universität
wien

MASTERARBEIT

Titel der Masterarbeit

"Integrating Natural Language Processing with Semantic-based Modeling"

verfasst von

Martin Bergner, BSc

angestrebter akademischer Grad

Diplom-Ingenieur (Dipl.-Ing.)

Wien, 2015

Studienkennzahl lt. Studienblatt:	A 066 926
Studienrichtung lt. Studienblatt:	Masterstudium Wirtschaftsinformatik
Betreut von:	Univ.-Prof. Mag. Dr. Hans-Georg Fill

"If something is important enough,
even if the odds are against you,
you should still do it."

— Elon Musk

Acknowledgements

My sincere appreciation I give to my supervisor Univ.-Prof. Mag. Dr. Hans-Georg Fill or the useful comments, remarks and engagement through the learning process of this master thesis. Special thanks to Univ.-Prof. i.R. Dr. Wilfried Grossmann for the valuable expertise and interesting discussions.

I also like to thank the participants in my survey, who have willingly shared their precious time.

Special thanks to Andreea Dumitrescu, who patiently revised and corrected my thesis.

I am also grateful for all of my friends, who were very supporting and forgiving during this exhausting time
– you will never hear "*I'm sorry - I don't have time - I have to write my thesis*" again!

Last but not the least, I would like to thank my family.
My parents Michael & Gertraud and both of my sisters, Barbara & Irmgard, for encouraging and supporting me with one of the greatest gifts throughout writing this thesis and my life in general – family.

Abstract

The evolution of the world wide web paved the way for the distribution of social media platforms throughout the world, changing the way people interact with each other as well as companies with their customers. Globalization and this new way of interaction puts enormous pressure on companies staying competitive. One strategy is the focus on business process improvement initiatives, reducing costs and raising efficiency throughout the company using knowledge of customers as well as employees. While the social media environment is well developed for companies in terms of marketing, the vast amount of data, generated by the users in return is merely processed and turned into insights supporting these business process improvement initiatives. This work proposes a method of processing data contained in social media platforms using natural language processing techniques in combination with semantic-based modeling, acting as an extension of the SeMFIS modeling methods based on the ADOxx metamodeling platform and includes the technical implementation in Java as well as a subsequent evaluation of the gained results from the posts of Facebook users on the page of an airline in terms of objectiveness and validity using the knowledge of human participants.

Keywords: natural language processing, business process improvement, social media, semantic-based modeling, conceptual models, adoxx, semfis

Contents

List of Figures

List of Tables

Part 1

Preface

1.1 Introduction

Globalization and new technologies, based on the world wide web, provide higher market transparency for customers and new ways of interaction and information retrieval (Kaplan and Haenlein, 2010). The emergence of Social Media, connecting millions of users around the world, increased the exchange of content and points of intersections between users and companies. In addition to managing marketing opportunities and customer care, companies are forced to reduce costs and raise efficiency due to this mix of globalization and increasing market pressure (Davis, 2013). It is not surprising that worldwide spending on business process management software is growing and customer expectations are changing increasingly fast nowadays (Moore, 2015). These rapidly changing customer requirements, the already mentioned, increasing market transparency, and the need of continuous improvement of processes implies huge effort in documentation of current and future customer needs (Johannsen and Fill, 2014a, p. 2). To meet the requirement of satisfying this high information demand, often knowledge from within the company combined with various different business process improvement approaches (Shtub and Karni, 2010) (Siha and Saad, 2008) is used. Commonly used BPI approaches mainly put their focus on process segmentation, employee knowledge, benchmarking etc. (Zellner, 2011) ignoring direct knowledge from customers. Other fields, regarding business processes, e.g. Quality management (Fundin and Bergman, 2003) in contrary, already use knowledge from customers directly and integrate it in the product development process to increase customer satisfaction. On the other hand, the vast amount of data collected in Social Media environments remains unused due to the lack of systems coping with the unstructured nature of user generated content.

This paper therefore investigates an approach of collecting and aggregating knowledge from social media sources and connecting this collected customer knowledge (Johannsen and Fill, 2014b) in business process improvement initiatives, using an integrated approach of semantic-based modeling and natural language processing.

1.1.1 Motivation

The possibilities and forms of usage of social networks have increased over the past years. From satisfying the human need of connecting and sharing (Viswanath et al., 2009), online apps & gaming (Nazir et al., 2008) to complex marketing structures, social media allows companies to target their potential customers in a more efficient way compared to traditional marketing (Holzner, 2008).
Excluding the creator and operator of the social network site, two stakeholders can be identified. The private users of the SNS, with the interest of getting connected, informed or entertained and companies with the interest of informing and acquiring potential customers as well as to collect all kinds of data as collecting information (e.g. feedback) is very valuable for a company. Following Fundin & Bergman, the importance of avoiding disappointed customers increases due to the enhanced competition on the market (Fundin and Bergman, 2003). One approach of coping with these kinds of challenges is Customer Relationship Management. Active Customer Relationship Management is not a new topic in the business world, defining systematic processes and techniques to actively maintain a healthy connection between company and customer. In this topic, Customer feedback is not only used to improve existing products and services or for the development of new products. It also delivers an important information base for Business Process Improvement. Given a Facebook page with a million followers, the amount of data generated by users can easily exceed the capability of a human collecting, aggregating and interpreting the information. Frequently, the efficient collection and interpretation of this data amount lacks. While data from Twitter is used for predicting flu or monitoring earthquakes, described at later chapters, the connection between social media data and business process improvement is still missing. In this environment two points of focus arise: The social network, which generates valuable data that needs to be transformed into insights and business process models, which describe the process landscape of a company, forming the base of business process improvement.

1.1.2 Problem Statement

Although there is a huge amount of data from social media users that has a huge potential, capitalizing on the ability to analyze the data, turning it into information and in the next step into insights that impact the bottom line is not very common according to (Gillan, 2010). Data and information provided by the huge amount of users is left unused in the social media channel (figure 1.1.1). Since the data is mostly unstructured and based on natural language, analysing and transforming the

data into insights is a very complex task.

Analyzing data implies a system which is able to understand these consumer conversations, which consist mainly of natural language. (Manning and Schütze, 1999) & (Jurafsky and Martin, 2000) describe several methods in the field of Natural Language Processing (NLP) that can be used to turn natural language data into valuable information. They describe the minimum requirements of an NLP system as follows: "An NLP system needs to determine something of the structure of text normally at least enough that it can answer 'Who did what to whom?' " (Manning and Schütze, 1999, p. 17). To turn this information into insights, the system must be able to understand the consumer data in a way that corresponds to the structure of the business of the company. "Key techniques go beyond text analytics to include opinion mining, sentiment analysis, topic modeling, social network analysis, trend analysis, and visual analytics" (Fan and Gordon, 2014, p. 74). The information provided by the customer has to be somehow mapped to the relevant business processes, revealing improvement potential. To achieve this requirement in a user-friendly and efficient way the information must be translated into a model which allows the user to compare and identify the relevant process.

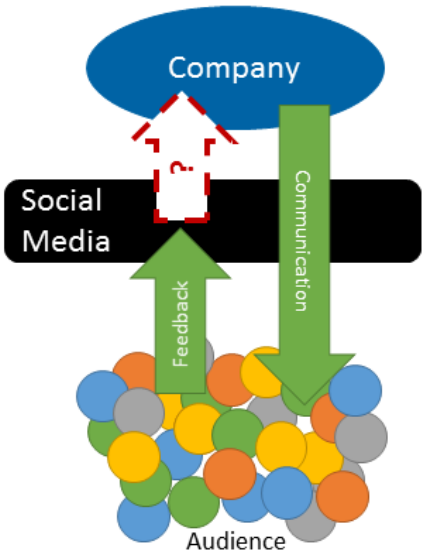

Figure 1.1.1: *Social Media Feedback*

1.1.3 Research Questions

The primary goal of this thesis is the research of natural language processing, focusing on social media data and the connection to conceptual modeling in order to gain insights that drive business process improvement as voice of the customer. This goal can be divided into three research objectives, which include the following aspects:

- Researching concepts and modeling methods
  Several methods for Natural Language Processing and Semantic-based Modeling have been researched. What concepts and methods of Natural Language Processing and Semantic-based Modeling can be used to provide the optimal foundation for linkage?

- Connecting natural language with conceptual modeling
  How can natural language, focused on social media, be actually linked with conceptual modeling, focused on Business Process Modeling, using a new modeling method?

- Technical Implementation
  How is a practical implementation realized, based on the knowledge gained?

In order to achieve the objectives stated above, this work is divided into three main parts:

- Research of related work, foundations and literature in order to achieve the knowledge needed (part 2).

- Implementing a practical implementation using the knowledge gained (part 3).

- Evaluation of the results of the implementation (part 4).

Furthermore this work should build on the effort and research achieved in developing the SeMFIS toolkit (OMILAB, 2015) & (Fill, 2012) based on the ADOxx platform (ADOxx.org, 2015) in order to extend the given functionality of the platform. To ensure a high level of compatibility, Java should be used as a platform for the natural language processing.

Part 2

Foundations

2.1 Introduction

In order to provide a solid knowledge base for the implementation of aforementioned concepts, comprehensive preliminary research, providing the required foundations regarding several fields of science had to be done. The following part describes the gained knowledge and separates the fields of research into four chapters:

- Chapter 2.2 "Social Media and Web": The evolution of social media along with the evolution of the world wide web, changing the way people interact with each other, increasingly determining the content presented and generating vast amounts of data which is available throughout the world.
- Chapter 2.3 "Social Media and Business": The relation of business and social media, transforming the way businesses and potential customers interact with each other, focusing on the increase of pressure on companies due to the seemingly uncontrollable customer relation in the social media environment and approaches to integrate customer feedback and business processes in order to drive business process improvement initiatives.
- Chapter 2.4 "Natural Language Processing": Approaches in the field of natural language processing focusing on their foundations and origins as well as current research. Since the field of natural language processing extends the capacity of this work, the chapter focuses especially on methods applicable for this research as well as fundamental techniques mandatory.
- Chapter 2.5 "Semantic-based Modeling": Conceptual modeling as a source and storage of knowledge, reducing complexity and increasing flexibility in order to support business process improvement initiatives. The chapter especially focuses on the ADOxx metamodeling framework and the SeMFIS toolkit as platform, as prior defined.

2.2 Social Media and Web

The term ”social media” is widespread throughout the world already as more and more users join this essential part of the world wide web, pushing petabytes of data into this environment. This chapter describes the origin of social media and its evolution along the evolution of the world wide web.

2.2.1 Definition of Social Media

Following Kaplan and Haenlein, social media describes “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content” (Kaplan and Haenlein, 2010, p. 61). Two terms in this definition need further explanation. Web 2.0, which is described in a later section of this paper and User Generated Content. According to the Organisation for Economic Cooperation and Development (OECD) (Vickery and Wunsch-Vincent, 2007), User Generated Content (or User Created Content) needs to meet three central characteristics:

- Publication requirement
  The created content has to be published in an online environment, which can either be public or restricted to specific groups.
- Creative effort
  A certain amount of creative effort is necessary. For example, publishing photographs, expressing thoughts etc.
- Creation outside of professional routines and practises
  In general, some of the main motivation factors are self-expression, achieving a certain level of fame, connecting with peers etc.

“..., User Generated Content (UGC) can be seen as the sum of all ways in which people make use of Social Media. The term, which achieved broad popularity in 2005, is usually applied to describe the various forms of media content that are publicly available and created by end-users” (Kaplan and Haenlein, 2010, p. 61). The exchange of this user generated content therefore is a central point regarding Social Media. By itself it can be seen as everything creative a human being fabricates outside of professional routines and practises. But human beings have always been creative throughout history, why does this term appear? As the world wide web

spreads across the world, so did the possibilities of users sharing their creations. A major part of this rise can be accounted to one evolutionary step of the world wide web, called Web 2.0.

2.2.2 Web 2.0 and Social Media

Shortly after the millennium a new era of the world wide web summarized by the term "Web 2.0" began. "Web 2.0" popularized by Tim O'Reilly at O'Reilly Media Web 2.0 Conference in 2004, does not describe a certain technology used or a new version of the Internet after the burst of the dot-com bubble in 2001. Web 2.0 represents the evolution of the world wide web from one-way networked communication to a platform allowing users to contribute to the information offered. Wikipedia, an online encyclopedia based on the idea that any web user is allowed to create or edit the information offered, can be presented as a successful adaptation of this evolution. Endless possibilities and a huge variety of new, functional online services lead web users to move more and more of their everyday activities into these online environments (O'Reilly, 2005). Boyd and Ellison date the beginnings of the first Social Networks back to 1997, but the breakthrough to the mass happened 2003 with the launch of Myspace, becoming a global phenomenon in 2006 (Boyd and Ellison, 2007). With the evolution of the Web from a one-way communication to an interactive platform with user-generated content, also the beginnings of Social Media attained a huge impulse.

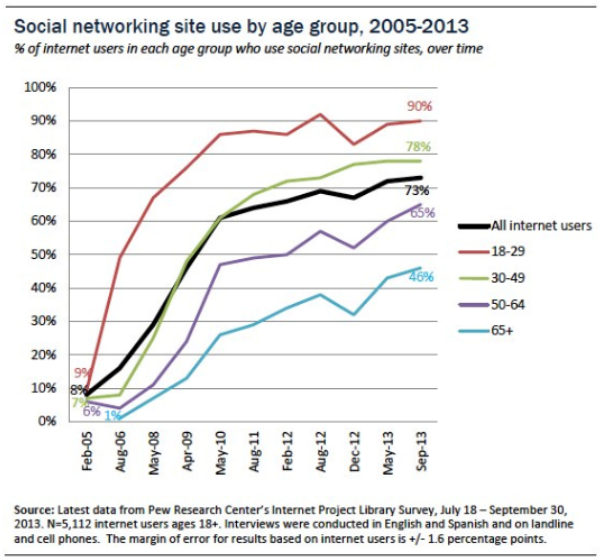

Figure 2.2.1: *Social Networking use in the US 2005-2013 (Duggan et al., 2015)*

2.2.3 Social Media Use

According to the Social Media Update 2014 from the Pew Research Center, 74% of all adults in the United States already use social networks. Figure 2.2.1 shows the use of social network sites from February 2005 until September 2013. Growth has slowed down, but engagement with the platform has increased. In August 2015 Facebook

reached 1 billion active users on a single day (Wattles, 2015) and 65% of Facebook users frequently or sometimes share content. Overall Facebook still dominates, while other platforms see higher rates of growth which implicates that the use of two or more sites gain popularity (Duggan et al., 2015). Besides Facebook, another platform has acquired huge popularity - Twitter. The microblogging platform has more than 200 million active and more than 500 million registered users (2012), underlining its rapid growth (O'Carroll, 2012).

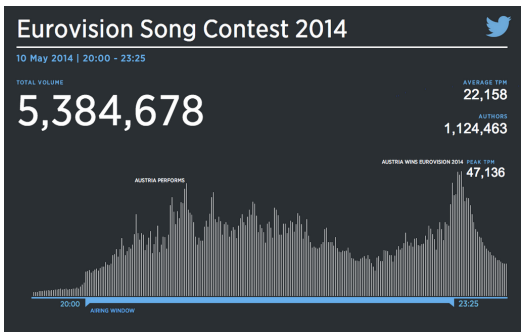

Figure 2.2.2: *Eurovision Song Contest 2014 (Twitter UK, 2014)*

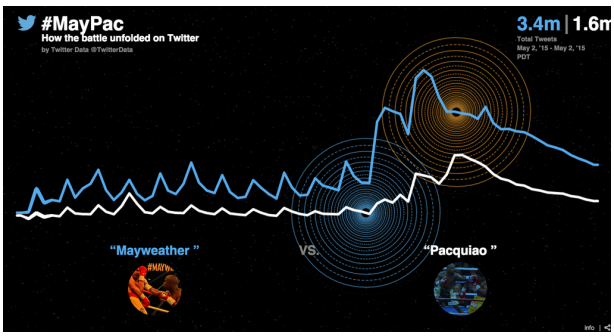

Figure 2.2.3: *Mayweather - Pacquiao Fight (Twitter Data, 2015)*

Figure 2.2.2 shows statistics from the Eurovision Song Contest Finals of 2014. During the airing window in total more than 5.3 million tweets with direct connection to the show have been sent, making an average of more than 22 thousand tweets per minute (Twitter UK, 2014). Figure 2.2.3 shows statistics of the Mayweather-Pacquiao Boxfight on May 2nd 2015, which led to a total of 5 million tweets worldwide, making an average of more than 147 thousand tweets per minute (Twitter Data, 2015). But not only entertainment events are generating data. Data from twitter has been successfully used for flu prediction (Culotta, 2010) or monitoring earthquakes (Sakaki et al., 2010) and especially Facebook users are following brands actively (figure 2.2.4). The use of Social Media has increased dramatically over the past years and so did the amount of data generated by the users.

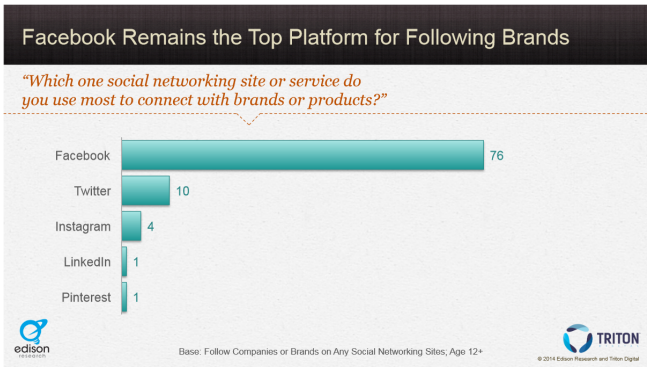

Figure 2.2.4: *Social Habit 2014 - users following brands actively on SNS (Edison Research, 2014, p. 23)*

2.2.4 Types of Social Media Platforms

Social media is often thought of as being synonymous with social network, which is not correct. Social media covers a huge number of different types of platforms including social networks. Brian Solis created the Conversation Prism, an attempt to classify popular and most promising social media Platforms. (Solis, 2015). Even though there are only popular Social Media Platforms classified, the amount is still impressive. Solis states that from version 3 to version 4 (current), 122 services were removed and 111 added. While Solis' attempt is very impressive, Jose van Dijck has found a more clear view. Jose van Dijck distinguishes between four major types of social media platforms (Van Dijck, 2013)[p. 8]:

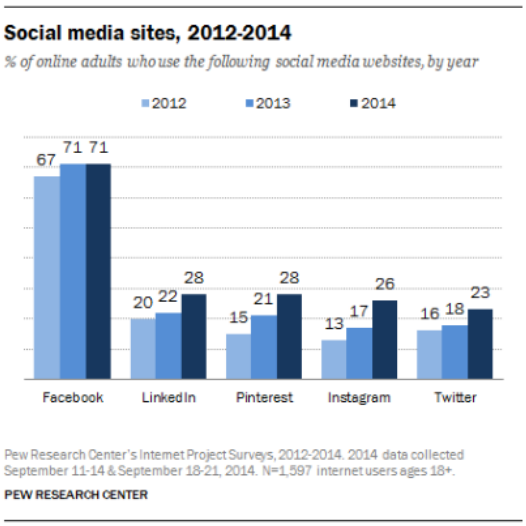

Figure 2.2.5: *Most popular networks in the US 2012-2014 (Duggan et al., 2015)*

- SNS - Social Network Sites
  Enabling interpersonal contact for individuals or groups. Popular examples are Facebook, Twitter, Google+, ...

- UGC - Sites for User generated content
  Enabling users to exchange creative content. Popular examples are youtube, Flickr, MySpace, ...

- TMS - Trading and Marketing sites
  Focus on the exchange or sale of products. Popular examples are eBay, Amazon, Groupon, ...

- PGS - Play and Game sites
  Online or mobile gaming platforms which provide social interactivity. Popular examples are FarmVille, The Sims Social, ...

While some big players in this ecosystem dominate their section and remain stable, a lot of small platforms and services emerge and disappear, making this system very volatile. The Application to use-case part of this paper sets the focus on Social Network Sites. Figure 2.2.5 shows the five most popular social networks for adults in the US from 2012, 2013 and 2014. The advent of Social Media is deeply connected

with the rise of Web 2.0 and experiences a new push with the emergence of Web 3.0.

2.2.5 Definition of Social Network Sites

Following Boyd and Ellison, a Social Network Site, as part of social media, is defined as "... web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system" (Boyd and Ellison, 2007, p. 211).

2.2.6 Web 3.0 and Social Media

Web 2.0 has accelerated the rise of a huge social media ecosystem which brought a new way of communication and interconnection to users all over the world. But things are changing fast in terms of the world wide web and the next evolutionary step is already on our doorstep: Web 3.0. Based on the semantic web vision T. Berners-Lee, J. Hendler and O. Lasilla expressed in 2001 (Berners-Lee et al., 2001), J. Hendler describes Web 3.0 as a set of "Semantic Web technologies integrated into, or powering large-scale Web applications" (Hendler, 2009, p. 111) and J. van Dijck states that "Futurists and information specialists consider Web 3.0 to be the Semantic Web, which will involve, among other developments, the rise of statistical, machine-constructed semantic tags and complex algorithms to enhance the personalization of information driven by conversational interfaces" (Van Dijck, 2013, p. 179).
Based on machine-enabled addition or connection of data and information, semantic web brings more and more data and information to the world wide web. While the amount of data on the world wide web explodes and terms like Big Data gain more and more popularity, the amount of finding useful information that transforms into action for businesses decreases.

2.3 Social Media and Business

Social media has not only established a new way of interaction between people across the world, it has also built a new channel between businesses and millions of their potential customers. Companies are finding themselves in an ecosystem, where customers are able to interact and communicate with each other more freely as well as with the company. Users are able to compare products and prices, get benefit from the experience of others or share their experience with others. Companies therefore tried to adopt quickly and use these new social media platforms for their public relations, especially focused on marketing (Barnes and Wright, 2013). In contrary to classical marketing and public relations management, which mostly consists of one-way communication, customers are now able to provide feedback either public or within the community, potentially reaching a huge audience (Kaplan and Haenlein, 2010). Companies thus have less and less control over the information provided or shared and become more and more observers. Especially negative feedback in this environment is able to harm significantly a once good reputation, hence forcing companies to monitor online conversation actively (States, 2009), focusing even more on enterprise social media (Davis, 2013) and analytics (Fan and Gordon, 2014). In the past, being able to control the information flow via active public relation management, or simply hiding issues, companies are now in a state, where social media forced them to establish an additional channel for customer care, providing quicker response times to keep their brand advocacy and customer satisfaction (Banfi et al., 2015). Key difference between the traditional, well-established channels and the new Social Media channel lies in higher public pressure on the company, because of different expectations of the customers and higher amount of feedback compared to the traditional channels. Customers expect 60 minutes or less response times in Social Media environments without caring about business hours or weekends (Forrester Consulting, 2015) & (Fan and Gordon, 2014). Companies often are not able to deal with the amount of feedback to meet those expectations (conversocial, 2013) and just provide First Responses (Forrester Consulting, 2015), not resolving the issue of the customer. In a traditional environment the company probably loses the affected customer, while in the social media environment, the issue can affect the opinion of uncountable customers. One step to reduce the increasing amount is using the knowledge gained from complaints to support business process improvement initiatives, eliminating re-occurrence of complaints, caused by processes itself.

2.3.1 Business Process Improvement

The improvement of business processes not only aims to reduce inefficiencies inside a companies processes. Business process improvement also aims to reduce time-to-market costs for products and services and to improve customer service. Globalization, high market transparency, increase of competition, new technologies and changes in stakeholder requirements are forcing companies to respond with constant improvement of their business processes to address these needs (Seethamraju and Marjanovic, 2009) & (Johannsen and Fill, 2014a). This requirement ties a lot of resources and won't be reduced in the near future (Johannsen and Fill, 2014b). A huge variety of different techniques and holistic approaches for Business Process Improvement are available, e.g. (Zellner, 2011), but "... are often perceived as over dimensioned or inefficient" (Johannsen and Fill, 2014b, p. 3) and overwhelming decision makers and project teams due to the huge variety and operational knowledge required. This combination leads to a point where the tacit knowledge of participants in BPI projects is not able to unfold effectively (Johannsen and Fill, 2014a).

2.3.2 Business Process Modeling

Making the complex structures and relationships of business processes in companies understandable and clear, even for regular users, is possible by creating a visual representation using business process modeling therefore, BPM should "... represent fragments of reality by reducing complexity without altering the underlying facts" (Fill, 2009, p. 16) and give users the possibility to "... manage complex issues such as for example the analysis of cause effect relationships of strategic goals, the re-engineering of interconnected business processes and assigned organisational structures or the realisation and control of IT service level agreements" (Fill, 2009, p. 16). Van der Aalst describes Business Process Modeling as a part of Business Process Management: "One of the main aspects and certainly an activity typically carried out in early phases of business process management projects is the design of business processes. There is a close relationship between business process design and business process modeling, where the former refers to the overall design process involving multiple steps and the latter refers to the actual representation of the business process in terms of a business process model using a process language. To this end, the term business process modeling is used to characterize the identification and (typically rather informal) specification of the business processes at hand. This phase includes modeling of activities and their causal and temporal relationships as well as specific business rules that process executions have to comply with" (van der

Aalst et al., 2003, p. 8).

Using different modeling methods or languages in business process management is state-of-the-art already, but seems to be stuck at the moment of initial design prior to implementation or as an informal tool 'specifying processes at hand'. This puts business process modeling into an observer state instead of the foundation, losing potential for the constant and efficient improvement of processes.

2.3.3 Business Process Modeling as base for BPI

Business process modeling takes an important part in business process management, as stated before and the integration of social media forces companies to constantly improve their processes in order to stay competitive. One common approach is to use business process modeling as base for business process improvement initiatives. As mentioned before, there is a huge variety of techniques and approaches (section 2.3.1) facing the same challenges. "In a BPI project, the project participants' tacit process knowledge (e.g. of process weaknesses, etc.) is transformed into explicit knowledge which needs to be codified, communicated and processed adequately" (Johannsen and Fill, 2014b, p. 2).

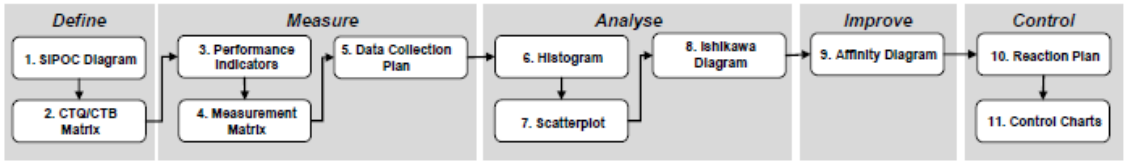

Figure 2.3.1: *Overview of the BPI roadmap showing the five phases and the techniques (Johannsen and Fill, 2014a, p. 8)*

Managing these requirements is very challenging for the participants and often exceeds the capacity in terms of expertise and time. (Johannsen and Fill, 2014b) propose an approach that uses a five-phased roadmap to identify key problems and defining concepts and solutions for improvements. For each phase appropriate and popular models are used to express the knowledge in an understandable and leveraging way to identify, categorize, visualize and control several performance indicators (figure 2.3.1). To connect the semantic relationships between each model, metamodels defining the underlying models are used, which then are connected via interrelations, bringing the expressed knowledge from a visual to a semantic context. An application of this modeling method based on the ADOxx metamodeling

platform (ADOxx.org, 2015) has been implemented and is freely available as part of the SeMFIS Toolkit (Johannsen and Fill, 2014b).

2.3.4 Integration of Feedback and BPI

In classical Quality Management, systematic processes work as feedback mechanisms, by transforming complaints of customers into knowledge for improvement of current products or design of future products or offerings (figure 2.3.2).

The origin of these complaints are the customers, who used several channels provided by the company to deliver their feedback e.g. Hotlines, E-Mail, etc. (Fundin and Bergman, 2003). Prepared companies used clearly defined processes to react quickly on this complaints. The changed environment forces companies to react and use Social Media as an additional channel and define strategies to effectively use and respond to this feedback (Grégoire et al., 2015) & (Forrester Consulting, 2015). Now the procedure of responding, resolving and using the knowledge of these complaints remains the same in principle and companies tried to implement their established processes on this new channel, resulting in the loss of customers and putting their reputation at risk (States, 2009). One problem in this environment is the re-occurrence of issues due to erroneous business processes. Taking into account, that re-occurring issues in a social media environment are able to reach a huge audience in a short period of time and can exaggerate quickly, companies experience high pressure, coping with these incidents. The RUPERT approach combines business process modeling with the Voice of the Customer (VOC) to codify, document and process knowledge created in BPI projects, giving companies the possibility to execute business process improvement initiatives more efficiently (Johannsen and Fill, 2014b).

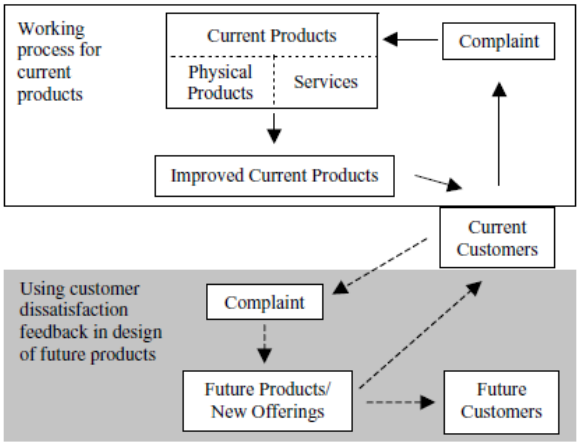

Figure 2.3.2: *Feedback mechanisms for Quality Management (Fundin and Bergman, 2003, p. 56)*

2.4 Natural Language Processing

The following chapter describes the challenges and techniques of Natural Language Processing. Most elements covered in this chapter refer to (Manning and Schütze, 1999) and (Jurafsky and Martin, 2000), which I most respectfully recommend to read for deeper insights.

Natural Language Processing (NLP) is one designation besides names like '... speech and language processing, human language technology, ... computational linguistics, and speech recognition and synthesis.'(Jurafsky and Martin, 2000, p. 1) This expanding field of science can be defined in different ways. Following the Stanford Encyclopedia of Philosophy, "Computational linguistics is the scientific and engineering discipline concerned with understanding written and spoken language from a computational perspective, and building artifacts that usefully process and produce language, either in bulk or in a dialogue setting" (Schubert, 2015, online). In accordance with (Uszkoreit, 2009), (Uszkoreit, 2002) & (Uszkoreit, 1996), Computational Linguistics is located between classic linguistics and computer sciences as an engineering-driven part of natural language processing. Computational linguistics puts the focus more on engineering than linguistic problems and is an important basis for the field of artificial intelligence (AI).

Natural Language, spoken and written, contains information not directly accessible to computing environments. Natural Language Processing aims to build a bridge between the human and computational domain.

"The goal of this new field is to get computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech" (Jurafsky and Martin, 2000, p. 1).

NLP faces a huge amount of challenges, bringing the complexity of Natural Language into computational perspectives. Following (Jurafsky and Martin, 2000), some applications deliver satisfactory results already, while other applications still need a lot of research and development.

2.4.1 Part-of-Speech Tagging (POS)

Part-of-Speech Tagging refers to "Assigning labels for grammatical classes of words through a computer program" (Mitkov, 2005, p. 751) or as Martin & Jurafsky define "the process of assigning a part-of-speech or other lexical class marker to each word in a corpus. Tags are also usually applied to punctuation markers; thus tagging for natural language is the same process as tokenization for computer languages, although tags for natural languages are much more ambiguous" (Jurafsky and Martin, 2000, p. 296). Words in texts processed with systems using this technique (also known as *taggers*) are marked with a tag usually describing the grammatical property usually, as this example shows.

> "The/DA man/NOUN went/VERB to/TO a/DT bar/NOUN nearby/JJ ./."

The context is very important, removing dis-ambiguities for example "She *flies*" = verb vs. "the *flies*" = noun.

This field of Natural Language Processing is a very important part of '... speech recognition, natural language parsing and information retrieval'(Jurafsky and Martin, 2000, p. 296) and there are many different tools available. These taggers assign labels to each token using a *tagset* contained in text corpora. One of the greatest challenges for these taggers is ambiguity. Words or tokens can have different meanings and the amount of these ambiguities varies from language to language. Following R.A. Amsler, who analyzed the structure of the Merriam-Webster Pocket Dictionary (MPD), "There are ten standard grammatical 'Parts Of Speech' (POS) in the MPD, namely: nouns (N), verbs (VB), adjectives (AJ), adverbs (AV), prepositions (PP), pronouns (PN), conjunctions (CJ), interjections (IJ), and the definite and indefinite articles (DA,IA)" (Amsler, 1980, p. 43). Since words can be ambiguous, Amsler states that of all parts of speech in the Merriam-Webster Pocket Dictionary only "3 parts of speech ... collectively account for 98% of all the main entries in the MPD: nouns, verbs and adjectives" (Amsler, 1980, p. 44). G. Hirst analyzed the results of Amsler and ordered words by their number of senses. Table 2.4.1 shows an excerpt of these results.

WORD	CATEGORY	SENSES
go	verb	63
fall	verb	35
run	verb	35
turn	verb	31
way	noun	31
work	verb	31
do	verb	30

Table 2.4.1: *English words with the greatest number of senses (excerpt) (Hirst, 1987, p. 5)*

Hirst differenciates 2 ambiguity problems (Hirst, 1987, p. 4ff):

- Word sense and case slot disambiguation
  Words can have more than one meaning as stated before.

- Syntactic disambiguation
  Not only words can have more than one meaning - even whole sentences "... can have parses with hundreds of meanings if semantic constraints are not considered" (Hirst, 1987, p. 9).

In linguistics the term ambiguity is a rather inaccurate description for a quite complex topic. Faced with Natural Language Processing problems, the differences of the terms *Homonymy* vs. *Polysemy*, *Hyponymy* vs. *Hypernymy* and *Synonymy* vs. *Antonymy* are very important to understand. They are all part of a field called Lexical Semantics, which is "... the study of what words mean and how their meanings contribute to the compositional interpretation of natural language utterances" (Pustejovsky and Boguraev, 1996, p. 1).

- Homonymy
  ... a lexical form with several meanings.
  A1: "There is constant vacuum in *space*."
  A2: "She needs more *space* for her shoes."
  B1: "I read a very interesting *book* yesterday."
  B2: "Please *book* that flight to Cuba."
  "... it is usually clear that the distinct senses associated with a word are unrelated. For this reason, it is assumed that homonyms are represented as separate lexical entries within the organization of the lexicon" (Pustejovsky and Boguraev, 1996, p. 8).

- Polysemy
  "... is the relationship that exists between related senses of a word, rather than arbitrary ones" (Pustejovsky and Boguraev, 1996, p. 8).
  A1: "He is reading the *newspaper* on the bus stop."
  A2: "She works for the *newspaper* for 7 years now."
  B1: "He makes a fishing rod out of this piece of *wood*."
  B2: "The bear lives in the *wood* now."
  Same as homonyms, polysemes have several meanings, but they are related to each other.

- Hyponymy and Hypernymy
  "Hyponymy is a relation of inclusion or entailment. A superordinate term

(or 'hypernym') includes a set of cohyponyms" (Brinton, 2000, p. 135). Figure 2.4.1 shows a hierarchy of Fish Hyponyms where "fish" is the Hypernym of "salmon" which has the co-hyponyms "sole", "bass", "trout" and "snapper". The relation can be transitive, so every word can be both a Hypernym and a Hyponym. The word "salmon" in figure 2.4.1 shows this transitivity. These relations are also asymmetric which means that a Hypernym always includes the Hyponym but not the other way around: A salmon will always be a fish, but a fish will not always be a salmon.

- Synonymy
  "Generally speaking, synonymy denotes the phenomenon of two or more different linguistic forms with the same meaning. Those linguistic forms are called synonyms, e.g. *peace* and *tranquility* can be substituted with one another in certain contexts" (Stanojević, 2009, p. 1). Synonymy therefore stands in contrast to Homonymy and Polysemy (one word has several meanings) - several words describe one meaning.

- Antonymy
  "Antonyms are words with opposite meanings: *hot* and *cold* or *long* and *short*" (Manning and Schütze, 1999, p. 110).

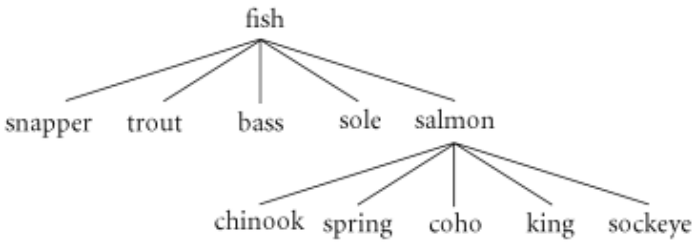

Figure 2.4.1: *A Hierarchy of Fish Hyponyms (Brinton, 2000, p. 135)*

More examples, demonstrating the ambiguity of language acutely, can be found in the work by Jeff Gray, Ph.D. at the University of Alabama who has created a large selection of ambiguous statements (Gray, 2003).

2.4.2 Named Entity Recognition (NER)

"Named Entity Recognition (NER) is the information extraction task of identifying and classifying mentions of people, organisations, locations and other named entities (NEs) within text. It is a core component in many natural language process-

ing (NLP) applications, including question answering, summarisation, and machine translation" (Nothman et al., 2013, p. 2).

These named entities are usually tagged within the text to enable further processing. In this short example a person, a location and a time specification are tagged:

"[Bill Gates]/PER plans to arrive at [Vienna Airport]/LOC at [2 pm]/TIME."

Correct tokenization is crucial for this technique to achieve good results. For example "Vienna" vs. "Vienna Airport", which result in slightly different locations.
Early systems, mostly built on rule-based algorithms, are making way for machine learning techniques. Three major factors have influence on the results (Nadeau and Sekine, 2007):

- Language
  Languages have different grammar and structure and so the amount of research is not evenly spread. English, German, Spanish, Dutch, Japanese, French, Chinese, Greek and Italian are represented well. (Nadeau and Sekine, 2007) Although German tends to be one of the more researched languages "... the level of performance of NER for German is still considerably below the level for English although German is a well-researched language" (Reimers et al., 2014, p. 104).

