so pogosto poceni in enostavne za objavo, povečujejo njihovo gledanost (branost) bolj kot pozitive novice, s čimer medijem omogočajo dober zaslužek (Ho, Chen & Sim, 2013; Trussler & Soroka, 2014; Vinkers, Tijdink & Otte, 2015; Kätsyri, Kinnunen, Kusumoto, Oittinen & Ravaja, 2016). Nekateri mediji skrbijo za uravnoteženo objavo pozitivnih in negativnih novic. V Romuniji so leta 2008 sprejeli zakon, ki je od medijev zahteval, da objavljajo petdeset odstotkov pozitivnih novic. Zakon je naletel na neodobravanje romunskega nacionalnega sveta za avdiovizualno oddajanje, ki je vztrajal, da morajo novice odražati resničnost, bodisi pozitivne ali negativne novice, neodvisne od vseh zakonov (International Journalists' Network, 2008).

V letu 2012 je Kovačič izvedel raziskavo, v kateri je v obdobju 35 mesecev - od decembra 2008 do oktobra 2011 - naključno zbral vzorec 2.386 RSS naslovov in kratkih povzetkov (do 250 znakov) iz 8 različnih slovenskih medijev (Kovačič, 2012). Ocenjevalci so ocenjevali RSS vsebine na podlagi vprašanja: *"Kako se počutite po tem, ko ste prebrali to novico?"*. Da bi izboljšali zanesljivost svojih raziskav, je bil vsak naslov in povzetek ocenjen s strani dveh neodvisnih ocenjevalcev. Po podatkih avtorja, na odločitve ocenjevalcev ni vplival drug ocenjevalec ali podporna ekipa. Rezultati so pokazali, da je v vseh obravnavanih medijih delež negativnih vsebin izrazito prevladujoč (glej Tabelo 1.1).

Hipoteze

Namen naše raziskave je ustvariti lastne jezikovne vire za analizo sentimentov, oceniti uspešnost različnih tehnik pri klasifikaciji sentimenta v dva (pozitiven in negativen) in tri (pozitiven, negativen in nevtralen) razrede, oceniti delež pozitivnih, negativnih in nevtralnih novic v medijih ter spremljati dinamiko sentimenta, vse z namenom, da prispevamo svoj delež k učinkovitejši računalniški analizi besedil v slovenskem jeziku. Raziskali smo naslednje hipoteze:

1. **Hypothesis 1 (H_1):** Ustrezna izbira klasifikatorja in predprocesnih nastavitev pripomore k učinkovitejši klasifikaciji dokumentov.

2. **Hypothesis 2 (H_2):** Granulacija besedil na manjše dele, kot na primer stavke, pripomore k doseganju boljših rezultatov pri klasifikaciji dokumentov.

3. **Hypothesis 3 (H_3):** Aplikacije razvite z lastnimi jezikovnimi viri, orodji in metodologijo je mogoče uporabiti v realnem življenju.

4. **Hypothesis 4 (H_4):** V pridobljenih novicah s politično, gospodarsko, poslovno in finančno vsebino petih slovenskih spletnih medijev je delež negativnih novic večji od deleža pozitivnih novic.

Cilji in znanstveni doprinos

Cilje te raziskave, ki vsebujejo enega ali več podciljev, lahko uvrstimo v štiri skupine:

1. Celovit pregled literature.
   - Poglobljen pregled obstoječih pristopov analize sentimenta;
   - Pregled relevantnih korpusov in leksikonov za analizo sentimenta;
   - Pregled relevantnih študij in učinkovitosti s področja analize sentimenta.

2. Razvoj jezikovnih virov v slovenskem jeziku.
   - Izdelava spletnih pajkov in pridobivanje besedila novic iz digitalnega arhiva

petih slovenskih spletnih medijev;
- Čiščenje podatkov;
- Izdelava spletne aplikacije za ročno označevanje, označevanje besedil na treh nivojih granulacije besedil s pomočjo več označevalcev;
- Izračun stopnje ujemanja med označevalci;
- Izdelava označenih korpusov in leksikona za analizo sentimenta v slovenskem jeziku;
- Zagotovitev javnega dostopa do razvitih jezikovnih orodij in virov, določitev pogojev uporabe in distribucije razvitih jezikovnih orodij in virov.

3. Izbor najprimernejših klasifikacijskih metod in predprocesnih nastavitev za učinkovito klasifikacijo spletnih besedil v slovenskem jeziku.
   - Izbor najprimernejših klasifikatorjev za analizo sentimenta spletnih besedil v slovenskem jeziku;
   - Izbor najprimernejših predprocesnih nastavitev za analizo sentimenta spletnih besedil v slovenskem jeziku;
   - Raziskava vpliva granulacije dokumentov na učinkovitost klasifikacije sentimenta;
   - Ocena učinkovitosti najprimernejših metod klasifikacije in predprocesnih nastavitev na anlizo sentimenta spletnih besedil v slovenskem jeziku.

