Leipzig University
Institute of Computer Science
Digital Humanities, M.Sc.

Quantifying Luhmann: A Semi-Supervised Approach to the Automatic Detection of Social Systems

Master's Thesis

Lukas Gienapp
Born Jun 12, 1997 in Mühlhausen

Matriculation Number 3707883

1. Referee: Jun.-Prof. Dr. Martin Potthast
2. Referee: Jun.-Prof. Dr. Manuel Burghardt

Submission date: February 11, 2022

Abstract

Social systems theory is a sociological framework that differentiates society based on functional aspects: Each subsystem of society, such as *politics, economy,* or *science*, performs a specific and well-delineated function. Primarily developed by German sociologist Niklas Luhmann, it is one of the most influential theories in modern (qualitative) sociology. Yet, it is notorious for its complexity and has rarely been applied in quantitative research thus far. The central question of this thesis therefore is whether it is possible to operationalize Luhmanns' social systems theory using computational methods to automatically recognize and attribute systems to text data. The proposed solution follows a two-step process: first, a set of descriptive terms is obtained from a corpus of Niklas Luhmann's primary literature. Each social system is associated with a particular set of twenty terms that semantically describe the nature of a system. Then, a generative semi-supervised classification method, the Seed-guided Multi Topic Model by Zha & C. Li (2019), is applied to categorize documents according to Luhmann's proposed taxonomy for social systems. A full ablative study investigates the problem at hand on four different datasets. This includes a replication of previous work, a novel scalability assessment, and the detection of social systems in two different setups: in-domain, i.e., the detection of systems in Luhmann's own texts, and cross-domain, i.e., the transfer of the seed term set to detect systems in a subset of the German Wikipedia. While the in-domain setup yields excellent results, both in classification performance and in the usability of the model for downstream tasks, demonstrating the validity and suitability of the method, the results in the cross-domain setup are of mixed quality, suggesting future investigations into the behavior of the model for domain transfer.

Contents

List of Figures

List of Tables

List of Abbreviations

20NG	Twenty Newsgroups
BoW	Bag of Words representation
DC	Dataless Classification (Chang et al., 2008)
ESA	Explicit Semantic Analysis
HITL	Human-in-the-Loop
IG	Information Gain
IR	Information Retrieval
LDA	Latent Dirichlet Allocation (Blei et al., 2001)
MNB	Multinomial Bayes Classifier
NB	Naïve Bayes
NLP	Natural Language Processing
PMI	Pointwise Mutual Information
SMTM	Seed-guided Multi-label Topic Model(Zha & C. Li, 2019)
SOTA	state-of-the-art
STM	Seeded Topic Model (C. Li et al., 2016)
SVC	Support Vector Classification
TFIDF	Term Frequency / Inverse Document Frequency (Spärck-Jones, 1972)

Chapter 1

Introduction

Classification is a core task across nearly all scientific disciplines, and comes in many different types, forms, and purposes. By arranging objects or phenomena into orderly categories, trends can be identified, relevant things can be separated from irrelevant things, groups can be compared, or recommendations can be made.

In the social sciences, especially in the field of sociology, which has a long-standing tradition of quantitative research and statistical methods, one mode of classification finds widespread application: the assignment of observations, for example, survey responses, text transcripts, or cultural artifacts, to a set of categories, which can then be compared with statistical tools. In this way, classification serves to quantify naturally occurring (social) phenomena and is the foundation for operationalizing and addressing a wide range of research questions. In this context, classification is commonly referred to as *coding*, based on a predefined *codebook* of possible categories, and guidelines describing how and when each category (*code*) is applicable. Traditionally, a human assessor is faced with the task of assigning the appropriate categories.

At the same time, given the digital convergence, social communication is increasingly mediated through digital media, and is thus observable and processable computationally. Large text corpora, in particular, have become a valuable resource for sociological analyses because they promise direct and low-barrier access to a wide range of latent variables of interest to sociology (Müller-Hansen et al., 2020; Roberts et al., 2016). This creates new possibilities to analyze societal phenomena quantitatively, which have previously only been addressed through qualitative analysis. But with ever more complex and sizable sets of observations, this task quickly becomes insurmountable through human effort only, necessitating the development of computer-aided approaches to support social science research.

The structural approach to sociological theory, with prominent advocates being sociologists Emil Durkheim, Talcott Parsons, and Niklas Luhmann (Heydebrand, 2001, p. 15230), encompasses theoretical frameworks that directly relate to the notion of categorization: they grasp society by means of differentiation, i.e. they conceptualize it as the sum of different parts, and analyze how these parts constitute themselves, what they consist of, and how they interact. One of the most prominent theories of this kind is the theory of social systems set forth by Niklas Luhmann. According to Luhmann, society consists of different (sub-)systems that are each categorized by their function—for example, *Politics, Art, Education,* or *Economy.*

However, Luhmanns' theory is not only notoriously complex (Albert, 2019, p. 2), but has so far almost exclusively been applied in a qualitative way. Despite the considerable value algorithmic methods for text analysis promise to deliver for practitioners in the field of the social sciences, many analyses base their insight on rather basic methodological approaches (Jacobs & Tschötschel, 2019), be it because of unawareness of other technical possibilities, methodological prejudice, or simply the barrier of entry. This creates issues with regard to scaling, repetition, and systematization of experiments (Grimmer & Stewart, 2013; Jacobs & Tschötschel, 2019). The quantitative analysis of social system theory is no exception here: it has so far been restricted to comparing the relative frequency of keywords describing each system in a large multilingual text corpus (Roth et al., 2017).

Yet, recent advancements in computer science would allow social science research to tackle increasingly complex issues with the help of algorithmic solutions (Roberts et al., 2016). Especially semi-supervised learning gives impetus for promising new applications, circumventing one of the basic problems of computational classification: the need for a large amount of labeled training data, which is often not available to cater to specific research questions in sociology, such as the automatic classification of social systems. Here, seed-guided classification, also referred to as Dataless Classification (DC, Chang et al., 2008), introduces a specific type of weakly supervised learning on text data that learns a classifier without any labeled data. Instead, each class is described by a set of representative seed terms, which are presumed to express the "nature" of the class, i.e. characteristics that a document should exhibit to be deemed part of said class. As Chang et al. argue, this approach incorporates the semantics of the label into the classification process, akin to how a human might categorize text. A human tasked with deciding whether a newspaper article fits the ressort of *politics* does so by relating the semantic meaning encoded in the class name to the content of a to-be-classified document, not by comparing a set of labeled examples to the candidate.

This thesis aims to explore the use of such a seed-guided classification model as a quantitative operationalization of social system theory. By extracting a set of descriptive seed terms directly from relevant social science literature, and subsequently fitting a weakly-supervised classification model using these terms as prior information, for the first time, automatic classification of social systems in text data becomes possible. While promising a scalable, flexible, and accurate quantitative grounding for novel insight in social systems research, this warrants a pilot study to critically evaluate the suitability of the method.

To this end, this thesis sets forth four main contributions: (1) a replication study of previous baseline evaluations of the SMTM, verifying both previous claims on classification performance, and conducting a comparative evaluation of different seed term extraction methods; (2) an investigation of the model's ability to scale to large text corpora, applying it to categorize a large subset of German Wikipedia articles; (3) a dataset of labeled documents for describing each social system mined directly from the publications of Niklas Luhmann, on which an in-domain classification experiment is conducted as first foray into automatic classification of social systems; and finally (4) a cross-domain classification experiment transferring seed terms mined on Luhmannian literature onto the Wikipedia dataset, establishing a first quantitative grounding for social system analysis.

In the following, Chapter 2 reviews related work on social system theory (Section 2.1), topic modeling in general (Section 2.2), and seed-guided classification in particular (Section 2.3). Based on this, Chapter 3 describes in detail the SMTM model (Section 3.2), as well as different methods for seed term mining (Section 3.3). The two datasets that were constructed as part of this thesis, derived from Niklas Luhmanns' monographs and from Wikipedia, are introduced in Chapter 4, and the experimental setup is detailed. Finally, Chapter 5 presents the four aforementioned experiments and discusses the suitability of the SMTM model for social systems classification. Chapter 6 concludes this thesis, and provides outlook on future applications of the method.

Chapter 2

Background

This chapter is divided into three parts: first, Section 2.1 provides background information on the theory of social systems, notably developed by German sociologist Niklas Luhmann. Then, Section 2.2 continues with an introduction to topic modeling, encompassing basic concepts, advantages, and drawbacks of the method, with a focus on applications in social science research. Finally, Section 2.3 gives an overview on the state of the art in semi-supervised, seed-guided classification, reviewing different approaches to both classification and seed term mining.

2.1 Social System Theory

In order to clarify the problem that the system theory approach is trying to address, it is first necessary to briefly outline the fundamental issues that drive the field of sociology. Sociology established itself as an independent area of research only in the middle of the 19th century during which massive changes in the entire society occurred (Saalmann, 2016, p. 14). Most notably, the first and second industrial revolutions took place, leading to urbanization and fundamental changes in education, family, working conditions and political organization. This in turn raised interest in questions regarding the inner workings of society and social structures, which soon led to the development of social theories, theoretical frameworks offering an explanatory approach (ibid., p. 14). In essence, all sociological theories try to address one central question, prominently formulated by Max Weber: "[wie] soziales Handeln deutend [zu] verstehen und [...] in [...] seinem Ablauf und seinen Wirkungen ursächlich [zu] erklären [sei]"(Weber & Winckelmann, 1972, p. 1)—how social action/society is to be understood and explained causatively in its course and effects.

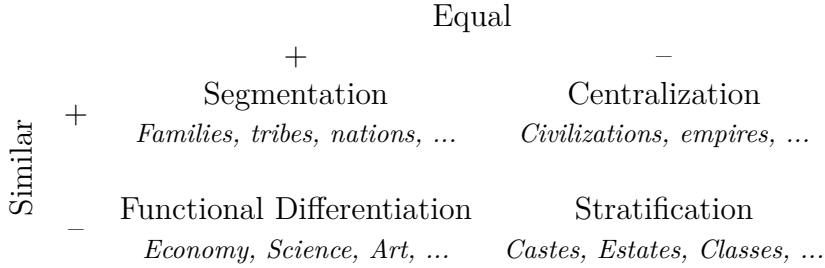


Figure 2.1: Social Differentiation (adapted from Roth, 2015, p. 113)

One particular group of theories approaches this question from a *structural* perspective (Saalmann, 2016, p. 42), with prominent advocates of this approach being sociologists Emil Durkheim, Talcott Parsons, and Niklas Luhmann (Heydebrand, 2001, p. 15230). "Structuralism is an intellectual tendency that seeks to understand and explain social reality in terms of social structures." (ibid., p. 15230). From this perspective, societal processes are characterized primarily as increasing differentiation, a key aspect of Luhmanns' theory of society.[1] Here, differentiation refers to the "intrasystem process of subsystem formation" (Luhmann, 1977), i.e. how a whole thing can be internally divided into parts, or, more specifically: how society can be decomposed into aspects.