- Textual genre or domain factor
  "... textual genre (journalistic, scientific, informal, etc.) and domain (gardening, sports, business, etc.)" (Nadeau and Sekine, 2007, p. 2) have also influence on the results. "... experiments demonstrated that although any domain can be reasonably supported, porting a system to a new domain or textual genre remains a major challenge" (Nadeau and Sekine, 2007, p. 2).

- Entity type factor
  "Overall, the most studied types are three specializations of 'proper names': names of 'persons', 'locations' and 'organizations'. These types are collectively known as 'enamex' since the MUC-6 competition" (Nadeau and Sekine, 2007, p. 3). These types can be divided into sub-types such as "city", "state" or "country" for locations or "politician", "musician" etc. for persons. Nadeau & Sekine state that beside these three popular specializations many other types exist, depending on the domain.

Recent studies (Derczynski et al., 2015) show that the performance of the various systems depend highly on the domain they are trained on or developed for. Systems trained on News-based corpora are performing poorly in web domains (e.g.

Twitter) and vice-versa. Still the domain of web texts such as tweets and Facebook posts are still not performing robustly. "Some Twitter-specific methods reach F1 measures[1] of over 80%, but are still far from the state-of-the-art results achieved on newswire" (Derczynski et al., 2015, p. 47). The reason for this lies in the fact that NER is highly vulnerable to incorrect tokenization, incorrect capitalisation, typographic errors and shortenings, which are less likely to occur on news-data than in social media or microblog environments. Beside these data-specific errors also genre-specific characteristics appear, such as hyperlinks in texts, multilingual posts, noisy content (emoticons, idiosyncratic abbreviations, etc.), social context. Derczynski defines several countermeasures against these errors:

- Language Identification Algorithms
  Ensuring the correct NER component is used.
- Part-of-Speech Tagging
  For error reduction – low success and impact.
- Normalisation
  For reducing spelling errors – low success and impact.

2.4.3 Sentence Boundary Disambiguation

"Segmentation is the process of taking an undifferentiated sequence of symbols and 'segmenting' it into chunks. For example sentence segmentation is the problem of automatically finding the sentence boundaries in a corpus" (Jurafsky and Martin, 2000, p. 178). Considering a human being, finding the start and end point of a sentence won't be a big challenge. (Manning and Schütze, 1999, p. 134f) state that a punctuation (. ? ! : , ; etc.) is a sentence boundary in 90% of the cases, regarding the English language. In fact, especially the period is a very ambiguous element (Grefenstette and Tapanainen, 1994, p. 4), taking a closer look to the following example:

> CELLULAR COMMUNICATIONS INC. sold 1,550,000 common shares at $21.75 each yesterday, according to lead underwriter L.F. Rothschild & Co. on Friday, 12.07.2000 at 15:00:00.[2]

It is quite obvious that language-specific particularities like abbreviations, dates, decimal separators etc. introduce errors on correct sentence segmentation. A lot

[1] Combination of recall & precision (Jurafsky and Martin, 2000, p. 576)
[2] Modified version of the example from (Kiss and Strunk, 2006, p. 486).

of approaches deal with this problem of NLP, since it marks one of the basic steps prior further processing. Following, (Read et al., 2012, p. 987ff) those approaches can be categorized into three classes:

- Rule-Based SBD
  Handcrafted lists are used to treat abbreviations (e.g. (Stanford University, 2015)).
- Machine Learning (Supervised)
  Learning against a pre-defined training set in a supervised setup.
- Machine Learning (Unsupervised)
  Learning with unannotated corpora (e.g. (Kiss and Strunk, 2006)).

(Read et al., 2012, p. 989ff) analysed the performance of representative approaches in this field and showed a significant drop of performance (up to 5%) between standardized corpora and user-generated content (texts from blogs, wikipedia, etc.).

2.4.4 Lemmatization

"A lemma is a canonical form - usually the base form - taken as the representative for all the different forms of a paradigm" (Mitkov, 2005, p. 38). The task of lemmatization therefore relates to "... finding the corresponding dictionary form for a given input word ..." (Mitkov, 2005, p. 38). Lemmatization should not be mixed up with the related method of stemming, whereas stemming does not include the context and is therefore, not able to link words less related correctly.

For example:

- Lemma 'walk' for: walking, walked, walks
- Lemma 'good' for: better, best

In the first example, the stemmer and the lemmatizer are able to find *walk* as a base form, while the lemma *good* can only be found by the lemmatizer. (Manning and Schütze, 1999, p. 132) state that, especially within the field of Information Retrieval (IR), the contribution of Lemmatization to the performance is not clearly positive, even negative in some cases. Three reasons are responsible for that:

- Loss of information due to too general terms
- Split of tokens into several and losing the actual meaning

- Lower level of inflection and derivation in the English language

Considering the German language, especially the level of inflection and derivation is far more complex and still makes sense. This step is essential in order to be able to remove stop words (described in a later section) efficiently and to increase the overall accuracy of the system. Therefore many approaches on this topic appear especially for the German language e.g. (Perera and Witte, 2005) or (Lezius et al., 1998). One of the fastest and most accurate for the German language are the mate-tools (Bohnet, 2010a) by Bernd Bohnet and Anders Björkelund. "For the languages of the 2009 CoNLL Shared Task[3], the parser could reach higher accuracy scores on average than the top performing systems" (Bohnet, 2010b, p. 96).

2.4.5 Word Splitting

Compound words are a phenomenon across nearly all languages. While there are no uncontroversial definitions (Scalise and Vogel, 2010, p. 5) & (Bauer, 2005), one can say compound words are basically words that consist of more than one stem. It is not clear whether compounding can be regarded as universal (Bauer, 2001). Still they are not very common in the English language and therefore are often ignored in NLP applications, while there are a lot of compounds in the German language. Following (Berton et al., 1996, p. 1165) there are two types of compounds:

- Semantic
  The semantic content cannot be derived from the individual parts. And the individual parts change the semantic context after separation
  e.g. *Skyscraper / Wolkenkratzer* or *bittersweet*.
- Orthographic
  The semantic content stays mostly the same after separation for the individual parts e.g. *School-bus / Schulbus*.

"German orthography requires that both types of compounds be written together, thus resulting in an exceedingly large and potentially unlimited set of compounds" (Berton et al., 1996, p. 1165). The well-known German compound word *Donaudampfschiffahrtselektrizitätenhauptbetriebswerkbauunterbeamtengesellschaft*[4] is a prominent proof for this unlimited compounding of nouns.

[3] A conference by the Linguistic Data Consortium to promote NLP applications and evaluate their performance. `https://catalog.ldc.upenn.edu/LDC2012T03` last access: 30.08.2015

[4] Association for subordinate officials of the head office management of the Danube steamboat electrical services.

Several approaches regarding NLP deal with this topic:

- TAGH (Geyken and Hanneforth, 2006)
  Based on weighted Finite State Automata, MORPA (Heemskerk et al., 1993) segmenting the word into atomic units and then testing it for morphological correctness.

- De Rijke & Monz (Monz and De Rijke, 2002)
  A recursive approach dividing the word into substrings based on a lexicon.

- Alfonseca, Bilac & Pharies from Google Inc. (Alfonseca et al., 2008)
  Creating a universal compounding model across state-of-the-art procedures.

While the decomposition of compound words makes sense for the *Donaudampf-schiff...* example, (Berton et al., 1996) state that the accuracy of the processing can decrease if very frequent compounds are divided, i.e. especially for semantic compounds the dividing is counterproductive. Suppressing the decomposition of frequent and semantic compounds seems to be the most promising approach.

2.4.6 Query Expansion

"Query expansion (QE) is a technique commonly used in information retrieval (IR) to improve retrieval performance by reformulating the original query – either adding new terms or reweighting the original terms" (Vechtomova and Wang, 2006, p. 324), (Beaulieu and Jones, 1998) & (Rocchio, 1971). Following (Carpineto and Romano, 2012) & (Jurafsky and Martin, 2000, p. 653) the most prominent usage for this technique are modern web search engines (e.g. google, bing, yahoo, etc.). Users are already used to type two or three keywords into the search field and, most of the times, retrieve the information they searched for. Therefore query expansion plays an important role in this field, adding synonyms or hypernyms (section 2.4.1) to expand the possible matches and subsequently extend the recall of documents. This is in most cases done by using "Thesauri" which are basically collections of words grouped by their similarity of meaning. This approach does not include the context of the query and is often criticised of reducing the precision in IR due to the ambiguity of words, which extends the searched term with wrong synonyms and distorts the original meaning. Another point is that most thesauri are often not suitable for the intended application. In the last years new data sources e.g. wikipedia, new methods e.g. ontologies (Bhogal et al., 2007) and hybrid methods especially, are improving the results and reliability of this technique (figure 2.4.2).

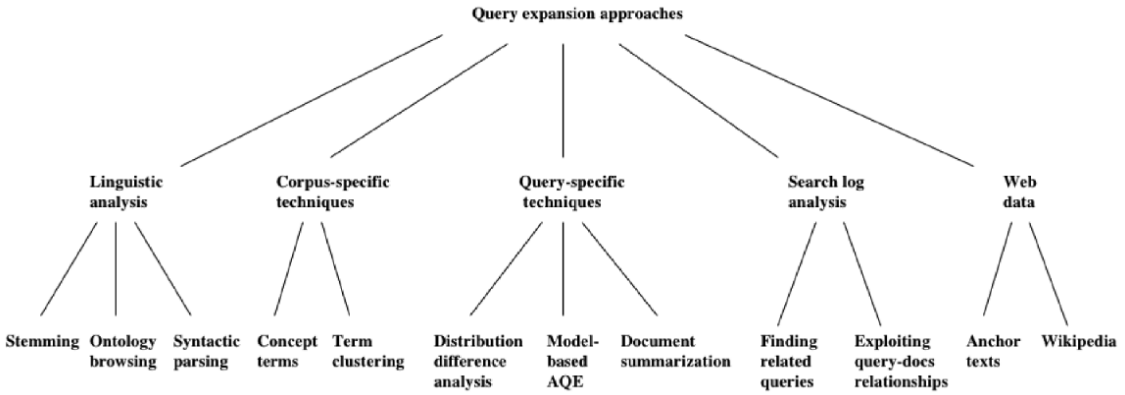

Figure 2.4.2: *A taxonomy of approaches to AQE (Carpineto and Romano, 2012, p. 25)*

2.4.7 Term Weighting

Having large amounts of text, NLP faces the challenge of categorizing those texts according to their content. Now to classify these texts or documents, significant words have to be extracted. A very common guess is that the most frequent words in a text are more significant. The famous *Luhn Assumption* states that the weight of a term that occurs in a document is simply proportional to the term frequency (Luhn, 1957). Term frequency is defined as "A measurement of the frequency of a word or term within a particular document. Term frequency reflects how well that term describes the document contents" (Mitkov, 2005, p. 759).
Given a collection of N documents, f_{ij} defines the number of occurrences of word i in document j. The term frequency TF_{ij} is then calculated as:

$$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}}$$

where the term frequency is normalized by the maximum number of occurrences of any term (Leskovec et al., 2014). To get a better understanding of these calculations a detailed look at the complete works of Shakespeare is very useful. A text file containing all works can be downloaded on the page of the *Projekt Gutenberg* (Projekt Gutenberg, 2015). Prior to the analysis all words have been converted to lowercase and all special characters (except "-") have been removed. A detailed look at the occurrences proves that *Stop Words* are very likely to become the most common words in any text. Table 2.4.2 shows the 10 most frequent words in all of Shakespeares works (Projekt Gutenberg, 2015) and all of them are stop words. However removing those stop words (table 2.4.3), does not bring better results. Given Shakespeares works, words like "lord", "good", "now" appear very often, but won't

be able to differentiate one document from the other. Significant words, defining the topic or category of a text are relatively rare. But also turning the occurrences the other way around (table 2.4.4) returns words without a particular meaning. The term frequency can be normalized through dividing the occurrences in table 2.4.2 by the maximum amount of occurrences, in this case 27536. However this normalization still returns the most frequent words of the documents. An appropriate filter is created by calculation of the inverse document frequency (idf). Using the term frequency, always the most frequent word in a document will get a value of 1 and those words are in most cases less useful stop words.

word	occurrences	share
the	27536	3,94%
and	26660	3,82%
i	20630	2,96%
to	19046	2,73%
of	18098	2,59%
a	14570	2,09%
you	13528	1,94%
my	12463	1,79%
that	11090	1,59%
in	10944	1,57%

Table 2.4.2: *10 most frequent words in all of Shakespeare's works (with stop words in red)*

word	occurences	share
thou	5471	0,78%
thy	4034	0,58%
shall	3574	0,51%
thee	3147	0,45%
lord	2989	0,43%
king	2841	0,41%
good	2804	0,40%
now	2750	0,39%
o	2593	0,37%
come	2481	0,36%

Table 2.4.3: *10 most frequent words in all of Shakespeare's works (without stop words)*

Spärck-Jones proposed that "terms should be weighted according to collection frequency, so that matches on less frequent, more specific, terms are of greater value than matches on frequent terms" (Spärck-Jones, 1972, p. 493). and laid the foundation for the inverse document frequency. "The measure of term specificity first proposed in that paper later became known as inverse document frequency, or IDF; it is based on counting the number of documents in the collection being searched which contain (or are indexed by) the term in question" (Robertson, 2004, p. 503). The formula is defined as:
Given a collection of N documents, a given term t_i appears in n_i documents of N.

word	occurences
self-substantial	1
buriest	1
niggarding	1
gazed	1
all-eating	1
proving	1
feelst	1
renewest	1
unbless	1
uneared	1

Table 2.4.4: *10 of the least frequent words in all of Shakespeares works (without stop words)*

$$idf(t_i) = log\frac{N}{n_i}$$

(Robertson, 2004, p. 504)

Multiplied with the term frequency, the inverse document frequency forms the tf-idf weighting algorithm. The TF-IDF (Term Frequency - Inverse Document Frequency) Weighting uses the prior described algorithms to provide a more useful weighting value. Following (Salton and Buckley, 1988), the TF-IDF is calculated as:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

Where t marks a term, d the document and D the document collection.
In section D.2 I made a detailed TF-IDF analysis on the works *Hamlet* and *Henry the 8th* by William Shakespeare in order to find significant words.

2.4.8 Similarity and Dissimilarity Measures

"Informally, the similarity between two objects is a numerical measure of the degree to which the two objects are alike. Consequently, similarities are *higher* for pairs of objects that are more alike. Similarities are usually non-negative and are often between 0 (no similarity) and 1 (complete similarity). The dissimilarity between two objects is a numerical measure of the degree to which the two objects are different. Dissimilarities are *lower* for more similar pairs of objects. Frequently, the term distance is used as a synonym for dissimilarity, ... Dissimilarities sometimes fall in the interval [0,1], but it is also common for them to range from 0 to ∞" (Tan et al., 2005, p. 66).

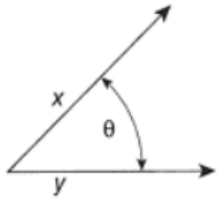

Figure 2.4.3: *Geometric illustration of the cosine measure (Tan et al., 2005, p. 76)*

Several measures are common, depending on the type of objects to be compared by using Euclidean, Minowski or Hamming Distance, Confusion Matrices, Simple Matching Coefficient, etc. (Tan et al., 2005, p. 67ff).

One promising measure for similarity in combination with the prior described TF-IDF weighting is the Cosine Similarity. "Documents are often represented as vectors, where each attribute represents the frequency with which a particular term (word) occurs in the document" (Tan et al., 2005, p. 74). Therefore it is possible to calculate the similarity of two documents (or sentences), by using the Cosine Similarity formula (Tan et al., 2005, p. 74ff):

$$cos(x, y) = \frac{x}{||x||} \cdot \frac{y}{||y||}$$

Dividing x and y by their lengths, normalizes the result to a length between 0 and 1. Figure 2.4.3 shows the geometric basis of this measure, where x and y are the vectors of two documents and the angle between those vectors, calculated by the cosine describe the similarity. Hence, the smaller the angle, the more similar the two vectors are.

2.4.9 Other Applications, Techniques & Concepts

In this section an excerpt of other Applications, Techniques & Concepts of NLP are mentioned. For deeper insights I most respectfully recommend to read (Manning and Schütze, 1999) and (Jurafsky and Martin, 2000).

- Tokenization
  "The process of segmenting text into linguistic units such as words, punctuation, numbers, alphanumerics etc." (Mitkov, 2005, p. 754). Tokenization is part of the first steps in any NLP application, which includes the prior described sentence boundary disambiguation and divides text into processable *tokens* (Grefenstette and Tapanainen, 1994).
  Directing the attention to prior described compound words, tokenization plays an important role for further processing considering semantic compound words e.g. *Skyscraper* split to *Sky* & *Scraper*. Depending on the domain, tokenization has a vast influence on performance of NLP applications, e.g. Biomedical Document Retrieval (Trieschnigg et al., 2007) or porting linguistic grammars to free text (Forst and Kaplan, 2006).

- Regular Expressions
  The foundation for regular expressions were laid by mathematician Stephen Cole Kleene in 1951 (Kleene, 1951), who characterized "regular sets as the smallest class of sets of strings which contains all finite sets of strings and which is closed under the operations of union, concatenation and Kleene closure" (Friedl, 2006, p. 260). Applied and further developed, regular expressions now are an integral component of computer science and linguistics from compilers to simple text processing programs. In a computing environment words are basically sets of symbols, called *strings*. A regular expression is able to characterize a set of strings based on a single string defined as *pattern*. Regular expressions follow a clear syntax with special characters, for example brackets "[ ]" specifying a disjunction for characters inside, an asterisk "*" (called "Kleene star") specifying *zero or more occurrences of the previous symbol* and

many others, turning regular expressions into a powerful and very important tool of computation and linguistics (Jurafsky and Martin, 2000, p. 22ff).

- Text Corpora
  Following the Oxford Handbook of Computational Linguistics a corpus (pl. corpora) is defined as "A body of linguistic data, usually naturally occurring data in machine readable form, especially one that has been gathered according to some principled sampling method" (Mitkov, 2005, p. 734). Text corpora are very useful for evaluating the performance of NLP applications. The origin of the corpora data varies from newspaper texts, to classics of literature and user generated content (e.g. blogs, social media). There are several corpora for different languages and domains available, for example the Penn Treebank corpus (Computer and Information Science Department, 1999) or the Wikipedia Corpus (Davies, 2014). A very detailed list of available corpora can be found on the page of the International Linguistics Community (International Linguistics Community, 2014).

- N-Grams
  'N-Gram grammar is a representation of an N-th order Markov language model in which the probability of occurrence of a symbol is conditioned upon the prior occurrence of N-1 other symbols.'(Brown et al., 2001) Depending on the order, three main types of n-grams can be distinguished: Unigram (1-gram), Bigram (2-gram) & Trigram (3-gram). According to the chosen or required unit (character or word), different n-gram sequences result. In this regards table 2.4.5[5] shows the n-gram sequences of the quote:

  "*An unexamined life is not worth living.*"-Socrates

  The main advantage of the use of n-grams in NLP applications is the reduction of data and thus reduction of processing time. N-grams are also used in other fields of computer science e.g. compression algorithms, speech recognition etc. (Jurafsky and Martin, 2000, p. 189ff).

[5]Character bigram & trigram were truncated due to lack of space.

	character as unit
sample	an_unexamined_life_is_not_worth_living
unigram	a,n,_,u,n,e,x,a,m,i,n,e,d,_,l,i,f,e,_,i,s,_,n,o,t,_,w,o,r,t,h,_,l,i,v,i,n,g
bigram	an,n_,_u,un,ne,ex,xa,am,mi,in,ne,ed,d_,_l,li,if,fe, ... ,rt,th,h_,_l,li,iv,vi,in,ng
trigram	an_,n_u,_un,une,nex,exa,xam,ami,min, ... ,t_w,_wo,wor,ort,rth,th_,h_l,_li,liv,ivi,vin,ing

	word as unit
sample	an unexamined life is not worth living
unigram	an,unexamined,life,is,not,worth,living
bigram	an unexamined,unexamined life,life is,is not,not worth,worth living
trigram	an unexamined life,unexamined life is,life is not,is not worth,not worth living

Table 2.4.5: *N-Gram examples with quote of Socrates*

First proposed informally by (Weaver, 1955, p. 20f) and based on Shannon's advances in information theory (Shannon and Weaver, 1959), the term *n-gram* was first introduced by (Solomonoff, 1957, p. 4) and is now one of the basic techniques in NLP applications.

Several training corpora are available for different languages. The most prominent and comprehensive example is the Google N-Gram database[6] or the Microsoft Web N-gram Services[7]

- Stop Words
  Stop Words are needed for correct grammar but deliver less useful semantic information. Common examples are "the", "and", "a", "an", "not", etc. for the English language. They are often removed before further processing of texts to increase performance by reducing the processed text. Stop words are criticised for having no clear definition, leading to a vast amount of differing stop word lists and for reducing the semantic information in texts when removed. A common example for this problem is the sentence "*to be or not to be*" from Hamlet, which is composed entirely out of stop words. Another example with the text "*the* meal *was not* good", which implies "meal was bad" turns after stop word removal to the opposite semantic value: "meal good". Research also shows that the removal of stop words can reduce retrieval performance depending on the domain and application. (Zaman et al., 2011)

- Language Identification
  Especially for language specific algorithms (e.g. Tokenization, Lemmatization,

[6] `http://googleresearch.blogspot.co.at/2006/08/all-our-n-gram-are-belong-to-you.html` last access: 30.08.2015

[7] `http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx` last access: 30.08.2015

Named Entity Recognition etc.) knowing the language is important (Grefenstette, 1995). Following (Dunning, 1994) & (Grefenstette, 1995) early approaches were using unique letter combinations, common words or N-grams. Current research concentrates on N-grams as stable base e.g. (King et al., 2014), (Shuyo, 2010), (Lui and Baldwin, 2012) or (Vatanen et al., 2010).

- Sentiment Analysis
  Following (Cambria et al., 2013, p. 15f), common tasks of Sentiment Analysis, along with Opinion Mining includes *polarity classification* (e.g. like or dislike) & *Agreement detection* even extended with the degree of positivity to enable several applications e.g. troll filtering or cyber-issue detection. Four main categories for existing approaches are pointed out (Cambria et al., 2013, p. 18f):

  – Keyword Spotting
    Classification of text using unambiguous keywords e.g. happy, sad, afraid, bored etc.

  – Lexical Affinity
    Including the context of those keywords training probability from linguistic corpora.

  – Statistical Methods
    Including elements of machine-learning algorithms, support vector machines etc.

  – Concept-based Techniques
    Use of web ontologies or semantic networks to detect even subtly expressed sentiments.

  Driven through the evolution of the web, including more and more user generated content, Sentiment Analysis is in the focus of current research, "... distinguishing itself as a separate field, falling between NLP and natural language understanding" (Cambria et al., 2013, p. 19).

  Current approaches are facing several difficulties due to dependency on other NLP applications (e.g. "... coreference resolution, negation handling, anaphora resolution, named-entity recognition, and word-sense disambiguation" (Cambria et al., 2013, p. 15)) or the use of Acronyms & Sarcasm. These difficulties are able to reduce the performance of Sentiment Analysis, compared to educated human beings, by 23% (Paragon Poll, 2013). Especially sarcasm, which is basically "the use of words that mean the opposite of what you really want

to say especially in order to insult someone, to show irritation, or to be funny" (Merriam-Webster, 2015) is a well-known issue. In contrary to texts of newspapers, books, official documents etc. where sarcasm is not very popular, the processing of user generated content often faces the challenge of sarcastic statements. Identifying sarcasm with an automated system seems to be impossible with the current possibilities and even human beings are struggling at the correct identification of sarcastic statements in some domains (González-Ibánez et al., 2011).

- Coreference Resolution
  "Coreference resolution is the problem of identifying which noun phrases (NPs, or mentions) refer to the same real-world entity in a text or dialogue" (Ng, 2009, p. 575). The term *Coreference* includes *Anaphora* which are "... the linguistic phenomenon of pointing back to a previously mentioned item in the text ..." (Mitkov et al., 2000, p. 1) and *Cataphora* which are mentioned before the item they are referring to (Jurafsky and Martin, 2000, p. 669). Linguists argue a distinction between coreference and anaphora (Holler-Feldhaus, 2004, p. 2). "Anaphora normally operates within a document (e.g. article, chapter, book), whereas coreference can be taken to work across documents" (Mitkov et al., 2000, p. 1).
  A very vivid example by (Hirst, 1981) shows the challenges coreference resolution faces:

  "*The chinchilla ate my portrait of Richard Nixon last night. It devoured it so fast, I didn't even have a chance to save the frame*" (Hirst, 1981, p. 3).

  For most humans reading this example it is obvious that *Richard Nixon* most likely mentions Nixon, the former president of the United States of America. It will also be obvious that the first *it* in the second sentence refers to the chinchilla and the second to the portrait. At last, it is also obvious, that *the frame* refers to the frame of the prior mentioned portrait, according to (Hirst, 1981).

  Research in coreference resolution was put into the spotlight by the CoNLL-2011[8] shared task. A team of the Stanford NLP group was able to deliver the best performance (Pradhan et al., 2011).

[8] A conference by the Linguistic Data Consortium to promote NLP applications and evaluate their performance. `http://conll.cemantix.org/2011/` last access: 30.08.2015

2.5 Semantic-based Modeling

In this chapter, the connection and extension of conceptual models based on a meta modeling approach via the SeMFIS (OMILAB, 2015) (Fill, 2012) toolkit developed on the ADOxx platform (ADOxx.org, 2015) is considered more closely.

2.5.1 Definitions

"Semantic Business Modeling describes a set of activities including their functional, behavioral, organizational, operational as well as non-functional aspects. These aspects are not only machine-readable, but also 'machine-understandable' which means that they are either semantically annotated or already in a form which allows a computer to infer new facts using an underlying ontology" (Lautenbacher et al., 2008, p. 61).

2.5.2 Conceptual Modeling

"Conceptual modelling is the activity of formally describing some aspects of the physical and social world around us for purposes of understanding and communication" (Mylopoulos, 1992, p. 3). In comparison to natural language or diagrammatic notations, conceptual modeling is able to capture the semantics of the application with its formal notation. Also in comparison to mathematical or other formal notations, conceptual modeling has an advantage by supporting "structuring and inferencial facilities that are psychologically grounded" (Mylopoulos, 1992, p. 3). Still, all advantages that come with conceptual modeling are intended to be used by humans, not machines (Mylopoulos, 1992). Analogous to Business Process Modeling (2.3.2), the semantics of conceptual models are often given as labels typically in natural language and meta models, as a semi-formal specification, only formalize syntax and type semantics. Business process modeling goes one step further and formalizes the execution or simulation but does not consider inherent semantics. To make models machine-processable formal semantics are needed e.g. making information about the labels of classes in models explicit (Fill, 2012).

2.5.3 Meta-modeling Framework

Models, which are basically an abstract representation of the reality, are also instances of meta-models, defining modeling techniques and mechanisms & algorithms (figure 2.5.1). The modeling technique is divided into a modeling language defining notation, syntax and semantics and a modeling procedure, containing steps to apply the modeling language and produce results. The syntax represents the grammar and consists of rules which defines the model creation while the semantics define the meaning of the syntactic elements. The notation of this framework is not static and allows to define the visualisation of an element based on the syntax and by considering the attached semantics (Fill and Karagiannis, 2013) & (Karagiannis and Kühn, 2002). The specifications of notations follow formal specifications e.g. the elements & their relations, and informal specifications e.g. descriptions in natural language, which can be defined as semi-formal specification (Bork and Fill, 2014).

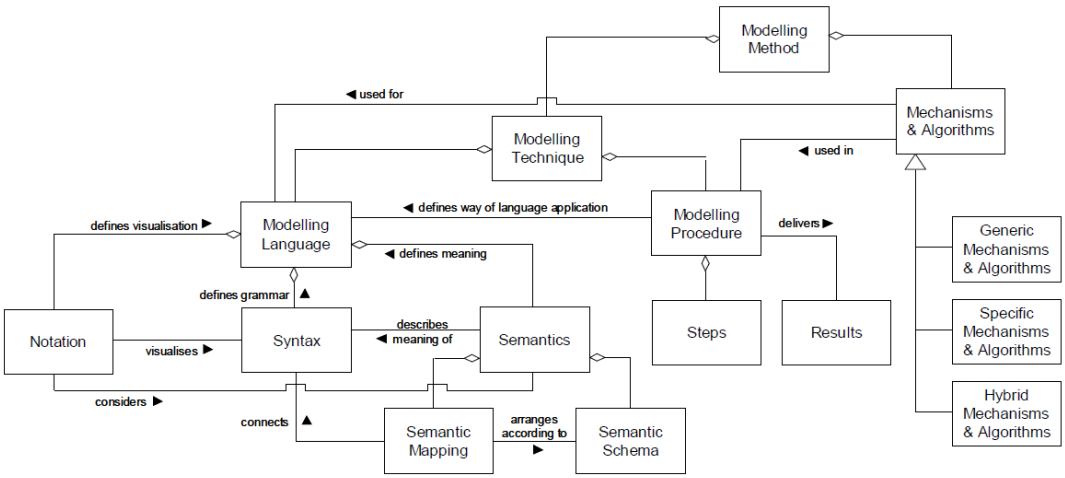

Figure 2.5.1: *Components of modelling methods (Karagiannis and Kühn, 2002, p. 3)*

The meta-model itself is based on a meta-meta-model (meta2-model) defined through a metamodeling language. Figure 2.5.2 shows this modeling hierarchy between the different levels and models.
Using higher levels of abstraction in form of meta-models, entails enhanced flexibility and interoperability. Rapid changes in business requirements, which increase the effort of maintaining the overview and growing complexity in developing applications are challenges that are alleviated using the meta-model approach (Fill and Karagiannis, 2013) & (Karagiannis and Kühn, 2002).

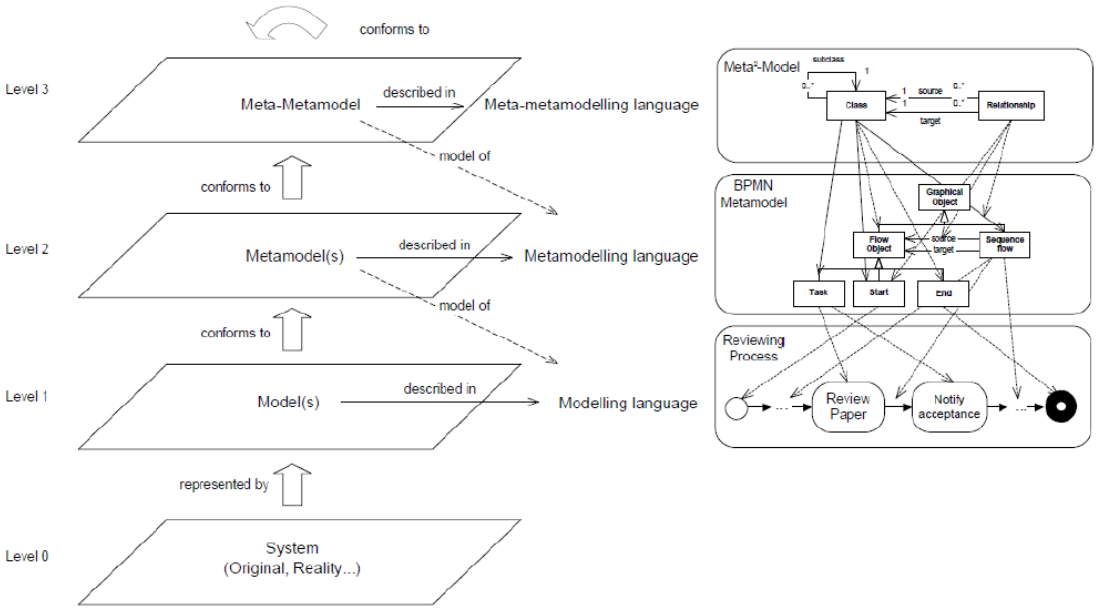

Figure 2.5.2: *Modeling Hierarchy (Karagiannis and Kühn, 2002)*

2.5.4 Ontologies

"An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an ontology is a systematic account of Existence. For knowledge-based systems, what "exists" is exactly that which can be represented" (Gruber, 1993, p. 2). Regarding the field of Information Science "Ontologies are used by people, databases, and applications that need to share domain information (a domain is just a specific subject area or area of knowledge, like medicine, counterterrorism, imagery, automobile repair, etc.) Ontologies include computer-usable definitions of basic concepts in a particular domain and the relationships among them. Therefore, ontologies encode knowledge in a domain and also knowledge that spans domains. So, they make that knowledge reusable" (Obrst, 2003, p. 366). Expressed in a logic-based language usually, ontologies enable automated reasoning, which then again enable intelligent applications (e.g. conceptual/semantic, search and retrieval (non-keyword based), software agents, decision support, speech and natural language understanding, knowledge management, intelligent databases, and electronic commerce) (Obrst, 2003). The term Ontology stands for a huge variety of models with different degrees of semantic richness and complexity and this context can be visualized by the Ontology Spectrum (figure 2.5.3).

Depending on the degree of semantic richness, several ontology languages are available and are divided into Traditional Ontology Specification Languages (e.g.

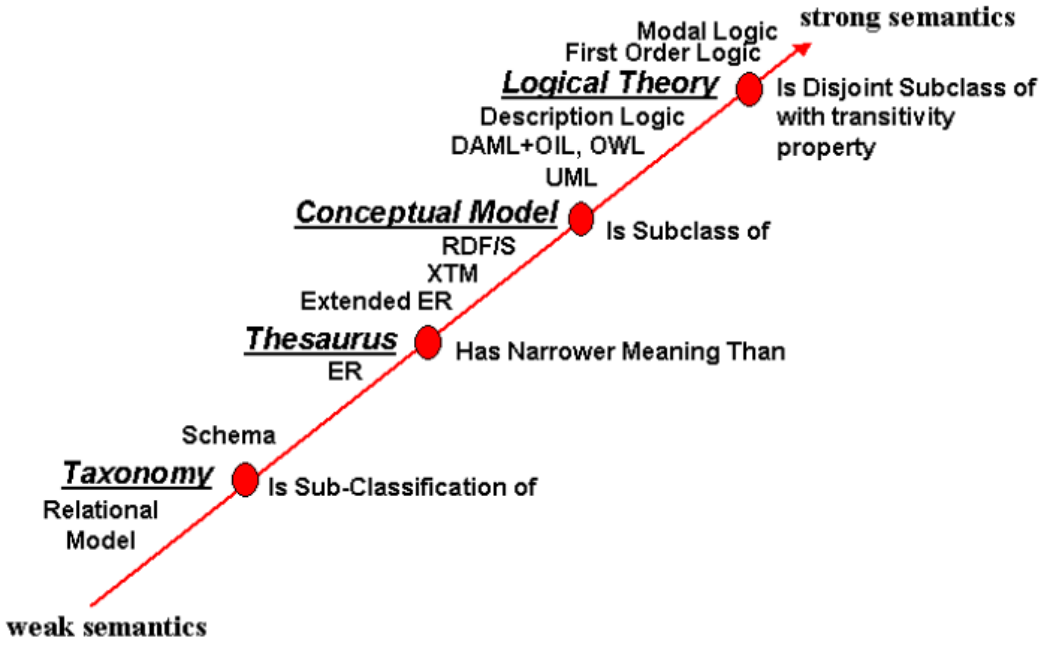

Figure 2.5.3: *Spectrum of Ontologies (Obrst, 2003, p. 367)*

Ontolingua, OCML, FLogic, LOOM), Web Standards (e.g. XML, RDF) and even Web-based Ontology Specification Languages (e.g. XOL, SHOE, OIL) (Corcho and Gómez-Pérez, 2000). Based on the Resource Description Framework (RDF), the Web Ontology Language (OWL) has been developed by the W3C[1] and is now one of the most important ontology languages (Cardoso and Pinto, 2015).
While the variety of Ontology Specification Languages widely stays steady, the variety of tools has increased rapidly. One can distinguish between three different environments in which the tools are located:

- Domain-specific environments
- Integrated environment for ontology development
- Supporting system for ontology development based on various techniques

While the majority of tools are freely available, also commercial tools for ontology development have increased (Mizoguchi and Kozaki, 2009) and several lists of tools are available varying from about 250 entries (W3C, 2014) up to 800+ (Bergman, 2015).
One popular tool used in the sphere of this paper is Protégé[2] (Tudorache et al., 2013) & (Gennari et al., 2003), free available, developed open source and maintained by

[1] World Wide Web Consortium `http://www.w3.org/` last access: 30.08.2015
[2] `http://protege.stanford.edu/` last access: 30.08.2015

Stanford University. The tool is used for "... knowledge acquisition, merging and alignment of existing ontologies, and plug-in new functional modules to augment its usability" (Mizoguchi and Kozaki, 2009, p. 328). It supports many ontology languages, is highly extensible due to a very sophisticated plugin structure and provides advantages in collaboration, standards compliance, server architecture or refinement (Mizoguchi and Kozaki, 2009).

2.5.5 The SeMFIS Approach

The SeMFIS approach and the toolkit aims at providing "an open platform for describing the semantic aspects of multiple conceptual modeling languages and models" (Fill, 2012, p. 6). Using the SeMFIS approach, the enrichment of semi-formal conceptual models with semantic aspects, e.g. an ontology model, leads to enhanced flexibility in terms of processing and the possibility to incrementally improve the model with semantic information according to the individual needs (Fill, 2012). Based on the ADOxx Modeling Platform and building upon concepts of meta modeling (figure 2.5.1), SeMFIS provides Semantic Annotation Models, which provide the ability to create semantic mappings between models of various modeling languages on a meta model perspective, using inter-model-references (figure 2.5.4). By doing so, users of the system have the possibility to use their familiar and intuitive conceptual models and add semantic and processable information by mapping semantic information e.g. ontology models or thesaurus models via annotation models, obtaining the highest level of flexibility whereas the model becomes machine-processable. This concept creates a bridge between Conceptual Modeling of regular users and the Spectrum of Ontologies, which categorizes the needed semantic methods and languages (Fill and Burzynski, 2009).