4. Izdelava (v vsakdanjem življenju) praktično uporabnih aplikacij na podlagi razvitih jezikovnih virov.
   - Ocena deleža pozitivnih, negativnih in nevtralnih novic v petih slovenskih spletnih medijih;
   - Spremljanje dinamike sentimenta petih slovenskih spletnih medijev z različnih vidikov (dinamika sentimenta v okviru dokumentov, dinamika sentimenta skozi čas, dinamika sentimenta glede na različno temo in vrsto vsebine, dinamika sentimenta glede na avtorja novic).

Glavni znanstveni doprinosi, opisani v tej disertaciji so:

1. Pregled in podroben opis postopkov za izdelavo označenega korpusov novic v sloven-

skem jeziku. Več kot deset tisoč pridobljenih novic smo ročno označili kot pozitivne, negativne in nevtralne na treh nivojih (na nivoju dokumentov, odstavkov in stavkov). Razvita jezikovna orodja in viri so prosto dostopni pod določenimi pogoji uporabe (glej Sekcijo 4.3).

2. Ocena učinkovitosti devetih tehnik strojnega učenja pri klasifikaciji sentimenta slovenskih spletnih besedil v dva (pozitivni in negativni) in tri (pozitivni, negativni in nevtralni) razrede. Rezultati kažejo, da Naivni (večrazsežnostni) Bayesov klasifikator dosega najboljše rezultate z vidika točnosti, F1-ocene in časovne zahtevnosti. Na uravnoteženih podatkih dosega F1-oceno nad 97% pri klasifikaciji dokumentov v dva razreda, pri klasifikaciji dokumentov v tri razrede pa nad 77%. Rezultati prav tako kažejo, da granulacija besedil na stavke, pripomore k doseganju boljših rezultatov pri klasifikaciji dokumentov.

3. Različne praktično uporabne aplikacije na razvitih jezikovnih virov. Analiza dinamike sentimenta na razvitih jezikovnih virih. Rezultati so pokazali, da je v povprečju sentiment močneje izražen na začetku dokumenta in izgublja svojo izraženost proti koncu dokumenta.

4. S strani petih slovenskih spletnih medijev (24ur, Dnevnik, Finance, Rtvslo, Žurnal24) smo med 1. septembrom 2007 in 31. januarjem 2016 pridobili več kot 250 tisoč novic s politično, gospodarsko, poslovno in finančno vsebino. Približno polovica vseh novic je nevtralnih, medtem ko je delež negativnih novic približno dvakrat večji od deleža pozitivnih novic.

ZAKLJUČEK

Hitra rast informacij na spletu (kot na primer mnenj uporabnikov, informacij o konkurentih, elektronskih sporočil strank, komentarjev na družbenih omrežjih, sporočil za javnost, pravnih dokumentov, dokumentov o izdelkih in storitvah itd.), je pospešila zanimanje za analizo spletnih besedil. Še posebej industrija, ki se ukvarja z nudenjem spletnih rešitev in storitev, je hitro zaznala pomen in potencial s sentimentom izraženih besedil. Danes te

dragocene informacije uporabljajo za izvajanje kontrole kakovosti, tržne raziskave, analizo potreb in interesov strank ter analizo trendov prodaje lastnih izdelkov in storitev.

V tej disertaciji je opisan proces pridobivanja lastne zbirke označenih korpusov spletnih novic (besedil) v slovenskem jeziku in leksikona za analizo sentimenta v slovenščini. Podrobno je razložen postopek ocenjevanja učinkovitosti klasifikacije besedil glede na uporabo različnih klasifikacijskih metod, opisan je proces ocenjevanja deleža pozitivnih, negativnih in nevtralnih novic v spletnih medijih ter spremljanje dinamike sentimenta na označenem korpusu spletnih novic.

S strani petih slovenskih spletnih medijev (24ur, Dnevnik, Finance, Rtvslo, Žurnal24) smo med 1. septembrom 2007 in 31. januarjem 2016 pridobili več kot 250 tisoč novic s politično, gospodarsko, poslovno in finančno vsebino. Poleg tega smo jih več kot deset tisoč ročno označili kot pozitivne, negativne in nevtralne na treh nivojih (na nivoju dokumentov, odstavkov in stavkov). Za vrednotenje procesa označevanja smo uporabili pet različnih mer. V splošnem vse mere kažejo na dobro notranjo usklajenost označevalcev na vseh treh nivojih, vendar pa se njihove vrednosti zmanjšujejo z nivojem granulacije. Tako se vrednosti teh mer zmanjšujejo od nivoja označenih dokumentov (preko nivoja odstavkov) proti nivoju označenih stavkov. Lastno razviti jezikovni viri so prosto dostopni pod licenco Creative Commons copyright license Attribution-ShareAlike 4.0 (glej Sekcijo 4.3). Z našim delom podpiramo odprtokodno skupnost, da bi prihodnjim raziskovalcem omogočili nadaljnji razvoj pri računalniškemu razumevanju (slovenskega) jezika.