To arrive at the differentiation of society that Luhmann ultimately proposes, it is first necessary to trace its historical development (Luhmann, 1987, 161f.) through four possibilities of social differentiation, which are depicted in Figure 2.1 along the two axes of equality (whether there is an inherent order between segmented units) and similarity (whether segmented units share the same characteristics). The path of historic development can be retraced by starting in the upper left cell and continuing clockwise, ultimately arriving at functional differentiation.

Starting with early societies, subsystems were characterized by notions of clans, tribes, and families. This is referred to as differentiation by *segmentation*, and society is divided into subsystems that are similar and equal (ibid., 161f. Roth et al., 2017, p. 3). Over the course of the neolithic revolution, as more advanced civilizations emerged, this division has been replaced by one motivated by *centralization*: first empires develop, and society could be differentiated in the center-periphery spectrum, which forms subsystems that are similar, but

[1] As Albert (2019, p. 3) remarks, while Luhmanns' *Gesellschaftstheorie* is often translated as *social theory*, the translation *theory of society* is more fitting, as Luhmann deliberately breaks with the notion of the *social* in classical sociology, which is often a variant of the theme of community; for Luhmann, differentiation and the resulting systems constitute society, not a shared set of norms and values, which traditionally defines what is social.

unequal. Soon, this was replaced by differentiation by *stratification*, as hierarchical societal systems emerge, like cast systems, or estates of the realm. Here, division is achieved through ranked (sub-)communities, or unequal, dissimilar subsystems (Roth et al., 2017, p. 4). Finally, *functional* social differentiation emerged in the modern age from its historically antecedent forms (Holmström, 2007, p. 257; Albert, 2019, p. 4,). Here, each system is defined by its function to society. For example, the system of *science* has the function of observation, the system of *economy* has the function of distribution, the system of *politics* has the function of exerting control. These functional systems are dissimilar, but equal.

Functional differentiation divides society into units based on function, by applying a so-called binary *code* to all inter-societal communication. For example, all communication that applies the code *true/untrue* is placed within the system of *science*; all communication that applies the code *innovative/imitative* is placed within the system of *art*. Each function system applies one single code only and exclusively. However, one act of communication might be alluding to (or categorized as belonging to) different systems at the same time (Roth & Schütz, 2015, p. 17). Besides a code and function, systems encompass a *medium*, i.e. the subject of communication, the primary method through which the system is communicated, and a program, deciding which of the two binary alternatives of the code is applicable to the communication at hand (Albert, 2019, p. 8). Within the aforementioned example, the system of *science* communicates about *truth* (its medium), and scientific *theory* (its program) dictates *truth* (its code). Similarly, the system of *art* is about *style* (its medium), and *fashion* (its program) distinguishes between *innovative* and *imitative* (its code, Roth & Schütz, 2015, p. 25; Roth et al., 2017, p. 5).

A second way of characterizing the system-constituting process is through their effect of reducing complexity, which is the key difference between a system and its environment: "[A] system is always less complex than its environment." (Ritzer, 2010, p. 335). Modern supply chain management could be used as a concrete example for this (see also ibid., p. 335): sourcing raw materials is reduced to information about their price and quality. All other information (such as the political situation in countries materials are sourced from) is ignored to reduce the complexity of the process, and becomes part of the environment; the supply chain as its own system thus constitutes a simplified view, i.e. component or sub-system, of society.

Reducing complexity induces decision processes (Holmström, 2007, p. 258). This increases uncertainty and contingency (ibid., p. 258; Ritzer, 2010, p. 335): "Contingency means that being depends on selection which, in turn, implies the possibility of not being and the being of other possibilities." (Luhmann, 1976, p. 509). To manage contingency, "systems develop new subsystems [...] in order to deal effectively with their environment" (Ritzer, 2010, p. 335).

Table 2.1: Overview social systems (adapted from Roth & Schütz, 2015, p. 24).

System	Code	Medium	Program	Function
Political	Government/Opposition	Power	Ideology	Limitation
Economy	Non-/Payment	Money	Price	Distribution
Science	Un-/True	Truth	Theory	Verification
Art	Innovative/Imitative	Style	Fashion	Creation
Religion	Immanent/Transcendent	Faith	Confession	Revelation
Legal	Un-/Lawful	Norm	Law	Standardization
Sport	Success/Failure	Achievement	Goal	Mobilization
Health	Ill/Healthy	Illness	Diagnosis	Restoration
Education	Un-/Placeable	Vita	Curriculum	Formation
Mass Media	Non-/Informative	Medium	Topic	Multiplication

Returning to the supply chain example from before, this could mean that a company creates a department to monitor the political situation, in order to include it in their decision making processes.

Luhmann further notices the problem of *double contingency* (Luhmann, 1995, p. 103): as social systems are established through communication, "every communication must take into consideration the way that it is received. But we also know that the way it is received will depend on the receiver's estimation of the communication" (Ritzer, 2010, p. 339). This forms a feedback loop: what the receiver understands depends on the communicator, who in turn depends on his expectation of the receiver (Vanderstraeten, 2002, p. 84). Social systems address this problem by defining a common reference frame for both parties, encoding societal expectations, norms, and values to ease communication taking place within the system (Ritzer, 2010, p. 340).

A third component to distinguish systems is *autopoiesis*: within Luhmanns' theory, social systems are *autopoietic*, i.e. "they produce all their elements within themselves" (Albert, 2019, p. 5). For example, for the economic system, its basic element (and the medium it is communicated by) is money. Yet, the system itself defines what is accepted as money, what money is worth, and what money can be used for (Ritzer, 2010, p. 336). "Within its boundaries, an autopoietic system produces its own structures. [...] [within the economic system] banks are established to store and lend money" (ibid., p. 336). Furthermore, autopoietic systems are self-referential and closed. This means that (1) they use themselves as a reference frame (the legal system is about laws, that refer to how laws can be enacted, applied, and interpreted; cf. ibid., p. 337), and (2) "there is no direct connection between a system and its environment" (ibid., p. 337), for example, the economic system only responds to needs of the surrounding societal environment to the extent they can be expressed in terms of money (the medium of the economic system).

To summarize, the notion of social systems as characterized by Luhmanns' system theory is generally derived by function; systems serve to simplify communication between different aspects (sub-systems) of society, and solve the double contingency problem within these communication processes. Furthermore, systems are autopoietic and thus all elements needed for their existence emerge from within themselves. While different sources provide different numbers and kinds of systems within society that fulfill these criteria, a general consensus of ten different functional systems exists (Roth & Schütz, 2015). Table 2.1 provides an overview on those ten. For each, the code they divide communication with, the medium they communicate through, the program they communicate by, and the societal function they serve is listed.

While Luhmanns' theory of social systems is notoriously complex (Albert, 2019, p. 2), it is also primarily applied in qualitative research as theoretical framework in which societal phenomena can be (qualitatively) analyzed and interpreted. Approaches trying to combine this framework with quantitative methods remain scarce. Roth et al. (2017) uses the Google n-gram corpus (Michel et al., 2011) to track the importance of systems over time, by comparing the relative frequency of system names as well as related terms occurring (i.e. the word "science" for the scientific system). Their analysis is repeated across different languages (German, English, French).

2.2 Topic Modeling

One specific method of computational text analysis that has gained widespread adoption in research is topic modeling: the decomposition of texts into *topics* – common themes, frames, or concepts, appearing as language patterns throughout and across the texts contained in a corpus. It aims to make the complexity of large text corpora manageable by representing each included text as a combination of such topics. It thus naturally lends itself to *coding*, i.e. classification in the social sciences, as its basic building block of identifying recurring patterns across many (textual) observations is inductive and theory-agnostic in nature and therefore can be applied to quantify text within many sociological theoretical frameworks (Jacobs & Tschötschel, 2019).

2.2.1 Generative Text Model

Conceptually, a topic model takes a term-document matrix, denoting which terms occur in which documents, and how often, as input. The algorithmic task to be solved is to create a decomposition of this matrix into two latent probability matrices, which constitute the output of a topic model: (1)

Algorithm 1: Generative process of the Latent Dirichlet Allocation (LDA, Blei et al., 2001) algorithm (Blei et al., 2001).

1 **foreach** $d \in D$ **do**
2 choose $N \sim \text{Poisson}(\epsilon)$, the number of terms in the document
3 choose $\theta \sim \text{Dir}(\alpha)$, the topic distribution of the document
4 **foreach** $t_n, n \in \{1, ..., N\}$*:* **do**
5 Choose a topic $z_n \sim \text{Multinomial}(\theta)$
6 Choose a term t_n from $p(t_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n
7 **end**
8 **end**

a document-topic probability matrix, which, for each document, indicates the relative likelihood of each topic being observed for this document. (2) a term-topic probability matrix, which, for each term, indicates the relative likelihood of this term being observed under the given topic.

The most widely used approach to solve for this decomposition is LDA. It is a generative text model, and can be best explained by following the generative process for a document, as each document in the to-be-analyzed collection is treated as the result of such. Given the two aforementioned latent probability distributions, a document is generated in three steps, depicted in Algorithm 1: (1) the number of terms in the document is sampled, dictating the length of the generated text (line 2); (2) a topic distribution is sampled, dictating the themes occurring in the generated text (3); (3) for position in the text, a term is sampled based on the documents topic distribution chosen, and the topic-word prior distribution (lines 5 and 6). With LDA, this generative process can be reversed. Given the observed documents, the two latent probability distributions that dictated the generation of the texts are estimated. This allows to identify trends and common themes in given text data with a conceptually simple and traceable manner, which gives rise to the popularity of the method in digital humanities studies.

2.2.2 Applications in Social Sciences

Roberts et al. (2014) showcase the use of a probabilistic topic model to analyze open-ended (i.e. free-text) survey responses with reduced human intervention. A topic model is used to quantify common themes among all responses to an open-ended question. These themes are then examined, interpreted and used to draw conclusions with respect to predetermined covariates (for example, political orientation of the respondent). A novel contribution is their *Structural*

Topic Model, which can adjust inference to such covariates for higher reliability. Maier et al. (2018) present an in-depth review of the application of standard LDA topic models in communications research, encompassing 20 publications ranging from the analysis of political speeches, over newspaper articles and tweets to public comments and blog posts. They place special emphasis on the preprocessing steps taken, parameter selection, as well as checks for interpretability, validity and reliability in each publication to develop a common methodological guideline for the application of LDA topic models in communications research.