Two particular advantages can be pointed out (Fill, 2014, p. 140f):

- The consistency to other models is ensured based on the original specification "It also means that any algorithms that were based on the original state of the language do not need to be adapted" (Fill, 2014, p. 141).

- All annotations and processing capabilities can also be added ex-post "Thus it is not necessary for the user to deal with formal semantic definitions at first hand and only add them as needed" (Fill, 2014, p. 141).

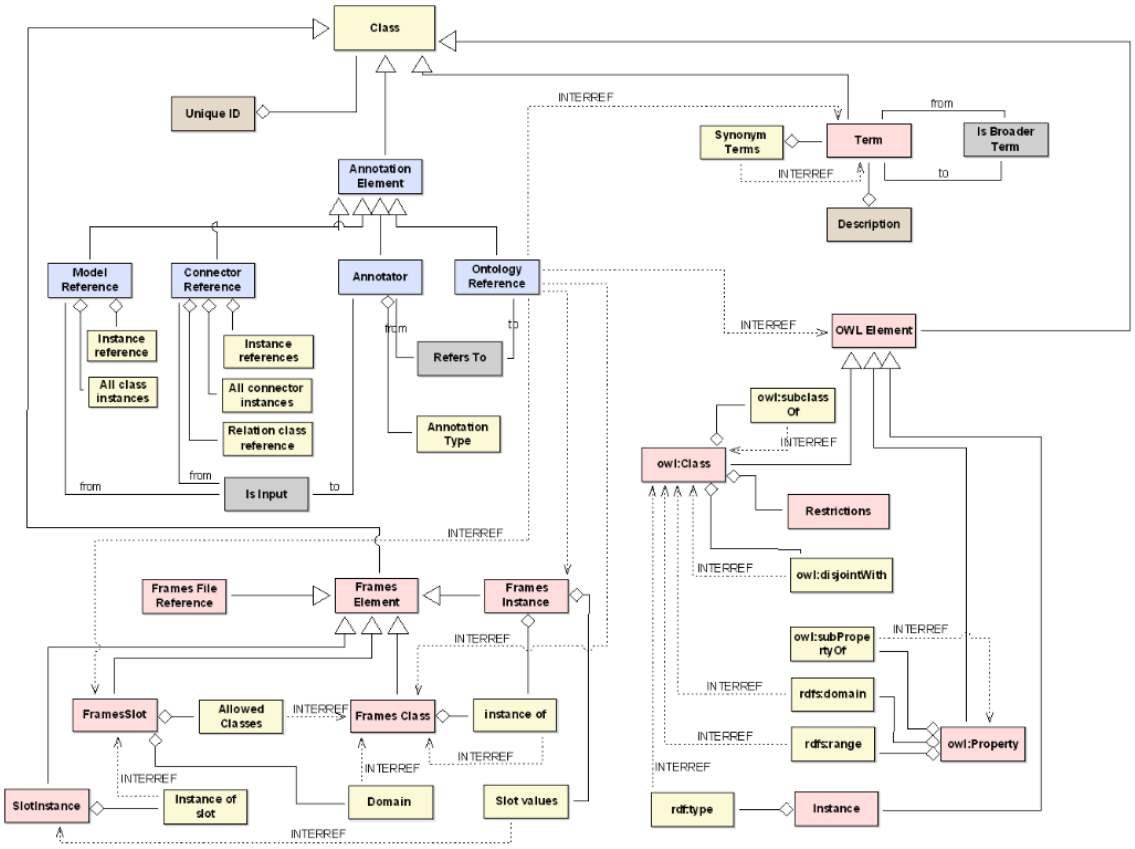

Figure 2.5.4: *SeMFIS Meta Model (Fill, 2015, p. 6)*

Overall, the toolkit combines analysis, simulation and evaluation components, HTML generation, import/export components, web services (e.g. REST), ontology management via a Protégé plugin (Fill et al., 2013) and is still in further development (Fill, 2012).

Part 3

Integrating Natural Language Processing and Semantic-based Modeling

3.1 Introduction

Based on the knowledge gained in the previous part, the concepts devised in the introduction were implemented using conceptual modeling and natural language processing based on the SeMFIS toolkit and Java as programming platform. The following part describes the development and implementation and is divided in six chapters:

- Chapter 3.2 ”Methodology”: The scientific methodologies used as foundation in order to gain new knowledge based on the preliminary research and under consideration of the field of study.
- Chapter 3.3 ”Concept & Requirements”: Based on the introduced motivation and problem statement the concept and requirements of the system are defined in this chapter.
- Chapter 3.4 ”Solution Approach”: Derived and refined from the first concepts an outright solution approach is described using conceptual modeling.
- Chapter 3.5 ”System Overview”: Description of the structure, components, interfaces and procedures of the developed system consisting of a conceptual modeling method, a Java-based application and a categorization service used to achieve prior defined requirements and provide a reliable basis for the evaluation approach.
- Chapter 3.6 ”Semantic-based Modeling”: Description of the modeling approach, extending the SeMFIS toolkit in order to combine natural language processing with conceptual modeling.
- Chapter 3.7 ”Integrating Natural Language Processing”: Based on the knowledge gained in the prior part of this thesis, the natural language processing methods are described in this chapter. Several measures have been taken and are described, facing obstacles that appeared during application of standardized methods and techniques on user-generated content loaded from social media platforms.

3.2 Methodology

Following (Howell, 2012), the methodology defines strategies and methods to be used that form the basis for the procedure and design of scientific research. Depending on the field of research, several methodologies are recommended. Focusing on recommendations in the field of business informatics (Wilde and Hess, 2007)[1], design-oriented information systems research (Österle et al., 2011) and under consideration of the research questions, the following four methodologies were applied during this work:

- Formal Deductive Research
  For this Thesis argumentative research and semiformal modeling as part of model-driven engineering are able to generate knowledge as well as to support the implementation. Model-driven engineering focuses on the creation of models in every development stage.

- Simulation
  "Illustrates the behaviour of the investigated system in a model and reenacts environmental states with the configuration of the model-parameters. The construction of the model and the observation of the endogenous modelfactors can provide new knowledge" (Wilde and Hess, 2007, p. 282). The simulation method was applied on the natural language processing part of the Thesis. Simulation can be categorized as a constructive, quantitative research method which fits to the requirements of the natural language processing.

- Prototyping
  "A pre-version of the system is developed and evaluated. Both steps are able to generate new knowledge" (Wilde and Hess, 2007, p. 282). A target of the Thesis was to develop a working implementation of the theoretical foundation. Prototyping definitely matches this intention and provided new knowledge.

- Survey
  Using Surveys, Questionnaires, Interviews etc. is able to generate new knowledge through the collection of data by a representative sample of a population. In this work the knowledge of participants has been collected to create reference data used for the evaluation of the system.

[1] Translated from German.

3.3 Concept & Requirements

Based on the considerations introduced in chapter 1.1, the gained knowledge in part 2 and by using the presented methodologies in chapter 3.2, the following concepts and requirements have been derived.

Figure 3.3.1 shows the first concept, connecting social media data with existing business process models. The social media data which is, in most cases, unstructured, based on natural language and contains a huge amount of information, should be transferred to a data model, focusing on German posts of users on companies' social media presentations in general. For the evaluation in chapter 4.1, a popular Austrian airline, with a seemingly high rate of user feedback on its social media presentation, has been used as test subject. On the other side, documented business processes of the company, which are usually structured and machine-processable, should be translated into a business topic ontology, considering interaction points between the company and the customers, and act as reference for the categorization in a later step. Both models loaded into a suitable data model contain natural language, which is processed using natural language processing techniques, depending on application-specific requirements. These requirements define the process steps of the natural language processing using a semantic-based modeling method. Both mentioned models are using the SeMFIS toolkit (Fill, 2014) as a platform and extend its functionality. The subsequent categorization of the social media data model under consideration of the business ontology and defined by the prior mentioned semantic-based modeling method, results in a data set of clustered entries, which are then transformed into a model integrated in the SeMFIS platform, providing valuable insights about the aforementioned business processes contained in the business ontology in a clean and clear way. Users of the system should be able to retrieve a detailed and clear overview of the defined business topics contained in the business ontology combined with the social media data, revealing trends, identifying occurring problems and opportunities using the resulting insights gained out of the massive amount of social media data. Two main components arise out of this first concept already: (1) A Java-based application, responsible for the loading of social media data and the natural language processing and (2) the SeMFIS toolkit, which provides the basis for modeling the business ontology, defining the process and presenting the gained results.

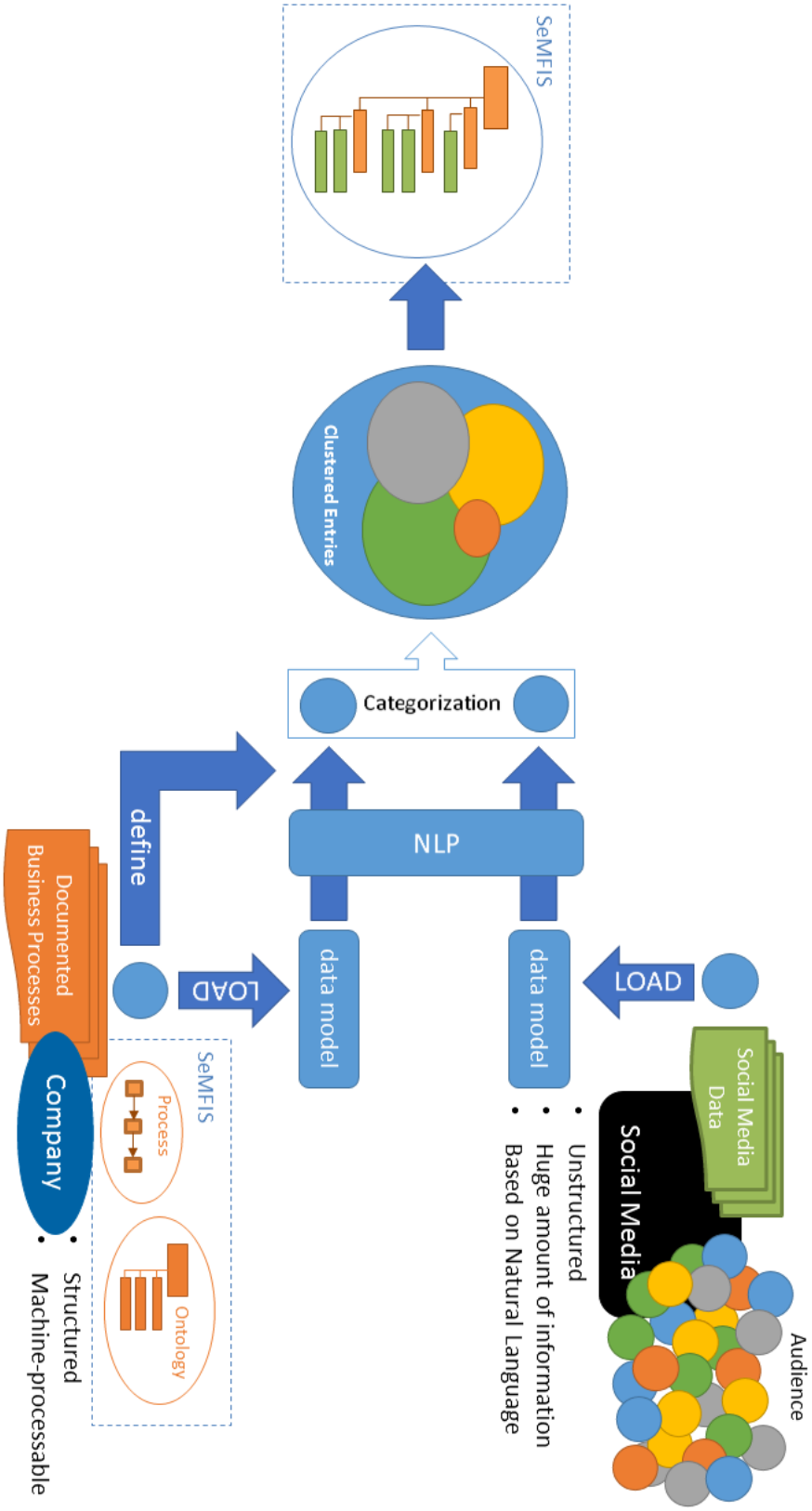

Figure 3.3.1: *First concept of the system*

Based on the knowledge gained and the prior defined concept, the primary requirements of the user had to be considered in order to develop a suitable conceptual modeling method. By developing a use-case diagram (figure 3.3.2), three main needed functionalities were derived. Defining a reference model, defining the processing and reviewing the results of the system. Defining the reference model includes the possibility to extend the model as needed. Defining the process of the system includes the definition of the loader criteria (e.g. which social media platform, time period etc.), being able to filter posts (e.g. by length, language etc.). It also includes the definition of the NLP process as well as the categorization criteria. After the posts are processed, the user should be able to review the results and gain valuable insights about trends, problems and possible opportunities.

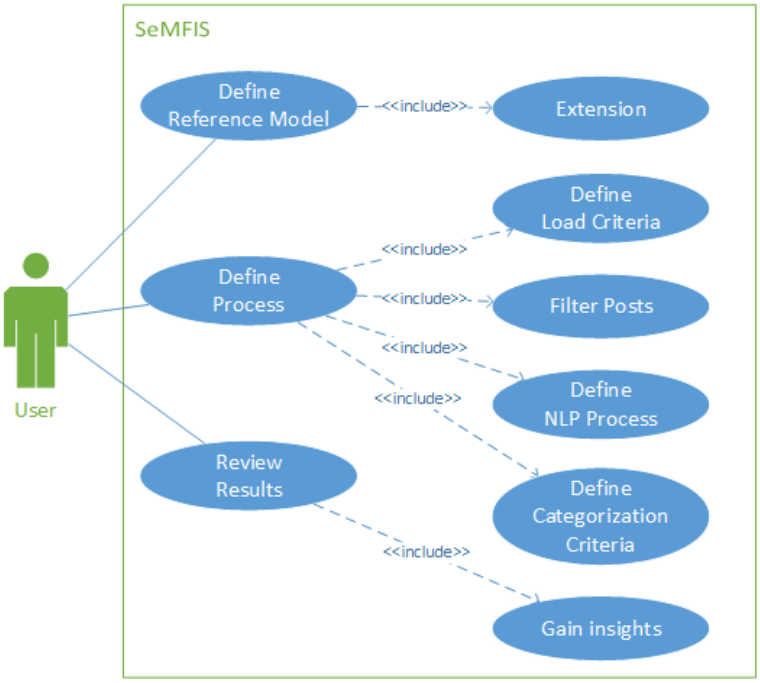

Figure 3.3.2: *Use Case Diagram of the System*

Based on this concept and the aforementioned requirements, an integrated system based on Java and the SeMFIS platform has been developed, which is described in the following chapter.

3.4 Solution Approach

Derived from prior described concepts, three models for the system need to be developed (figure 3.4.1):

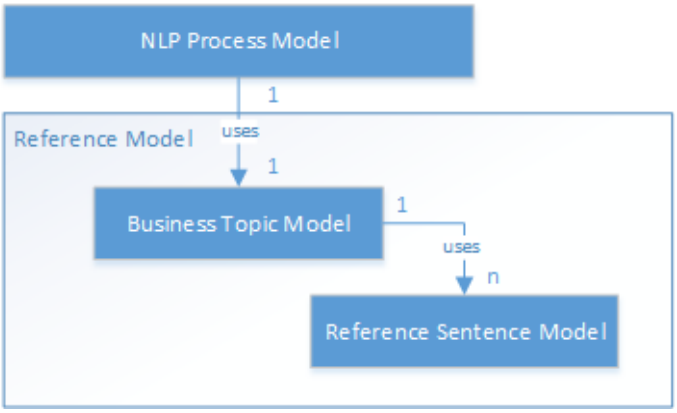

Figure 3.4.1: *First concept defining the models*

(1) NLP Process Model, which defines the process steps and settings of the system and references to a (2) Business Topic Model, which specifies the primary business ontology and references to several (3) Reference Sentence Models, which contain reference sentences necessary for the categorization. Business Topic and Reference Sentence Models virtually form the Reference Model of the system.

Combining concepts, requirements and platforms, results in a conceptual model (figure 3.4.2), depicting the components and relationships of the approach. The NLP process model, which contains the process and the settings and the reference model, built out of business topic and reference sentence model, are modeled by the by the user of the system under consideration of the documented business processes and application-specific requirements, using the extended SeMFIS platform. Depending on the structure, substance and availability of the business processes and the required target, the application requirements, defining the process steps and settings are derived. To start the processing, the NLP application is executed via SeMFIS, which provides the needed models for the application by exporting them to an XML file[1]. This model file contains the NLP process and reference model defining the business ontology, the process and the settings needed by the NLP application. Based on these settings the language data of a social media appearance, defined by the user, and the reference model is processed and categorized, resulting in a model based on the structure of the reference model and represented as XML file. The user is able to import the resulting model, using standard components of SeMFIS and gains insights about the social media data. A detailed sequence of the components (figure 3.5.2) can be found in a later section.

[1]Detailed information about the processes and data models of the nlp application can be found in chapter 3.7.

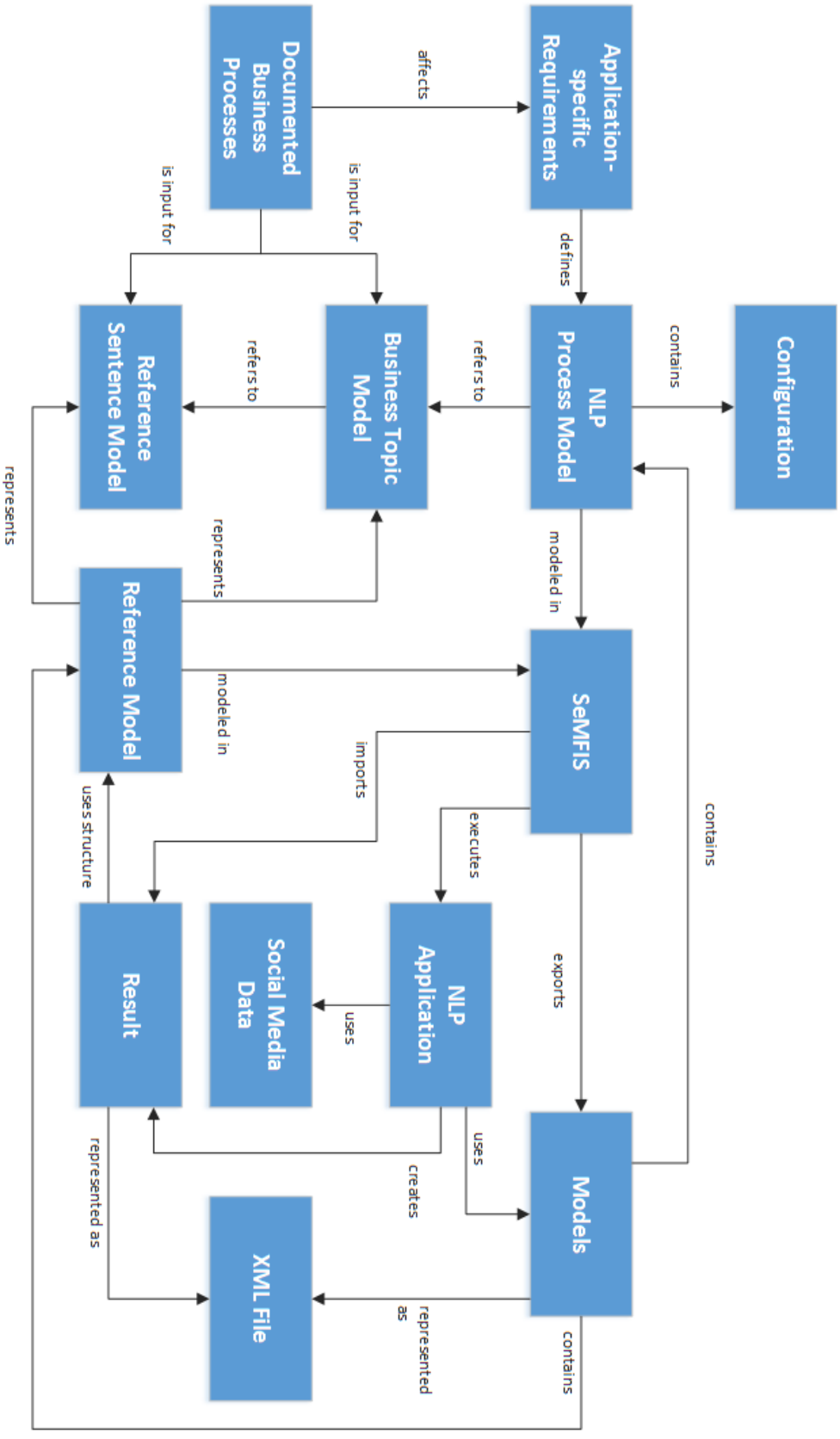

Figure 3.4.2: *Conceptual relationships of the approach*

3.5 System Overview

In this chapter, the main components and processes of the system are specified. This includes the conceptual modeling methods developed based on the SeMFIS toolkit, a Java-based application for the natural language processing and a categorization service. This categorization service is not part of the main system as it is only used to gain reference data for the evaluation part of this thesis.

3.5.1 Components and Interfaces

Figure 3.5.1 shows the four main components of the system: SeMFIS, NLP Application, File System and SurveyDB[1]. These are interacting through several interfaces: ADOScript to export the Reference Model as well as the packaged Java Application to the File System and to execute the App via CommandLine. The package java.io is used to load the Reference Model and the NLP-specific files (e.g. Stop Word Lists, etc.). After the processing is complete, java.io exports the Resulting Model to the File System, which is then imported by ADOxx. For the evaluation a connection to the Survey Database has been established, which is basically a simple SQL database located in the Microsoft Azure Cloud System[2].

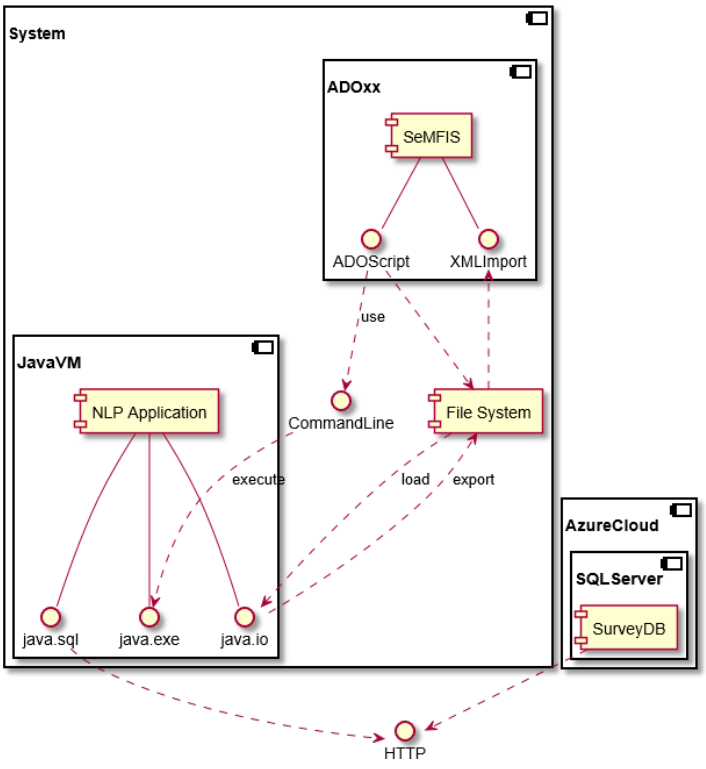

Figure 3.5.1: *Main Components*

[1]The Categorization Web Service (section 3.7.2) is not part of the main system and was used one-time only in order to gain reference data stored in the SurveyDB.

[2]For more information visit: `http://azure.microsoft.com` last access: 30.08.2015

3.5.2 Processes and Interfaces of the System

Figure 3.5.2 shows the main participants and processes of the system. The main interaction of the system with the user is limited to SeMFIS, which is therefore responsible for execution of the NLP Application. The interaction of the modeling toolkit and the NLP application is achieved, using ADOScript. The executable jar and all needed libraries are stored in the SeMFIS database and are exported via ADOScript to the working directory. This working directory is retrieved from the publisher classes of the NLP process model. Publisher classes contain an attribute "Publish Path", to define where the results of the process should be stored as XML. This path is also used for this purpose. Following this stage, the ADOScript code enables to retrieve the publish path, export the selected models in XML format, copy the NLP files to the working directory and execute the application. The following code samples show the retrieval of the publish path, the export of the reference model into an XML file, the copying of a file to the path location and the execution of the Java application. The complete source code can be found in chapter E.

```
  ...
SET curModel:(VAL fmodel)
# we get the working directory path from the settings class
CC "Core" GET_ALL_OBJS_OF_CLASSNAME modelid:(curModel)
    classname:(setClassName1) #-->RESULT ecode:intValue objids:list

  ...

IF (objids!="")
{
   SET id_obj:(VAL token(objids,0," ")) # we just take the first
       (there should be only 1 setting object
   CC "Core" GET_ATTR_VAL objid:(id_obj) attrname:(setAttrName)
       as-string #attrid:(attrID) as-string #-->RESULT
       ecode:intValue val:anyValue
   SET str_path:(val)
   ...
}
```

Listing 3.5.1: *ADOScript for the retrieval of the publish path*

```
  ...
SETL sADOXMLExportFile:(str_path+"\\SetBO.xml")
# Export ADOxml to the file space for further processing
CC "Documentation" XML_MODELS modelids:(selModels) mgroupids:("")
   attrprofs:("") apgroups:("")
CC "AdoScript" FWRITE file:(sADOXMLExportFile) text:(xml) binary:0
...
```

Listing 3.5.2: *ADOScript exporting models to an XML file*

```
...
CC "AdoScript" DELETE_FILES (str_path+"\\social.jar")
CC "AdoScript" FCOPY from:("db:\\NLP_final.jar")
   to:(str_path+"\\social.jar")
...
```

Listing 3.5.3: *ADOScript copying file from database to path location*

```
...
SYSTEM ("java -Xmx3072m -jar \""+str_path+"\\social.jar\"
   \""+str_path+"\\SetBO.xml\"") with-console-window
...
```

Listing 3.5.4: *ADOScript executing Java jar file*

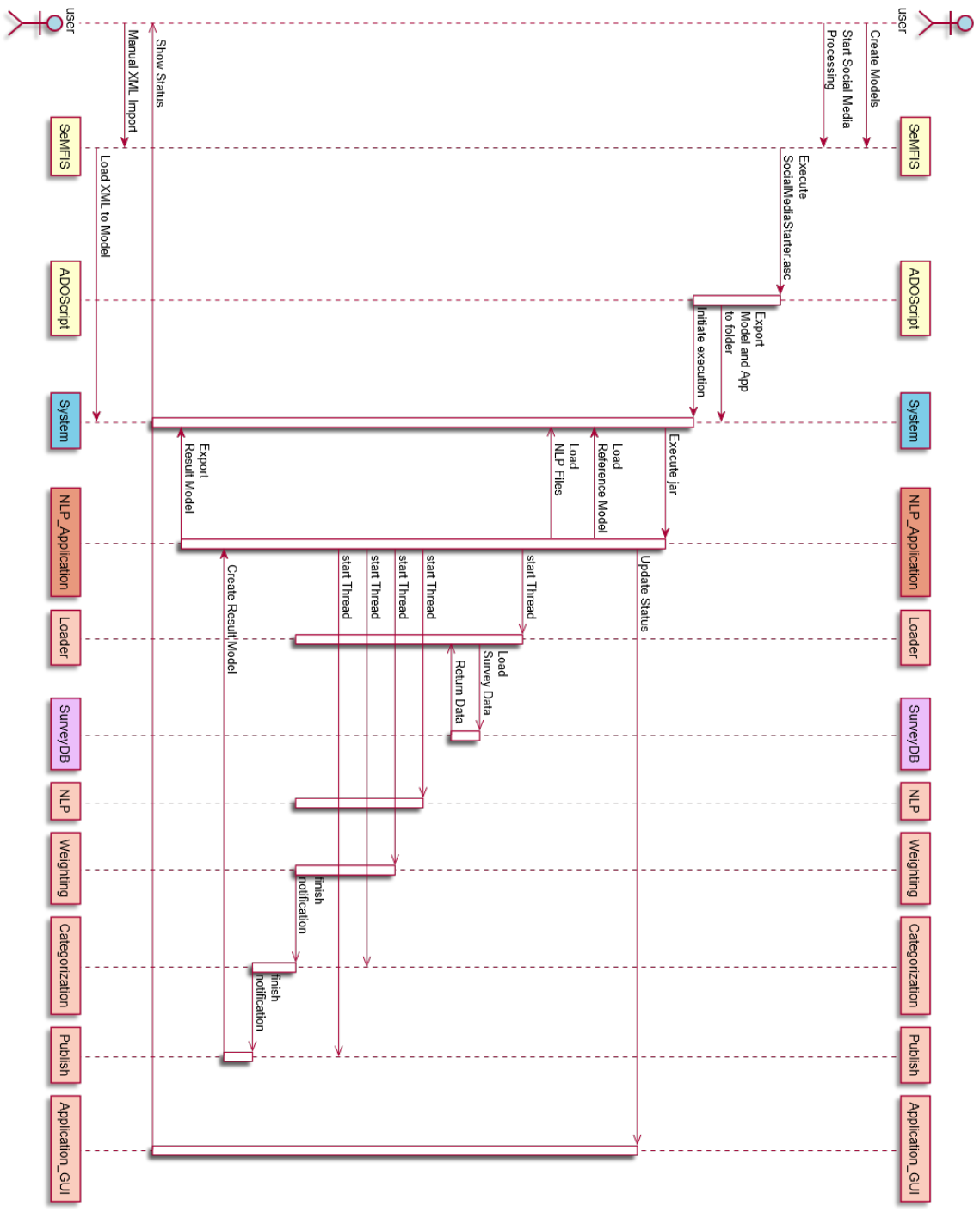

Figure 3.5.2: *Sequence Diagram of the main processes*

3.5.3 Categorization Service & Survey Database

To evaluate the accuracy of the natural language processing component, a survey with a sample of 300 posts has been implemented[3].

Participants were able to categorize the sample posts according to categories and their personal belief. The data of this manual categorization was stored in a database, which then provided the data to the NLP system for detailed comparison. Figure 3.5.3 shows the components of this Social Media Categorization Service. Participants were able to access the service using any Web browser at `http://socialcat.azurewebsites.net`. An Azure Web App based on php[4] and implemented on an invisible virtual server located in the Azure Cloud provided the service to the client. The created data during the run-time of the service was stored in an Azure SQL database implemented on a SQL Server in the Azure Cloud. To communicate with the database AJAX[5], provided by the jquery[6] Java Script library was used. One basic requirement regarding the usability was to develop a very intuitive and fast interface, which could be achieved by using a responsive design with asynchronous server requests. In order to be available on every device by a responsive design, the library jquery mobile[7] was used, so that participants were able to reach the service with their mobile phone as well as with their computers anywhere. Figure 3.5.4 displays the interaction of a user and main participants for this service. Switching pages, displaying or updating a post from the database was done by using Ajax requests, which are handled asynchronously.

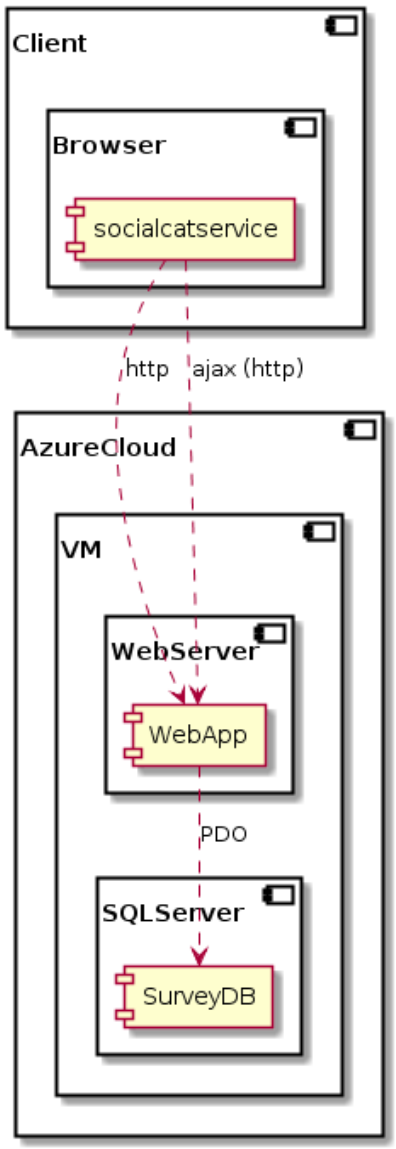

Figure 3.5.3: *Components of the Categorization Service*

[3]Detailed information about the setting and the results in chapter 4.1.
[4]An open-source server-side scripting language.
[5]Asynchronous Java Script and XML
[6]`https://jquery.com/` last access: 30.08.2015
[7]`https://jquerymobile.com/` last access: 30.08.2015

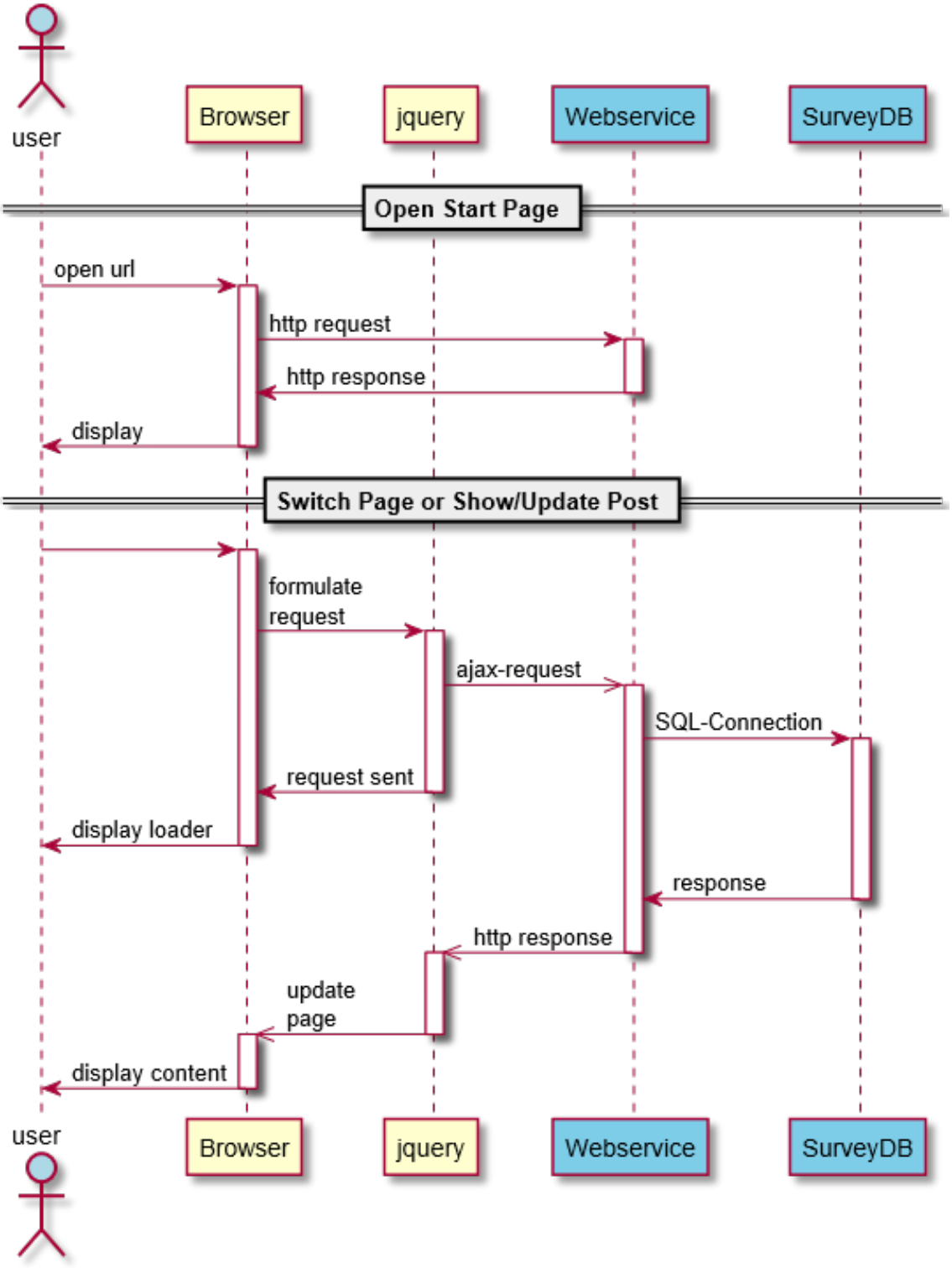

Figure 3.5.4: *Sequence Diagram of the Categorization Service*

Figure 3.5.5 shows the Database Diagram of the Survey Database. Posts are stored into table *tab_Messages* with their original id and the message text (pre-processed for better readability). The categories of the system are stored into table *tab_Categories* and the results of the user input (the Survey data) is stored into *tab_Survey*. To differentiate between the users, the session variable of the http-session was stored additionally.