Predstavili smo učinkovitost ocenjevanja metod klasifikacije na razvitih jezikovnih virih. Najprej smo empirično ovrednotili najpogosteje uporabljene klasifikacijske metode v kombinaciji z izbiro predprocesnih nastavitev na označenih besedilih v dva (pozitivni in negativni) in tri razrede (pozitivni, negativni in nevtralni). Pokazali smo, da se v naših eksperimentih pri klasifikaciji dokumentov Naivni (večrazsežnostni) Bayesov klasifikator in Metoda podpornih vektorjev izkažeta kot najbolj učinkoviti metodi z vidika časovne zahtevnosti in različnih mer točnosti. Naivni (večrazsežnostni) Bayesov klasifikator doseže najboljšo F1-oceno 97,85% pri klasifikaciji dokumentov v dva razreda, pri klasifikaciji dokumentov v tri razrede pa 77,76% (glej Tabelo 5.7). V obeh primerih so bili najboljši rezultati doseženi z uporabo TF-IDF in seznama blokiranih besed, brez lematizacije besedil ter z izbiro 3.000 značilk z metodo *Gain ratio*. V splošnem klasifikatorji dosegajo boljše rezultate na uravnoteženih podatkih (po 1.000 dokumentov na razred),

zlasti pri klasifikaciji dokumentov v tri razrede. Rezultati prav tako kažejo, da granulacija besedil na manjše dele, kot na primer stavke, pripomore k doseganju boljših rezultatov pri klasifikaciji dokumentov. V zvezi s predprocesnimi nastavitvami je najboljša možnost tista, ki ne uporablja lematizacije. Vse razen ene ali dveh takih možnosti uporabijo tudi transformacijo velikih črk v male in uporabo seznama blokiranih besed. Zdi se, da je vpliv drugih možnosti mešan.

Za ocenjevanje deleža pozitivnih, negativnih in nevtralnih novic v spletnih medijih smo uporabili Naivni (večrazsežnostni) Bayesov klasifikator, ki se je v predhodnih eksperimentih izkazal za najbolj učinkovitega. Ugotovili smo, da Finance (s 37%) objavijo največ pozitivnih novic, medtem ko 24ur in Rtvslo objavita največji delež negativnih novic (24ur: 42%, Rtvslo: 37%). Ugotovili smo, da je približno polovica izmed vseh pridobljenih novic nevtralnih, ter da je delež negativnih novic približno dvakrat večji od deleža pozitivnih novic (glej Tabelo 6.1).

Nazadnje smo spremljali dinamiko sentimenta na označenem korpusu spletnih novic. Pri spremljanju dinamike sentimenta na nivoju dokumenta smo odkrili zanimiv vzorec. Opazili smo, da je pri vseh opazovanih medijih sentiment (pozitiven ali negativen) na začetku novic bolj izrazit, pri čemer njegova intenzivnost (pozitivna ali negativna) enakomerno pojenja skozi novico, ter se proti koncu približuje nevtralnemu sentimentu (glej Slike 7.1, 7.2, 7.3 in 7.4). To odkritje bi lahko pomembno vplivalo na analizo sentimenta v novicah v prihodnosti.

Nadaljnje delo

Jezikovne vire je relativno enostavno zgraditi, zato si s svojim delom prizadevamo spodbuditi tudi druge raziskovalce, zlasti predstavnike drugih manjših jezikovnih skupin, da na podoben način izdelajo nove jezikovne vire. V prihodnosti nameravamo obogatiti, razvijati in povečevati velikost že zgrajenih jezikovnih virov, razširiti izbor spletnih medijev, in če bo mogoče, nadaljevati z obsežnim ročnim označevanjem pridobljenih novic.

V zadnjih desetletjih smo priča izjemnemu razvoju jezikovnih tehnologij, virov in aplikacij. Vendar pa se znanstveniki v okviru svojih raziskav še vedno srečujejo z nekaterimi izzivi. Le-ti so najpogosteje povezani z zaznavanjem čustev in vsebin, klasifikacijo besedil, odkrivanjem špekulacij in negacij, ustrezno zaznavo in obravnavo citatov, sarkazma in hu-

morja ter vprašanj, povezanih s semantiko in slovnico. Ravno v tem delu vidimo velik potencial za nadaljnje raziskave.

Trenutno je naše delo povezano z metodami modeliranja, kjer s spremljanjem sentimenta skozi daljša časovna obdobja še naprej raziskujemo skrite povezave med temami in dogodki, kar pričakujemo, da bo vodilo do novih rezultatov. V prihodnosti si želimo uporabiti jezikovne vire tudi v tržno usmerjenih projektih, ki bi spodbudili povpraševanje po storitvah, kot so sledenje ažurnim informacijam o izdelkih, storitvah in dogodkih, analiziranje in ocenjevanje rezultatov za potencialne stranke. Prav tako je naš namen nadaljnje preučevanje različnih metod in tehnik, ki lahko izboljšajo dosežene rezultate, zlasti v okviru nadaljnjega izboljšanja učinkovitosti klasifikacije dokumentov v tri razrede.