Jacobs & Tschötschel (2019) conduct two case studies to showcase the feasibility of topic modeling as a tool for discourse analysis. First, they analyze 11 744 political speeches in the European Parliament from 1999 to 2016 to gain insight on the extent of hegemony within trade policy-making. They apply standard LDA with 120 topics and conduct a qualitative interpretation of the resulting topics. In their second case study, they analyze Austrian news paper articles with regard to the concept of economic growth and economic crises with a similar approach. R. Huang (2019) analyze 51 288 postings from the Chinese social media platform Weibo using LDA in conjunction with social network analysis techniques to reveal the link between the structure and the focal topics of communities with regards to labor issues.

2.2.3 Limitations

Jacobs & Tschötschel (2019) conduct a meta-analysis on the feasibility and practicality of utilizing topic models in discourse analysis. They remark that topic modeling is agnostic to the underlying theoretical framework it is employed to operationalize, and therefore exhibits high methodological polyvalence. Another benefit is the context-dependency of word semantics that is reflected in topic models: as LDA derives topics from word cooccurrences, it is equipped to handle polysemy, and does not impose a precognition on the language use present in the corpus. Another benefit of this inductive process is the feedback the method gives to the researcher: instead of imposing a predefined categories on text, as would be the case for supervised classification approaches, the unsupervised nature of topic modeling allows to refine the categories during interpretation, dependent on what patterns actually occur in the target data.

In this regard, Roberts et al. (2014) note, too, that a key benefit of topic models is their explorative nature: in standard social science coding practices, a predefined set of labels is assumed to describe survey responses, which may or may not fit the actual obtained data. Here, topic models can be helpful, as they model the latent information in the text directly, and in a case of

mismatch between prior expectation and data, the theoretical framework or the coding practices can be revised. Wesslen (2018) highlights the intuitive interpretation and possibilities for visual interfaces to topic models.

Ramage et al. (2009) identify trust in the output of computational methods as a main problem for the application of such techniques for support in social science research. For (unsupervised) topic-model-based approaches, three challenges are apparent: (1) *characterization* of topics: latent topics have no inherent canonical descriptions, and characterizing a topic by its word distribution after inference is prone to poorly representing the actual meaning of a topic. Single-word descriptions ("names") are problematic in particular. (2) *meaning* of topics: comparing the output of topic models trained on different data (subsets) can be misleading, as the topic context and meaning may change. This further aggravates the characterization problem. (3) *trustworthiness* of topics: a models output must be judged by human assessors working in the intended domain, to gauge the trustworthiness and usefulness of a model for social science research. This places additional emphasis on result presentation and accessibility, as tools for interactive exploration of the topic space are deemed an important aid in model interpretability.

Wesslen (2018) identifies two technical problems that hinder the widespread use of topic-model-based approaches in social sciences: first, the word and topic probabilities inferred by the generative process are point probabilities that lack the confidence intervals required for statistical testing. Second, topic models are multi-modal since topic model inference is an NP-hard problem. Thus, a solution cannot be deemed globally optimal and depends on the model initialization. As this can lead to different results being generated from the same data, the legitimacy of such results is questionable. A similar technical problem is remarked by Jacobs & Tschötschel (2019), who identify the number of topics the model is initialized with as drawback, as with most traditional approaches to topic modeling, this is an a-priori parameter chosen by the practitioner, which may or may not align with the actual number of topics present in the corpus. They further note that topic modeling works best in a setting with well-delineated topics and text data from a coherent domain.

2.3 Seed-guided Classification

Seed-guided classification is a specific type of weakly supervised learning on text data that learns a classifier without any labeled data. Instead, each class is described by a set of representative seed terms, which are presumed to express the "nature" of the class, i.e. characteristics that a document should exhibit to be deemed part of said class. As Chang et al. (2008) argue, in this approach,

the meaning of the label is incorporated into the classification process, similar to how a human would categorize a text. A human who is to decide whether a newspaper article fits into the politics section relates the meaning encoded in the class name to the content of a document to be classified, rather than comparing a set of labeled examples to the candidate. More formally, the problem can be described as follows: given an unlabeled document set D and a set of classes C, assign each document $d \in D$ to one more more classes $c \in C$ using only the set of descriptive seed terms S_c associated with each class as prior information.

2.3.1 Types of Seed-guided Classification

Zha & C. Li (2019) differentiate three types of seed-guided classification: (1) *classification-based* approaches focus on deriving pseudo-labels for documents based on the given seed terms and apply a supervised classification algorithm using the bootstrapped training data. In the following, this is referred to as *pseudo-label approaches* to avoid unclear usage of the term *classification*. (2) *semantic* approaches aim to embed a classes' seed term representations and the unlabeled documents into a shared vector space. Classification is then achieved by nearest-neighbor search among the classes. (3) *probabilistic* approaches utilize probabilistic models to derive latent classes in text data in a generative fashion.

Pseudo-label Approaches. One of the first approaches for seed-guided classification was proposed by Liu et al. (2004). They aim to bootstrap a set of pseudo-labels from D to apply established supervised classification methods to. A representative document is built for each class in C, consisting of its seed words. This allows to compare each document in D to the representative document of a class using the cosine similarity of both document representations. A document is then assigned to the class it is most similar to. The most similar documents for each class are retained as pseudo labels for a NB classifier. Ko & Seo (2004) bootstrap document labels for training a supervised classifier from category labels. First, term co-occurrence with the category label is used to build an extended set of descriptive seed terms, forming clusters. A NB classifier is then trained on these term sets to extend the classification to the document level.

Semantic Approaches. For the group of semantic approaches, methods based on Explicit Semantic Analysis (ESA) are widespread: Chang et al. (2008) use nearest-neighbor classification in an ESA vector space derived from Wikipedia, calculating the distance between the document (centroid) and each cat-

egory label. One upside of the vector-space based approach is that they can be extended to multi-lingual classification problems, as long as a shared vector space can be constructed (Song et al., 2016). While inferior to modern probabilistic approaches, given recent advances in phrase embedding techniques (Cer et al., 2018; Reimers & Gurevych, 2019), revisiting this method might be worthwhile.

Probabilistic Approaches. Most of the probabilistic approaches to seed-guided classification extend unsupervised topic modeling schemes such as LDA to incorporate prior information, steering the topic inference into the desired direction. Hingmire & Chakraborti (2014) & Hingmire et al. (2013) first infer latent topics using unsupervised LDA. Then, these topics are labeled by a human assessor and training is continued, incorporating the new label information to collapse topics with the same label into one category. To perform classification without human intervention, X. Li et al. (2018) propose the Seeded Topic Model (STM, C. Li et al., 2016), that adopts Generalized Pólya Urn sampling instead of the standard Gibbs-sampling based inference procedure to incorporate seed term supervision for a single-membership classification task. Zha & C. Li (2019) extend the STM model to mixed-membership classification.

2.3.2 Seed Term Mining

The mining of seed terms (also referred to as keywords) aims to extract a lexicon of descriptive terms to aid in a classification problem. Early approaches utilize such methods to build lexicons for tasks like sentiment analysis or opinion mining (Hai et al., 2012; Hu & Liu, 2004; Remus et al., 2010). These lexicons were then often used to directly infer the classification of a document based on term occurrences, and thus needed to be quite large. More recent approaches, notably generative seed-guided classification, have considerably reduced the required size of seed term sets, by expanding the semantic information encoded in such a set through a generative process. Two processes of mining such terms can be discerned: (1) fully automated approaches, requiring no interaction apart from initialization; and (2) human-in-the-loop approaches, where a human judge is part of the seed selection process.

Independent of the method seed terms are derived with, Jin et al. (2020) set forth three criteria for a high-quality seed term set: (1) seed terms should be representative for their associated categories; (2) seed terms should not be rare words; (3) seed terms for different categories should have little to no overlap. These criteria characterize two underlying properties of seed terms: they should be *descriptive* of the category they are attributed to, and *contrastive* to all other categories in the category set.

Automated Approaches. The fully automated approach, which performs seed term selection without human intervention, can be characterized similarly to the problem of keyword extraction in natural language processing (NLP): construct a set of words which reflect the topics and information contained a document. However, the notable difference between both concepts is the level of granularity: keyword extraction usually happens at the document-level, i.e., for each document, such a set of descriptive terms is found; seed term extraction happens at the level of classes, i.e. a common set of terms is found describing a coherent group of documents. Yet, a large body of work in keyword extraction can be applied to seed term selection as well, primarily stemming from two categories of methodological approaches (C.-C. Huang et al., 2015; Zhang et al., 2020): (1) graph-based methods, identifying important terms by imposing a graph structure on the words in a text, for example by analyzing the cooccurrence of terms; and (2) statistical methods, identifying important terms based on their statistical properties.

Human-in-the-Loop Approaches. In this type of approach, a human is part of the selection process, either with, or without the help of a computer. For the former, domain experts are tasked with building a seed term set based on their intrinsic knowledge. In the latter hybrid process, a computer performs an initial selection based on a (focused) collection of relevant documents, which is then refined by the human judge. For example, Liu et al. (2004) extract seeds words from a document set D with entropy-based feature selection, clustering D using an unsupervised clustering algorithm (k-means) and ranking all words in D by their discriminative power. The human judge is then tasked to assign each of these words to a class. In practice, human-in-the-loop approaches remain scarce, as human interference is costly and complicated. Yet, for social science research in particular, having a domain expert judge incorporated in hybrid process is paramount for a high-quality classification result.

Ko & Seo (2004) start out with a single word (title) for each category and automatically derive a set of descriptive seed words based on co-occurrence patterns. Using term occurrence vectors over the document set D, they select seed words by calculating the cosine similarity between each term and title vector, selecting those terms as seed words for a category that exhibit a high similarity with its title, but a low similarity to other categories' title vectors. Similarly, Jin et al. (2021) also base their seed word selection on the category title only, yet propose a two-step process: first, a noisy set of candidate words is mined by calculating the Pointwise Mutual Information (PMI) for every category title and term over the document set D; then, after selecting the top-scoring terms associated with each title, the seed set is refined by training interim classifiers to gauge the impact of each individual seed term, and subsequently select a final set of terms which yields the lowest model error.

When supervised data is available to bootstrap seed terms from, other automated methods become possible: Jin et al. (2020) utilize a PMI-based measure to select highly descriptive terms for each label. They propose adjusting the PMI score using the term frequencies, to not only mine words that are semantically appropriate to represent a category, but also have the necessary frequency and coverage to successfully serve as prior for seed-guided classification. In a similar information-theoretic approach, Jagarlamudi et al. (2012) use the Information Gain (IG) of each term-category combination to identify seed terms. Additional measures for classic keyword extraction such as TF/IDF (Spärck-Jones, 1972) or term saliency (Chuang et al., 2012) are applicable in the case of supervised training data as well.