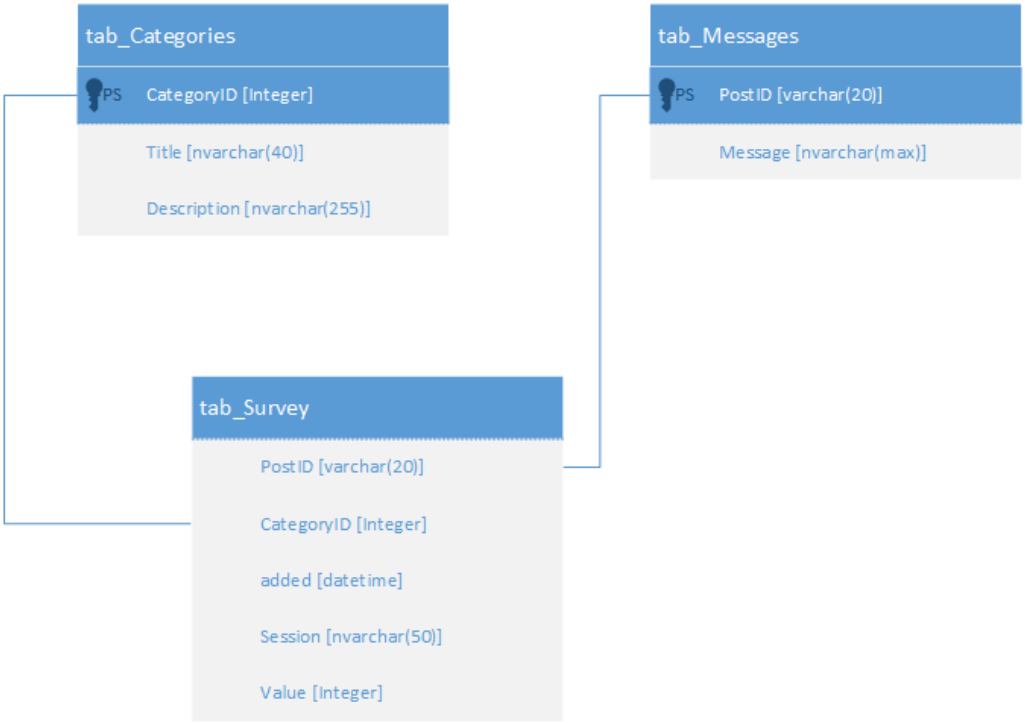

Figure 3.5.5: *Database Diagram of SurveyDB*

Posts and categories of the survey were stored manually into the tables. After creating an acceptable amount of data by the users (*tab_Survey*), the data in the tables was used to be compared to the results of the NLP processing[8]. Hence the Loader Component of the system is able to add the survey data to the Original Post objects directly[9].

[8]Detailed information about the setting and the results in chapter 4.1
[9]chapter 4.1 shows detailed information.

3.6 Semantic-based Modeling

Based on the knowledge gained in chapter 2.5, a modeling approach based on the SeMFIS toolkit has been developed, combining Natural Language Processing with Conceptual Modeling. Prior sections (3.3) were used to derive models, which are able to fulfil the necessary requirements and are described in this chapter. The NLP Process Model, enabling the user to define the process and settings and two models (Business Topic & Reference Sentence) acting as Reference Model. As the NLP process model should contain the settings for the categorization, it must also use the business topic model containing the topics and reference sentence models. In order to be able to review the results of the process, the two reference models act as presentation layer. Figure 3.6.1[1] shows the conceptual relationships of the three models.

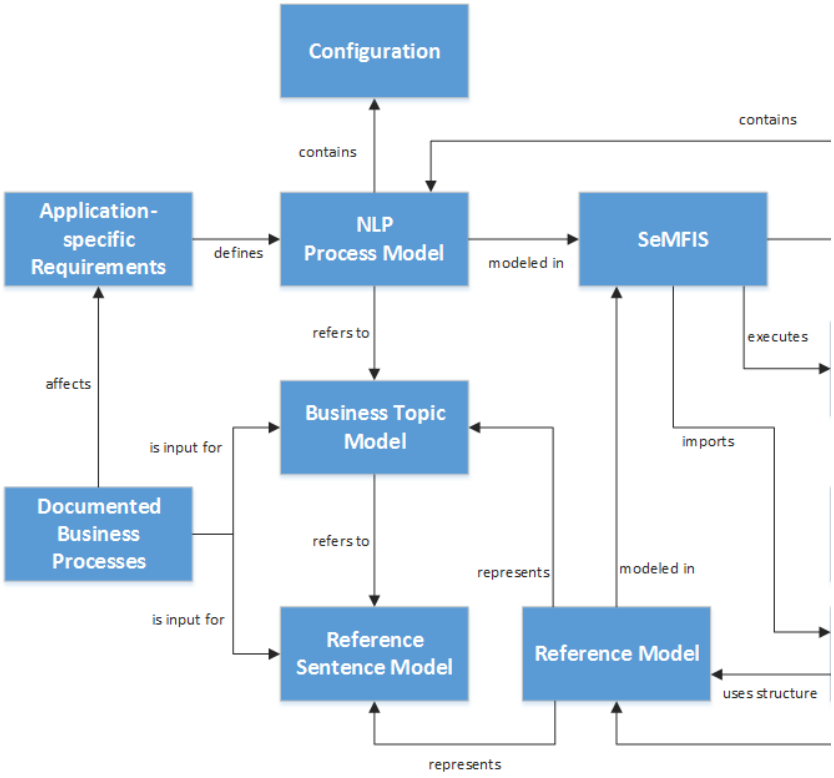

Figure 3.6.1: *Conceptual relationships of the approach (excerpt)*

[1]Excerpt of figure 3.4.2.

3.6.1 Metamodel

Further development of the first concept led to a metamodel (figure 3.6.2), defining the three models used in the system. Three inter-model references (INTERREF) have been defined. Business Topics contain one Reference Sentence Model, the Social Media Loader holds a reference to a Business Topic Model (the reference model has to be processed by the NLP classes as well as the data) and Categorization classes.

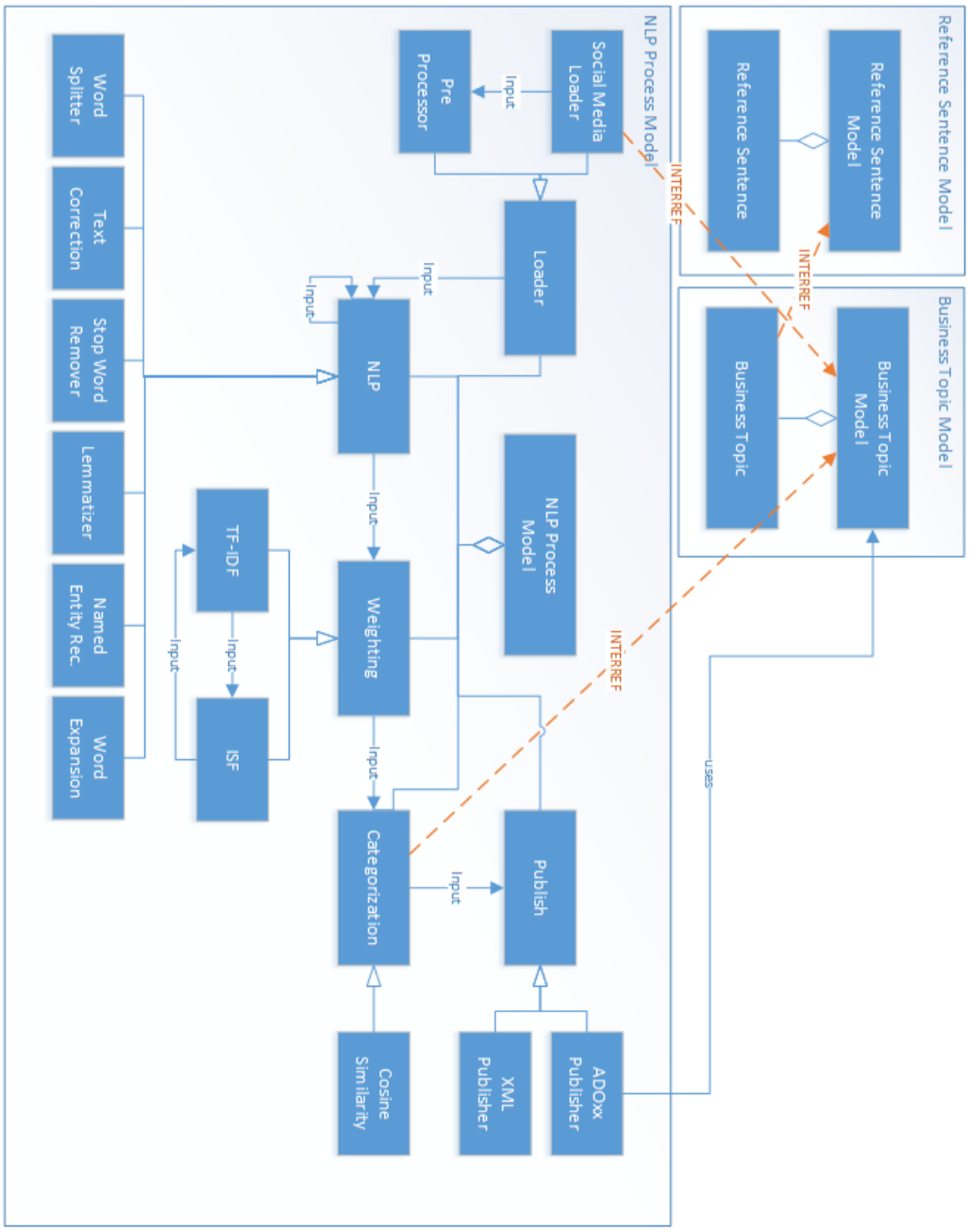

Figure 3.6.2: *Metamodel of the approach extending the SeMFIS metamodel*

3.6.2 Business Topic & Reference Sentence Model

Containing the topics derived from the documented business process models, the business topic model defines the reference categories for the categorization process (figure 3.6.3 shows an example).
Finding these topics can be a real challenge, since they have to define a point at which business processes meet the customer or have a strong influence on the customer's experience. Sometimes the business process models are well documented, but very complex or, on the other hand, are not documented properly at all.

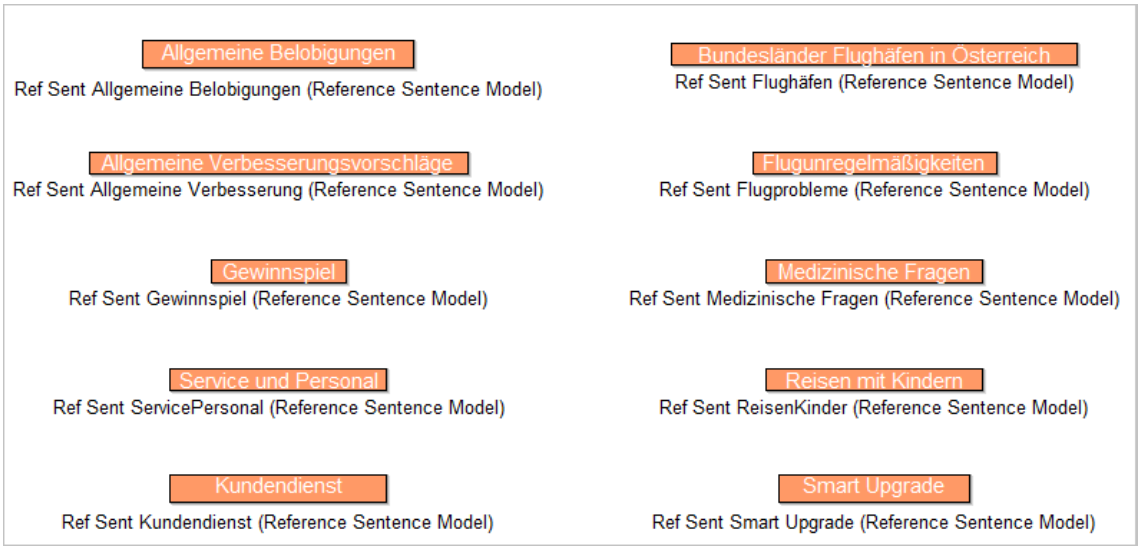

Figure 3.6.3: *Example of a Business Topic Model*

A practical way to face this challenge is the investigation of the customer journey (Richardson, 2010) to identify the processes and topics from a customers perspective. An additional way is the use of Frequently Asked Question (FAQ) sections on the official web appearance. FAQ sections evolve out of re-occurring customer complaints and resemble interaction points of the company with the customers in a very structured way. After the topics are defined, the reference sentence models for each topic are created and linked via an inter-model-reference. Figure 3.6.4 shows an example of a reference sentence model. This model has two uses: (1) It enables to add reference sentences to a certain business topic used for the categorization of the social media data. (2) After the data is processed and categorized, the ADOxx Publisher class of the NLP process model uses the structure of the business topic and reference sentence model to create an XML file, which can be imported to SeMFIS.

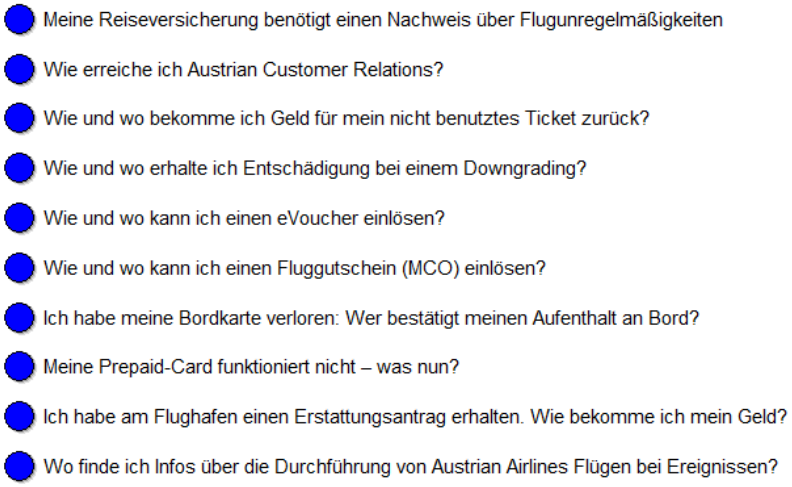

Figure 3.6.4: *Example of a Reference Sentence Model*

The following tables describe the created classes used in the business topic and the reference sentence model.

Business Topic	**Description**
<name> <Reference Sentence Model>	The main class of the model, showing the title and the reference.
Attribute Name	**Type**	**Description**
Name	String	The name displayed also defines the title of the topic.
Reference Sentence Model	INTERREF	Inter-model-reference to the corresponding Reference Sentence Model.

Table 3.6.1: *Class Business Topic*

Reference Sentence	**Description**
<Sentence>	The main class of the model, used as reference sentence or as categorized post entry.
Attribute Name	**Type**	**Description**
Sentence	Longstring	The reference sentenc or the pre-processed text of the social media post.
MatchValue	Double	The match value of the categorization.
Post URL	Longstring	URL to the original post.

Table 3.6.2: *Class Reference Sentence*

3.6.3 NLP Process Model

Based on the knowledge gained in chapter 2.4 the NLP process model was developed, enabling the user to define the process steps and all settings availabe. Figure 3.6.5 shows an example. The model follows five process groups: *Load*, *NLP*, *Weight*, *Categorize* & *Publish*. The order of these groups cannot be changed, but the order within is customizable. To avoid errors in advance the relation between the objects of these classes follow cardinality rules.

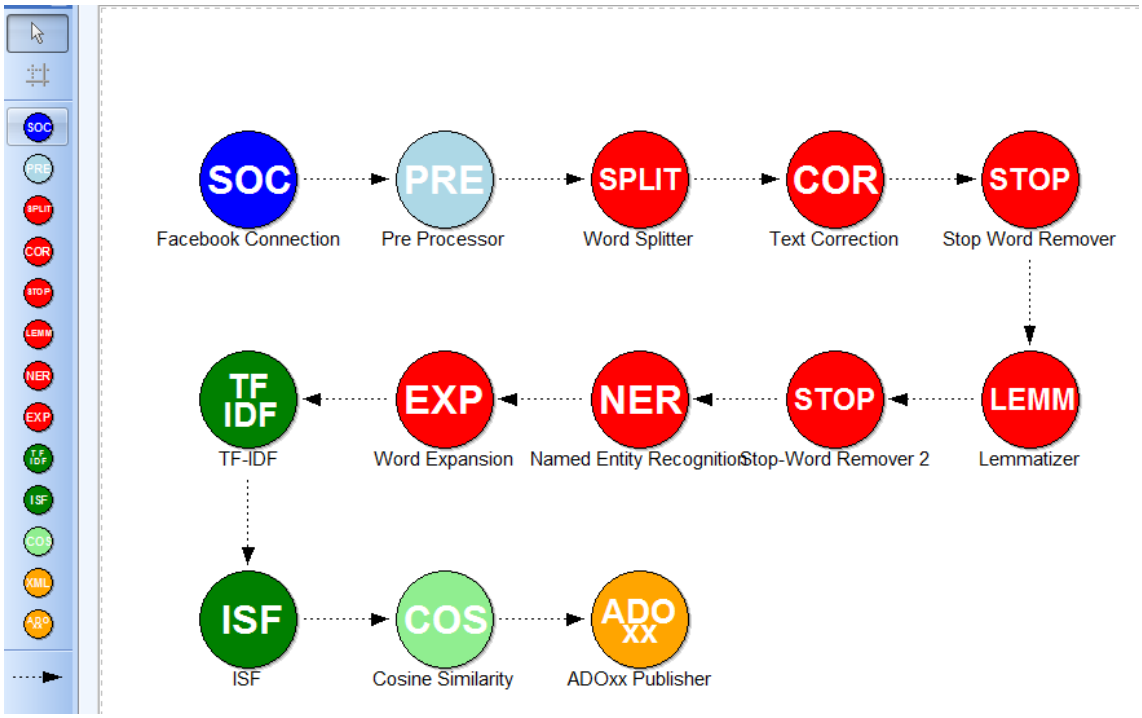

Figure 3.6.5: *Example of a NLP Process Model*

The following tables describe important classes used in the NLP process model. All other tables can be found in chapter F.

Social Media Loader	Description
	This class is responsible for the settings of the data loading process. It enables the user to choose between several social media platforms and to define the corresponding settings.

Attribute Name	Type	Description
Name	String	The name displayed.
ClassGroup	String	This attribute is used to disambiguate the different groups. Users cannot change this value. Default 'Loader'.
Access Token	Longstring	Social Media platforms use an access token to access the data. Most of the times an active user on the platform is mandatory.
Start Time	Datetime	Starting date and time of the load.
End Time	Datetime	Ending date and time of the load.
Platform	Enumeration	List of social media platforms to select.
Target Page	String	Name of the target page on the social media platform.
Social Media URL	Longstring	Base url of the platform. Used for the link to the original post.
Reference Model	INTERREF	Inter-model-reference to the business topic model. The reference model has to be loaded and processed too.
Survey DB Connection URL	Longstring	Connection url to a fl server database containing the survey data (optional).
Survey Amount	Enumeration	Load all data (Full) or only assigned posts (Assigned).

Table 3.6.3: *Class Social Media Loader*

Pre Processor	Description
	This class is responsible for the splitting of posts into sentences and several filterings prior to processing. Users are able to filter posts by length & language, replace abbreviations, hyperlinks etc.

Attribute Name	Type	Description
Name	String	The name displayed.
ClassGroup	String	This attribute is used to disambiguate the different groups. Users cannot change this value. Default 'Loader'.
Extract Messages	Enumeration	Yes or No selection for splitting of posts into sentences.
Minimum Sentence Length	Integer	Minimum length of post or sentence to be loaded.
Minimum Words	Integer	Minimum count of words necessary to be loaded.
Remove Hyperlinks	Enumeration	Yes or No selection. Hyperlinks are removed entirely.
Replace Abbreviations	Enumeration	Yes or No selection to replace abbreviations by using a customizable list.
Abbreviation List Path	Longstring	Path to the text file containing the abbreviations.
Replace Mail Addresses	Enumeration	Yes or No selection. Mail addresses are replaced with the tag [Email].
Replace Emoticons	Enumeration	Yes or No selection to replace emoticons by using a customizable list.
Emoticon List Path	Longstring	Path to the text file containing the emoticons.
Replace Numericals and Dates	Enumeration	Yes or No selection. Identified dates and numericals are replaced with tags.
Filter Language	Enumeration	Yes or No selection.
Preferred Language	Enumeration	List of available languages. German & English currently.
Language Models Path	Longstring	Path to the language models of the language detection library.
Include Page Posts	Enumeration	Yes or No selection wheter posts of the target page should be included or not.

Table 3.6.4: *Class Pre Processor*

Word Lemmatizer	Description
	This class is responsible for the lemmatization and is also able to do a capitalization correction.

Attribute Name	Type	Description
Name	String	The name displayed.
ClassGroup	String	This attribute is used to disambiguate the different groups. Users cannot change this value. Default 'NLP'.
Lemmatizer Model Path	Longstring	Path to the lemma-model file.
POS Model Path	Longstring	Path to the POS-model file.
Capitalization Correction	Enumeration	Yes or No selection.

Table 3.6.5: *Class Word Lemmatizer*

Named-Entity Recognition	Description
	This class is responsible for processing the Named-Entity Recognition. It is possible to identify locations and persons using a corpus.

Attribute Name	Type	Description
Name	String	The name displayed.
ClassGroup	String	This attribute is used to disambiguate the different groups. Users cannot change this value. Default 'NLP'.
German Corpus Path	Longstring	Path to the corpus file used for the NER.

Table 3.6.6: *Class Named-Entity Recognition*

Word Expansion	Description
	This class is responsible for addition of synonyms to each word using a thesaurus file.

Attribute Name	Type	Description
Name	String	The name displayed.
ClassGroup	String	This attribute is used to disambiguate the different groups. Users cannot change this value. Default 'NLP'.
Count of Additions	Integer	Number of synonyms added to each word.
Thesaurus Path	Longstring	Path to the thesaurus file used

Table 3.6.7: *Class Word Expansion*

Cosine Similarity	Description
	This class is responsible for categorizing the messages depending on the prior weighting values.

Attribute Name	Type	Description
Name	String	The name displayed.
ClassGroup	String	This attribute is used to disambiguate the different groups. Users cannot change this value. Default 'Categorizer'.
Minimum Inverse Sentence Frequency	Double	Minimum inverse sentence frequency needed to be processed.
Minimum Similarity	Double	Minimum similarity needed to be categorized.
Reference Model	INTERREF	Inter-model-reference to the business topic model needed for the categorization.

Table 3.6.8: *Class Cosine Similarity*

ADOxx Publisher	Description	
	This class exports an ADOxx compatible XML file, containing the results of the categorization in a Business Topic and Reference Sentence Model.	
Attribute Name	**Type**	**Description**
Name	String	The name displayed.
ClassGroup	String	This attribute is used to disambiguate the different groups. Users cannot change this value. Default 'Publisher'.
Publish Path	Longstring	Path for the XML file. This path is also used by the ADOScript for export and acts as a working directory.
File Name	String	Name of the file exported.
version	String	Version attribute for the header of the XML file. Default '3.1'.
ADOversion	String	ADOxx version attribute for the header of the XML file. Default 'Version 5.1'.

Table 3.6.9: *Class ADOxx Publisher*

3.7 Integrating Natural Language Processing

In this chapter the methods and approaches of chapter 2.4 are put into practice in a Java-based application.

3.7.1 Approach

As already stated, the main GUI and the modification of the settings as well as the definition of the different process steps is done in the SeMFIS toolkit based on the ADOxx modeling platform. The JAVA application, developed to connect to the Social Media Platforms and to process the data via Natural Language Processing (Chapter 2.4), is executed after the modeling by the user via SeMFIS and displays only the current state of the processing (Figure 3.5.2 & Figure 3.7.1). Due to the amount of data used in the processing (loaded data, dictionaries, libraries etc.), the application is executed with the parameter *-Xmx3072m*, which sets the maximum Java heap size to 3 GB. Since 32bit Java installations are not able to use this amount, it is mandatory to install the 64bit version of Java.

Figure 3.7.1: *Screenshot of the status GUI*

Three of the five main processing steps are mandatory for the system and their order cannot be changed[1]:

- Loading
  Loading the data to be processed from Social Media Platforms or file data.
- NLP
  Natural Language Processing of the prior loaded data.

[1]See chapter 3.6 for more details

- Weighting
  Weighting Algorithms for the processed data[2].

- Categorization
  Categorization Algorithms against the Reference Model[3].

- Publishing
  Publishing the results to XML format.

3.7.2 Application Structure

In this section the most important classes of the NLP Application are described, while all other class diagrams can be found in chapter G. Class main, shown in figure 3.7.2, contains the JFrame for the status GUI and uses packages from all the five main processing steps. The path to the exported model delivered by SeMFIS (XML file) is obtained as argument on the execution[4]. To retrieve the models and their settings, the *UnmarshalHandler* transfers the models to the instances of the model classes of package *models.xmlModels*, which are based on the ADOxx Data Model. The application starts right away with the loading and processing of the social media data. The settings, defining which platforms, credentials (e.g. Access Token), requested page etc. are located in the model as well as all other defined process steps. Main uses objects of Handler classes out of each processing group, instantiated with the NLP process model from the UnmarshalHandler. Depending on the process steps and settings defined in the NLP process model, the Handler classes instantiate objects from a sub-package techniques that loads or processes the data. All objects of the handler classes are threaded and executed simultaneously. To follow the same structure, threaded classes implement the ThreadMethods class, which contains the used methods. Two methods are used to process data, startProcess and postProcess, while two methods are used to report the status back to the main class, finish_init_notify and finish_process_notify. To enable the threads accessing the same data, StatusHolder provides an important contribution, holding all variables, processed data and controls the status updates on the GUI. To reduce the needed memory space, the handler objects are set to null after finishing and a garbage collection is started.

[2]Can be skipped.
[3]Can be skipped.
[4]`http://docs.oracle.com/javase/7/docs/technotes/tools/solaris/java.html` last access: 30.08.2015

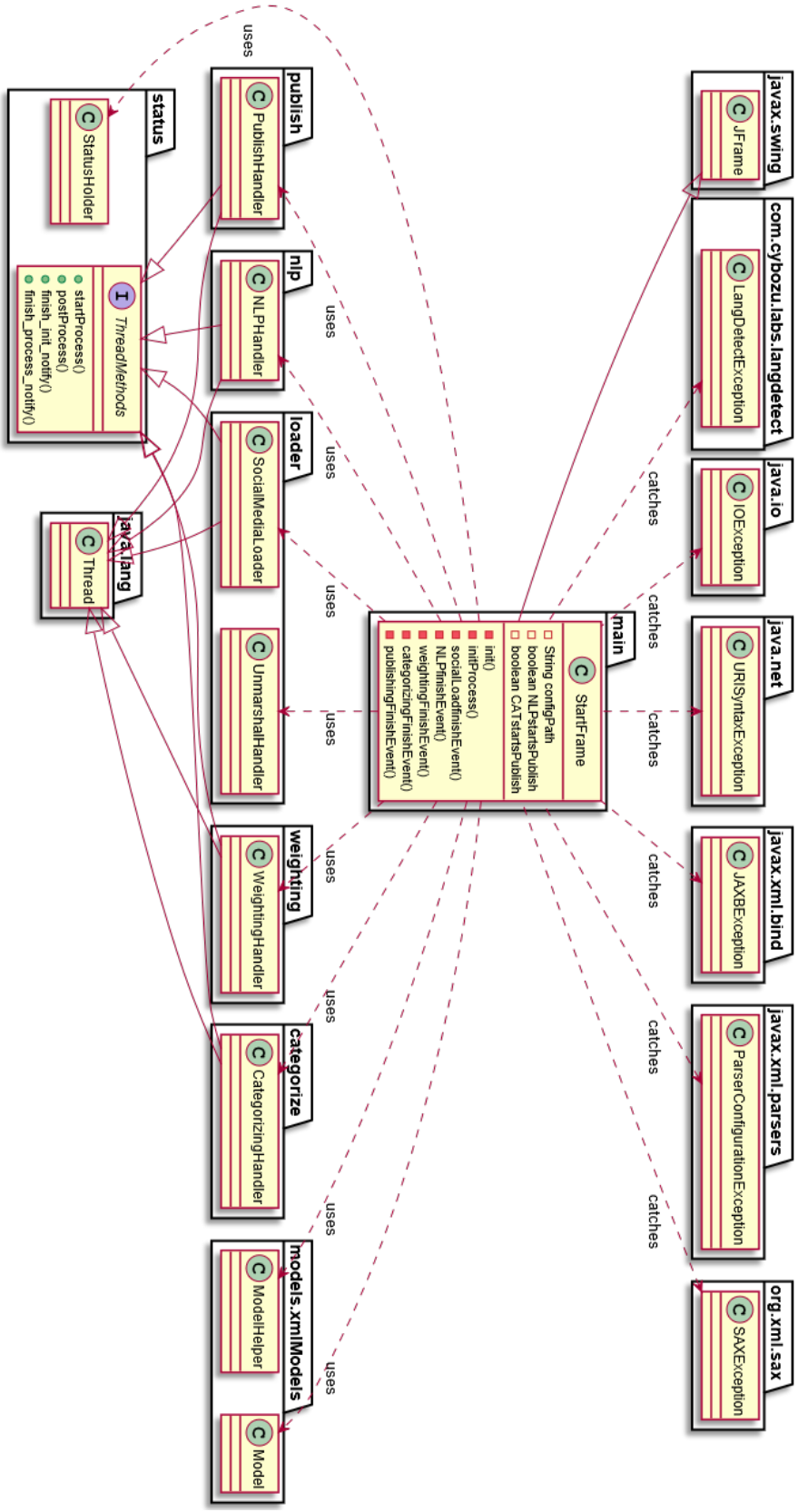

Figure 3.7.2: *Classes of package main*

The data exchange between the NLP Application and SeMFIS is implemented vian XML. Best functionality is achieved by using the XML Schema[5] ADOxx provides in the detailed documentation.
Figure 3.7.4 shows the translated ADOxx XML-Schema to a JAVA class-based data model. To access the values stored in this model more efficiently, a Helper Class (*ModelHelper*) has been created. Figure 3.7.3 shows the structure.

The package *models* (Figure 3.7.5) contains two data model structures for the application:

- OriginalPostHolder
  Based on the structure of the *Post* class of the RestFB library[6] the *OriginalPost* class has been created, which is used to contain the original data from the Social Media Platforms as well as the Survey Results for the post and the assigned matches from the Categorization algorithm. The contained data in the *OriginalPostHolder* is used as base for the publish process.

- MessageHolder
  Users are able to split the text content of the posts into separate messages[7]. In order to increase the processing speed and decrease the complexity, the *Message* class has been created, which contains the sentence of a post (Reference to post via *PostID*), the Reference Sentences (because they have to be processed by the NLP process as well) and several other calculation variables.

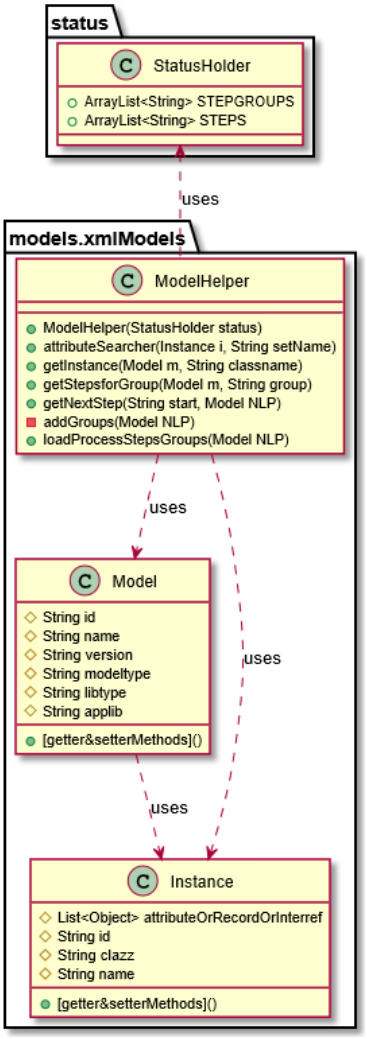

Figure 3.7.3: *Classes of package models.xmlModels - detail of ModelHelper*

The two objects of those classes are stored in the *StatusHolder* class, which is the central point for variables, data and updating the GUI in the application.

[5]Schema used from the documentation `https://www.adoxx.org/AdoScriptDoc/files/Message_Ports/Component_APIs/Documentation/XML_MODELS-js.html` last access: 30.08.2015
[6](Allen and Bartels, 2015)
[7]Section 2.4.3 for more details

Creation of the class structure shown in figure 3.7.4 was achieved by processing the aforementioned ADOxx XML-Schema using JAXB[8].

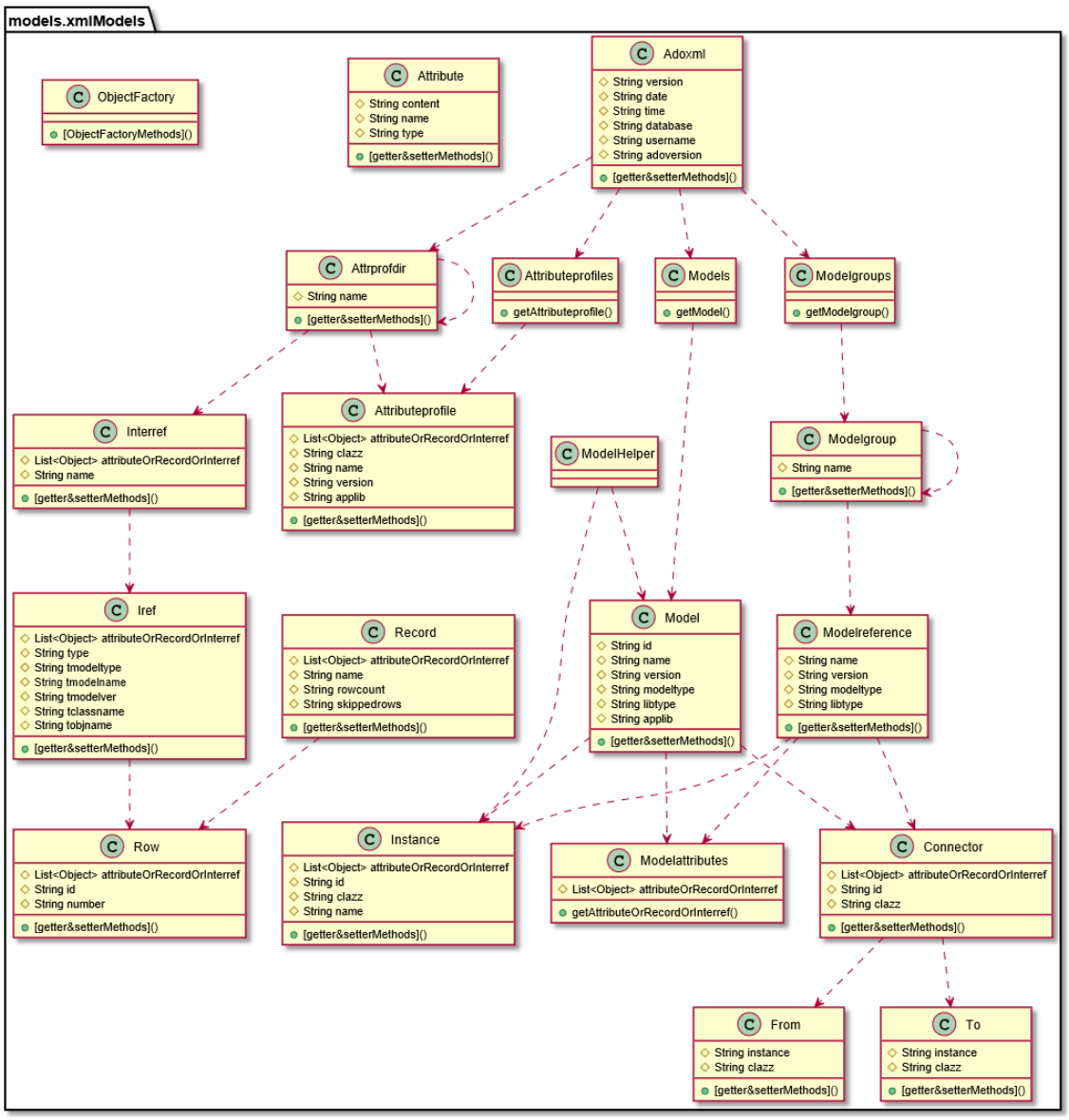

Figure 3.7.4: *Classes of package models.xmlModels*

[8]For more information visit `http://www.oracle.com/technetwork/articles/javase/index-140168.html` last access: 30.08.2015

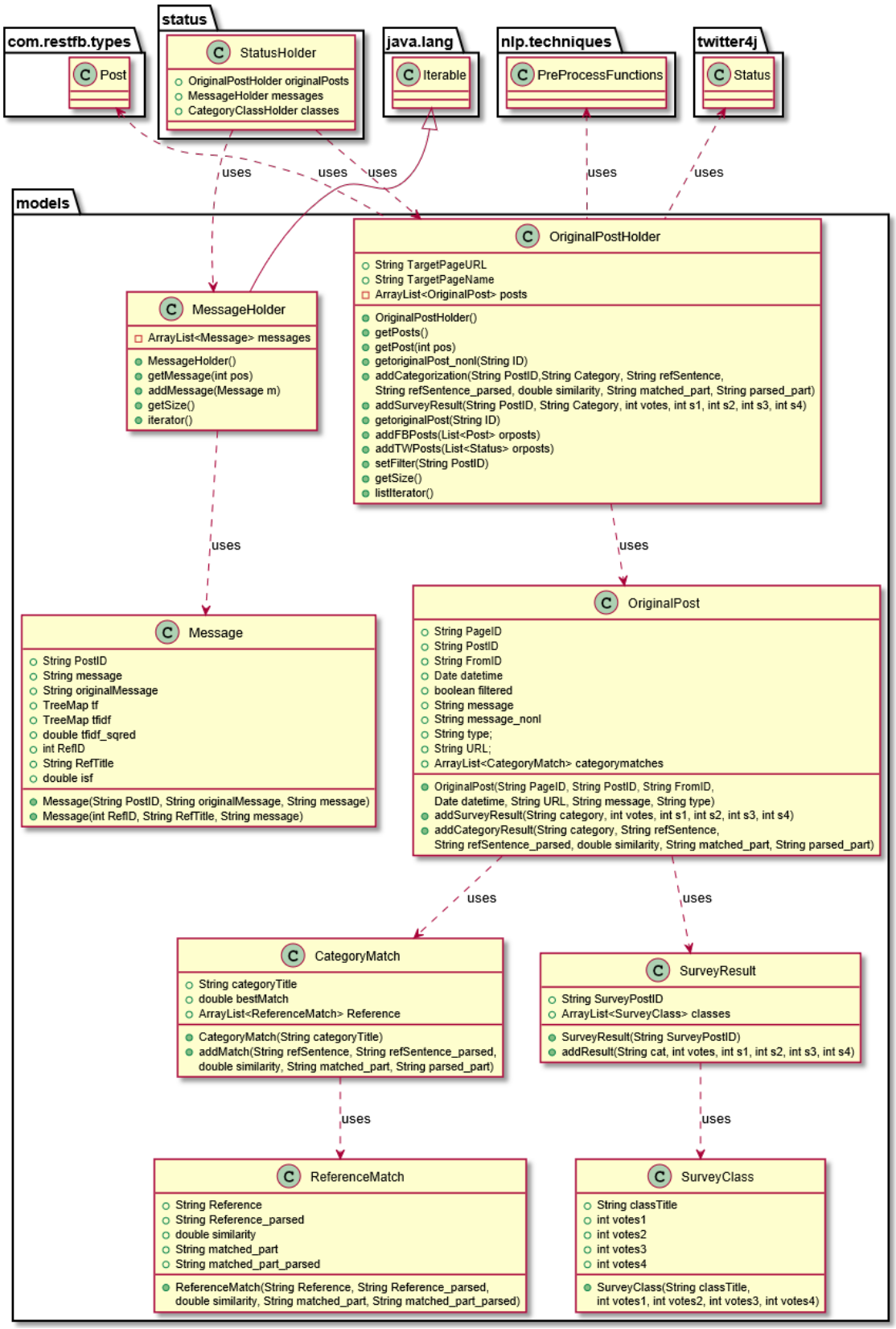

Figure 3.7.5: *Classes of package model*

Figure 3.7.6 & 3.7.7 show the two data sources (*SeMFIS* & *Social Media*), the different Functions (e.g. *RefLoader*, *Pre Processor*, etc.) and the used data models (e.g. *OriginalPostHolder*, *MessageHolder*, etc.) of the application.

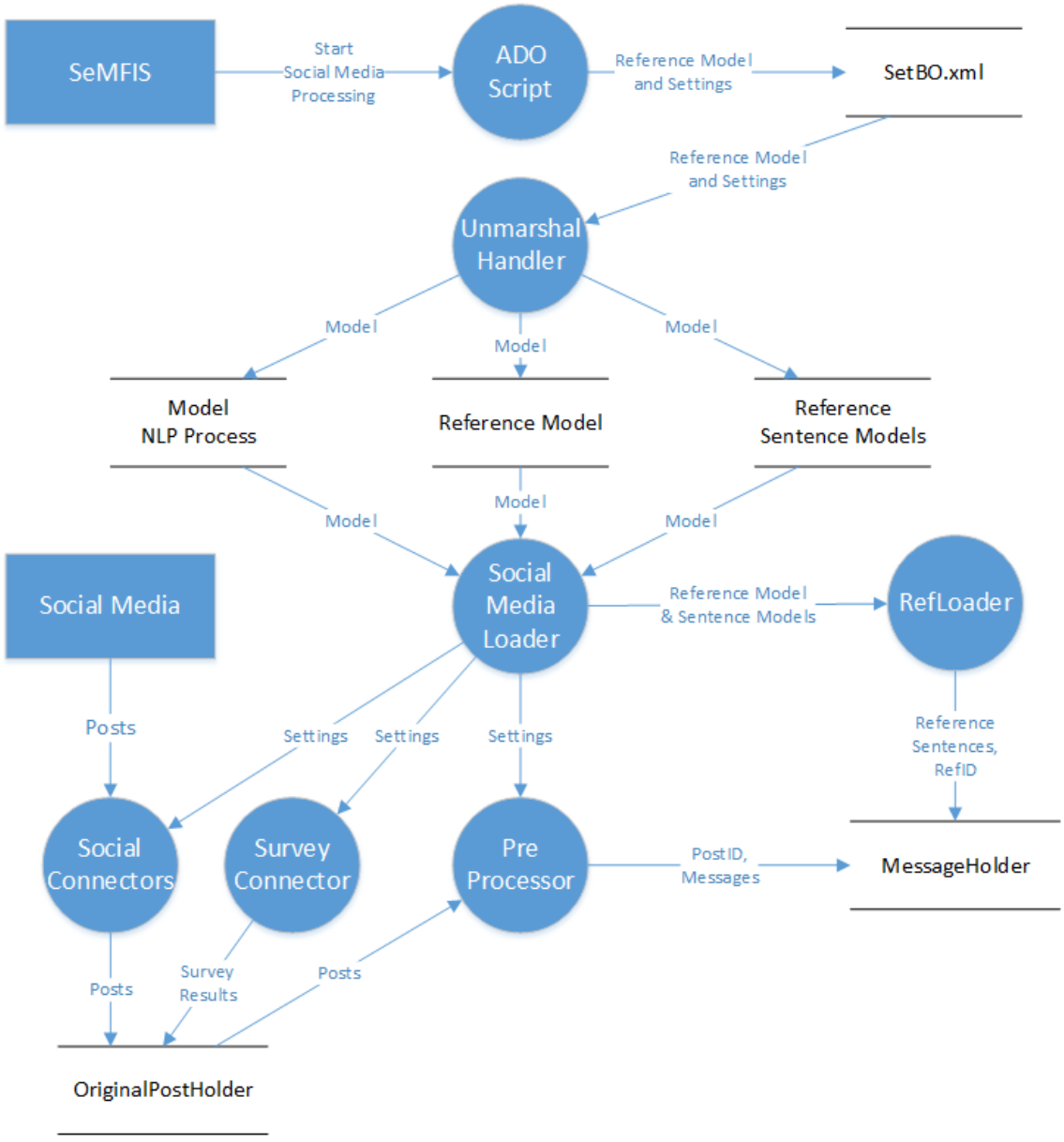

Figure 3.7.6: *Data Flow of the System (1/2)*

The diagram has been split into two parts, therefore figure 3.7.6 does not end with an output and figure 3.7.7 does not start with an input.

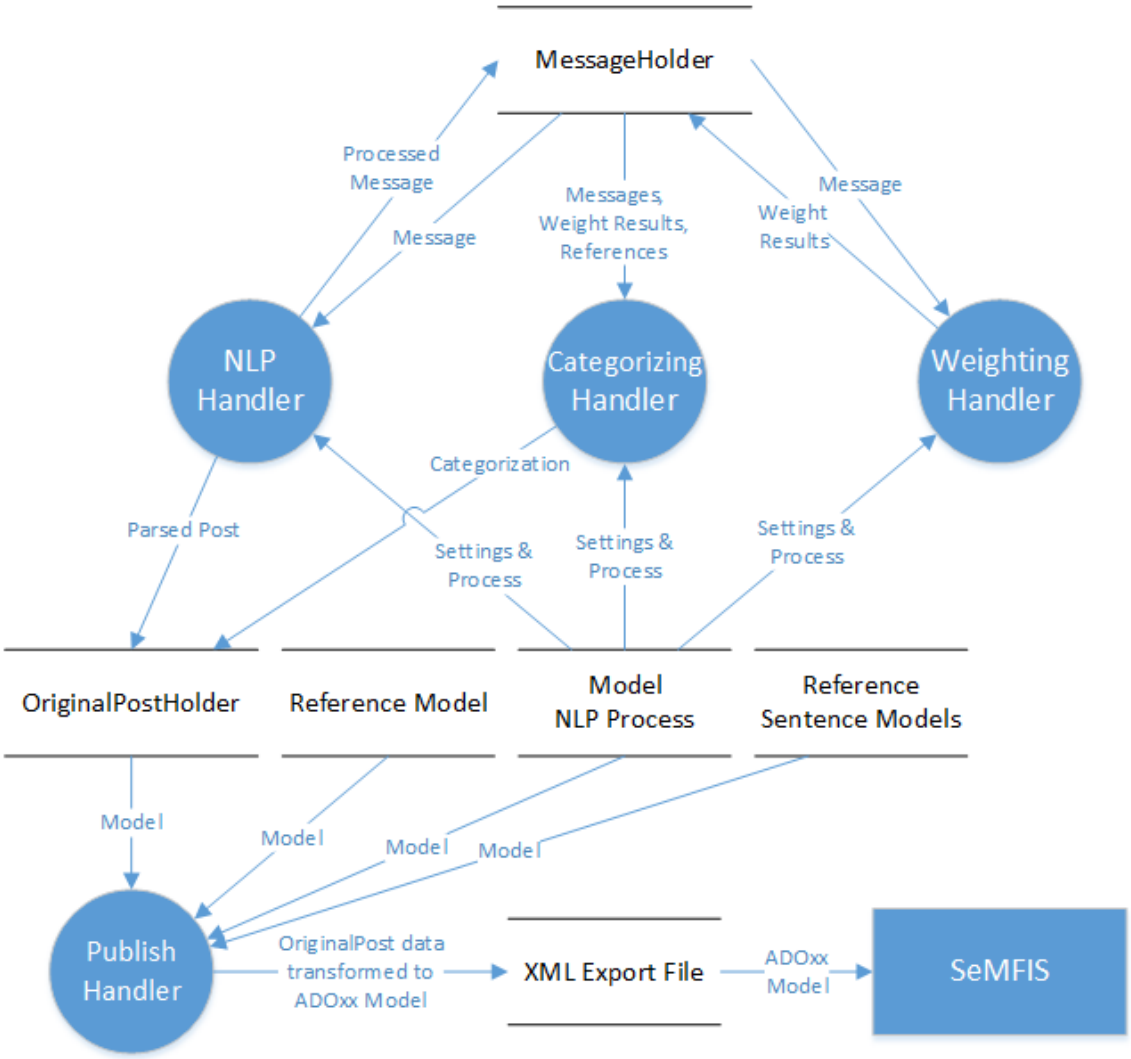

Figure 3.7.7: *Data Flow of the System (2/2)*

3.7.3 Processing User-generated Content

Chapter 2.4 described the main applications and techniques of NLP, which provided the base for the NLP process in the system. Many of these techniques rely on the syntactic correctness (grammar and spelling) of the processed data. During the development several obstacles descending from user-generated content arose.

Prior starting actual processing, it is recommended to filter the posts for the preferred language. Even if the appearance of the company on the social media platform can clearly be recognized as German, a certain amount of users will still post in an-

other language (most of the time English) on the page. To filter social media data for the preferred language, the Language Detector of Nakatani (Shuyo, 2010) calculating probabilities from features of spelling using Naive Bayes with character n-gram was used. Stating 99.8% precision for 49 languages, it delivered satisfactory results used on social media data.

In order to gain the relevant information out of social media posts more efficiently it is recommended to split the posts into the contained sentences. Some parts of social media posts have no semantic value and hence, disturb the correct categorization. Another argument to split the posts into sentences is the occurrence of posts which describe various problems e.g. a person with a delayed flight also faced unfriendly service personnel. Sentence Boundary Disambiguation (described in section 2.4.3) provided by the Stanford NLP (Stanford University, 2015) library was used to split the social media posts into the contained sentences. Created for standardized texts and the English language, the library did not deliver satisfying results being used on social media data. The German date formatting (DD.MM.YYYY), different decimal separators, abbreviations, e-mail addresses & Hyperlinks, emoticons or simple particularities of the users decreased the performance. (Manning and Schütze, 1999) describes several approaches to get a reliable result, but in the environment of standard texts in English language, for example using the appearance of a capital letter after a punctuation as indicator. In a standardized text this is a very reliable sign for the end of a sentence while in the social media world, correct orthography is not important for daily users. To retrieve the best results in a German social media environment, these steps should be considered prior using one of the available tools for segmentation:

- Remove Hyperlinks
  In case of categorization, hyperlinks won't deliver useful information and can therefore be removed from the sentence, using regular expressions.

- Remove Mail Addresses
  The information to whom or by whom is not necessary and can therefore be removed as well.

- Remove New Lines
  To get a more efficient way of processing the sentence.

- Replace Emoticons
  Emoticons can be removed or replaced by the textual expression of their meaning (e.g. :-) equals 'happy' or :-(equals 'sad')[9].

[9]A list can be found on wikipedia: `https://en.wikipedia.org/wiki/List_of_emoticons` last

- Numericals and Dates
  Several formats of date values exist throughout the world, but are used slightly different. For instance, using dots as a separator of day, month and year is common in German speaking countries as well as separating decimal numbers in English speaking countries and should be removed. For the correct categorization of the posts or sentences it is not necessary to know when something happened or how many and can therefore also be replaced or removed using regular expressions.

- Abbreviations
  Abbreviations should either be replaced by their full textual representation or completely removed.

- Remove Special Characters
  At last step, all special characters, except the punctuations should be removed as they don't deliver any necessary contribution to the categorization.

After processing these steps, the punctuation is a very reliable identificator for sentence segmentation.

Having posts segmented correctly into the contained sentences, another particularity emerged. Social media posts normally do not follow spelling or grammar conventions as stated earlier, which induces a lot of issues at analysing and interpreting the data. Depending on the user group, following and posting on the companies social media appearance, the level of spelling or grammatical mistakes vary. Users, capable of correct spelling and grammar are likely to stick to correct capitalization in social media and some are not. This brings advantages and disadvantages. In the German language, nouns start with a capital letter, which seems to be a reliable distinctive feature in comparison to verbs or adjectives (Engel, 2008). Especially for words that share the same appearance for example:

- "arm"
  Either a part of the body (*Arm*) or being poor.

- "essen"
  Either a meal (*Essen*) or the action of eating.

It would be impossible to distinguish between those words without knowing the context or looking at the capitalization. Another problem arises with the capitalization of the first letter at the beginning of a sentence. This rule transforms any word

access: 30.08.2015

to a possible noun and reduces the precision for the system. The last, but not so common problem are sentences with all letters capitalized or sporadic capitalized. To reduce this error, two measures proved to be successful. First measure was the creation of a Text Correction Handler based on the wikipedia list of commonly misspelled words[10], capable of replacing common spelling errors in the data. The second measure was the correction of wrong capitalization using POS (section 2.4.1). In order to lemmatize words correctly, the used Lemmatizer of the mate-tools (Bohnet, 2010b) uses POS, including the context of the sentence. With capitalization correction enabled, every word except nouns is transformed to lowercase and the first letter of every noun is capitalized. After these two measures the precision increased significantly as other process steps e.g. stop word removal relied on the correct capitalization.

Another interesting aspect, which reduces accuracy is the general structure of posts created by users on a companies social media wall. Social media posts often consist of an introductory sentence e.g. "Hello social media team of company xy!", a simple "Hello company xy!" or similar, as well as a closing sentence e.g. "Best Regards Person xy.", "Thank you very much!" or similar (figure 3.7.8). An analysis[11] of posts[12] on the facebook page of 'Austrian Airlines' (the national airline company of Austria, which has a very active user interaction on facebook), revealed that about 26% of the posts include such an introductory sentence. Due to the length restrictions, the majority of twitter posts do not include those sentences. Overall these sentences contain no useful information and reduce accuracy.

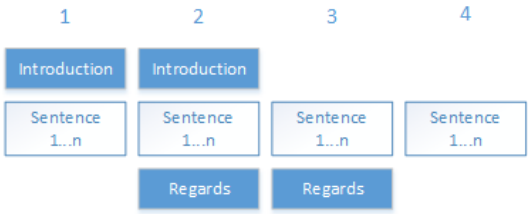

Figure 3.7.8: *Social Media Post Structure*

To identify such sentences, a modified form of the prior explained Inverse Document Frequency was used. Analogous to the IDF, the ISF not focuses on removing the most frequent occurring sentences (in order to keep appearing social media trends), therefore a logarithm is used to level off the results. Given a collection of N documents, a given sentence s_i appears in n_i documents of N.

[10] English: `https://en.wikipedia.org/wiki/Commonly_misspelled_English_words` last access: 30.08.2015
German: `https://de.wikipedia.org/wiki/Kategorie:Wikipedia:Falschschreibung` last access: 30.08.2015

[11] Results in section D.1

[12] 200 random posts from January until June 2015. Only German posts without posts of the company and (clear recognizable) spam.

$$isf(s_i) = log\frac{N}{n_i}$$

Using the ISF in the test system and ignoring those introductory sentences, increased the accuracy[13] of the categorization by 4%.

The last obstacle, processing user generated content, appeared at the use of Query Expansion (section 2.4.6) to increase the recall value of the system. The approach was to use the data provided by the openthesaurus project[14], which is an open source thesaurus for the German language. First usage attempts revealed that the differences between German and Austrian-German could not be ignored. Furthermore the expansion relied heavily on correct spelling (which has been coped with the text correction and lemmatizer). One particularity of user-generated content could not be solved: Slang words, which are quite popular. Despite these problems, the query expansion increased the performance of the system, especially the recall.

3.7.4 External Libraries

Especially for the Natural Language Processing, several external libraries have been used, which are listed in table 3.7.1.

Library	**File Name**	**Usage**	**Source**
Mate-Tools	anna-3.61.jar	Lemmatizer	(Bohnet, 2010a)
Apache Commons Lang	commons-lang-2.6.jar	Pre Processing	(The Apache Software Foundation, 2015)
Composition Splitter	jwordsplitter-4.1.jar	Word Splitter	(Naber, 2015)
Language Detection Library for Java	langdetect.jar & jsonic-1.2.0.jar	Language Detection	(Shuyo, 2010)
Microsoft JDBC Driver for SQL	fljdbc41.jar	Survey DB Connection	(Microsoft Corp., 2015)
Stanford NLP	stanford-corenlp-3.5.2.jar	Named Entity Recognition & Sentence Boundary Detection	(Stanford University, 2015)
RestFB	restfb-1.11.0.jar	Facebook Connection	(Allen and Bartels, 2015)
twitter4j	twitter4j-core-4.0.4.jar	Twitter Connection	(Yamamoto, 2015)

Table 3.7.1: *External library list of the NLP application*

[13]Evaluation of the accuracy described in chapter 4.1
[14]For more information visit `https://www.openthesaurus.de/` last access: 30.08.2015

3.7.5 External Datasets

Especially for the Natural Language Processing, several external datasets, corpora etc. have been used, which are listed in table 3.7.2.

Dataset	**Used by**	**Source**
Lemma model	Lemmatizer	(Bohnet, 2010a)
POS Model	Lemmatizer	(Bohnet, 2010a)
Stop Word List	Stop-Word Remover	`https://code.google.com/p/stop-words/`
Incorrect Word List	Text Correction	`https://de.wikipedia.org/wiki/Kategorie:Wikipedia:Falschschreibung`
Emoticon List	Pre Processor	`https://en.wikipedia.org/wiki/List_of_emoticons`
Language Library	Pre Processor	(Shuyo, 2010)
WaCky Corpus	Named Entity Rec.	(Baroni et al., 2009)
Thesaurus	Word Expander	`https://www.openthesaurus.de/`

Table 3.7.2: *External dataset list of the NLP application*

Part 4

Evaluation & Results

4.1 Evaluation Approach

In this chapter the used approach of evaluating the performance of the system using a web-based categorization survey and the gained results are described. The accuracy of the categorization results had to be evaluated in order to recognize malfunctions and space for further improvement.

To obtain valuable data for comparison, a web-based survey[1] has been conducted. For this evaluation, the Facebook page of Austrian Airlines[2], the national airline company of Austria which has a very active user interaction on Facebook, has been chosen.

Figure 4.1.1: *Screenshot of the Categorization Service*

From January 2015 to June 2015 about 1.000 to 1.200[3] user posts had been available on the Facebook page. To ensure a sampling error below 5%, a sample size of 303 posts has been selected randomly. The reference model for the Austrian Airlines contained 18 categories, based on the FAQ-section of their official web presence. These 18 categories were used for the survey as well as the NLP application. Participants of the survey rated the fit of the post to one or more of these categories, using a simplified Likert-Scale (Figure 4.1.3). For simplification reasons the textual notation has been replaced with red - symbols and green + symbols. Being accessible via the web, no information about the participants was stored, except the session variable to prevent double responses. The participants were 55 students of the University of Vienna. To retrieve a meaningful result, at least three opinions per post and category were aimed.

[1] Technical details in section 3.7.2.
[2] `https://www.facebook.com/austrianairlines` last access: 30.08.2015
[3] Number of posts vary due to deletions and subsequently addings.

Based on the FAQ topics on the Austrian Airlines web presence[4], 18 categories have been created for the Reference Model:

- Allgemeine Belobigungen* (*General Commendations*)
- Allgemeine Verbesserungsvorschläge* (*Suggestions for Improvement*)
- Gewinnspiel* (*Competitions*)
- Service und Personal (*Service and Personnel*)
- Kundendienst (*Customer Service*)
- Spotting und Fotos* (*Spotting and Photos*)
- Webseite, AirManager und App (*Website, AirManager and App*)
- Umbuchung und Storno (*Rebooking and Cancellation*)
- Gepäck (*Baggage*)
- Buchung (*Booking*)
- Fliegen und mehr (*Travel by Air and more*)
- Check-in und Mobile Services (*Check-in and Mobile Services*)
- Wunschsitz (*Preferred Seat*)
- Bundesländer Flughäfen in Österreich (*Local Airports in Austria*)
- Flugunregelmäßigkeiten (*Flight Irregularities*)
- Medizinische Fragen (*Medical Questions*)
- Reisen mit Kindern (*Traveling with Children*)
- Smart Upgrade (*Smart Upgrade*)

The reference sentence models were created, using the subcategory questions on the FAQ section. The first test runs showed clearly a lack of certain categories[5] to fit the needs of the social media world. *General Commendations* has been added as a category due to the high amount of commendations in general. Another topic *Suggestions for Improvement* was necessary, because posts with issues often also contained suggestions. Furthermore, *Competitions* was reasonable, because the Austrian Airlines occasionally posted competitions[6], where the users had to reply a certain message in order to win. The last topic, *Spotting and Photos* was created in order to categorize the posts of hobby photographers, posting photos of planes onto the page.

[4] Austrian Airlines German FAQs:
`http://www.austrian.com/Info/FAQ.aspx?sc_lang=de&cc=AT`
Austrian Airlines international FAQs:
`http://www.austrian.com/Info/FAQ/Rebooking_Storno.aspx?sc_lang=en&cc=ZZ`
last access: 30.08.2015

[5] Categories marked with asterisk have been added manually.

[6] At the end of the evaluation no competition was active anymore.

The integration of the survey data into the system can be described by adapting the conceptual model (figure 3.4.2) from chapter 3.4. In order to reduce the complexity of the graphic, the SeMFIS platform has been removed as it does not affect this concept, but is still part of the system. Figure 4.1.2 now shows the integration of the survey data into the system. The 300 posts as a sample and their unique IDs were stored in the survey database as well as a copy of the used business topic model, containing the, prior mentioned, 18 categories used. The categorization service[7], enables participants to categorize the sample posts to the different topics based on the likert scale and their personal beliefs. The results of this categorization are stored in the survey database, which are used by the NLP application by linking and adding the survey data to the processed posts from the social media platform using the unique ID of the post.

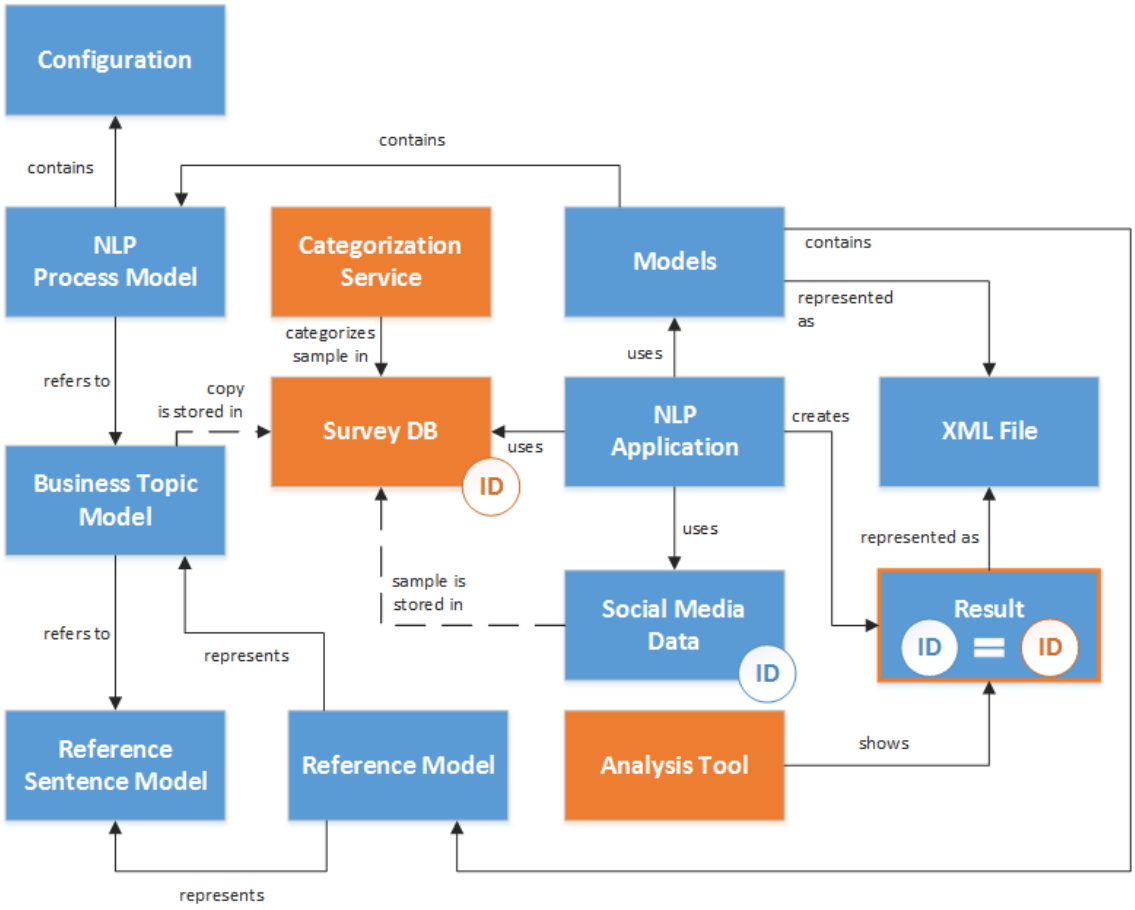

Figure 4.1.2: *Conceptual relationships of the approach including survey data*

After this linking, the result[8] stored in an XML file contains the categorization

[7]Technical details about the categorization service in section 3.5.3.
[8]Detailed information about the data models and data flow in section 3.7.2.

and the survey result. For detailed insights of this data and the behaviour of the system, a JavaScript-based analysis tool has been created, which is described in a later section.

4.1.1 Measure of Speed and Categorization Performance

To measure the speed of the performance, the total time of processing 500 Social Media Posts and the performance in each process group was timed. By dividing the times in seconds by the 500 posts a new KPI, showing the *Process time per post* emerges. Before executing the survey, an adequate way of comparison for the categorization performance had to be selected. Since the posts can be assigned to more than one category, Recall and Precision seemed to be a very reliable and understandable measure to evaluate the performance and to find space for improvement.

"Recall or Sensitivity ... is the proportion of Real Positive cases that are correctly Predicted Positive. ... Conversely, Precision or Confidence ... denotes the proportion of Predicted Positive cases that are correctly Real Positives" (Powers, 2011, p. 2). In a binary contingency table (table 4.1.1 [9]), the counts of returned statements are classified in True Positives (tp), True Negatives (tn), False Positives (fp) and False Negatives (fn).

- True Positives (tp)
  describes correct classified items, which positively meet the given criteria.
- True Negatives (tn)
  describes correct classified items, which negatively meet the given criteria.
- False Positives (fp)
  describes incorrect classified items, which positively meet the given criteria.
- False Negatives (fn)
  describes incorrect classified items, which negatively meet the given criteria.

This classification leads to two formulas, calculating Recall and Precision (Olson and Delen, 2008, p. 138):

$$Recall = \frac{tp}{tp+fn} \qquad Precision = \frac{tp}{tp+fp}$$

[9] "... rp, rn and pp, pn refer to the joint and marginal probabilities, and the four contingency cells and the two pairs of marginal probabilities each sum to 1" (Powers, 2011, p. 2). R relates to Recall and P to Precision.

	+R	-R	
+P	tp	fp	pp
-P	fn	tn	pn
	rp	rn	1

Table 4.1.1: *"Systematic and traditional notations in a binary contingency table. Colour coding indicates correct (green) and incorrect (red) rates or counts in the contingency table" (Powers, 2011, p. 2).*

Two scales were used for the assignment of a post to a category.

- Weak
  A post is assigned to the category if 50 or more than 50 percent of the participants voted with *+* or *++*.
- Strict
  A post is assigned to the category if 50 or more than 50 percent of the participants voted with *++*.

Figure 4.1.3: *Categorization Scale*

4.2 Evaluation Results

The current chapter describes the results obtained through the prior mentioned survey and the analysis of the system's performance in three sections: (1) Analysing the survey data by itself to find insights about the distribution of categories, participation etc. (2) Analysis of the speed of the system in order to identify improvement potential and (3) Recall & Precision of the system, compared to the data contained in the prior described survey database.

4.2.1 Analysis of the Survey Data

The categorization service was available for the participants for one month and has been turned off after this period.

Table 4.2.1 shows the statistics of the participation. More than 263 posts have 3 or more than 3 opinions on each of the 18 categories. Each of the 55 participants voted on 17 posts on average with the 18 categories, making a total sum of 16650 data entries. This high number could be achieved due to the cross-device availability of the service (participants were able to vote on mobile phones too) and the fast and intuitive web-interface.

more than 3 opinions	26
exact 3 opinions	263
less than 3 opinions	14
Votes Overall	16650
Categories	18
Votes on Post	925
Participants	55
Votes per Participant	303
Posts per Participant	17

Table 4.2.1: *Participation Statistics*

Taking a look at the data by itself (figure 4.2.1) reveals slightly different results depending on the assignment. Only 5.3% (weak) and 16.8% (strict) of the Posts have none of the 18 categories assigned. The majority of posts in both scales have been assigned to one or two categories. In comparison to the strict scale, the weak scale is flatter and goes up to a maximum of six categories per post. Especially the assignment of a third category is drastically reduced with the strict assignment.

Another interesting analysis is the distribution of posts assigned to the different categories (figure 4.2.2), showing that the first five categories are responsible for nearly 60% of the posts on the Facebook page (on average). The category *Kundendienst (Customer Service)*, which holds about 16% of the posts, deals with issues regarding customers waiting for a response of the customer service or similar and confirms the assumptions of the introduction, stating that the customer relation-

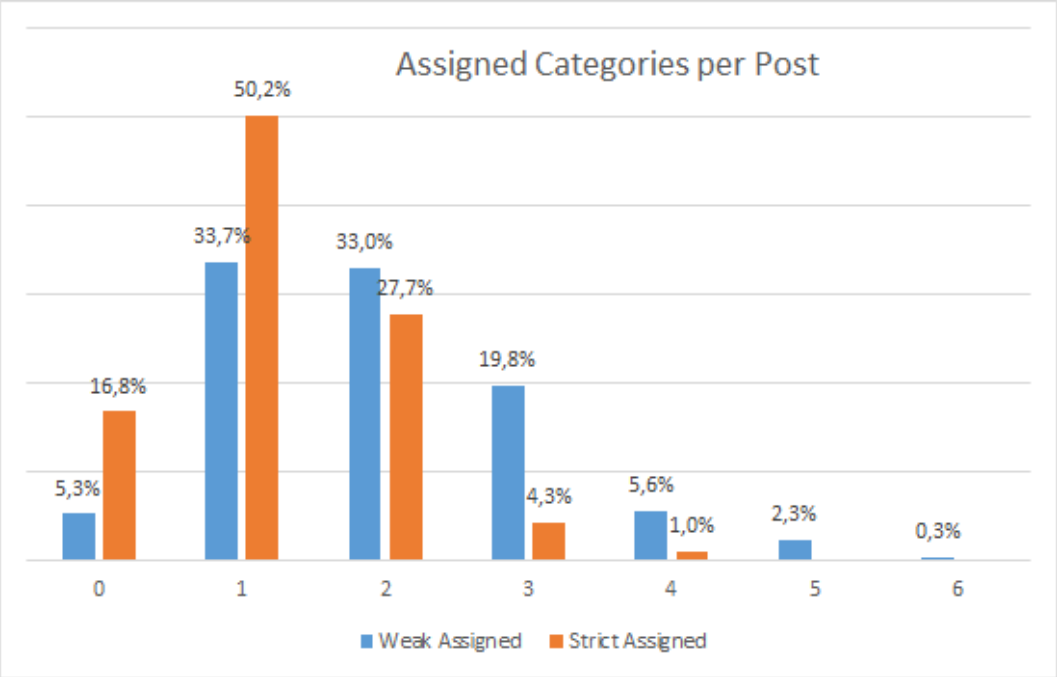

Figure 4.2.1: *Statistics of assigned posts*

	- -	-	+	++	Sum
Votes (All)	14646	750	1230	**16875**	
Percentage	87%	1%	4%	7%	**100%**

	- -	-	+	++	Sum
Votes (Cleaned)	1432	249	750	625	**3056**
Percentage	47%	8%	25%	20%	**100%**

Votes sure	13819	82%

Table 4.2.2: *Distribution of votes*

ship management faces a lot of pressure in the social media environment. Overall these five categories describe popular issues in terms of an Airline company e.g. *Gepäck (Baggage)*, *Flugunregelmäßigkeiten (Flight Irregularities)*, *Service und Personal (Service and Personnel)*. Due to the, already mentioned, sarcastic manner in the social media environment, the category *Allgemeine Belobigungen (General Commendations)* also contains a remarkable amount of posts.

Table 4.2.2 shows the votes on all posts and categories. In the first analysis the great majority of votes (87%) show no connection to a category. This is misleading as there were 18 categories and the majority of Posts had just one to maximum six categories assigned. Therefore another analysis was made to clean the votes where

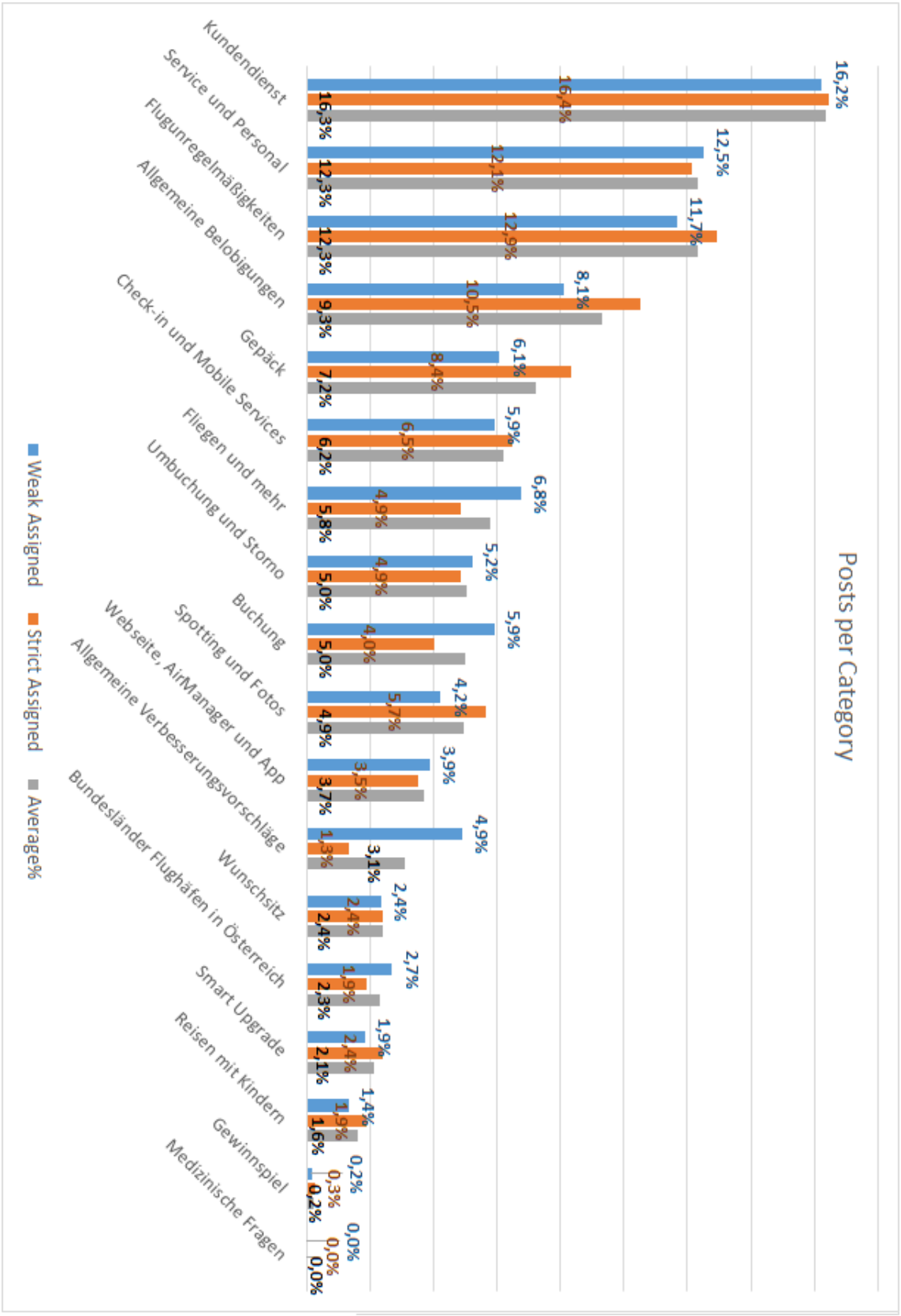

Figure 4.2.2: *Statistics of assigned posts per category*

the participants were 100% sure (Votes where the count of - - or count of ++ is the same as the count of total votes). After cleaning those votes (82%), the distribution shows that still nearly half of all votes (47%) showed no connection to the category. This resembles the fact that only one or two categories fit to the post on average. If we want to calculate this for the 303 posts of the survey we can identify the following assumption for posts being 100% sure:

- Posts containing votes of ++ the same amount as total votes on the category
  These posts fall to at least one category for sure and therefore are not difficult to categorize.
- Posts containing no votes on the categories -, + & ++
  These posts don't fall in any category for sure and therefore also not difficult to categorize.