2.3.3 Domain Adaption

For any classification problem, an unknown classification function h has to be found that translates from an input space (training samples represented by features) to an output space (class labels). This classification function is attentive to the probability distributions of features in the training samples (Kouw & Loog, 2019). *Domain adaption* is defined as a change in classification setting, where input space and output space remain the same, while the probability distributions of features change – a phenomenon also known as 'covariate shift' (ibid.). Consider a generic text classification problem in NLP: the input space are documents encoded in a bag-of-words (BoW) representation. The output space are labels, for example positive (1), neutral (0) and negative (-1) sentiment. One particularly widespread training source for sentiment classification are movie reviews (Maas et al., 2011). As such, a sentiment classifier trained on this corpus will be attentive to the particular feature (i.e. term) frequencies occurring in the text domain of movie reviews. However, using an example given by Blitzer et al. (2007), the word 'useful' will probably not occur often to denote positive sentiment in this particular setting. Yet, when the classification context (i.e. text domain changes) to reviews of kitchen appliances, the word 'useful' becomes much more frequent in positive descriptions of items: the underlying feature distributions have changed, degrading the performance of the classifier when it is not adapted to the new domain.

Jin et al. (2020) have proposed the use of dataless classification algorithms to tackle the issue of domain adaption: first, a set of seed terms is learned from a labeled source domain; this seed term set is then used to initialize a dataless classifier in the target domain. The idea is that, if the seed term set contains general terms that are descriptive of a class in both domains, the dataless classifier will learn to be attentive to new features in the target domain co-occurring with the transferred general features, achieving domain adaption.

Returning to the example of sentiment classification adapting from movie reviews to kitchen appliance reviews, the seed term set for positive sentiment learned on the source domain might include the terms *good, intriguing*, and *perfect*. In the target domain, the terms *good, perfect* are descriptive of positive sentiment too, which allows the dataless classifier to perform domain adaption. It picks up on *useful* being descriptive of positive sentiment as well (since it co-occurs frequently with *good* and *perfect*), and is less attentive on *intriguing* (which likely exhibits a much lower frequency in the target domain).

Jin et al. (2020) demonstrate that dataless classification, specifically the method of C. Li et al. (2016), exhibits a vastly superior classification performance in a domain adaption setting, compared to supervised learning methods such as support vector classification (SVC) or multinomial bayes classification (MNB). They train (or extract seed terms, respectively) the classifiers on a source dataset of crawled newspaper articles, labeled with their respective article categories. The classifiers are then evaluated both on the source and on a target dataset, which also consists out of newspapers articles, yet is crawled 4 years later, such that a shift in vocabulary can be presumed and domain adaption becomes necessary. While supervised methods outperform dataless ones on the source data test set, the inverse is true for the target data test set: here, dataless classification performs similar to, and sometimes better than itself the original data, yet the performance of supervised classifiers degrades heavily down to half of their original scores. Thus, the dataless methods outperforms the supervised classifiers both in absolute scores in the target domain, as well as stability across domains.

Chapter 3

Methodological Approach

This chapter describes the methodological approach to seed-guided classification used in this thesis. It is structured as follows: Section 3.1 formally defines the to-be-solved problem, Section 3.2 introduces the Seed-guided Multi-label Topic Model (SMTM, Zha & C. Li, 2019) model by Zha & C. Li (ibid.) as one of the state-of-the-art solutions to the task. Finally, Section 3.3 describes different methods of deriving a seed term set from labelled data, most notably that of Jin et al. (2020), which has previously been demonstrated to be one of the best-performing approaches.

3.1 Problem Statement

Let D be a set of size $|D|$, and d a document in D composed of terms t from a vocabulary T of size $|T|$. Further, let C be in a predetermined set of categories c, where each category is characterized by a set of seed terms $\mathbb{S}_c \subseteq T$, encoding prior knowledge about the content of each category. The goal is to (1) assign each document $d \in D$ to one or more categories $c \in C$ such that the assignment corresponds to the prior information given by $\mathbb{S}_c$ sufficiently well; and (2) assign each term $t \in T$ a category probability, such that the assignment corresponds to the prior information given by $\mathbb{S}_c$ sufficiently well. These two requirements can be interpreted as decomposing the document-term matrix $D \times T$ into a document-category matrix $D \times C$ and a category-term matrix $C \times T$, which together optimally model the latent category space given the prior information encoded in $\mathbb{S}$. Thus, two components are needed: an algorithmic approach that is able to accurately perform such a decomposition, and a suitable method of constructing a seed term set $\mathbb{S}_c$ for every category.

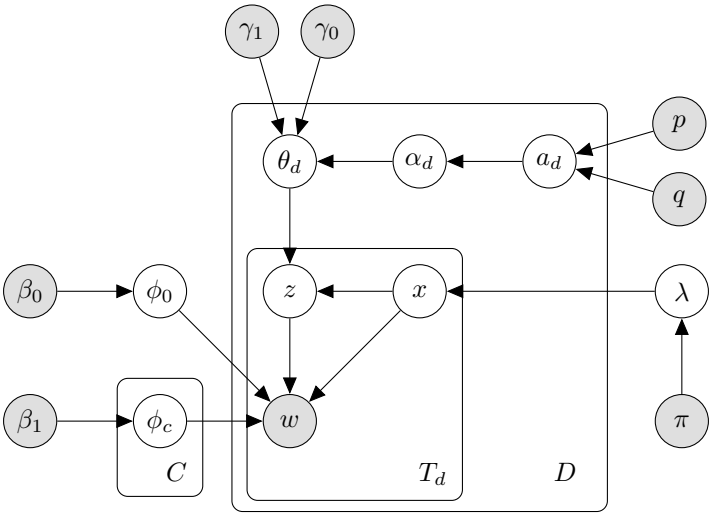


Figure 3.1: Plate notation for the SMTM model, taken from Zha & C. Li (2019).

3.2 Seed-guided Multi-Topic Model

The SMTM model, proposed by Zha & C. Li (2019) is a multi-class extension to the STM introduced by C. Li et al. (2016). It encompasses two kinds of topics: *category topics*, in one-to-one correspondence to the desired categories of the underlying classification problem, and a general *background topic*, capturing general semantic information in the to-be-classified texts that are not indicative of a certain category. Each term in each document can either be associated with one category topic, or the background topic. In this way, both document-category and term-category probability distributions can be estimated through the fitting process and used for downstream tasks, most prominently text classification. This section formally defines the model: first, the generative process is described, then the concept of supervision for topic model inference is introduced, and finally, the model inference and prediction process is detailed. All equations, unless otherwise noted, are taken from Zha & C. Li (2019), but were adapted to be consistent with the nomenclature.

3.2.1 Generative Process

The basic idea of generative text modeling across all topic models, such as LDA, is also present in SMTM: each topic has a term distribution, and each document has a topic distribution. Each term in a document is then the result of combining these distributions in a generative manner. However, two auxiliary variables are introduced allowing the SMTM model to differentiate *category* topics from the *background* topic.

First, at the term level, a binary indicator variable $x_{d,t}$ for each term t and document d is set equal to 1 if the term is associated with a category topic, and equal to 0 if associated with the background topic. This binary indicator is the result of a Bernoulli process with parameter λ, that is globally uniform and derived from a Beta prior, and can thus be iteratively estimated. Together with a selector variable $z_{d,t}$, which denotes which category c is currently associated with t, these variables jointly model if, and to which category each term in each document belongs. Secondly, at the document level, an auxiliary Bernoulli variable $\alpha_{d,c}$ is introduced, equal to 1 if category c is 'selected' for the document, and 0 otherwise. The Bernoulli distribution of $\alpha_{d,c}$ is sampled from is dependent on a document-specific parameter a_d, which depends on a Beta prior with hyperparameters p and q, and can thus be estimated according to a documents' contents. Given these selection mechanisms, together with the classic Dirichlet prior term distributions, each with its respective parameters, the generative process for a text as assumed by the SMTM model is depicted in Figure 3.1 and can be summarized as follows:

1. Sample a background word distribution $\phi_0 \sim Dirichlet(\beta_0)$
2. Sample $\lambda \sim Beta(\pi)$
3. For each category $c \in C$:
   (a) Sample a word distribution $\phi_c \sim Dirichlet(\beta_1)$
4. For each document $d \in D$:
   (a) Sample $a_d \sim Beta(p, q)$
   (b) For each category $c \in C$:
      i. Sample selector $\alpha_{d,c} \sim Bernoulli(a_d)$
   (c) Determine selected category set $A_d = \{k : \alpha_{d,k} = 1\}$
   (d) Sample category distribution $\theta_d \sim Dirichlet(\gamma_0 \alpha_d + \gamma_1 \mathbf{1})$
   (e) For each position $i \in \{1, ..., |d|\}$:
      i. Sample $x_{d,t} \sim Bernoulli(\lambda)$
      ii. If $x_{d,t} = 0$:
         - Sample $w_{d,i} \sim Multinomial(\phi_0)$
      iii. If $x_{d,t} = 1$:
         - Sample $z_{d,i} \sim Multinomial(\{\theta_{d,k} : k \in A_d\})$
         - Sample $w_{d,i} \sim Multinomial(\phi_{z_{d,i}})$

3.2.2 Incorporating Supervision

The model as devised so far is unsupervised: its convergence is purely dictated by the word occurrences in the document collection. To incorporate the prior information given by the seed terms, and thus influence the model convergence to produce the desired classification result, Zha & C. Li (2019) propose a guided Gibbs-sampling-like procedure based on the Generalized Pólya urn model. In normal Gibbs sampling, in each iteration, the counts for every observation are subject to uniform modification, i.e., every count is increased or decreased by the same score (=1) every time. By applying the Generalized Pólya urn model, the promotion can be biased, as count modification is not assumed to be uniform anymore. This allows to influence the inference process by incorporating supervision based on seed term occurrence.

This can be achieved at two levels: document level and term level. At the document level, individual promotion scores can be chosen in such a way that a document shows greater affinity towards a category of which it contains seed terms, dictated by a $D \times C$ matrix that indicates the category promotion associated with each document/category combination. Similarly, at the term level, a term shows greater affinity towards a category with seed terms it co-occurs with, dictated by a $T \times C$ matrix that indicates the category promotion scores associated with each term. Both promotion matrices are static and can be precomputed for efficient inference.