Filtering these posts reveals that 55% of the posts were categorized without any divergence while 45% of the posts showed divergences being categorized (figure 4.2.3). That means more than half of the posts were categorized without any difficulty, while 45% showed inhomogeneities in the votes.

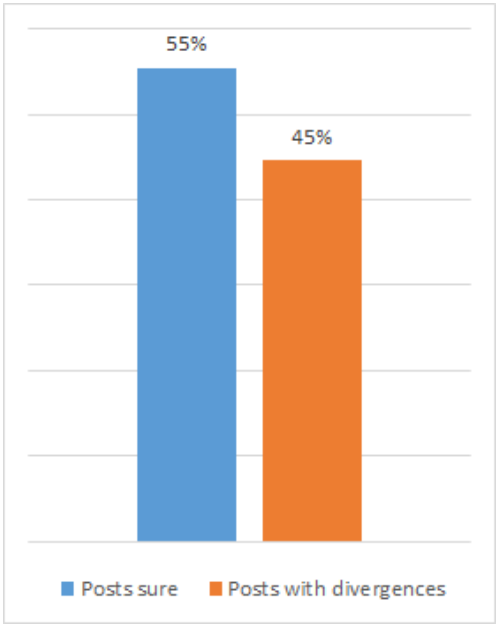

Figure 4.2.3: *Comparison of posts sure and posts with divergences*

In summary this exemplary insights are very valuable already, identifying improvement potential and a clear overview of the issues concerning the customers in the social media environment. Still, this data is retrieved by using human knowledge and is therefore very time consuming. In the next steps the data of this survey is used to evaluate (described in section 4.2.3) the performance of the implemented system (part 3), which is able to deliver the data used for gaining these insights in an automated and efficient way.

4.2.2 Speed of the System

To measure the speed according to the prior described setting, two workstations were used in order to obtain an average value.

- Station 1
  AMD Phenom 9750 Quad-Core 2.40GHz 4GB RAM
- Station 2
  Intel Core 2 Duo 3.16GHz 4GB RAM

500 Facebook posts from January 2015 to June 2015 from the page of Austrian Airlines were used. All possible NLP processes were activated. Measured times are in seconds.

	Station 1	Station 2	Average	Time per Post
Loader	20	18	**19**	**0,038**
NLP	79	68	**73,5**	**0,147**
Weighting	3	2	**2,5**	**0,005**
Categorization	3	2	**2,5**	**0,005**
Publishing	2	1	**1,5**	**0,003**
Overall Time	85	74	**79,5**	**0,159**

Table 4.2.3: *Timing Results of the System*

Overall the system takes a bit more than a minute to process those 500 posts, which were split into approximately 1700 messages. As described in part 3, all of the process steps are threaded in order to improve the performance. Table 4.2.3 shows very obvious that the Natural Language Processing takes the biggest amount of the overall time, with about 74 seconds on average. The reason for this is that this part of the system uses several external libraries, dictionaries and corpora which have to be loaded and initiated prior to execution. This initiation phase takes about 20-30 seconds before the process even starts.

4.2.3 Recall & Precision

As prior described, Recall and Precision was used to measure the processing performance of the system. To get detailed insights of the behaviour and the processing steps, the results were exported to XML and analyzed, by creating a web based analysis tool based on the jquery[1] Java Script library. The XML Publisher class in the SeMFIS NLP Process Model exports the processed data in the right XML structure for this analysis tool. The tool shows a detailed overview of the categories and the assigned posts (Figure 4.2.4) and it is possible to display the survey posts

[1]Visit `https://jquery.com/` last access: 30.08.2015

only and switch between the two scales (Strict & Loose).

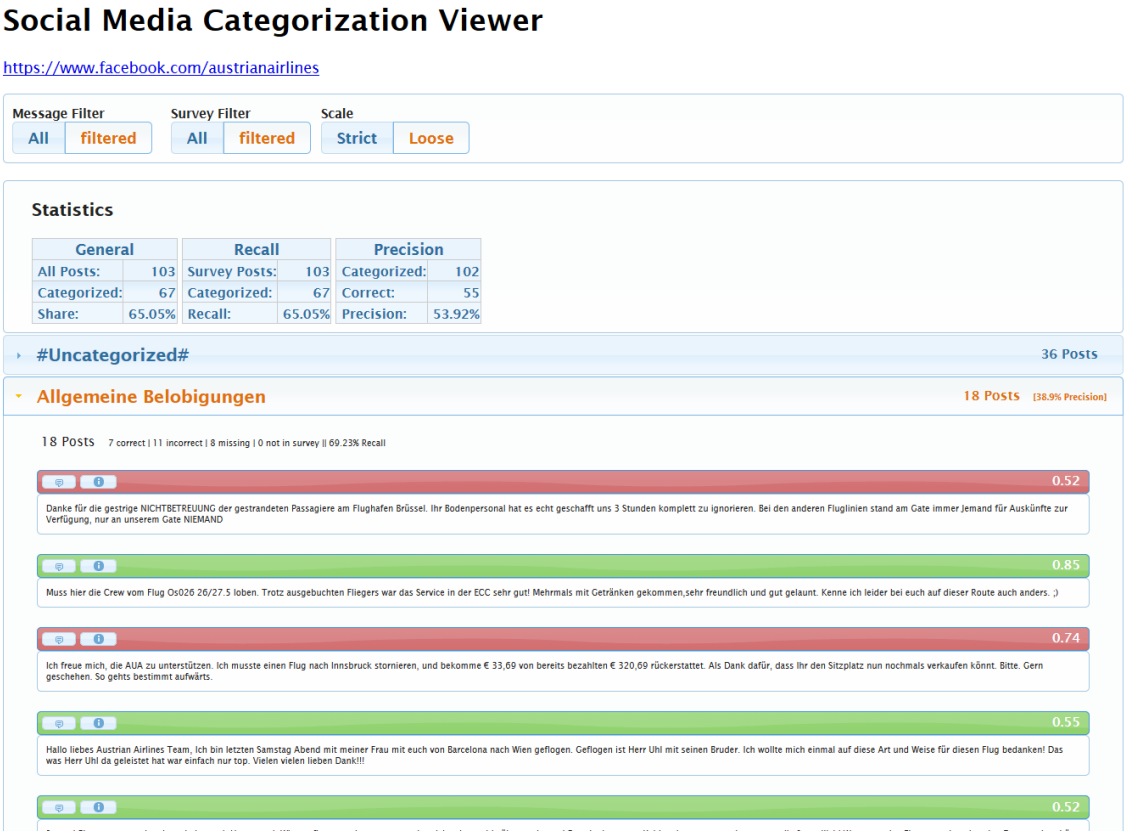

Figure 4.2.4: *Start Screen Analysis Tool*

Correct categorized posts (depending on the survey) are displayed in green and incorrect posts in red. Showing the details of a certain post shows the original message, the parsed representation, the best match of the category (with details which sentence of the post, highlighted), details about the other matches and a comparison between the survey and the categorization algorithm (Figure 4.2.5). These insights were very valuable for improving, understanding and debugging the algorithm.

General		Recall		Precision	
All Posts:	103	Survey Posts:	103	Categorized:	102
Categorized:	67	Categorized:	67	Correct:	55
Share:	65.05%	Recall:	65.05%	Precision:	53.92%

Figure 4.2.6: *Results of Recall & Precision*

Depending on the process and the settings (e.g. Minimum Similarity or Inverse Sentence Frequency), the system reached a Recall value of 65% and Precision of 54% overall (Figure 4.2.6). From the 303 survey posts 103 remained accessible on the page[2]. A look at the different categories reveal big divergences in precision. Two

[2]The Facebook Administrators of the page regularly delete posts.

10152901503184842

Original Message

Auf meinem Rückflug von Paris nach Wien war es nicht möglich Web Check-In zu machen? Beim Service am Schalter geht unnötig viel Zeit drauf, abgesehen vom sehr unfreundlichen Personal...

Parsed Message

Rück Flug Luftreise Luftbeförderung Flugpassage Flugreise [Location] möglich zu machen gangbar viabel realisierbar denkbar erreichbar Web World Wide Web Internet Netz WWW Check-in Abfertigung Checkin Check In Webcheckin Web Check In CheckIn machen walten tun handeln agieren wirken tätig sein Service Dienst Dienstleistung Tafelgeschirr Schalter Einschalteinrichtung Anschalteinrichtung Auskunft Information Auskunftsschalter unnötig unnütz nutzlos fruchtlos zu nichts nütze unnotwendig Zeit Uhrzeit Tempus Zeitform Intervall drauf drogenberauscht zu stoned bekifft zugedröhnt dicht absehen bedenken vorhersehen einplanen voraussehen aufhören beherrschen unfreundlich abweisend distanziert ablehnend unnahbar pampig nicht nett Personal privat persönlich Personalbestand Angestellte Arbeitnehmer

Matched Reference

Check-in und Mobile Services 0.82

Reference	Ist ein Web Check-in möglich wenn ich mit einer Waffe reise
Reference parsed	Web World Wide Web Internet Netz WWW Check-in Abfertigung Checkin Check In Webcheckin Web Check In CheckIn möglich zu machen gangbar viabel realisierbar denkbar erreichbar Waffe Kampfgerät reise
Matched part	Auf meinem Rückflug von Paris nach Wien war es nicht möglich Web Check-In zu machen?
Matched part parsed	Rück Flug Luftreise Luftbeförderung Flugpassage Flugreise [Location] möglich zu machen gangbar viabel realisierbar denkbar erreichbar Web World Wide Web Internet Netz WWW Check-in Abfertigung Checkin Check In Webcheckin Web Check In CheckIn machen walten tun handeln agieren wirken tätig sein

Other Matches +

Survey Results

Survey Category	--	-	+	++	Votes	Strict	Loose	Categorization	
Allgemeine Belobigungen	3	0	0	0	3	0.0%	0.0%	Allgemeine Belobigungen	✓
Allgemeine Verbesserungsvorschläge	3	0	0	0	3	0.0%	0.0%	Allgemeine Verbesserungsvorschläge	✓
Buchung	3	0	0	0	3	0.0%	0.0%	Buchung	✓
Bundesländer Flughäfen in Österreich	3	0	0	0	3	0.0%	0.0%	Bundesländer Flughäfen in Österreich	✓
Check-in und Mobile Services	0	0	0	3	3	100.0%	100.0%	Check-in und Mobile Services	✓
Fliegen und mehr	3	0	0	0	3	0.0%	0.0%	Fliegen und mehr	✓
Flugunregelmäßigkeiten	3	0	0	0	3	0.0%	0.0%	Flugunregelmäßigkeiten	✓
Gepäck	3	0	0	0	3	0.0%	0.0%	Gepäck	✓
Gewinnspiel	3	0	0	0	3	0.0%	0.0%	Gewinnspiel	✓
Kundendienst	2	0	0	1	3	33.3%	33.3%	Kundendienst	✓
Medizinische Fragen	3	0	0	0	3	0.0%	0.0%	Medizinische Fragen	✓
Reisen mit Kindern	3	0	0	0	3	0.0%	0.0%	Reisen mit Kindern	✓
Service und Personal	1	0	1	1	3	33.3%	66.7%	Service und Personal	✓
Smart Upgrade	3	0	0	0	3	0.0%	0.0%	Smart Upgrade	✓
Spotting und Fotos	3	0	0	0	3	0.0%	0.0%	Spotting und Fotos	✓
Umbuchung und Storno	3	0	0	0	3	0.0%	0.0%	Umbuchung und Storno	✓
Webseite, AirManager und App	1	0	2	0	3	0.0%	66.7%	Webseite, AirManager und App	✓
Wunschsitz	2	1	0	0	3	0.0%	0.0%	Wunschsitz	✓

Figure 4.2.5: *Detail of a Post*

categories, marked with an asterisk had no posts assigned nor from the system neither from the survey and therefore are not considered. The first category *Allgemeine Belobigungen* (General Commendations), has a high recall value of nearly 70% but a poor precision value. The reason for this is the use of sarcasm in the social media world, which is not recognized by the system, therefore a lot of posts describing a problem in a sarcastic manner are assigned to this category.

Shortcut	Category	Recall	Precision
ABE	**Allgemeine Belobigungen**	69.2%	38.9%
AVB	**Allgemeine Verbesserungsvorschläge**	37.5%	33.3%
BUC	**Buchung**	20.0%	100.0%
FLU	**Bundesländer Flughäfen in Österreich**	66.7%	0.0%
CHE	**Check-in und Mobile Services**	54.6%	100.0%
FLI	**Fliegen und mehr**	53.3%	12.5%
FUN	**Flugunregelmäßigkeiten**	61.5%	75.0%
GEP	**Gepäck**	50.0%	60.0%
GEW	**Gewinnspiel***	100.0%	100.0%
KUN	**Kundendienst**	61.9%	61.5%
MED	**Medizinische Fragen***	100.0%	100.0%
RMK	**Reisen mit Kindern**	33.3%	0.0%
SUP	**Service und Personal**	48.0%	58.3%
SMU	**Smart Upgrade**	50.0%	100.0%
SPF	**Spotting und Fotos**	80.0%	100.0%
UST	**Umbuchung und Storno**	50.0%	33.3%
WAA	**Webseite, AirManager und App**	69.2%	55.6%
WUN	**Wunschsitz**	80.0%	0.0%
	Average	**55.3%**	**51.8%**
	Median	**54.0%**	**57.0%**

Table 4.2.4: *Recall and Precision of the Categories*

Overall the results vary from 0% to 100% in precision and from 20% to 80% in recall. Putting Recall and Precision into a Matrix (Figure 4.2.7) shows clearly which categories are performing good (in Quadrant B), bad (in Quadrant C), while there are categories with good precision but bad recall (in Quadrant A) and reverse (Quadrant D).

Category *Spotting und Fotos SPF* is performing very well. It is mainly categorizing posts of amateur photographers who usually add the hashtag *#austrianmoment* and the specifications of the used camera as text. Another reason for the good performance of the categories in quadrant B lies in the Reference Model. All categories in this quadrant have a Reference Model containing at least 18 relatively long sentences. Compared to *Allgemeine Verbesserungsvorschläge AVB*, which is performing

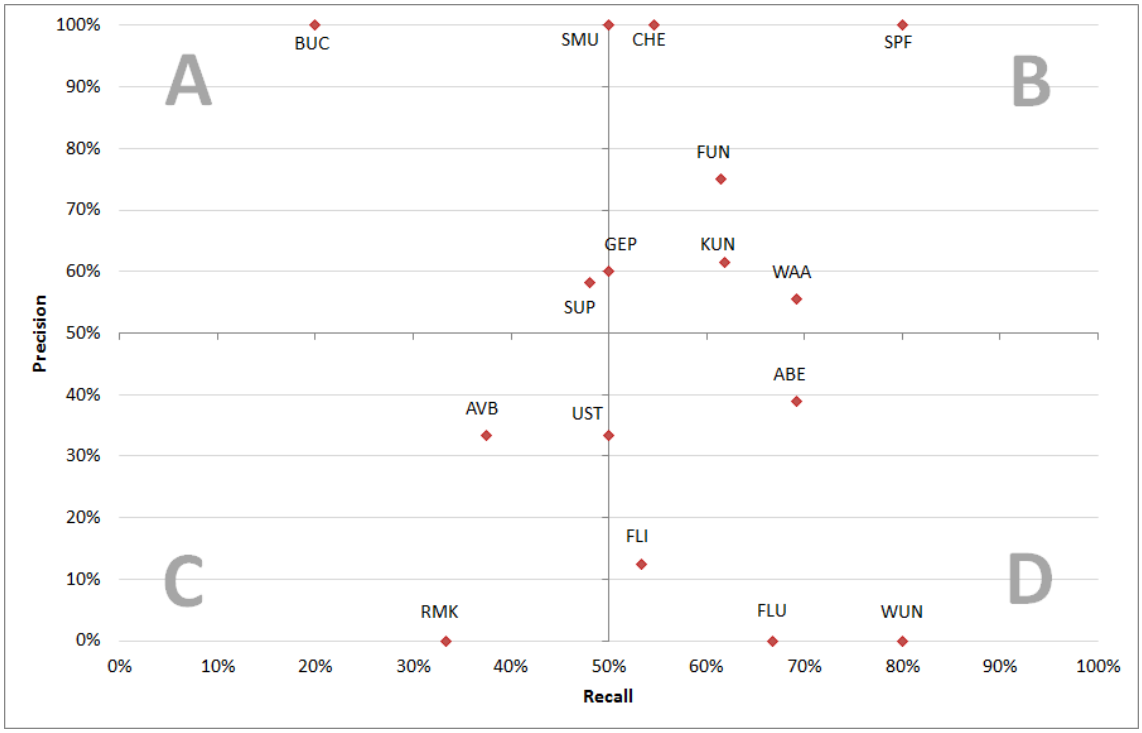

Figure 4.2.7: *Recall and Precision Plot*

poorly, having a reference set of 8 fairly short sentences.

Category *Wunschsitz WUN* has a high recall value but poor precision. The reason for that also lies in the Reference Model. The Reference Model contains 22 fairly long sentences, but they are not really dealing with the posts on the social media data. Another problem is, that the reference sentences have similarities with other categories like *Buchung*, *Check-in und Mobile Services* etc.

The missing precision in category *Fliegen und mehr* also arises because of too short sentences (e.g. sentences with only 5 words). After removal of Stop Words there is not enough semantic information left to weight and compare properly.

The poor recall value of the category *Buchung BUC* occurs because of missing reference sentences dealing with the social media topics and misspelling of the users on several posts.

4.3 Analysis of exemplary Results

This chapter shows two exemplary categorized posts using the models and data from the prior described evaluation.

4.3.1 Successful Categorization

Figure 4.3.1 shows a post of a user who asks about hand luggage regulations, regarding his motorcycle helmet. The post contains an introduction sentence: "*Hallo Austrian-Airlines Team,*" which was not considered due to the inverse sentence frequency. Looking at the results of the survey data (figure 4.3.2 - left part) shows that this post fits to category *Gepäck (Baggage)* and *Kundendienst (Customer Service)* only. The categorization (right part) shows the results of the correct categorization process. The matched sentence for *Gepäck* was "*Meine kurze Frage dazu, ist es erlaubt im Handgepäck einen Motorradhelm mitzunehmen, insofern die Größe das Gewicht des Handgepäcks eingehalten wird?*", which fits very good in this category. The second match, regarding category *Kundendienst* was sentence "*Freue mich auf eine rasche Antwort*" implying the expectation of a response of the customer service (figure 4.3.3 & figure 4.3.4).

Figure 4.3.1: *Successful post - example (original)*

Survey Category	--	-	+	++	Votes	Strict	Loose	Categorization	
Allgemeine Belobigungen	3	0	0	0	3	0.0%	0.0%	Allgemeine Belobigungen	✓
Allgemeine Verbesserungsvorschläge	3	0	0	0	3	0.0%	0.0%	Allgemeine Verbesserungsvorschläge	✓
Buchung	3	0	0	0	3	0.0%	0.0%	Buchung	✓
Bundesländer Flughäfen in Österreich	3	0	0	0	3	0.0%	0.0%	Bundesländer Flughäfen in Österreich	✓
Check-in und Mobile Services	3	0	0	0	3	0.0%	0.0%	Check-in und Mobile Services	✓
Fliegen und mehr	3	0	0	0	3	0.0%	0.0%	Fliegen und mehr	✓
Flugunregelmäßigkeiten	3	0	0	0	3	0.0%	0.0%	Flugunregelmäßigkeiten	✓
Gepäck	0	0	0	3	3	100.0%	100.0%	Gepäck	✓
Gewinnspiel	3	0	0	0	3	0.0%	0.0%	Gewinnspiel	✓
Kundendienst	1	0	2	0	3	0.0%	66.7%	Kundendienst	✓
Medizinische Fragen	3	0	0	0	3	0.0%	0.0%	Medizinische Fragen	✓
Reisen mit Kindern	3	0	0	0	3	0.0%	0.0%	Reisen mit Kindern	✓
Service und Personal	3	0	0	0	3	0.0%	0.0%	Service und Personal	✓
Smart Upgrade	3	0	0	0	3	0.0%	0.0%	Smart Upgrade	✓
Spotting und Fotos	2	1	0	0	3	0.0%	0.0%	Spotting und Fotos	✓
Umbuchung und Storno	3	0	0	0	3	0.0%	0.0%	Umbuchung und Storno	✓
Webseite, AirManager und App	3	0	0	0	3	0.0%	0.0%	Webseite, AirManager und App	✓
Wunschsitz	3	0	0	0	3	0.0%	0.0%	Wunschsitz	✓

Figure 4.3.2: *Survey and categorization result of successful example*

Matched Reference

Gepäck 0.45

Reference	Kann ich als Handgepäck mitnehmen
Reference parsed	können fähig sein im Stande sein drauf haben vermögen beherrschen mächtig Hand Kralle Pranke Pfote Greifhand Flosse gepäck mitnehmen auflesen aufsammeln vergessen zu bezahlen klauen einfach nehmen beklauen
Matched part	Meine kurze Frage dazu, ist es erlaubt im Handgepäck einen Motorradhelm mitzunehmen, insofern die Größe das Gewicht des Handgepäcks eingehalten wird?
Matched part parsed	kurz mini- stummelig lütt Miniatur- im Westentaschenformat mini Frage Chose Kiste Ding Punkt Issue Aufgabe erlauben Zustimmung geben einwilligen Erlaubnis erteilen billigen Segen geben verabschieden Hand Kralle Pranke Pfote Greifhand Flosse gepäck Motorrad Ofen Töff Bock Krad Feuerstuhl Kraftrad helm mitnehmen auflesen aufsammeln vergessen zu bezahlen klauen einfach nehmen beklauen insofern von daher dieserhalb und desterwegen insoweit somit deswegen anders gesagt Größe Ausmaß Liga Kaliber Magnitude Größenordnung Format Gewicht Sprengkraft Bedeutung Hantel Ballast Inertia Hand Kralle Pranke Pfote Greifhand Flosse [Person] einhalten beherzigen annehmen halten akzeptieren erfüllen befolgen

Figure 4.3.3: *Match details of successful example (1/2)*

Kundendienst 0.49

Reference	Danke für die schnelle Antwort
Reference parsed	danke danke schön danke sehr besten Dank man dankt verbindlichsten Dank wir haben zu danken für zu Händen z Hd das per für jedes je schnell zügig eilig rapide flott geschwind rasch Antwort Erwiderung Entgegnung Rückäußerung Widerrede Gegenrede Auskunft
Matched part	Freue mich auf eine rasche Antwort
Matched part parsed	freue rasch zügig eilig rapide flott geschwind behänd Antwort Erwiderung Entgegnung Rückäußerung Widerrede Gegenrede Auskunft

Figure 4.3.4: *Match details of successful example (2/2)*

4.3.2 Unsuccessful Categorization

Figure 4.3.5 shows an example of a post categorized into *Allgemeine Belobigungen (General Commendations)*, which is not correct. A look at the details reveals a common problem regarding social media and natural language processing - sarcasm. The person in the example is thanking for the bad service in a sarcastic manner. The natural language processing and categorization is not able to identify this sarcastic statement and therefore categorizes the post as *General Commendations*. A closer look at the survey result (figure 4.3.8) reveals that participants would also categorize the post as *Service und Personal (Service and Personnel)* and *Flugunregelmäßigkeiten (Flight Irregularities)*. The first reason for this missed categories is the word *gestrandet (stranded)*, which only retrieves one synonym out of the thesaurus. The second reason is the compound word *Bodenpersonal (Ground Crew*, which is divided into *Boden* and *personal*, but the meaning is distorted through the word *Boden*. This distortion also leads to the problem that the capitalization correction does not recognize *personal* as noun which, then again, leads to the problem that the query expansion does not work (figure 4.3.7).

Figure 4.3.5: *Unsuccessful post - example (original)*

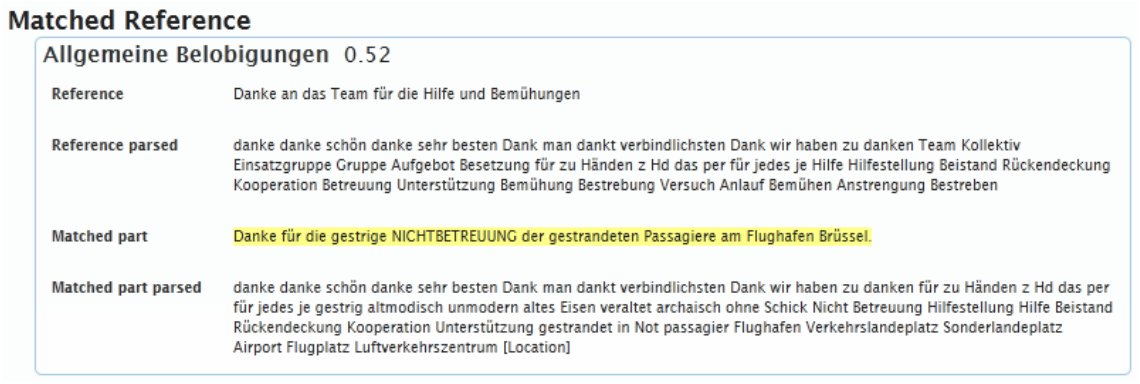

Figure 4.3.6: *Match details of unsuccessful example*

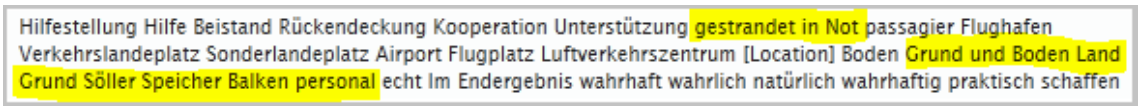

Figure 4.3.7: *Problematic processing of unsuccessful example*

Survey Category	--	-	+	++	Votes	Strict	Loose	Categorization	
Allgemeine Belobigungen	3	0	0	0	3	0.0%	0.0%	Allgemeine Belobigungen	✗
Allgemeine Verbesserungsvorschläge	2	1	0	0	3	0.0%	0.0%	Allgemeine Verbesserungsvorschläge	✓
Buchung	3	0	0	0	3	0.0%	0.0%	Buchung	✓
Bundesländer Flughäfen in Österreich	3	0	0	0	3	0.0%	0.0%	Bundesländer Flughäfen in Österreich	✓
Check-in und Mobile Services	3	0	0	0	3	0.0%	0.0%	Check-in und Mobile Services	✓
Fliegen und mehr	3	0	0	0	3	0.0%	0.0%	Fliegen und mehr	✓
Flugunregelmäßigkeiten	1	0	1	1	3	33.3%	66.7%	Flugunregelmäßigkeiten	✗
Gepäck	3	0	0	0	3	0.0%	0.0%	Gepäck	✓
Gewinnspiel	3	0	0	0	3	0.0%	0.0%	Gewinnspiel	✓
Kundendienst	3	0	0	0	3	0.0%	0.0%	Kundendienst	✓
Medizinische Fragen	3	0	0	0	3	0.0%	0.0%	Medizinische Fragen	✓
Reisen mit Kindern	3	0	0	0	3	0.0%	0.0%	Reisen mit Kindern	✓
Service und Personal	0	0	0	3	3	100.0%	100.0%	Service und Personal	✗
Smart Upgrade	3	0	0	0	3	0.0%	0.0%	Smart Upgrade	✓
Spotting und Fotos	3	0	0	0	3	0.0%	0.0%	Spotting und Fotos	✓
Umbuchung und Storno	3	0	0	0	3	0.0%	0.0%	Umbuchung und Storno	✓
Webseite, AirManager und App	3	0	0	0	3	0.0%	0.0%	Webseite, AirManager und App	✓
Wunschsitz	3	0	0	0	3	0.0%	0.0%	Wunschsitz	✓

Figure 4.3.8: *Survey and categorization result of unsuccessful example*

4.4 SeMFIS Import

In this chapter the prior described and analyzed results are exported in the structure of the ADOxx XML format in order to import the resulted insights back to SeMFIS using the standard import mechanism.

4.4.1 Export & Import Requirements

As prior described, it is possible to export the resulting data of the NLP application in two formats. The XML Publisher class, used for the evaluation and analysis tool (described in the following part), exports the object of the *OriginalPostHolder* class to an XML file using Jaxb. The ADOxx Publisher class also exports the resulting data to XML, but uses the Adoxml data model as structure (section 3.7.2). In order to be sure to meet the requirements of the import, I would personally recommend to use the built-in export function of the modeling toolkit and to export an example file to a folder. The standard XML export function of the platform also stores a dtd file[1], which can be used to adjust the export.

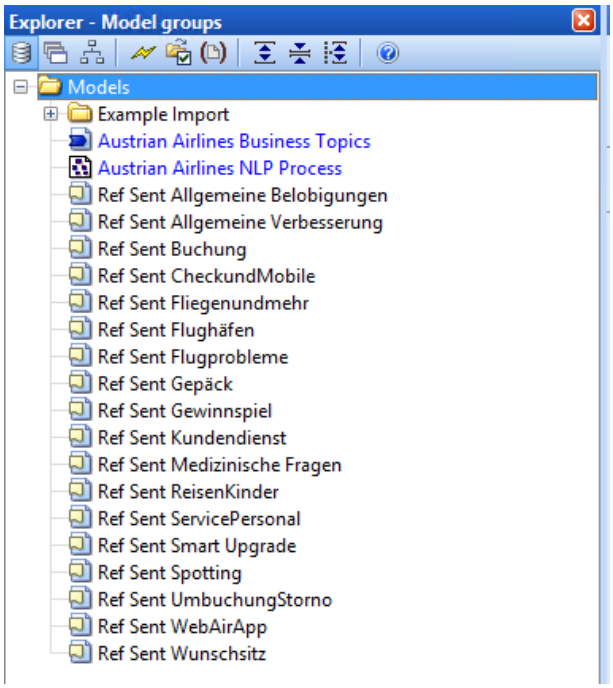

Figure 4.4.1: *NLP Process and Reference Model in the SeMFIS toolkit*

In this system, the *PublishHandler* class of the NLP application uses the *ADOxxXML* class in package *publish.techniques* to translate the data from the *OriginalPostHolder* to the ADOxx data model. Section 3.7.2 describes the data flow of the system in detail.

[1]Document type definition. Defines the structure of an XML document.

4.4.2 Import Approach

Based on the structure of the reference model, the results exported by the ADOxx Publisher are imported by using the standard import mechanism of the platform as defined in the concept (chapter 3.4). Figure 4.4.1 shows the Business Topic Model with its various Reference Sentence Models acting as Reference Model for the system and the NLP Process Model, defining the settings and the process. The models contain the already mentioned business topics of the Austrian Airlines described in prior sections. A detailed look at the different models can be obtained in chapter 3.6. After successful import of the exported XML file of the NLP application (figure 4.4.2), the results appear in the modeling toolkit using the structure of the reference model. Figure 4.4.3 shows posts categorized in *Webseite, AirManager und App (Website, AirManager and App)*. Opening the notebook of a post (figure 4.4.4), shows details like the whole text of the post, the match value, which is basically the the cosine similarity value used for the categorization and the original post url in order to view the original post (figure 4.4.5).

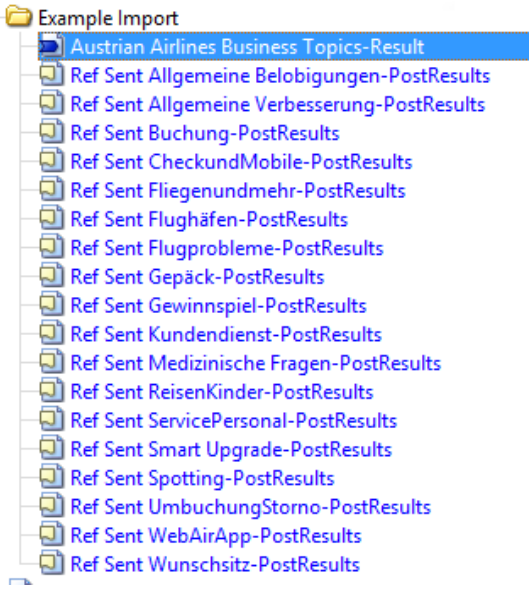

Figure 4.4.2: *Imported result using the SeMFIS toolkit*

Figure 4.4.3: *Imported posts representations of category (excerpt)*

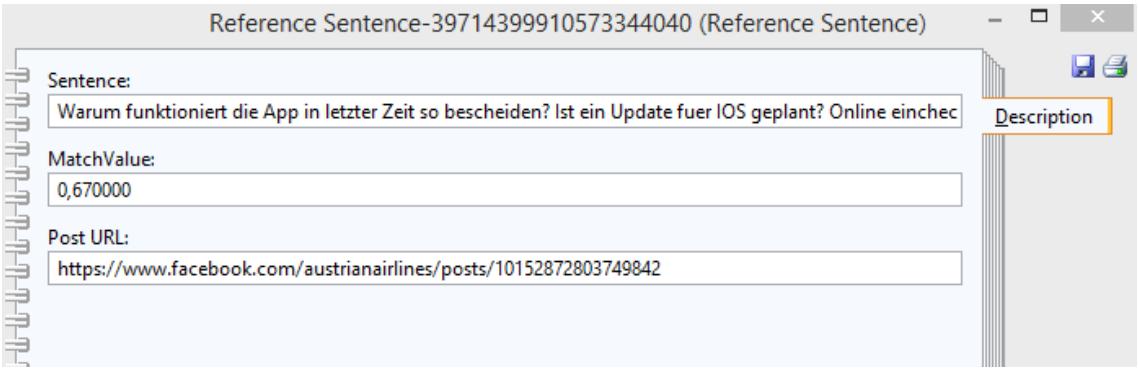

Figure 4.4.4: *Detail of categorized post*

Figure 4.4.5: *Original Post of the example*

Part 5

Conclusion

5.1 Summary

The emergence of social media platforms has changed the way businesses interact with potential customers enormously. After marketing established an additional channel to the customer in order to communicate their message, companies were forced to react to the enormous amount of feedback produced. The next logical step was establishing strategies and methods guiding the customer relationship management on this environment. While companies are still overwhelmed easily by the vast amount of feedback, the knowledge of the customer contained is still not reached efficiently due to the unstructured nature of the data. This thesis therefore researched an approach of collecting and aggregating knowledge from social media sources and connecting this collected customer knowledge in business process improvement initiatives, using an integrated approach of semantic-based modeling and natural language processing. In the field of natural language processing a tremendous amount of research is carried out, delivering a great number of language-specific approaches, techniques and applications due to the complexity and linguistic distinctions. This circumstances often lead to a situation where approaches, focusing on the German language, are not present or built on applications initially developed for the English language. More obstacles arise from emoticons, hyperlinks or similar and the low priority social media users devote to correct grammar and spelling, since most of the natural language processing approaches are developed for accurate orthography. In particular, sarcasm seems to be one of the unsolved challenges, since users on social media platforms tend to express issues in a sarcastic manner like they would do in person. Apart from these obstacles the segmentation of post texts into sentences improved the performance of the system in the realized approach through the possibility to filter unnecessary parts by using the derived inverse sentence frequency. The conceptual modeling methods, extending the SeMFIS platform, enabled an acutely flexible and convenient way of managing and customizing the system, without the need of changing the code of the Java application. Through the flexible structure, any user is able to alter the natural language processing and adapt it to the specific purpose as well as the reference model. In order to produce a suitable reference model, the use of the frequently asked question section of the company's web appearance offered a reliable, initial outline. But still, the social media environment stays in an ever-changing state and the constant adaption of the models is necessary. Overall the developed system of this approach could achieve valuable results, especially in clearly distinguished categories with a large reference sentence set.

5.2 Future Work

"Constant development is the law of life" –Mahatma Gandhi

Even if the proposed approach is able to achieve valuable results already, there is always room for improvement and further development. To stay within the limits of a Master Thesis, the following ideas are just suggestions and were not developed during this research.

After the categorization of the system is finished, the information gained from the resulting model is very valuable already, delivering insights about the prior defined business topics of the reference model in a clean and clear way. To use these insights more effectively, these clustered entries should be summarized and integrated as VOC (Voice of the Customer) into the CTQ (Critical to Quality) and CTB (Critical to Business) model of the RUPERT modeling tool (Johannsen and Fill, 2014b). In this way, the voice of the customer would directly drive business process improvement initiatives (figure 5.2.1).