Supervision at Document Level

Supervision is integrated at the document level by constructing a category-document scoring matrix P, where the value $P_{c,d}$ for category c and document d is estimated based on seed terms occurring in the document. First, $u_{c,d}$ is calculated, denoting the importance of seed term occurrence in documents:

$$u_{c,d} = \begin{cases} 1 & \mathbb{S}_c \cap \{t \in T | t \in d\} \neq \emptyset \\ \mu & \text{otherwise.} \end{cases}, \mu \in [0, 1]. \tag{3.1}$$

Here, μ is a parameter to weigh the importance of observing seed terms. If μ is small, seed term occurrences are valued highly, if μ is large, seed term occurrences are less important for promotion. $u_{c,d}$ is then normalized across all categories for each document and multiplied by the number of categories:

$$P_{c,d} = \frac{u_{c,d}}{\sum_{c'} u_{c',d}} \cdot |C|. \tag{3.2}$$

Supervision at Term Level

Similarly to the document level, supervision is integrated at the term level by constructing a category-term scoring matrix Q, where the value $Q_{c,t}$ for category c and term t is estimated based on cooccurrence with seed terms, as formerly proposed by C. Li et al. (2016). First, $u_{c,t}$ is calculated, which is the average probability of cooccurrence for term t across all seed terms $s \in \mathbb{S}_c$ for category c:

$$u_{c,t} = \frac{1}{|\mathbb{S}_c|} \sum_{s \in \mathbb{S}_c} p(t|s) = \frac{1}{|\mathbb{S}_c|} \sum_{s \in \mathbb{S}_c} \frac{df(t,s)}{df(s)}. \tag{3.3}$$

Here, $df(s)$ denotes the number of documents containing seed term s, and $df(t,s)$ denotes the number of documents containing both term t and seed term s. This average is then normalized across all categories for a term to obtain $\tilde{u}_{c,t}$, with ϵ being a very small value[1] to avoid 0:

$$\tilde{u}_{c,t} = \max\left(\frac{u_{c,t}}{\sum_{c'} u_{c',t}}, \epsilon\right). \tag{3.4}$$

Finally, $Q_{c,t}$ is obtained by normalizing $\tilde{u}_{c,t}$ across all terms for a category, and multiplying by the number of terms in the vocabulary:

$$Q_{c,t} = \frac{\tilde{u}_{c,t}}{\sum_{t'} u_{c,t'}} \cdot |T|. \tag{3.5}$$

3.2.3 Inference

The model parameters are inferred iteratively. In every iteration, first $x_{d,t}$ and $z_{d,t}$ are sampled for each document d and term t:

$$P(z_{d,t}, x_{d,t}|t, z_{\neg d,t}, x_{\neg d,t}, \alpha, \beta_0, \beta_1, \gamma_0, \gamma_1, \pi) \propto$$
$$\begin{cases} \frac{n_0^{\neg d,t}+\pi}{n_0^{\neg d,t}+n_1^{\neg d,t}+2\pi} \times \frac{n_{0,t}^{\neg d,t}+\beta_0}{\sum_{t'}^{T}(n_{0,t'}^{\neg d,t}+\beta_0)} & x_{d,t} = 0 \\ \frac{n_1^{\neg d,i}+\pi}{n_0^{\neg d,i}+n_1^{\neg d,i}+2\pi} \times \frac{n_{c,t}^{\neg d,i}+\beta_1}{\sum_{t'}^{T}(n_{c,t'}^{\neg d,t}+\beta_1)} & \\ \quad \times \frac{\alpha_{d,c} n_{d,c}^{\neg d,t}+\alpha_{d,c}\gamma_0+\gamma_1}{\sum_{c'}^{C}(\alpha_{d,c'} n_{d,c'}^{\neg d,t}+\alpha_{d,c'}\gamma_0+\gamma_1)} & z_{d,t} = c, x_{d,t=1} \end{cases} \tag{3.6}$$

All variables noted with n are counting occurrences: n_0 is the total number of terms assigned to the background topic, $n_{0,t}$ is the number of times a term t is assigned to the background topic, n_1 is the total number of documents assigned to categories, $n_{c,t}$ is the number of times term t is assigned to an individual

[1] $\epsilon = 0.01$ as per Zha & C. Li (2019)

category c. The superscript $n^{\neg d,t}$ denotes that the assignment of the current term t in document d is excluded from the count. Then, $\alpha_{d,c}$ is sampled for every document and category:

$$P(\alpha_{d,c}|t, z, x, \alpha_{\neg d,c}, \beta_0, \beta_1, \gamma_0, \gamma_1, \pi) \propto \begin{cases} \Gamma(n_{d,c} + \gamma_0 + \gamma_1) \times \Gamma(|\alpha_d^{\neg c}|\gamma_0 + C\gamma_1 + n_{d,\cdot}^{\neg c}) \\ \times\Gamma(|\alpha_d^{\neg c}|\gamma_0 + \gamma_0 + C\gamma_1) \times (p + |\alpha_d^{\neg c}) & \alpha_{d,c} = 1 \\ \Gamma(\gamma_0 + \gamma_1) \times \Gamma(|\alpha_d^{\neg c}|\gamma_0 + C\gamma_1 + n_{d,\cdot}^{\neg c}) \\ \times\Gamma(|\alpha_d^{\neg c}|\gamma_0 + C\gamma_1) \times (q + C - |\alpha_d^{\neg c} - 1) & \alpha_{d,c} = 0 \end{cases} \tag{3.7}$$

Here, $n_{d,c}$ denotes the number of times category c is assigned to document d, and $n_{d,\cdot}$ is its sum over all categories. Further, $|\alpha_d^{\neg c}|$ is the number of categories a document is assigned to, with the superscript denoting that the current category is excluded from the count.

The sampling process of one iteration of the SMTM model is shown in Algorithm 2. In the first block (lines 1–23), for each term t in each document d, the counts of the current topic assignment as indicated by $x_{d,t}$ and $z_{d,t}$ are decreased (lines 4–11), a new topic assignment for t is sampled (line 12, see Equation 3.6), and the counts for the new assignment are increased (lines 13–20). In the second block, for each document d and category c, the α-selector is sampled similarly, by first decreasing the counts for the current value (line 26), sampling $\alpha_{d,c}$ (line 27, see Equation 3.7), and increasing the counts for the new assignment.

While this procedure is very similar to the standard Gibbs sampler, the notable difference is that counts are not increased/decreased in a uniform manner, but rather by the individual promotion scores as given in $P_{c,d}$ for document assignment counts, and $Q_{c,t}$ for term assignment counts. This results in a sampling procedure as formalized by the generalized Pólya urn model.

3.2.4 Prediction

One fundamental shortcoming of the SMTM model, as well as topic-model-based classification in general is the inability to predict classes for a new, previously unconsidered document. Multiple approaches have been proposed to address this shortcoming (Yao et al., 2009), yet few can operate without the need of refitting the entire model to the augmented data.

A basic, yet very efficient approach that can be applied to infer class probabilities for a new document is a NB-like approach (ibid.). Given a term-category probability matrix ϕ that can be estimated from the count matrices derived during the fitting process, and a BoW document vector d that holds

Algorithm 2: SMTM sampling procedure

foreach $d \in D$ **do**
  foreach $t \in T$ **do**
    if $t \in d$ **then**
      if $x_{d,t} = 0$ **then**
        $n_0 \leftarrow n_0 - 1$
        $n_{0,t} \leftarrow n_{0,t} - 1$
      else
        $n_1 \leftarrow n_1 - 1$
        $n_{d,z_{d,t}} \leftarrow n_{d,z_{d,t}} - P_{z_{d,t},d}$
        $n_{t,z_{d,t}} \leftarrow n_{t,z_{d,t}} - Q_{z_{d,t},t}$
      end
      sample $x_{d,t}, z_{d,t}$
      if $x_{d,t} = 0$ **then**
        $n_0 \leftarrow n_0 + 1$
        $n_{0,t} \leftarrow n_{0,t} + 1$
      else
        $n_1 \leftarrow n_1 + 1$
        $n_{d,z_{d,t}} \leftarrow n_{d,z_{d,t}} + P_{z_{d,t},d}$
        $n_{t,z_{d,t}} \leftarrow n_{t,z_{d,t}} + Q_{z_{d,t},t}$
      end
    end
  end
end
foreach $d \in D$ **do**
  foreach $c \in C$ **do**
    $n_d \leftarrow n_d - \alpha_{d,c}$
    sample $\alpha_{d,c}$
    $n_d \leftarrow n_d + \alpha_{d,c}$
  end
end

the number of occurrences in the document for each term t, the category distribution $\hat{\theta}_d$ is given by Bayes' rule:

$$\hat{\theta}_d = \sum_{t \in T} \phi_t \times d. \tag{3.8}$$

Usually, $\hat{\theta}_d$ is normalized to $\sum \hat{\theta}_d = 1$ to be usable as probabilities for downstream tasks. As noted by Yao et al. (2009), this method is expected to perform rather poorly, and represents a general baseline method with sub-par accuracy. Yet, no specialized prediction methods have so far been developed for SMTM.

3.3 Seed Term Selection

In order to train the SMTM model, first, a suitable set $\mathbb{S}_c$ of seed terms has to be inferred for each category. This section introduces four methods for seed term extraction from a given corpus of labeled training documents, three automated approaches (TF-IDF, PMI, Term Saliency) and one human-in-the-loop approach.

3.3.1 TFIDF

TFIDF (Spärck-Jones, 1972) is a common baseline for keyword extraction and important measure in Information Retrieval (IR) to estimate the relative importance of a term to a document in a document collection. It can be intuitively extended to measure the importance of terms for categories, by treating each category as the sum of its documents. It is thus defined as:

$$tfidf(t,c) = tf(t,c) * idf(t) = tf(t,c) \cdot \left(\log \frac{|C|}{df(t)} + 1 \right) \tag{3.9}$$

where $tf(t,c)$ is the number of times term t occurs in documents of category c, and $df(t)$ is the total number of documents containing t.

3.3.2 Pointwise Mutual Information

PMI (Church & Hanks, 1990) has been proposed by Jin et al. (2020) for seed term selection, measuring the degree of association of each term t to each category c from an information-theoretic perspective. PMI is defined as:

$$\begin{aligned} pmi(t,c) &= \log \frac{P(t,c)}{P(t) \cdot P(c)} = \log \frac{P(c|t)}{P(c)} \\ &= \log \frac{df(t,c) \sum_{c \in C} df(c)}{df(t) \cdot df(c)}, \end{aligned} \tag{3.10}$$

where $df(t,c)$ is the number of documents of category c term t appears in, $df(c)$ is the number of documents of category c, and $df(t)$ is the number of documents term t appears in overall. However, PMI tends to predict high degrees of association for very rare terms, which is opposed to the second criterion for a a good seed term set defined by Jin et al. (ibid.).

Thus, to combat this issue, Jin et al. (2020) introduce *frequency-adjusted* PMI by multiplying each terms' PMI value with its log-document-frequency, and setting a threshold θ below which the PMI is set to zero:

$$pmi_{adj}(t,c) = \begin{cases} \log df(t) \cdot pmi(t,c) & \text{if } df(t) \geq \theta \\ 0, & \text{otherwise.} \end{cases} \tag{3.11}$$

3.3.3 Saliency

Term saliency (Chuang et al., 2012) is usually used to estimate the importance of a term for each topic post-inference in the general LDA setting. Yet, it can also be applied prior to infer the importance of terms for categories. First, a category-term probability matrix is constructed, which indicates the likelihood $P(c|t)$ that the term t is observed in category c. Further, the marginal probability $P(c)$ of this matrix indicates the likelihood of any random term being observed in category c. The saliency of a term t is then defined as:

$$saliency(t) = P(t) \cdot \sum_{c \in C} P(c|t) \cdot \log \frac{P(c|t)}{P(c)}, \tag{3.12}$$

where $\sum_{c \in C} P(c|t) \cdot \log \frac{P(c|t)}{P(c)}$ is the Kullback-Leibler divergence between $P(c|t)$ and $P(c)$, referred to in this context as the *distinctiveness* of a term. For seed term selection, the saliency is estimated from observed term counts:

$$saliency(t) = \frac{df(t)}{\sum_{t \in T} df(t)} \cdot \sum_{c \in C} \frac{df(t,c)}{df(t)} \cdot \log \frac{df(t,c) \cdot \sum_{c \in C} df(c)}{df(t) \cdot df(c)} \tag{3.13}$$

This rationale behind term saliency can be motivated from an information-theoretic standpoint (ibid.): the more dissimilar the category likelihood of a specific term is in comparison to the likelihood of any term, the more informative it is. The distinctiveness is multiplied with the overall probability of a term occurring, placing an emphasis on more frequent words.

3.3.4 Human-in-the-Loop Selection

As outlined in Section 2.3.2, several Human-in-the-Loop (HITL) approaches to seed term selection have been proposed. While the objective of this thesis is to develop a fully automated process, one manual approach is considered as reference. Both the method of X. Li et al. (2018) and Zha & C. Li (2019) rely on the same three-step process, initially proposed by Chen et al. (2015): first, standard unsupervised LDA is applied to infer latent topics; then, latent topics are mapped to the closest corresponding category; and finally, for each category, up to 10 of the most probable terms are chosen as seed terms, based on their topic probability.

Chapter 4

Data & Experimental Setup

This chapter describes the data and experimental setup used throughout the subsequent study on social system classification. First, Section 4.1 provides information on the creation and provenance of the datasets used or constructed as part of this thesis, first describing each in detail; then, Section 4.2 describes the setup of the experiments that have been conducted.

4.1 Data

Four data collections are used or created for experimentation. The first two, Twenty Newsgroups (20NG) and Ohsumed, are existing datasets chosen due to their popularity in prior work on seed-guided classification; here, they are relied upon as baseline to verify the correctness of the model implementation, to conduct a replication study of previous papers, and to perform parameter optimization for later steps. The third dataset encompasses a large set of articles from the German Wikipedia, and is used as a massive resource of real-world text data to conduct social system classification on, as well as an ablation study investigating the scalability of seed-guided classification. The fourth dataset is newly introduced and derives from a set of monographs published by Niklas Luhmann on each of his theorized social systems. It is intended to mine seed terms from and thus gain first insight into how feasible it is to detect social systems in text data using seed-guided classification.

In the following, each of these four datasets is described in detail, including from where data was obtained, and which processing steps were taken to transform it for experimental use. Alongside that, Table 4.1 includes detailed information about each datasets' characteristics: document count, vocabulary size, category count, category size, and document size.

Table 4.1: Document count, vocabulary size, class count, mean class size, and mean document length for the datasets used throughout the experiments.

Dataset	$\|D\|$	$\|T\|$	$\|C\|$	Category Size†		Doc. Size‡	
				Mean	Std.	Mean	Std.
20NG	18 846	134 098	20	942.3	96.98	65.24	182.44
Ohsumed	13 929	43 486	23	605.6	606.34	97.40	32.53
Wikipedia	95 222	1 986 831	9	17 632.6	24 888.44	247.78	257.17
Luhmann	2653	22 645	9	294.8	186.47	88.78	5.74

† in number of included documents ‡ in number of included terms

4.1.1 Twenty Newsgroups

The 20NG dataset is a multi-label, single-membership dataset that comprises about 18 000 postings from Usenet newsgroups, spanning 20 topics of discussion. It is split into train and test subsets, based on postings published before and after a specific date. Each posting features different annotated parts.

For the scope of this experiment, headers, footers and quotes were stripped from the 20NG data. This reduces the obtainable evaluation scores in comparison to prior publications testing on the 20NG data, yet provides a more realistic classification setting and inhibits overfitting on meta-text. Before the model is trained on the data, it is converted to lowercase, accents are stripped and characters are mapped to ASCII encoding. All terms occurring in less than 5 documents were removed from the vocabulary, together with stopwords from the curated english stopword list supplied by `sklearn`.

4.1.2 Ohsumed

The Ohsumed dataset[1] consists of abstracts of publications from the medical domain, taken from the MEDLINE database. In line with the evaluation experiments of Zha & C. Li (2019) and Joachims (1998), the texts of 13 929 unique abstracts are considered for classification among 23 disease categories. Opposed to 20NG, the Ohsumed dataset allows mixed membership, i.e. one document can belong to multiple categories. The same text preprocessing approach as with the 20NG data was taken.

4.1.3 Wikipedia

Wikipedia data was obtained from the German Wikipedia dump of November 1st, 2021. Three separate dump files were combined to retrieve the necessary

[1] `http://disi.unitn.it/moschitti/corpora.htm`

data. The plain text dump[2] contains clean plain text of each German Wikipedia article as processed by the Elasticsearch search backend of Wikipedia. It is used instead of the commonly used `wikitext` data dump featuring in-text markup, as this markup is notoriously hard to parse, and no structural information from the text is required for the downstream task. Therefore, the pre-parsed dump supplied by Wikipedia itself is used. Further, the metadata dump[3] includes, for each Wikipedia page, the namespace (indicating if the page is an article, a category page, a discussion page, etc.), the title, and the ID. Finally, the categorization dump[4] includes category membership in the form of `(page_id, category)` tuples. Note that only the lowermost categories in the Wikipedia category hierarchy are included in this resource, and all parent categories have to be reconstructed.

As a first step, a joint data table is build by merging the plain text data (uniquely identified by an articles' title) with the metadata (containing title and ID) and corresponding categories (specified by article ID and category title). However, each page is only annotated with its immediate corresponding categories (leaf categories) of the Wikipedia category system. Yet, since a total of 476 607 categories exists in the German Wikipedia, a more coarse-grained categorization is desired for filtering and classification of the data. A promising approach for this is utilizing the 40 root categories of the factual classification of the German Wikipedia[5][6] (Appendix A). To obtain the root categorization of Wikipedia articles, first, the complete Wikipedia category graph has to be reconstructed, in order to establish a mapping between each of the 476 607 total categories and the 40 root categories. This graph is not readily available and has to be induced from the category links present in the categorization data dump. As each of the categories is its own Wikipedia page, which links to its respective parent categories, these links constitute the directed edges of the category graph. A mapping between a leaf category and a root category is then established if a path from leaf to root exists. Note that the category graph is not a strict tree: it can contain multiple linkages and cycles. Therefore, each category can be mapped to more than one root. The final categorization of each article then is the union of all root categories corresponding to an articles' leaf categories.

[2] `https://dumps.wikimedia.org/other/cirrussearch/20211101/dewiki-20211101-cirrussearch-content.json.gz`

[3] `https://dumps.wikimedia.org/dewiki/20211101/dewiki-20211101-page.sql.gz`

[4] `https://dumps.wikimedia.org/dewiki/20211101/dewiki-20211101-categorylinks.sql.gz`

[5] `https://de.wikipedia.org/wiki/Kategorie:Sachsystematik`

[6] Besides factual classification, spatial classification and temporal classification are available. Additionally, meta classifications such as disambiguation, lists and Wikipedia-specific tags can be obtained. However, none of these alternatives prove useful in filtering a comprehensive subset of the German Wikipedia.

In order reduce the dataset to a manageable size, since the resource requirements of the SMTM model are quite high and a focused version of the dataset is desirable to increase model accuracy, the dataset is sampled using the root category information. As indicated in Appendix A, required and rejected categories are defined. Each article has to be a member of at least one required category and cannot contain any rejected category. Rejection categories are chosen to remove articles about persons, places, and events, as these contain little content relevant to the system classification. Conversely, root categories that closely relate to any of the social systems are made a requirement. Each articles' category set is reduced to only the 9 required categories (*Wirtschaft, Gesundheit, Bildung, Kommunikation und Medien, Politik, Religion, Recht, Kunst und Kultur, Wissenschaft*). Note that there is no corresponding category to the social system of *Moral*; instead, *Gesundheit* is added, as *Health* has been suggested as additional social system (Roth & Schütz, 2015, p. 24).

Based on this filtered set, only those articles are retained that belong to between 1 and 4 categories. Furthermore, articles are filtered by token count, only including those who consist of between 50 and 5000 tokens. Since the vocabulary size still exceeds the compute resources available, the text of each article was reduced to only contain terms with a minimum document frequency of 10. In addition to the standard `sklearn` stop-words, a custom stop-word list was curated to filter out Wikipedia-specific language and meta-text (see Appendix B). This results in a total 95 222 articles with multi-class, multi-label annotations for further consideration.

4.1.4 Luhmann

As primary source for the characterization of social systems, and to mine respective seed terms from, the main work that constitutes Luhmanns' formulation of system theory was retrieved, encompassing nine monographs published between 1990 and 2008. Each book used, alongside its corresponding social system and year of publication is listed in Table 4.2. All books were obtained in PDF form and converted to machine-readable plain-text using GROBID (Lopez & Romary, 2015), a state-of-the-art toolkit for extracting text and structural information from PDF files.

In order to obtain the seed terms for each social system, plain text is first extracted from the monographs' PDF files, omitting all structural information included in GROBID. Then, the text is tokenized and in addition to the standard stop-word set for German as supplied by `sklearn`, a custom stop-word list is curated to combat common extraction errors. For example, the publishers name occurs frequently in the page headers of some books and is therefore omitted. The complete custom stop-word list is given in Appendix B.

Table 4.2: Overview of Luhmanns' monographies on Social System Theory.