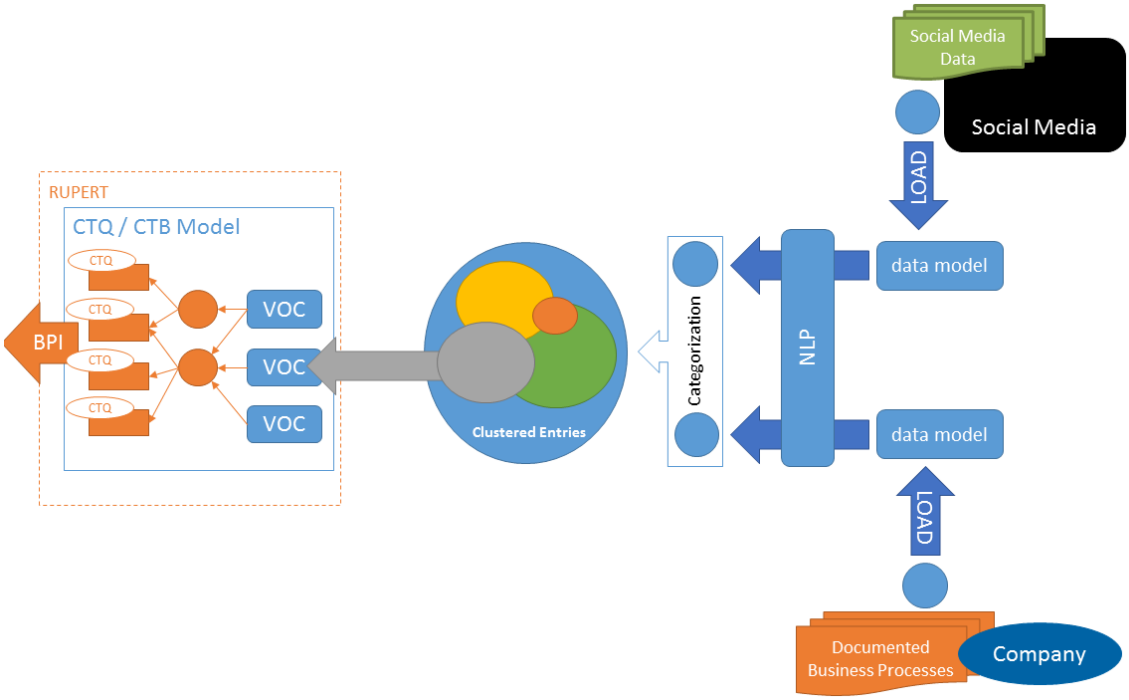

Figure 5.2.1: *Integration to RUPERT*

Bibliography

ADOxx.org (2015). ADOxx metamodelling platform. Online accessible at `http://www.adoxx.org/` last access: 11.08.2015.

Alfonseca, E., Inc, G., Bilac, S., and Pharies, S. (2008). S.: Decompounding query keywords from compounding languages. In *Proceedings of ACL-08: HLT, Short Papers*, pages 253–256.

Allen, M. and Bartels, N. (2015). Restfb - a lightweight java facebook graph api and old rest api client. Online accessible at `http://restfb.com/` last access: 16.08.2015.

Amsler, R. A. (1980). The structure of the merriam-webster pocket dictionary. Technical report, Austin, TX, USA. Online accessible at `ftp://ftp.cs.utexas.edu/pub/techreports/tr80-164a.pdf` last access: 27.08.2015.

Banfi, F., Gbahoué, B., and Schneider, J. (2015). Higher satisfaction at lower costs: Digitizing customer care. Online accessible at `http://www.mckinsey.com/~/media/McKinsey/dotcom/client_service/Telecoms/PDFs/Digitizing_customer_care.ashx` last access: 16.06.2015.

Barnes, N. G. and Wright, S. (2013). Fortune 500 are bullish on social media: Big companies get excited about google+, instagram, foursquare and pinterest. Online accessible at `http://www.umassd.edu/media/umassdartmouth/cmr/studiesandresearch/2013_Fortune_500.pdf` last access: 10.08.2015.

Baroni, M., Bernardini, S., Ferraresi, A., and 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.

Bauer, L. (2001). Compounding. In Haspelmath, M., editor, *Language Typology and Language Universals*. Mouton de Gruyter, The Hague.

Bauer, L. (2005). and compounding. In *Morphology and its demarcations: Selected papers from the 11th Morphology meeting, Vienna, February 2004*, volume 264, page 97. John Benjamins Publishing.

Beaulieu, M. and Jones, S. (1998). Interactive searching and interface issues in the okapi best match probabilistic retrieval system. *Interacting with computers*, 10(3):237–248.

Bergman, M. K. (2015). Sweet tools — a comprehensive listing of semantic web technologies and tools. Online accessible at `http://www.mkbergman.com/sweet-tools/` last access: 12.08.2015.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, 284(5):34–43.

Berton, A., Fetter, P., and Regel-Brietzmann, P. (1996). Compound words in large-vocabulary german speech recognition systems. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 2, pages 1165–1168 vol.2.

Bhogal, J., Macfarlane, A., and Smith, P. (2007). A review of ontology based query expansion. *Information processing & management*, 43(4):866–886.

Bohnet, B. (2010a). mate-tools tools for natural language analysis, generation and machine learning. Online accessible at `https://code.google.com/p/mate-tools/` last access: 10.08.2015.

Bohnet, B. (2010b). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97. Association for Computational Linguistics.

Bork, D. and Fill, H.-G. (2014). Formal aspects of enterprise modeling methods: A comparison framework. In *47th Hawaii International Conference on System Sciences (HICSS)*, Proceedings of the 2014 47th International Conference on System Sciences, pages 3400–3409, USA. IEEE Computer Society. Nominated for Best Paper Award.

Boyd, D. M. and Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.

Brinton, L. J. (2000). *The structure of modern English: a linguistic introduction*. John Benjamins Publishing.

Brown, M. K., Kellner, A., and Raggett, D. (2001). Stochastic language models (n-gram) specification. *W3C Working Draft*, 3.

Cambria, E., Schuller, B., Xia, Y., and Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *Intelligent Systems, IEEE*, 28(2):15–21.

Cardoso, J. and Pinto, A. M. (2015). *The Web Ontology Language (OWL) and its Applications*, pages 754–766. Information Science Pub.

Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50.

Computer and Information Science Department (1999). Penn treebank project. Online accessible at `https://www.cis.upenn.edu/~treebank` last access: 29.05.2015.

conversocial (2013). Social customer service performance report. Online accessible at `https://cdn2.hubspot.net/hub/154001/file-214612143-pdf/Social_Customer_Service_Performance_Report_June_2013.pdf` last access: 09.08.2015.

Corcho, O. and Gómez-Pérez, A. (2000). A roadmap to ontology specification languages. In Dieng, R. and Corby, O., editors, *Knowledge Engineering and Knowledge Management Methods, Models, and Tools*, volume 1937 of *Lecture Notes in Computer Science*, pages 80–96. Springer Berlin Heidelberg.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. ACM.

Davies, M. (2014). Penn treebank project. Online accessible at `http://corpus.byu.edu/wiki` last access: 29.05.2015.

Davis, D. (2013). 3rd biennial pex network report: State of the industry–trends and success factors in business process excellence. Online accessible at `http://www.specialoperationssummit.com/media/8820/20008.pdf` last access: 27.08.2015.

Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., and Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32 – 49.

Duggan, M., Ellison, N., Lampe, C., Lenhart, A., and Madden, M. (2015). Social media update 2014. Online accessible at `http://www.pewinternet.org/2015/01/09/social-media-update-2014/` last access: 22.03.2015.

Dunning, T. (1994). Statistical Identification of Language. Technical Report Technical Report MCCS-94-273, Computing Research Laboratory, New Mexico State.

Edison Research (2014). Social habit 2014. Online accessible at `http://www.edisonresearch.com/social-habit-report/` last access: 09.08.2015.

Engel, U. (2008). *Deutsche Grammatik*. Technische Universität Dortmund. Online accessible at `http://hdl.handle.net/2003/25838` last access: 27.08.2015.

Fan, W. and Gordon, M. D. (2014). The power of social media analytics. *Commun. ACM*, 57(6):74–81.

Fill, H.-G. (2009). *Visualisation for Semantic Information Systems*. Springer Science & Business Media.

Fill, H.-G. (2012). SeMFIS: A tool for managing semantic conceptual models. In *Workshop on Graphical Modeling Language Development*, Lyngby, Denmark.

Fill, H.-G. (2014). On the social network based semantic annotation of conceptual models. In Buchmann, R. A., Kifor, C. V., and Yu, J., editors, *Knowledge Science, Engineering and Management - 7th International Conference, KSEM 2014*, Knowledge Science, Engineering and Management - 7th International Conference, KSEM 2014, pages 138–149.

Fill, H.-G. (2015). Metamodeling as an interface between syntax and semantics. In Glück, B., Lachmayer, F., Schefbeck, G., and Schweighofer, E., editors, *Elektronische Schnittstellen in der Staatsorganisation*, pages 43–50. books@ocg.

Fill, H.-G. and Burzynski, P. (2009). Integrating ontology models and conceptual models using a meta modeling approach. In *11th International Protégé Conference, Amsterdam*.

Fill, H.-G. and Karagiannis, D. (2013). On the conceptualisation of modelling methods using the ADOxx meta modelling platform. *Enterprise Modelling and Information Systems Architectures - An International Journal*, 8(1).

Fill, H.-G., Schremser, D., and Karagiannis, D. (2013). A generic approach for the semantic annotation of conceptual models using a service-oriented architecture. *International Journal of Knowledge Management*, 9(1).

Forrester Consulting (2015). The definitive guide to social customer service. Online accessible at `http://landing.conversocial.com/the-2015-definitive-guide-to-social-customer-service` last access: 09.08.2015.

Forst, M. and Kaplan, R. M. (2006). The importance of precise tokenizing for deep grammars. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 369–372.

Friedl, J. E. F. (2006). *Mastering Regular Expressions*. O'Reilly Media, 3rd edition edition.

Fundin, A. P. and Bergman, B. L. (2003). Exploring the customer feedback process. *Measuring Business Excellence*, 7(2):55–65.

Gennari, J. H., Musen, M. A., Fergerson, R. W., Grosso, W. E., Crubézy, M., Eriksson, H., Noy, N. F., and Tu, S. W. (2003). The evolution of protégé: an environment for knowledge-based systems development. *International Journal of Human-computer studies*, 58(1):89–123.

Geyken, A. and Hanneforth, T. (2006). TAGH: A Complete Morphology for German based on Weighted Finite State Automata. In *Finite State Methods and Natural Language Processing. 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1-2, 2005. Revised Papers*, volume 4002, pages 55–66. Springer.

Gillan, P. (2010). The new conversation: Taking social media from talk to action. *Harvard Business Review*. Online accessible at `http://www.sas.com/reg/gen/corp/1207823-pr` last access: 27.08.2015.

González-Ibánez, R., Muresan, S., and Wacholder, N. (2011). Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 581–586. Association for Computational Linguistics.

Gray, J. (2003). Collection of ambiguous or inconsistent/incomplete statements. Online accessible at `http://www.gray-area.org/Research/Ambig` last access: 30.08.2015.

Grefenstette, G. (1995). Comparing two language identification schemes. In *The proceedings of 3rd International Conference on Statistical Analysis of Textual Data (JADT 95)*. Online accessible at `http://www.xrce.xerox.com/content/download/23364/170614/file/Gref---Comparing-two-language-identification-schemes.pdf` last access: 27.08.2015.

Grefenstette, G. and Tapanainen, P. (1994). *What is a word, what is a sentence?: problems of Tokenisation*. Rank Xerox Research Centre.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220.

Grégoire, Y., Salle, A., and Tripp, T. M. (2015). Managing social media crises with your customers: The good, the bad, and the ugly. *Business Horizons*, 58(2):173 – 182.

Heemskerk, J. S., van Heuven, V., et al. (1993). Morpa, a morpheme lexicon based morphological parser. Online accessible at `https://openaccess.leidenuniv.nl/handle/1887/2571` last access: 27.08.2015.

Hendler, J. (2009). Web 3.0 emerging. *IEEE Computer*, 42(1):111–113.

Hirst, G. (1981). *Anaphora in Natural Language Understanding*. Springer-Verlag, Berlin.

Hirst, G. (1987). *Semantic interpretation and the resolution of ambiguity*. Cambridge University Press.

Holler-Feldhaus, A. (2004). Koreferenz in hypertexten: Anforderungen an die annotation. *Osnabrücker Beiträge zur Sprachtheorie (OBST)*, 68:9–29. Online accessible at `http://www.cl.uni-heidelberg.de/~holler/obst68_holler.pdf` last access: 27.08.2015.

Holzner, S. (2008). *Facebook marketing: leverage social media to grow your business*. Pearson Education.

Howell, K. E. (2012). *An introduction to the philosophy of methodology*. Sage Publishing.

International Linguistics Community (2014). Text and corpora list. Online accessible at `http://linguistlist.org:8888/sp/GetWRListings.cfm?wrtypeid=1` last access: 29.05.2015.

Johannsen, F. and Fill, H.-G. (2014a). Codification of knowledge in business process improvement projects. In *European Conference on Information Systems (ECIS)*, European Conference on Information Systems (ECIS).

Johannsen, F. and Fill, H.-G. (2014b). RUPERT: A modelling tool for supporting business process improvement initiatives. In *International Conference on Design Science Research in Information Systems and Technology (DESRIST)*, Advancing the Impact of Design Science: Moving from Theory to Practice.

Jurafsky, D. and Martin, J. (2000). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence. Prentice Hall.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59 – 68.

Karagiannis, D. and Kühn, H. (2002). Metamodelling platforms. In Bauknecht, K., Tjoa, A., and Quirchmayr, G., editors, *E-Commerce and Web Technologies*, volume 2455 of *Lecture Notes in Computer Science*, pages 182–182. Springer Berlin Heidelberg.

King, B., Radev, D., and Abney, S. (2014). Experiments in sentence language identification with groups of similar languages. *COLING 2014*, 473:146.

Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Kleene, S. C. (1951). Representation of events in nerve nets and finite automata. Technical report, DTIC Document. Online accessible at `http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA596138` last access: 27.05.2015.

Lautenbacher, F., Bauer, B., and Seitz, C. (2008). Semantic business process modeling-benefits and capability. In *AAAI Spring Symposium: AI Meets Business Rules and Process Management*, pages 71–76.

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge University Press.

Lezius, W., Rapp, R., and Wettler, M. (1998). A freely available morphological analyzer, disambiguator and context sensitive lemmatizer for german. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 743–748, Stroudsburg, PA, USA. Association for Computational Linguistics.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317.

Lui, M. and Baldwin, T. (2012). Langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 25–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.

Merriam-Webster (2015). Definition of sarcasm. Online accessible at http://www.merriam-webster.com/dictionary/sarcasm last access: 09.08.2015.

Microsoft Corp. (2015). Microsoft jdbc driver for sql server. Online accessible at https://www.microsoft.com/en-US/download/details.aspx?id=11774 last access: 16.08.2015.

Mitkov, R. (2005). *The Oxford Handbook of Computational Linguistics*. Oxford Handbooks Series. Oxford University Press. Online accessible at https://books.google.at/books?id=0aClhre-vW4C last access: 27.08.2015.

Mitkov, R., Evans, R., Orasan, C., Barbu, C., Jones, L., and Sotirova, V. (2000). Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, pages 49–58. Citeseer.

Mizoguchi, R. and Kozaki, K. (2009). Ontology engineering environments. In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 315–336. Springer Berlin Heidelberg.

Monz, C. and De Rijke, M. (2002). Shallow morphological analysis in monolingual information retrieval for dutch, german, and italian. In *Evaluation of cross-language information retrieval systems*, pages 262–277. Springer.

Moore, S. (2015). Gartner says spending on business process management suites to reach $2.7 billion in 2015 as organizations digitalize processes. Online accessible at http://www.gartner.com/newsroom/id/3064717 last access: 16.06.2015.

Mylopoulos, J. (1992). Conceptual modelling and telos. Online accessible at http://www.cs.toronto.edu/~jm/2507S/Readings/CM+Telos.pdf last access: 27.08.2015.

Naber, D. (2015). Jwordsplitter. Online accessible at http://www.danielnaber.de/jwordsplitter/ last access: 10.08.2015.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nazir, A., Raza, S., and Chuah, C.-N. (2008). Unveiling facebook: a measurement study of social network based applications. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, pages 43–56. ACM.

Ng, V. (2009). Graph-cut-based anaphoricity determination for coreference resolution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 575–583, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Obrst, L. (2003). Ontologies for semantically interoperable systems. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 366–369. ACM.

O'Carroll, L. (2012). Twitter active users pass 200 million. Online accessible at `http://www.theguardian.com/technology/2012/dec/18/twitter-users-pass-200-million` last access: 25.05.2015.

Olson, D. L. and Delen, D. (2008). *Advanced Data Mining Techniques*. Springer Publishing Company, Incorporated, 1st edition.

OMILAB (2015). SeMFIS platform for engineering semantic annotations of conceptual models. Online accessible at `http://www.omilab.org/web/semfis` last access: 10.08.2015.

O'Reilly, T. (2005). What is web 2.0 - design patterns and business models for the next generation of software. *O'Reilly Network*. Online accessible at `http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html` last access: 22.03.2015.

Österle, H., Becker, J., Frank, Ü., Hess, T., Karagiannis, D., Krcmar, H., Loos, P., Mertens, P., Oberweis, A., and Sinz, E. J. (2011). Memorandum on design-oriented information systems research. *European Journal of Information Systems*, 20(20):7–10.

Paragon Poll (2013). Case study: Advanced sentiment anaysis. Online accessible at `http://paragonpoll.com/sentiment-analysis-systems-case-study/` last access: 21.08.2015.

Perera, P. and Witte, R. (2005). A self-learning context-aware lemmatizer for german. In *Proceedings of the Conference on Human Language Technology and Em-*

pirical Methods in Natural Language Processing, HLT '05, pages 636–643, Stroudsburg, PA, USA. Association for Computational Linguistics.

Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 1–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Projekt Gutenberg (2015). The complete works of william shakespeare. Online accessible at `http://www.gutenberg.org/ebooks/100` last access: 31.05.2015.

Pustejovsky, J. and Boguraev, B. (1996). *Lexical semantics*. Oxford University Press, USA.

Read, J., Dridan, R., Oepen, S., and Solberg, L. J. (2012). Sentence boundary detection: A long solved problem? In *COLING (Posters)*, pages 985–994.

Reimers, N., Eckle-Kohler, J., Schnober, C., and Gurevych, I. (2014). Germeval-2014: Nested named entity recognition with neural networks. *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition, Hildesheim, Germany*.

Richardson, A. (2010). Using customer journey maps to improve customer experience. *HBR Blog Network, posted*, 8(5). Online accessible at `https://hbr.org/2010/11/using-customer-journey-maps-to/` last access: 27.08.2015.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, 60(5):503–520.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In Salton, G., editor, *The Smart retrieval system - experiments in automatic document processing*, pages 313–323. Englewood Cliffs, NJ: Prentice-Hall.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Scalise, S. and Vogel, I. (2010). *Cross-disciplinary issues in compounding*, volume 311. John Benjamins Publishing.

Schubert, L. (2015). Computational linguistics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2015 edition. Online accessible at `http://plato.stanford.edu/archives/spr2015/entries/computational-linguistics/` last access: 30.08.2015.

Seethamraju, R. and Marjanovic, O. (2009). Role of process knowledge in business process improvement methodology: a case study. *Business Process Management Journal*, 15(6):920–936.

Shannon, C. E. and Weaver, W. (1959). *The mathematical theory of communication*. University of Illinois Press. Online accessible at `http://raley.english.ucsb.edu/wp-content/Engl800/Shannon-Weaver.pdf` last access: 27.08.2015.

Shtub, A. and Karni, R. (2010). Business process improvement. In *ERP*, pages 217–254. Springer US.

Shuyo, N. (2010). Language detection library for java. Online accessible at `http://code.google.com/p/language-detection/` last access: 10.08.2015.

Siha, S. M. and Saad, G. H. (2008). Business process improvement: empirical assessment and extensions. *Business Process Management Journal*, 14(6):778–802.

Solis, B. (2015). The conversation prism version 4. Online accessible at `http://www.conversationprism.com/` last access: 02.04.2015.

Solomonoff, R. J. (1957). An inductive inference machine. In *IRE Convention Record, Section on Information Theory, Part*, volume 2, pages 56–62. Online accessible at `http://world.std.com/~rjs/indinf56.pdf` last access: 21.08.2015.

Spärck-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Stanford University (2015). The stanford nlp. Online accessible at `http://nlp.stanford.edu/` last access: 16.08.2015.

Stanojević, M. (2009). Cognitive synonymy: A general overview. *FACTA UNIVERSITATIS-Linguistics and Literature*, (Vol. 7/2):193–200.

States, K. (2009). Youtube and others expose you to the whole wide world. Online accessible at `http://www.insidetucsonbusiness.com/news/small_business/youtube-and-others-expose-you-to-the-whole-wide-world/article_86fc3913-bb3e-54da-aae7-b43ac1729ce9.html` last access: 27.08.2015.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

The Apache Software Foundation (2015). Apache commons lang. Online accessible at `https://commons.apache.org/proper/commons-lang/download_lang.cgi` last access: 16.08.2015.

Trieschnigg, D., Kraaij, W., and de Jong, F. (2007). The influence of basic tokenization on biomedical document retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 803–804, New York, NY, USA. ACM.

Tudorache, T., Nyulas, C., Noy, N. F., and Musen, M. A. (2013). Webprotégé: A collaborative ontology editor and knowledge acquisition tool for the web. *Semantic web*, 4(1):89.

Twitter Data (2015). Floyd mayweather vs. manny pacquiao boxing championship 2015. Online accessible at `https://twitter.com/TwitterData/status/594737292562010112` last access: 25.05.2015.

Twitter UK (2014). Eurovision song contest 2014 statistics. Online accessible at `https://twitter.com/TwitterUK/status/465427409660289024/photo/1` last access: 25.05.2015.

Uszkoreit, H. (1996). What is computational linguistics? Online accessible at `http://www.coli.uni-saarland.de/~hansu/what_is_cl.html` last access: 30.08.2015.

Uszkoreit, H. (2002). New chances for deep linguistic processing. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02), August 24 - September 1*. Morgan Kauffmann Press.

Uszkoreit, H. (2009). Linguistics in computational linguistics: observations and predictions. *EACL 2009*, page 22.

van der Aalst, W., ter Hofstede, A., and Weske, M. (2003). Business process management: A survey. In *Business Process Management*, volume 2678 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin Heidelberg.

Van Dijck, J. (2013). *The Culture of Connectivity: A Critical History of Social Media*. Oxford University Press.

Vatanen, T., Väyrynen, J. J., and Virpioja, S. (2010). Language identification of short text segments with n-gram models. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Vechtomova, O. and Wang, Y. (2006). A study of the effect of term proximity on query expansion. *Journal of Information Science*, 32(4):324–333.

Vickery, G. and Wunsch-Vincent, S. (2007). *Participative web and user-created content: Web 2.0 wikis and social networking*. Organization for Economic Cooperation and Development (OECD), Paris, France, France.

Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P. (2009). On the evolution of user interaction in facebook. In *Proceedings of the 2Nd ACM Workshop on Online Social Networks*, WOSN '09, pages 37–42, New York, NY, USA. ACM.

W3C (2014). Rdf - semantic web standards. Online accessible at `http://www.w3.org/RDF/` last access: 12.08.2015.

Wattles, J. (2015). Facebook hits one billion users in a single day. Online accessible at `http://money.cnn.com/2015/08/27/technology/facebook-one-billion-users-single-day/` last access: 28.08.2015.

Weaver, W. (1955). Translation. *Machine translation of languages*, 14:15–23. Online accessible at `http://htl.linguist.univ-paris-diderot.fr/biennale/et09/supportscours/leon/Leon.pdf` last access: 27.08.2015.

Wilde, T. and Hess, T. (2007). Forschungsmethoden der Wirtschaftsinformatik. *WIRTSCHAFTSINFORMATIK*, 49(4):280–287.

Yamamoto, Y. (2015). twitter4j - a java library for the twitter api. Online accessible at `http://twitter4j.org/en/index.html` last access: 16.08.2015.

Zaman, A., Matsakis, P., and Brown, C. (2011). Evaluation of stop word lists in text retrieval using latent semantic indexing. In *Digital Information Management (ICDIM), 2011 Sixth International Conference on*, pages 133–136. IEEE.

Zellner, G. (2011). A structured evaluation of business process improvement approaches. *Business Process Management Journal*, 17(2):203–237.

Acronyms

BPM Business Process Management

BPI Business Process Improvement

NLP Natural Language Processing

RUPERT Regensburg University Process Excellence and Reengineering Toolkit

SeMFIS Semantic-based Modeling Framework for Information Systems

INTERREF Inter-Model-Reference

VOC Voice of the Customer

CTQ Critical to Quality

CTB Critical to Business

RDF Resource Description Framework

OWL Web Ontology Language

W3C World Wide Web Consortium

HTTP Hypertext Transfer Protocol

GUI Graphical User Interface

KPI Key Performance Indicator

QE Query Expansion

AQE Automatic Query Expansion

POS Part-of-Speech

NER Named Entity Recognition

NE Named Entity

MT Machine Translation

SBD Sentence Boundary Disambiguation

IE Information Extraction

QA Question Answering

SNS Social Networking Site

CRM Customer Relationship Management

TF Term Frequency

IDF Inverse Document Frequency

ISF Inverse Sentence Frequency

tp True Positives

tn True Negatives

fp False Positives

fn False Negatives

Part 6

Appendix

A Zusammenfassung

Die Evolution des world wide webs bereitete den Weg für die Verbreitung von Social Media Plattformen über die ganze Welt. Dabei veränderte sich nicht nur die Art und Weise wie Menschen miteinander kommunizieren, sondern auch wie Unternehmen mit potenziellen Kunden interagieren. Globalisierung und dieser neue Weg der Interaktion übt enormen Druck auf Unternehmen aus, wettbewerbsfähig zu bleiben. Eine Strategie ist der Fokus auf Geschäftsprozessoptimierungs-Initiativen um, durchgehend Kosten zu senken und Effizienz zu steigern unter Nutzung von Mitarbeiter- und Kundenwissen. Während Social Media Plattformen für Unternehmen hinsichtlich des Marketings mittlerweile weit verbreitet sind, wird der enorme Umfang an Daten, generiert von den Usern und so enorm wichtig für diese Geschäftsprozessoptimierungs-Initiativen, kaum verarbeitet und in nutzbares Wissen verwandelt. Diese Arbeit schlägt eine Methode vor, welche die Daten aus Social Media Plattformen unter Verwendung von natürlicher Sprachverarbeitung und semantischen Modellierungsmethoden als Erweiterung der SeMFIS Modellierungsmethoden basierend auf der ADOxx Metamodeling Platform verarbeitet. Dies schließt die technische Implementierung in Java, als auch eine nachfolgende Evaluierung der gewonnenen Ergebnisse hinsichtlich Objektivität und Validität unter Verwendung von Daten aus dem Facebook Auftritts einer Airline und gewonnenen Daten, basierend auf dem Wissen von menschlichen Akteuren, ein.

B Curriculum Vitae

Name: Martin Bergner, BSc (WU)
Nationality: Austria
E-Mail: martin.bergner@live.at

Education:

2013–now University of Vienna
Field of study: Business informatics

2008–2012 Vienna University of Economics and Business (WU)
Field of study: Business informatics

2007–2008 Mandatory Military Service

2002–2007 HTBLA Mössingerstraße Klagenfurt (Higher Technical College)

1998–2002 Secondary School in Metnitz

1992–1998 Primary School in Grades

Research:

Master Thesis
"*Integrating Natural Language Processing with Semantic-based Modeling*"
Under the supervision of Univ.-Prof. Mag. Dr. Hans-Georg Fill.

Bachelor Thesis
"*Automatisierung eines Value Stream Dashboards zur besseren Darstellung produktionsrelevanter Kennzahlen*"
In collaboration with Flex Althofen and under the supervision of Univ.Prof. Dr. Sarah Spiekermann and Dipl.-Wirt.Inform. Marie Caroline Oetzel.

Professional Career:

2014–2015 Data Analyst at Microsoft Austria

2012–2014 Application Architect at Flex Althofen

2005–2012 Intern at Flex Althofen (Periodically)

C Schriftliche Versicherung

Ich versichere, die von mir vorgelegte Arbeit selbstständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben. Ich habe mich bemüht, sämtliche Inhaber der Bildrechte ausfindig zu machen und ihre Zustimmung zur Verwendung der Bilder in dieser Arbeit eingeholt. Sollte dennoch eine Urheberrechtsverletzung bekannt werden, ersuche ich um Meldung bei mir.

_____

Datum, Unterschrift

D Additional Results & Analysis

This chapter covers additional results and analyses not listed in the main part due to the limited amount of space and readability.

D.1 Introduction Sentences

Results of the analysis of 200 German, random posts on the facebook page of Austrian Airlines between January and June 2015. Posts of the page and obvious spam posts have been filtered. The tables show only the first 50 characters of the sentence. Results show, that 25% of the entries include an introduction sentence, which disturbs the accuracy of the system.
The ID of the post in the first column can be used to view the original:
`https://www.facebook.com/austrianairlines/posts/[postID]`

816124641828732	<3-liche Grüße aus Hallstatt & von Dirndl to go :	No Intro
10153142316329800	OSVIE201505434 vom 04.03.15 nicht mal eine Antwor	No Intro
10153140756284800	Vielen Dank für die Korrektur! Boarding Pass mit A	No Intro
10153140000894800	Liebe Austria Team ! Donnerstag dem 25.6.2015 sol	Intro
10153136553329800	Liebes Austrian Airlines Team! Seit nun zwei Monat	Intro
10153136216719800	Liebes Austrian Airlines-Team, meine Familie und i	Intro
10153135183999800	Sg Austrian, stimmt es wirklich, dass die Lautspre	Intro
10153133094619800	Liebes Austrian Airlines Team! Ich fliege eigentl	Intro
10153130495509800	Ich tippte "muslimische Stewardess + Austrian Airl	No Intro
10153130483304800	Eine interessante Geschichte: Ich war dumm genug e	No Intro
10153129647509800	Servus! Gut, jetzt gibt es also auch bei Austrian	Intro
10153126962644800	Mit Abstand die schlechteste Airline weit und brei	No Intro
10153374605569100	Muslimische AUA-Flugbegleiter agieren antisemitisc	No Intro
10153124923204800	S.g. Herr Strache: - wenn ich als Unternehmer in	Intro
10153124579234800	Danke, dass Sie meinen Flug so umgemodelt haben, d	No Intro
998252173532510	Welche Fluglinie ist die beste und beliebteste? Di	No Intro
10153497490598700	Sehr geehrte Damen und Herren, gerade am Weltflüch	Intro
10153494795543700	So sehr ich mich über die zumindest zeitweise Auss	No Intro
10153119373354800	Liebe Austrian Airlines, aufgrund einer nicht zus	Intro
10153492528438700	So sehr ich mich über die zumindest zeitweise Auss	No Intro
10153117763574800	Ich gratuliere zu Euren Mitarbeiterinnen - Hut ab	No Intro
10153117633484800	Herzlichen Dank an die Pilotin und Crew die die Ab	No Intro
10153117603119800	Mal die neue Uniform beiseite, die inneren Werte z	No Intro
10153112427614800	Eine Schande, dass sie heute auf dem Flughafen in	No Intro
10153111605364800	Ja ich war zu spät, aber niemand hilft einem auf e	No Intro

10153111480599800	Austrian Airlines, I hope to fly to Vienna in Octo	No Intro
10153110852479800	Liebes Austrian Airlines Team. Mal wieder ein Pr	Intro
10153110839444800	Danke für meine erste und unvergessliche Geschäfts	No Intro
10153110767829800	@ Cabin ok: IATA (International Association of Air	No Intro
994022557288805	Wird der geplante GDS-Zuschlag der Lufthansa ein S	No Intro
849875681770310	Austrian so weit das Auge reicht.	No Intro
10153106047429800	Liebe Austrian, ich hoffe sehr, ihr macht den IATA	Intro
466778786817824	My Austrian, oder My Estonian?!? :) :(	No Intro
10153104962174800	Hallo, an welchen Flugtagen fliegt ihr nächstes Ja	Intro
10153103039679800	Am 10.5.2015 wurde unser Chicago-Flug kurz vor dem	No Intro
10153102969249800	Hola! Kann ich irgendwo in Mexico City meinen Flug	Intro
10153102047539800	Sehr geehrte Damen und Herren, Allein seit Februa	No Intro
10153101156319800	Am Sonntag bei mein em Rückflug von Rom wollte mei	No Intro
10153100945034800	Warum wurden alle Flüge nach Male im August gestri	No Intro
10205959942471000	liebe aua, geht's noch? was sagt ihr so dazu?	Intro
10153100198659800	"Suche dauert an. Bitte sehen Sie zu einem spätere	No Intro
847612998663245	Und nach drei Versuchen kommen wir immerhin zum al	No Intro
847611905330021	Liebe Austrian, das ist jetz langsam sehr lästig!	Intro
10153099172739800	Sg. Damen und Herren, ich versuche nachträglich zw	Intro
10153098291494800	Austrians@Thailand hat Stammtisch heute ab 19 Uhr	No Intro
1446404555677980	Liebes Austrian-Team, Für einen Flug von Wien nach	Intro
10153097490989800	Ist denn jetzt alles rund um UMs deutlich mühsamer	No Intro
10153096744724800	Sehr geehrte Damen und Herren, ich komme gerade vo	Intro
10153096397464800	Hallo, leider kann ich bei unseren Flügen am 8. Ju	Intro
10153094064419800	Wieso werden die vom Euroairport startenden Austri	No Intro
714736221987971	in flight Airbus A-320 von Dubrovnik nach Wien übe	No Intro
714735948654665	in flight Dash 8 von Zagreb nach Wien über dem süd	No Intro
926396464086092	OE-LBI, Austrian Airlines, Airbus A320-200 at Berl	No Intro
10153093847599800	Liebe austrian Airlines, Ende Juni werde ich nac	Intro
10153093450954800	Austriann ??	No Intro
10153092804004800	Das Myholidaywebcheckin System ist ja nicht gerade	No Intro
1098714933477370	Servus Austrian ;)	Intro
1083692571660260	Frankfurt 04.05.2015 Austrian Airlines DHC-8-402	No Intro
10153087919689800	Vorabend Check-In ab Linz geht NUR für Ferienflüge	No Intro
10203010112893600	Die Email feedback@austrian.com scheint nicht zu e	No Intro
10153087030524800	Habe nun das erste Mal einen Flug bei austrian air	No Intro
986910904666637	Lufthansa sorgt für medizinische Zwischenfälle an	No Intro
1599892050291580	#austrianmoment Fokker 70 OE-LFQ im Steigflug vom	No Intro
10153078847809800	Eine super Airline kann man nur empfehlen *****	No Intro
1100104383349300	Viele Grüße aus Flughafen München 31.05.15 Austria	No Intro
843505922407286	Nichts als Fehlermeldungen erhält man bei eurem on	No Intro
1598993987048050	Dash 8 OE-LGM landet am Flughafen Graz (GRZ). http	No Intro
10153068373689800	Hallo Austrian Airlines, wie so oft in letzter Zei	Intro

10153066937249800	Sehr geehrtes Austrian Airlines Team, zunächst ein	Intro
1598664980414290	Dash 8 OE-LGF startet am Flughafen Graz (GRZ). htt	No Intro
1598664927080960	Fokker 70 startet vom Flughafen Graz (GRZ) nach Dü	No Intro
10153064434579800	ein großes D A N K E an die beste flugbegleiterin,	No Intro
10153058592919800	Danke für die gestrige NICHTBETREUUNG der gestrand	No Intro
10153055690029800	Muss hier die Crew vom Flug Os026 26/27.5 loben. T	No Intro
10204138666539200	Vielen vielen dank!!! Wir waren sprachlos....das i	No Intro
10153053763139800	Liebes AUA-Team! Haben alle Unterlagen für das Ei	Intro
1597844953829620	#austrianmoment Fokker 70 OE-LFQ wenige Sekunden n	No Intro
10153046291154800	Der dritte Flug in Folge der überbucht ist.... der	No Intro
1597438507203600	Dash 8 landet am Flughafen Graz (GRZ) im Wind & Re	No Intro
10153333506558600	Auf verschiedenen Seiten wollen Eure Banner ein Be	No Intro
1597080953906020	Dash 8-402 OE-LGL "Altenrhein" startet vom Flughaf	No Intro
10153039249719800	Kann es sein, dass seit gestern Nachmittag bei euc	No Intro
858827837535414	Ganz Europa schaut nach Wien. #ESC2015 #Eurovision	No Intro
10153035457154800	Also ich finde, diese fürchterlichen Uniformen sin	No Intro
599425420200749	Sehr geehrtes Austrian Airlines Team. Seit Septemb	Intro
10153032457894800	Bei meinem Flug morgen von Stockholm nach Wien ist	No Intro
10153031522709800	Ich freue mich, die AUA zu unterstützen. Ich musst	No Intro
10153031269724800	Lieber Mitarbeiter der AUA von Gate 26, der Sie vo	No Intro
10153029057959800	wie schaut es eigentlich aus mit Shark Lets an den	No Intro
10153028051364800	ich habe mehrfach versucht einen Flug auf der Aust	No Intro
10153027387759800	Statt neuer Uniformen wie wäre es mit der Wiederei	No Intro
987743201237962	;) Die neuen Uniformen	No Intro
10153024352484800	Hallo liebes Austrian Airlines Team, Ich bin letzt	Intro
10153024103569800	Welche Geschäftspolitik führt die AUA wenn sie Flü	No Intro
10153023729684800	Kommt man bei euch telefonisch auch irgendwann mal	No Intro
1594369380843850	Dash 8 OE-LGA startet am Flughafen Graz. https://y	No Intro
1594369257510530	Dash 8-402 OE-LGB startet vom Flughafen Graz (GRZ)	No Intro
465584196931216	Hallo, würde mich sehr über eure Unterstützung bei	Intro
10153022043844800	Wann stellt ihr die neue Uniform vor?	No Intro
10153021226064800	Es ist das Jahr 2015: ich buche selbst einen Flug	No Intro
Result	**27 with Intro**	**100**

Table D.1: *Introduction Sentence Analysis Results (first 100)*

10152862256744800	Endlich Miami !!! ich warte jetzt schon 10 Jahre d	No Intro
1559296854351100	#austrianmoment OE-LGE am Grazer Flughafen	No Intro
10152861512514800	Liebe Austrian Airlines, vielen Dank für einen bes	Intro
10152861367019800	Nach 15 Stunden Reise endlich in Hongkong angekomm	No Intro
946454538712306	bin bis jetzt immer gerne aua geflogen. da das man	No Intro
10152860219174800	bin bis jetzt immer gerne aua geflogen. da das man	No Intro
10152859945729800	Hallo liebes Austrian Airlines Team! Fliege morge	Intro

10152854820594800	Würde mal wer von euch das GPS in der OE-LFI überp	No Intro
1406149336363560	Hallo Austrian Airlines :)! Hier ein pic von eurer	Intro
10152850690334800	Ich frage mich wieder einmal, geht es dem fliegend	No Intro
10152847232364800	Wisst ihr was ? :D ich freu mich MEEEEEEGA auf die	No Intro
402771116550369	Ihr habt wirklich eine wirklich super Business Cla	No Intro
10152844274949800	Feedback zum neuen Webdesign, zumindest beim Check	No Intro
10152838681354800	Vor 3 Wochen habe ich Ihnen zusätzlich ein Einschr	No Intro
10152838228319800	Liebes Austrian-Team! Ich komme direkt vom Flug OS	Intro
1006689849358690	Flightsimulator X Foto, mit der PMDG 777 von LOWL(	No Intro
10152836065514800	Are you ready for take-off? Austrian fan Thorsten	No Intro
10204773561797400	Im Anflug auf Innsbruck mit der Dash8/400	No Intro
815630055140090	Mit Vollgas aus Innsbruck raus :D OS1402 Ferry nac	No Intro
897775280244847	Das neueste Flottenmitglied OE-LGQ Wilder Kaiser n	No Intro
10152826654634800	Wenn ich einen Austrian Flug buche, dann hätte ich	No Intro
418715684961068	Hallo an euch in Ägypten Eines der besten Mittel d	No Intro
10152825422809800	Hallo Austrian Team Mein Name ist Daniel und ich w	Intro
602050683264625	https://www.youtube.com/watch?v=u5Kf5m-6sIs Hier m	No Intro
1548540265426760	OE-LGH "Vorarlberg" am Flughafen Graz.	No Intro
10152820890679800	Die Austrian Arilines reißen sich ja in anderen Lä	No Intro
10152820451564800	Der Valentinstag ist zwar erst morgen, wir verrate	No Intro
10152820432969800	Liebes Austrian Airlines - Ich habe einen Fug gebu	Intro
10152820427919800	Liebes Austrian Team! Beim morgigen Valentins Spec	Intro
10152819337144800	Warum lässt die Austrian Airlines keine Name Chang	No Intro
10152817255784800	Wie lange dauert es denn im Normalfall, bis auf ei	No Intro
10152816649464800	Hallo, gibt es schon Informationen wie sich der ge	Intro
10152813840189800	Hi Austrian Airlines, Da ich kurzfristig erfahren	Intro
1021019687913560	Schneefall... ;)	No Intro
1020948387920690	Seitenwind...	No Intro
10205239453858300	#austrianmoment #winterspotter	No Intro
10152805898689800	Und wieder einmal ist die Anbindung DUS-LNZ heute	No Intro
10153097508762900	Guten Tag, leider funktioniert die Austrian app se	Intro
10152803460044800	Ich habe eine E-Mail an feedback@austrian.com vor	No Intro
889628291059546	OE-LGM Take-Off nach Wien mit dem Nordkettenpanora	No Intro
10152801518724800	das nenne ich Kundenorientierung und Umgang mit Ku	No Intro
808989862470776	OE-LDF im Anflug auf die 26 als AUA670 aus Kiev :)	No Intro
10152800895099800	Bin ich naiv? "Ihr Flug ist nun zum Einchecken ber	No Intro
10152623152724200	Habe heute um 19:00 Uhr angerufen weil ich umbuche	No Intro
10152799718699800	Hallo. Ich habe fuer Maerz 2 Austrian Airlines Flu	Intro
10152799386359800	Herzlichen Dank, dass die Austrian den oberösterre	No Intro
10152798882614800	Habe heute überprüft wie die Lage mit meinem Flug	No Intro
10152797005089800	Hallo liebes Austrian Airlines-FB-Team! Vielleicht	Intro
10152796947064800	Hallo! Kann man mit Baby nicht im Web eine boardka	No Intro
10152796234904800	Gerüchten zu Folge hat die Austrian Airlines berei	No Intro

10152793336849800	ich möchte mich bei Ihnen ganz herzlich für folgen	No Intro
10152791837119800	Hallo, ich fliege am Samstag um 17:15 nach Dubai u	Intro
10152791652474800	Hallo. Ich fliege am Samstag von Wien über Frankfu	Intro
10152791111039800	Ihr solltet euch SCHÄMEN ..Streik in Düsseldorf u	No Intro
10152791023134800	Austrian Airlines Danke vielmals :)	No Intro
10152787374434800	kann man eventuell nach einem Monat eine email bea	No Intro
806235122746250	OE-LGQ im Triebwerkcheck in Innsbruck ;)	No Intro
10152784820039800	Wie schauts mit den AUA-Flügen morgen aus New York	No Intro
10203440859925200	OS 802 ATH-VIE am 25. Jännerr 2015. Athen von ob	No Intro
10203440847204900	OS 801 VIE-ATH am 24. Jänner 2015. Anflug auf Athe	No Intro
10152780655509800	Liebe Austrian Airlines! ich würde dieses Jahr von	Intro
10152779173784800	Liebe Austrian Airlines, diese Destinationen seh	Intro
336310316566018	Austrian Airlines OE-LVG Dnepropetrovsk - (UKDD /	No Intro
10152778645494800	Hallo, ihr solltet bei der Auswahl von Personal et	No Intro
10152776920274800	Hallo Austrian ! Nehmt Euch an Eurem Schwesterunte	Intro
10152776537709800	danke für die rasche abwicklung, den anruf und das	No Intro
10152776451554800	warum kann man gerade ”aus technischen gründen” ni	No Intro
10152776125324800	schon wieder troubles mit os. zuerst wird ohne inf	No Intro
10205860971468400	Winter in Peking - Eislaufen der etwas anderen Art	No Intro
1535175653429890	Gestern auf Flug OS977 / OS978	No Intro
10152770556404800	Traumurlaub im Inselparadies: Ab Oktober 2015 hebe	No Intro
10152769774179800	Habe euch gestern eine PN gesendet ??	No Intro
10152765543279800	Habe euch eben eine PN geschickt :)	No Intro
1526600750936490	Ex-Austrian airlines	No Intro
10152765040139800	Wir finden es sehr bedauerlich, dass am ersten Fer	No Intro
10152763003309800	Liebe AUA, gilt es bei Ihnen als ”normaler Kunden	Intro
10152761851514800	Keine Zeitung mehr im Flugzeug? Nicht einmal mehr	No Intro
10152761807549800	<3 Austrian <3	No Intro
10152759893174800	Ein beschissener service von Austrian. Leider hat	No Intro
10152759675999800	Herrlich mit Niki zu fliegen und sogar noch mehrer	No Intro
10152759597919800	Liebes Austrian Kundenserviceteam! Am 14.11 habe i	Intro
10152759376534800	Immer schwächeres Service bei der Austrian, jetzt	No Intro
10152758785759800	Mein gepäck ist seit ca. 4 Monaten weg wann kann m	No Intro
10152758747249800	eine frage bitte: ich fliege OS 89 nach NY dann OS	No Intro
10152758509289800	Am 7.11. eine Anfrage geschickt, am 13.11. noch um	No Intro
10152758241909800	Hey. Ich fliege im Februar von Dubai nach Hause un	Intro
10152756274274800	Liebes Team von Austrian Airlines! Ich bin gester	Intro
485764714922355	AUA, das ist Peinlich!	No Intro
1530162070597910	#austrianmoment Die OE-LFK am Flughafen Graz (GRZ)	No Intro
10152748772914800	Nach 14 Stunden Verspätung von OS63 und verpassten	No Intro
10152748115659800	Es hilft nicht immer mit der gleichen Antwort Wir	No Intro
10152747981884800	Am 23. Oktober durfte ich mich mit einer Flugversp	No Intro
10152744850619800	Flug OS 802 Athen - Wien am 5.1.15 - Check In Scha	No Intro

10152744716249800	Liebe Austrian, für die Verspätung von 2 1/2 Stund	Intro
10152744171114800	Deshalb setzen Unternehmen bei ihren Radio- und Fe	No Intro
10152744051454800	”Es steht Ihnen frei eine andere Fluglinie zu buch	No Intro
10152743700624800	Liebe Austrian..... keine Zeitungen mehr in der Ec	Intro
10152743640909800	Hallo, Ich bin 15 Jahre alt, und möchte sehr gern	Intro
10152743623614800	Eine Katastrophe und schlechte Erfahrung	No Intro
894738177225919	ich hoffe es Ihnen gefallen! ^_^	No Intro
Result	**24 with Intro**	**100**

Table D.2: *Introduction Sentence Analysis Results (second 100)*

D.2 TF-IDF on Hamlet and Henry the 8th

Henry the 8th and Hamlet are two of the best known works of William Shakespeare. Prior to the analysis all words have been converted to lowercase and all special characters (except "-") have been removed. Together these documents hold an impressive amount of 51.981 words (6.799 unique).

word	occurences	share
the	905	3,52%
halberds	734	2,86%
to	601	2,34%
honest	562	2,19%
lose	561	2,18%
a	455	1,78%
alleged	414	1,61%
ignorance	372	1,45%
that	301	1,30%
emperor	268	1,17%

Table D.3: *10 most frequent words in Henry the 8th out of 25.701 (with stop words)*

word	occurences	share
the	1085	4,13%
and	964	3,67%
to	742	2,82%
of	673	2,56%
i	577	2,20%
a	554	2,11%
you	549	2,09%
my	521	1,99%
in	439	1,67%
it	417	1,59%

Table D.4: *10 most frequent words in Hamlet out of 26.280 (with stop words)*

Table D.3 and D.4 show the 10 most occurring words in each of the plays. As in the analysis of all of Shakespeares works, mostly stop words appear as the most frequent, bringing no clue of the content. Only the calculation of the inverse document frequencies of the unique words of both documents (D.5) shows a better picture in the first step. Words like "hamlet", "wolsey", "cardinal" do not appear in both documents but have a high number of occurrence in sum. Typical stop words with high occurrences in both documents like "the", "and", etc. obtain a idf score of 0. Bringing term frequency and inverse document frequency together, results in a highly promising picture (D.6 & D.7). In addition to recommendations to read both of this milestones in literature and theatre history, the words "wolsey", "buckingham" describe two of the main characters of Henry the 8th - *Cardinal Wolsey* and the *Duke of Buckingham* who gets framed by Wolsey for treason. Now quite everyone is aware of King Henrys relationship to women, so the word "wedding-day" in the top 20 is not much surprising. Also not surprising is the absence of the word "king". This particular word appears a lot of times in both documents and therefore gets an idf score of 0. The word "henry" also does not appear in the top 20, it appears on place 109, because of its small term frequency (only 13 appearances in King Henry the 8th).

word	occurences	idf
ham	358	0,30103
hamlet	116	0,30103
hor	109	0,30103
wolsey	95	0,30103
pol	86	0,30103
cardinal	66	0,30103
katharine	62	0,30103
...	...	...
the	1990	0,0000
and	1698	0,0000
to	1343	0,0000
of	1235	0,0000

Table D.5: *Inverse Document Frequencies of Hamlet and Henry the 8th (excerpt)*

word	term frequency	occurrences	idf	tf-idf
halberds	81,1%	734	0,301029996	0,244150
alleged	45,7%	414	0,301029996	0,137709
vanities	29,6%	268	0,301029996	0,089145
endeavours	24,9%	225	0,301029996	0,074842
him-which	23,6%	214	0,301029996	0,071183
unloosd	23,2%	210	0,301029996	0,069852
scribes	20,9%	189	0,301029996	0,062867
uncontemnd	20,8%	188	0,301029996	0,062534
wedding-day	20,8%	188	0,301029996	0,062534
july	20,6%	186	0,301029996	0,061869
buckingham	17,9%	162	0,301029996	0,053886
inventory	16,2%	147	0,301029996	0,048897
greatest	13,7%	124	0,301029996	0,041246
well-beloved	12,8%	116	0,301029996	0,038585
memusic	12,2%	110	0,301029996	0,036589
thought-i	10,6%	96	0,301029996	0,031932
admirer	10,5%	95	0,301029996	0,031600
butts	10,5%	95	0,301029996	0,031600
wolsey	10,5%	95	0,301029996	0,031600
haberdasher	9,7%	88	0,301029996	0,029271

Table D.6: *20 words with the highest TF-IDF weights in Henry the 8th*

word	term frequency	occurrences	idf	tf-idf
ham	33,0%	358	0,301030	0,099326
hamlet	10,7%	116	0,301030	0,032184
hor	10,0%	109	0,301030	0,030242
pol	7,9%	86	0,301030	0,023860
laer	5,7%	62	0,301030	0,017202
oph	5,3%	58	0,301030	0,016092
horatio	4,2%	46	0,301030	0,012763
laertes	4,2%	46	0,301030	0,012763
ros	4,1%	45	0,301030	0,012485
clown	3,3%	36	0,301030	0,009988
polonius	3,3%	36	0,301030	0,009988
ghost	3,1%	34	0,301030	0,009433
mar	2,9%	32	0,301030	0,008878
guil	2,7%	29	0,301030	0,008046
guildenstern	2,7%	29	0,301030	0,008046
ophelia	2,5%	27	0,301030	0,007491
rosencrantz	2,4%	26	0,301030	0,007214
osr	2,3%	25	0,301030	0,006936
denmark	2,1%	23	0,301030	0,006381
elsinore	2,1%	23	0,301030	0,006381

Table D.7: *20 words with the highest TF-IDF weights in Hamlet*

The calculation of Hamlet receives an even better result, revealing 10 of the main characters of the play[1]: *Hamlet, son of the former and nephew to the present king*; *Horatio, friend to Hamlet*; *Laertes, son to Polonius*; *Polonius, Lord Chamberlain*; *The Ghost of Hamlets Father*; *The two Clowns gravediggers*; *Guildenstern, a courtier*; *Ophelia, daughter to Polonius*; *Rosencrantz, a courtier*; *Marcellus, Officer* and two venues *Denmark* and *Elsinore*. Even without correct tokenization or normalization, the tf-idf algorithm delivers good results for the plays Henry the 8th and Hamlet from Shakespeare.

[1]Shakespeare abbreviated some of the names in the play like *Ham* for *Hamlet* or *Oph* for *Ophelia*.

E Interaction Code (ADOScript)

The following code, described in section 3.5.2, shows the interaction of ADOScript with the NLP application.

```
#CC "Modeling" GET_ACT_MODEL
CC "CoreUI" MODEL_SELECT_BOX multi-sel mgroup-sel

IF (endbutton = "ok") {
   SET selModels:(modelids)

   #SET id_actModelId: (modelid)
   SET setClassName1: "XML Publisher"
   SET setClassName2: "ADOxx Publisher"
   SET setAttrName: "Publish Path"

   # search for the model with the settings
   FOR fmodel in: (selModels)
   {
      SET curModel:(VAL fmodel)
      # we get the working directory path from the settings class
      CC "Core" GET_ALL_OBJS_OF_CLASSNAME modelid:(curModel)
          classname:(setClassName1) #-->RESULT ecode:intValue objids:list
      IF (objids="")
      {
         CC "Core" GET_ALL_OBJS_OF_CLASSNAME modelid:(curModel)
             classname:(setClassName2) #-->RESULT ecode:intValue
             objids:list
      }

      IF (objids!="")
      {
         SET id_obj:(VAL token(objids,0," ")) # we just take the first
             (there should be only 1 setting object
         CC "Core" GET_ATTR_VAL objid:(id_obj) attrname:(setAttrName)
             as-string #attrid:(attrID) as-string #-->RESULT
             ecode:intValue val:anyValue
         SET str_path:(val)
```

```
#CC "AdoScript" INFOBOX (str_path)

SETL sADOXMLExportFile:(str_path+"\\SetBO.xml")
# Export ADOxml to the file space for further processing
CC "Documentation" XML_MODELS modelids:(selModels) mgroupids:("")
    attrprofs:("") apgroups:("")

CC "AdoScript" FWRITE file:(sADOXMLExportFile) text:(xml) binary:0
CC "AdoScript" INFOBOX "Export completed successfully. Start
    Social Media Processing."
CC "AdoScript" DELETE_FILES (str_path+"\\social.jar")
CC "AdoScript" FCOPY from:("db:\\NLP_final.jar")
    to:(str_path+"\\social.jar")

CC "AdoScript" DIR_CREATE path:(str_path+"\\lib")

CC "AdoScript" DELETE_FILES (str_path+"\\lib\\anna-3.61.jar")
CC "AdoScript" FCOPY from:("db:\\anna-3.61.jar")
    to:(str_path+"\\lib\\anna-3.61.jar")

CC "AdoScript" DELETE_FILES
    (str_path+"\\lib\\commons-lang-2.6.jar")
CC "AdoScript" FCOPY from:("db:\\commons-lang-2.6.jar")
    to:(str_path+"\\lib\\commons-lang-2.6.jar")

CC "AdoScript" DELETE_FILES (str_path+"\\lib\\jsonic-1.2.0.jar")
CC "AdoScript" FCOPY from:("db:\\jsonic-1.2.0.jar")
    to:(str_path+"\\lib\\jsonic-1.2.0.jar")

CC "AdoScript" DELETE_FILES
    (str_path+"\\lib\\jwordsplitter-4.1.jar")
CC "AdoScript" FCOPY from:("db:\\jwordsplitter-4.1.jar")
    to:(str_path+"\\lib\\jwordsplitter-4.1.jar")

CC "AdoScript" DELETE_FILES
    (str_path+"\\lib\\jwordsplitter-4.1.jar")
CC "AdoScript" FCOPY from:("db:\\jwordsplitter-4.1.jar")
    to:(str_path+"\\lib\\jwordsplitter-4.1.jar")

CC "AdoScript" DELETE_FILES (str_path+"\\lib\\langdetect.jar")
```

```
        CC "AdoScript" FCOPY from:("db:\\langdetect.jar")
            to:(str_path+"\\lib\\langdetect.jar")

        CC "AdoScript" DELETE_FILES (str_path+"\\lib\\restfb-1.11.0.jar")
        CC "AdoScript" FCOPY from:("db:\\restfb-1.11.0.jar")
            to:(str_path+"\\lib\\restfb-1.11.0.jar")

        CC "AdoScript" DELETE_FILES (str_path+"\\lib\\fljdbc41.jar")
        CC "AdoScript" FCOPY from:("db:\\fljdbc41.jar")
            to:(str_path+"\\lib\\fljdbc41.jar")

        CC "AdoScript" DELETE_FILES
            (str_path+"\\lib\\stanford-corenlp-3.5.2.jar")
        CC "AdoScript" FCOPY from:("db:\\stanford-corenlp-3.5.2.jar")
            to:(str_path+"\\lib\\stanford-corenlp-3.5.2.jar")

        CC "AdoScript" DELETE_FILES
            (str_path+"\\lib\\twitter4j-core-4.0.4.jar")
        CC "AdoScript" FCOPY from:("db:\\twitter4j-core-4.0.4.jar")
            to:(str_path+"\\lib\\twitter4j-core-4.0.4.jar")

        SYSTEM ("java -Xmx3072m -jar \""+str_path+"\\social.jar\"
            \""+str_path+"\\SetBO.xml\"") with-console-window
    }
  }


}
```

Listing E.1: *ADOScript for the export and execution*

F Additional Model Classes

Additional descriptions of the classes of the models described in section 3.6.3.

Word-Splitter	Description
	This class is responsible for the splitting of compound words.

Attribute Name	Type	Description
Name	String	The name displayed.
ClassGroup	String	This attribute is used to disambiguate the different groups. Users cannot change this value. Default 'NLP'.

Table F.1: *Class Word-Splitter*

Text Correction	Description
	This class is responsible for replacing incorrect words using the common incorrect word list from wikipedia.

Attribute Name	Type	Description
Name	String	The name displayed.
ClassGroup	String	This attribute is used to disambiguate the different groups. Users cannot change this value. Default 'NLP'.
Word List Path	Longstring	Path to the txt file containing the common incorrect words from wikipedia.

Table F.2: *Class Text Correction*

Stop-Word Remover	Description
	This class is responsible for removing stop words from a message using a list.

Attribute Name	Type	Description
Name	String	The name displayed.
ClassGroup	String	This attribute is used to disambiguate the different groups. Users cannot change this value. Default 'NLP'.
Stop Word List Path	Longstring	Path to the txt file containing the stopwords.

Table F.3: *Class Stop-Word Remover*

TF-IDF	Description
	This class weights each term of the given message by using the TF-IDF algorithm.

Attribute Name	Type	Description
Name	String	The name displayed.
ClassGroup	String	This attribute is used to disambiguate the different groups. Users cannot change this value. Default 'Weight'.

Table F.4: *Class TF-IDF*

ISF	Description
	This class weights sentences of all posts, using the inverse sentence frequency.

Attribute Name	Type	Description
Name	String	The name displayed.
ClassGroup	String	This attribute is used to disambiguate the different groups. Users cannot change this value. Default 'Weight'.

Table F.5: *Class ISF*

XML Publisher	Description	
	This class exports an XML file, containing the results of the categorization structured by the OriginalPostHolder Java class. The result is compatible to be examined with the analysis tool.	
Attribute Name	**Type**	**Description**
Name	String	The name displayed.
ClassGroup	String	This attribute is used to disambiguate the different groups. Users cannot change this value. Default 'Publisher'.
Publish Path	Longstring	Path for the XML file. This path is also used by the ADOScript for export and acts as a working directory.
File Name	String	Name of the file exported.

Table F.6: *Class XML Publisher*

G Additional Class Diagrams

Additional Class Diagrams of the NLP Application described in chapter 3.5.

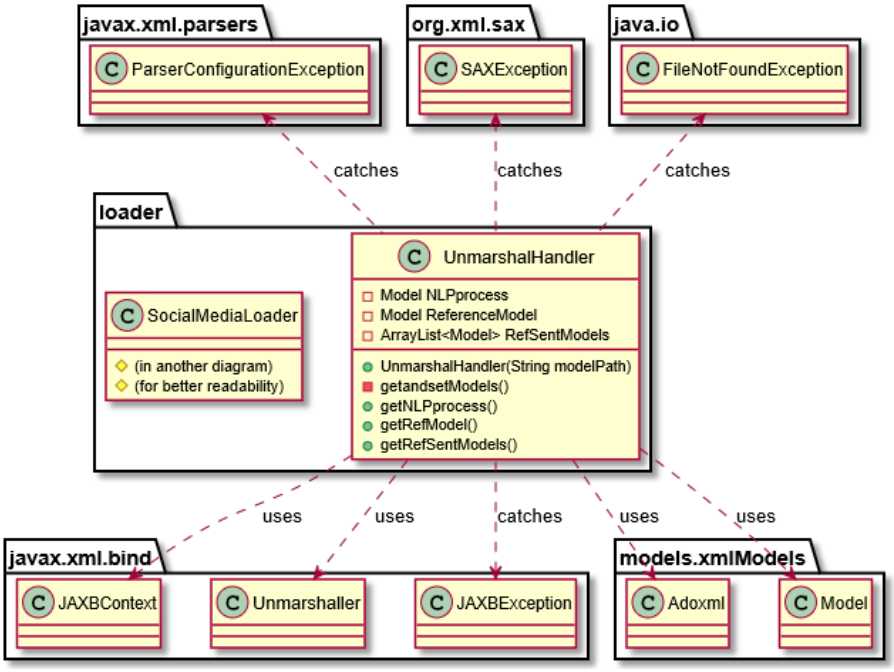

Figure G.1: *Classes of package loader (1/2)*

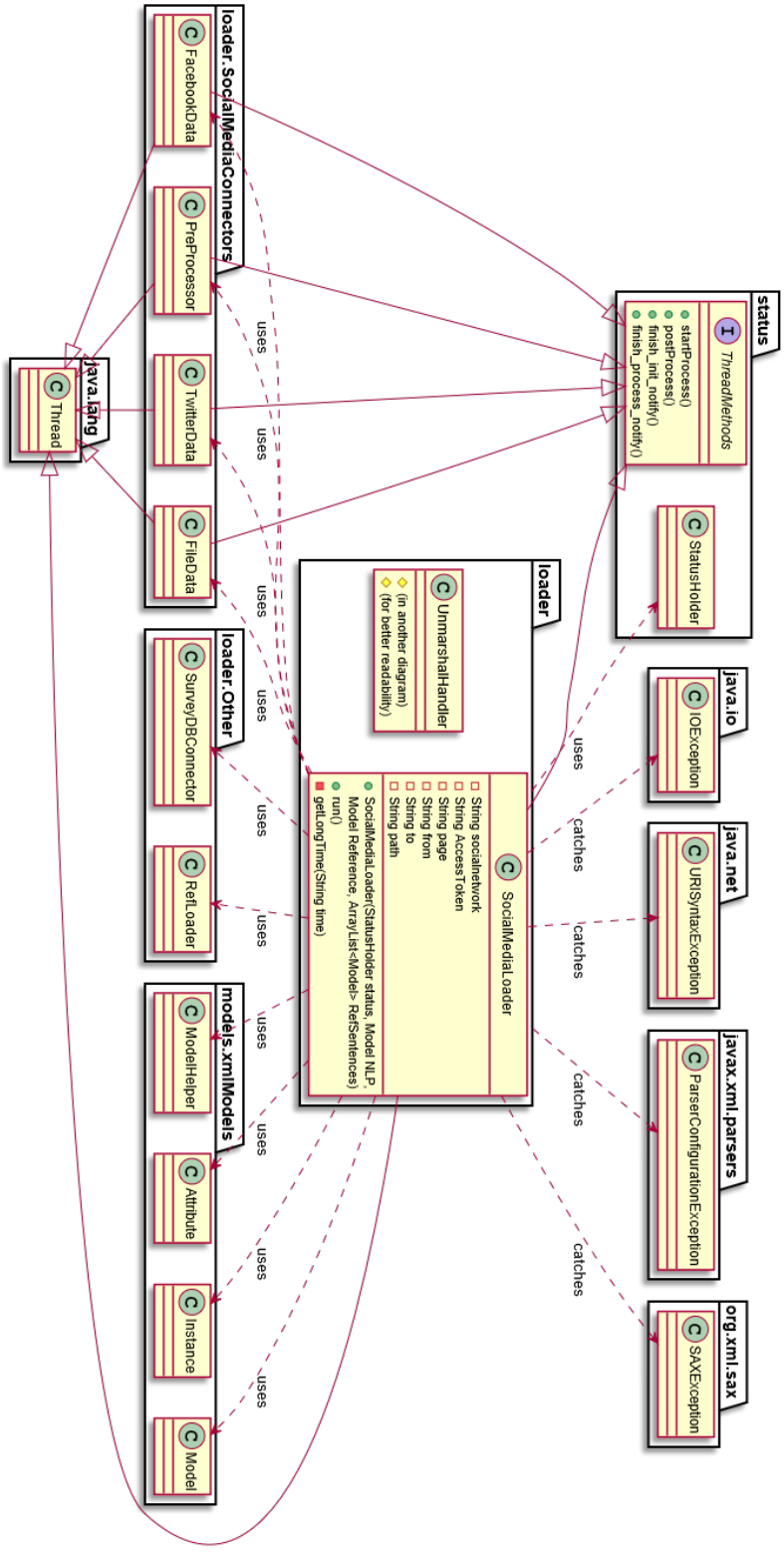

Figure G.2: *Classes of package loader (2/2)*

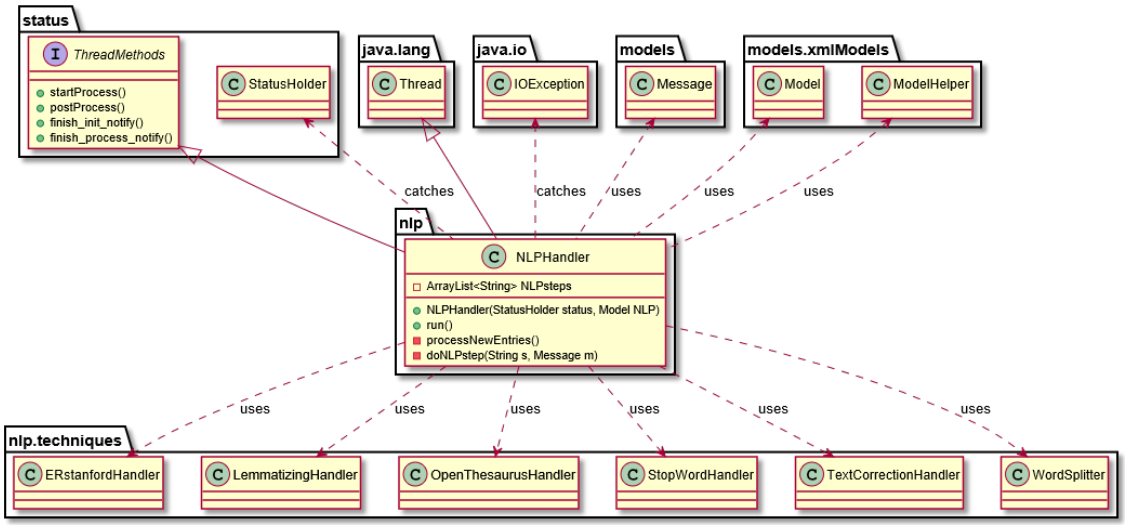

Figure G.3: *Classes of package NLP*

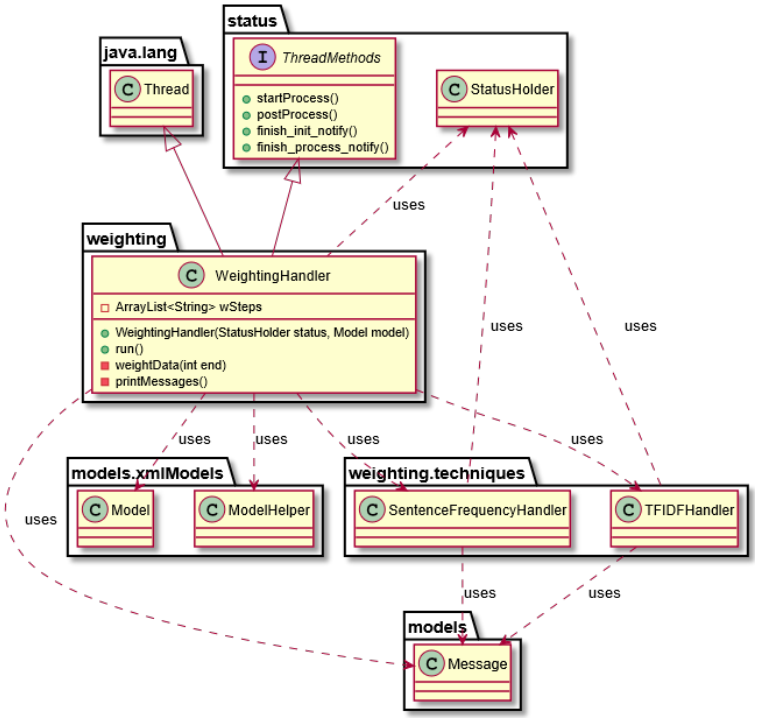

Figure G.4: *Classes of package weighting*

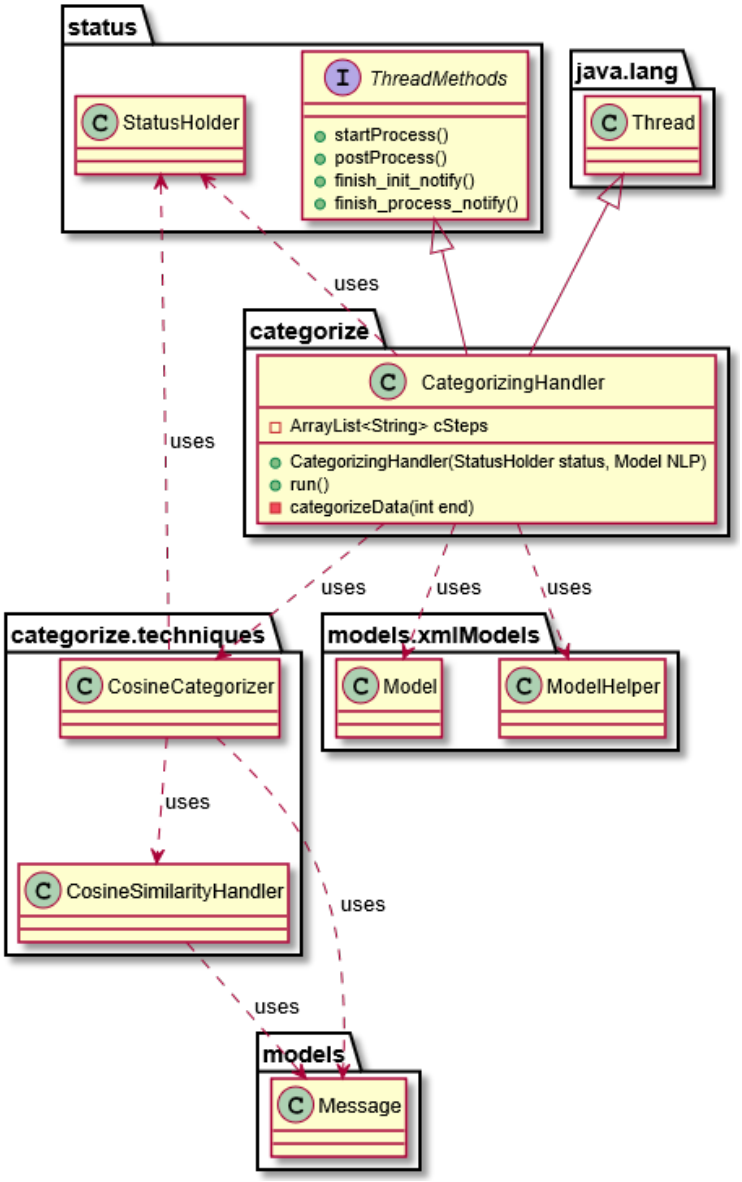

Figure G.5: *Classes of package categorize*

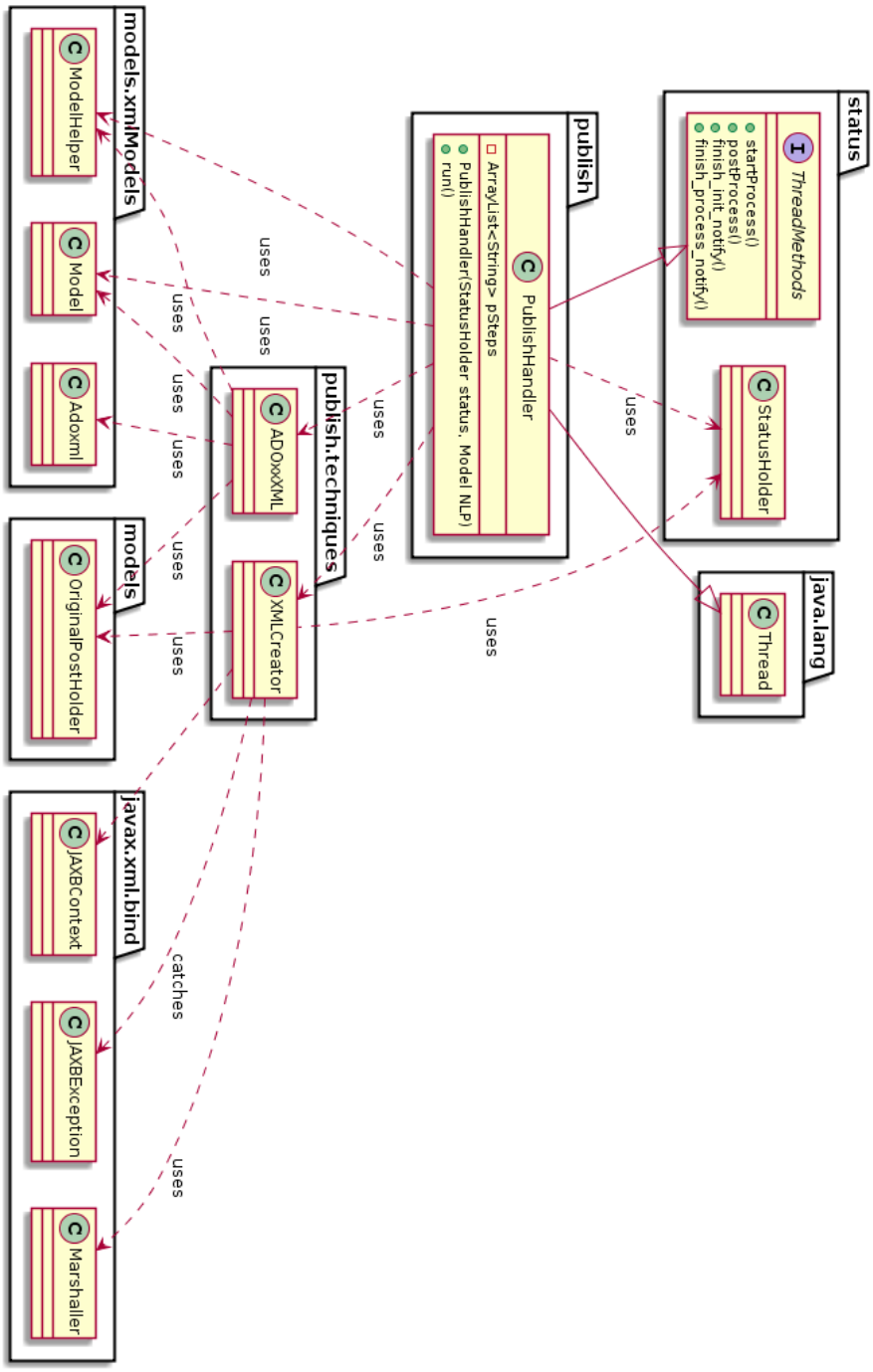

Figure G.6: *Classes of package publish*