System	Original Monography Title	Year
Science	Die Wissenschaft der Gesellschaft	1990
Law	Das Recht der Gesellschaft	1993
Economy	Die Wirtschaft der Gesellschaft	1994
Art	Die Kunst der Gesellschaft	1997
Politics	Die Politik der Gesellschaft	2000
Religion	Die Religion der Gesellschaft	2000
Education	Das Erziehungssystem der Gesellschaft	2002
Moral	Die Moral der Gesellschaft	2008

Since using the books directly would only yield eight very long documents, which diminishes their usability for classification, pseudo-documents are created by splitting each books' text into chunks of 90 (whitespace-separated) words, roughly corresponding to the mean document length of the Ohsumed data. All these chunks, labeled with the respective social system explained by the book each chunk stems from, together form the final Luhmann dataset. As each chunk stems from one book, it is a multi-class, single-label dataset.

4.2 Experimental Setup

Four different experiments are conducted as part of this thesis. First, a replication study of Zha & C. Li (2019) and Jin et al. (2020) is performed. It includes an in-depth evaluation of model parameters, applies four different seed term mining strategies, and compares the classification results of the SMTM model with supervised methods, contextualizing its performance. This allows to motivate an optimal parameter choice for all subsequent experiments.

In the second experiment, the scalability of the model is investigated. While the baseline evaluation operates on rather small datasets, the scope is now massively expanded to included a sizable subset of the German Wikipedia. Based on the Wikipedia categories obtained as described in the previous section, in-domain classification is performed by extracting seed terms, and classifying texts into their respective categories using the nearly 100 000 articles.

Third, a first step towards classifying social systems in text data is taken by applying the SMTM model to the Luhmann dataset, mining seed terms and conducting categorization on Luhmanns' books themselves. This is to ensure that the model is equipped to work on the domain-specific data and establish a first foray into social system classification to compare later results to.

In the final experiment, the goal is to detect social systems in a cross-domain setting, by mining seed terms for each social system from the corresponding monography published by Niklas Luhmann, and transferring them to a target domain by training the SMTM model on the Wikipedia dataset. The results are compared to the previously conducted in-domain classification.

In total, these four experiments provide a complete ablation study of all critical parts of the model and allow a first insight into the feasibility of using seed-guided classification for the detection of social systems in text data. All experiments were conducted on a large SLURM[7]-based compute cluster (*Webis Gammaweb*[8]). Each training job utilized a full cluster node equipped with a 40 core CPU and 430GB of RAM. However, due to the nature of the model, inhibiting parallel execution, only one of the available cores was under load.

While Zha & C. Li (2019) provide a model implementation alongside their paper, it is not equipped to scale up to the size of the Wikipedia data. Therefore, a custom implementation was written in Cython, a subset of the Python language that can be cross-compiled into native C code for maximum performance. The implementation adheres to the `sklearn` model API for interoperability with text processing tools widely established in the field of digital humanities. All code and data created as part of this thesis is openly available[9].

[7] https://slurm.schedmd.com

[8] `https://webis.de/facilities.html#gammaweb`

[9] `https://git.webis.de/code-teaching/theses/thesis-gienapp`

Chapter 5

Experimental Results & Discussion

Given the experimental setup described in Section 4.2, this chapter reports on each individual experiment, compares them, and discusses their outcome and impact. The baseline evaluation is conducted in Section 5.1, followed by the scalability evaluation in Section 5.2. Then, social system classification commences, first in the in-domain setting (Section 5.3), and finally in the cross-domain setting (Section 5.4).

5.1 Baseline Evaluation of SMTM

The first experiment replicates the evaluation studies of Zha & C. Li (2019) and Jin et al. (2020), quantifying the performance and behaviour of the SMTM model. First, four different methods of seed term selection are evaluated (Section 5.1.1), followed by an evaluation of the impact of training iterations (Section 5.1.2), model parameters (Section 5.1.3), and a comparative evaluation with supervised classification methods (Section 5.1.4). Throughout, each step is conducted once on the 20NG dataset and once on the Ohsumed dataset, with results reported for each. This is to quantify and compare the models' performance between the single membership and mixed membership setting.

5.1.1 Impact of Seed Term Selection

To gain insight on the impact the different seed term selection methods as well as the number of seed terms used have on the classification result, for each of the four methods described in Section 3.3, the SMTM model is fitted on each of the two datasets respectively, using the top k seed terms, with $k \in [1, ..., 50]$.

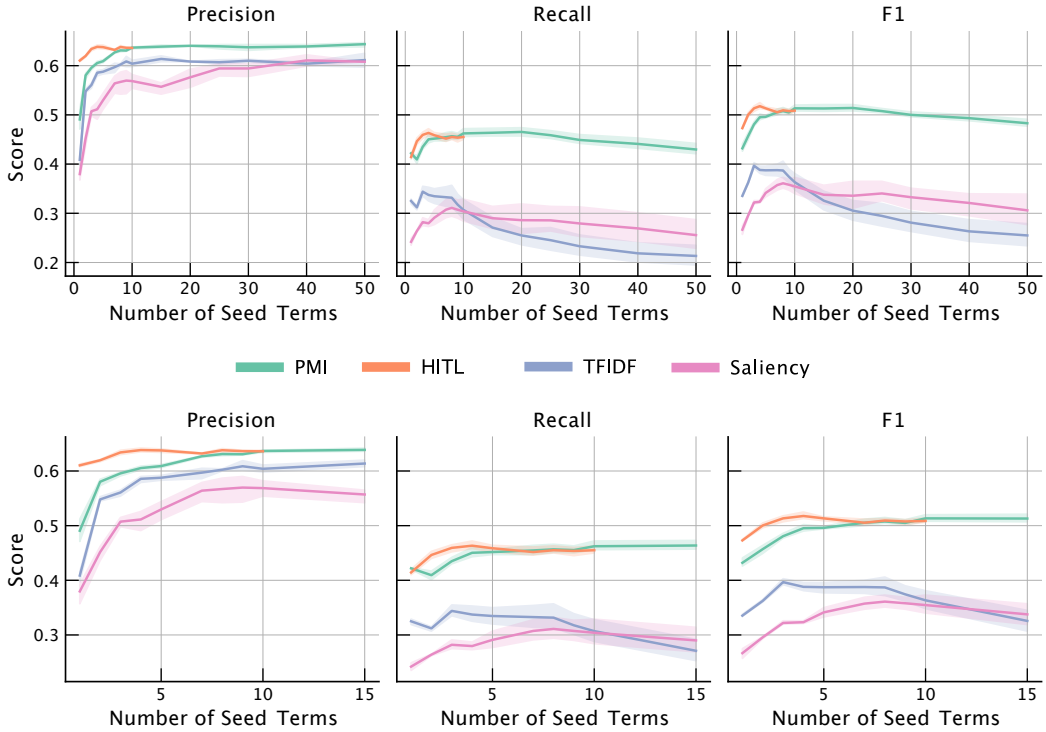


Figure 5.1: Number and method of seed terms and obtained precision, recall, and F_1 score on the 20NG dataset. Average and Standard Deviation over 25, 50, and 100 iterations each. Full (top) and zoomed to < 15 seed terms (bottom).

Both TFIDF and Saliency-based seed term selection are parameter-free. For PMI-based seed term selection, we use the parameter choice of Jin et al. (2020) and set $\theta = 5$. For HITL seed term selection, the seed term set as derived by X. Li et al. (2018) for 20NG and Zha & C. Li (2019) for Ohsumed is used. Note that both only include up to 10 seed terms per topic, thus the evaluation only includes models fitted up to this point for the HITL method. The model is initialized with the parameter recommendations by Zha & C. Li (ibid.), with $\mu = 0.3$, $\pi = 1$, $p = q = 1$, $\beta_0 = \beta_1 = 0.01$, $\gamma_0 = \frac{50}{|C|}$, and $\gamma_1 = 10^{-7}$. For each seed term method, separate runs are conducted at 25, 50, and 100 iterations, yielding varying results. Figure 5.1 shows the obtained Precision, Recall, and F_1 score by seed term count on the 20NG dataset for each method. The average and standard deviation across all three runs is shown. The lower portion of the figure shows a zoomed-in view for less than 15 seed terms. For all performance metrics, manual and PMI-based selection performs best, while TFIDF and saliency-based selection performs worse, with TFIDF being the better of the two up to 10 seed terms, and the saliency method for more. While PMI-based selection performs worse than HITL selection for very few seed terms, it performs better from 10 terms on.

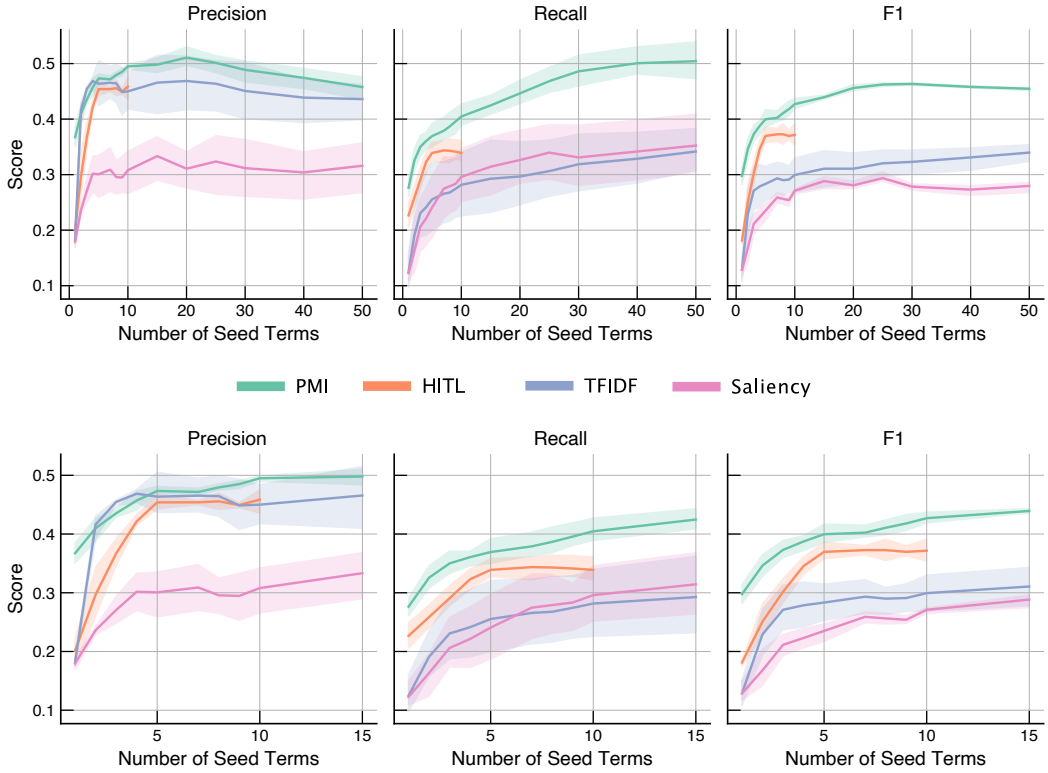


Figure 5.2: Number and method of seed terms and obtained precision, recall, and F_1 score on the Ohsumed dataset. Average and Standard Deviation over $25, 50$, and 100 iterations each. Full (top) and zoomed to < 15 seed terms (bottom).

A second trend observable for all methods is that higher seed term counts do not necessarily increase classification (F_1) performance. While scores are universally increasing up to 15 or 20 seed terms, depending on the method, a downwards trend for recall is observable beyond that point, while precision remains about the same. One possible explanation for this trend could be that a higher seed term count "broadens" the scope of each category up to a point where the sampling process does not select the corresponding categories anymore (increasing false negatives), thus yielding lower recall values, while selected topics are then chosen with a comparatively higher confidence (increasing true positives), thus explaining the precision remaining high throughout.

Crucially, the data depicted in Figure 5.1 describes a single-class classification setting. The model is inherently a multi-class process, so the final prediction is derived by choosing the label with the highest predicted probability, while other predictions are discarded, which is a precision-oriented choice.

Figure 5.2 shows the same evaluation conducted on the Ohsumed dataset. The first trend is also observable here: HITL and PMI-based selection outperform TFIDF and saliency-based selection. Additionally, in this multi-class setting, PMI outperforms all other methods by a larger margin than in the single-class setting examined on the 20NG data. Yet, the second observed trend is reversed: with higher seed term counts, the recall now keeps increasing, while the precision slightly drops beyond 20 terms. This can be explained following the same reasoning as before: higher seed term counts yield less distinct class boundaries, which in turn increases the model uncertainty.This leads to more classes per document being predicted by the model, increasing the recall at the cost of precision. In the single-class setting, this effect is canceled out by reducing the prediction to the choice of highest confidence, which recovers from the loss of precision (reducing the number of false positives), but may introduces false negatives, lowering recall.

Overall, in both settings the PMI-based method of selecting seed terms performs best out of all tested approaches, with respect to both precision and recall. For experiments where no data is available to derive labels from automatically, the HITL method proposed by Chen et al. (2015) achieves similar well-performing results. For all methods and settings, the ideal count of seed terms was around 20 terms, after which performance was deteriorating either in terms of precision, or recall, depending on the setting. Furthermore, across both settings, PMI-based seeding is the least sensitive method with respect to iteration count, as the variance in evaluation scores across the three different iteration counts is among the lowest observed for each run.

5.1.2 Impact of Sampling Iteration Count

A second influencing factor on model performance is the number of sampling iterations the model is fitted with. As in the previous experiment, the model is initialized with the parameter recommendations by Zha & C. Li (2019). For each number of iterations $n \in [10, 20, ..., 140]$, the model was fitted separately 10 times to eliminate variance from random initialization. PMI-based seed term selection is used, as it has been shown to perform best prior. Evaluation was conducted for both datasets. The average precision, recall, and F_1 score by iteration count are reported in Figure 5.3.

In the evaluation conducted by Zha & C. Li (ibid.), only the F_1 score up to 20 iterations on the Ohsumed data was reported and a fast convergence as well as stability beyond 10 iterations has been remarked. Yet, in Figure 5.3, a different trend can be observed: while the F_1 score remains relatively stable after 20 iterations (Ohsumed) and 40 iterations (20NG), the precision and recall scores reported additionally change significantly.

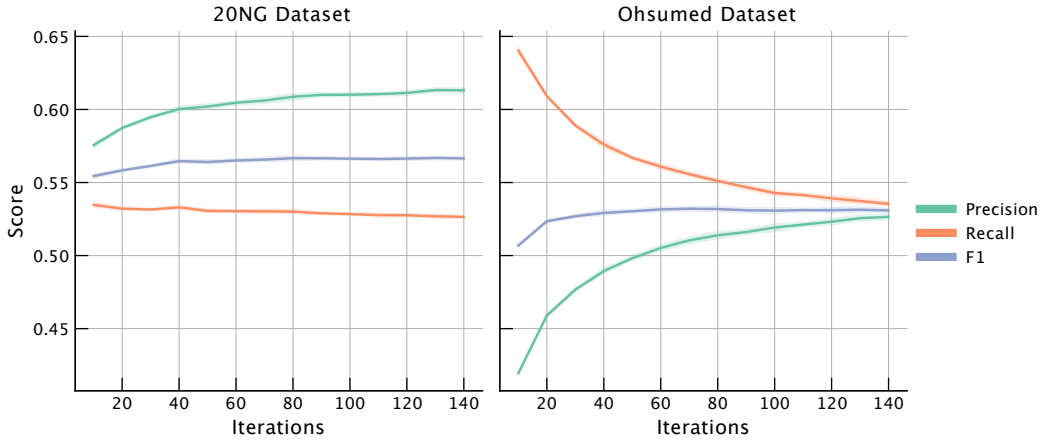


Figure 5.3: Number of iterations and obtained precision, recall, and F_1 score, averaged over 10 runs on the 20NG and the Ohsumed dataset.

For Ohsumed (mixed-membership classification), the recall starts out with 0.65 at 10 iterations, but decreases to 0.54 with higher iteration counts. This trend can be attributed to the sparsity of the models' sampling strategy: the model favors few predicted categories instead of many predicted categories, thus, with higher iteration counts, the number of predicted classes per sample tends to decrease. This in turn reduces the obtained recall, as the probability of selecting a relevant class by chance is lower. On the other hand, the precision increases, from 0.42 at 10 iterations to 0.53 at 140 iterations. Once again, the models' sparsity is contributing to this trend: while the number of predicted categories per sample decreases with higher iteration counts, the models' confidence in each prediction increases, in turn lowering the false positive rate, thus increasing precision.

For 20NG (single-membership classification), a similar, yet less pronounced effect can be observed: again, precision is steadily increasing with higher iteration counts, while the recall is slightly decreasing. Since for single-membership classification only the category with the highest model confidence is selected, it is to be expected that the relative changes in evaluation scores are less sensitive to iteration count. In both settings, recall and precision change at similar rates with respect to the iteration count, which leads to a stable F_1 score, corroborating the observation made by Zha & C. Li (2019). Yet, the additional insight gathered for precision and recall allows to define a more robust recommendation for iteration count: in precision-oriented settings, a high iteration count of > 100 is preferable, while for recall-oriented settings, the existing recommendation of < 50 iterations holds true.

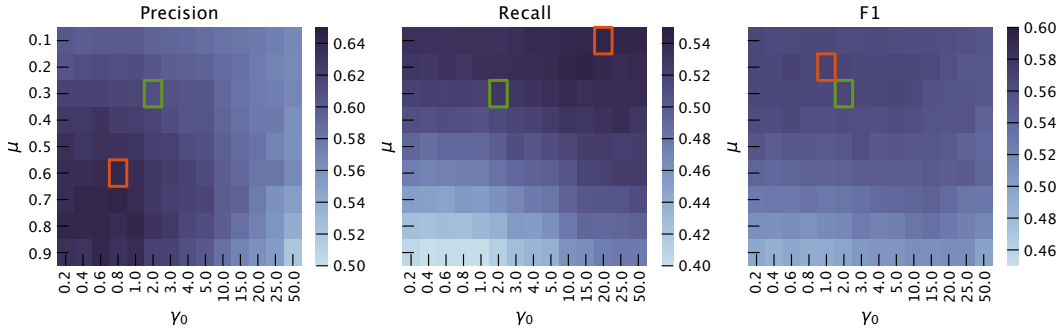


Figure 5.4: Precision, recall, and F_1 for different combinations of μ and γ_1 on the 20NG dataset. Maximum per metric marked orange. Parameter recommendation by Zha & C. Li (2019) marked green.

5.1.3 Impact of Model Parameters

To maximize the performance of the SMTM model, parameter tuning is carried out. In accordance with the recommendations of Zha & C. Li (2019), the primary parameters that are likely to have the largest impact on classification results μ and γ_0. μ controls the impact of document-level supervision, i.e. a low value for μ results in high importance being placed on observing seed(-related) terms in documents, while a high value for μ, seed term (co-)occurrence is less important for the sampling process. γ_0 controls the sparsity of the model, i.e. a low value for γ_0 result in fewer categories per sample being preferred, while a high value for γ_0 favors a less sparse category assignment.

The secondary parameters, which Zha & C. Li (ibid.) found the model to be generally not sensitive to are fixed to the default recommendations of $p = q = 1$, $\beta_0 = \beta_1 = 0.01$, and $\gamma_1 = 10^{-7}$. Parameter tuning of μ and γ_0 is carried out as a full grid search, evaluating the model for each unique combination of parameters. Evaluation was conducted at 100 iterations, which produces a model that exhibits balanced performance with respect to both precision and recall. Once more, the PMI-based seed term sets is used.

Figure 5.4 shows parameter evaluation results for the 20NG dataset. For precision, μ shows less influence, yet values around 0.6 perform most favorable. γ_0 has greater impact on resulting scores, with lower values yielding higher model performance, i.e. the more sparse the sampling process is, the higher the precision. For recall, the inverse trend is observable: higher values of γ_0 yield high scores. Also, unlike for precision, μ shows to be more influential, with higher importance for seed terms leading to higher model performance. For F_1, the parameter recommendations by Zha & C. Li (ibid.) ($\gamma_0 = \frac{50}{|C|}$, $\mu = 0.3$) are approximately reproduced, with the maximum F_1 score at $\gamma_0 = 1$, $\mu = 0.2$.

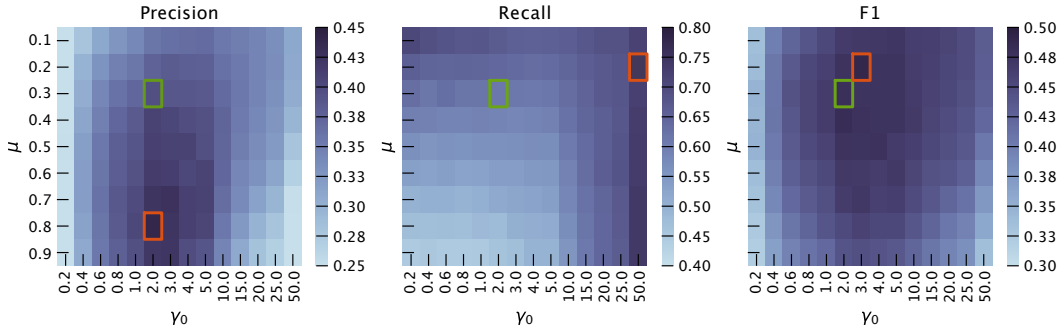


Figure 5.5: Precision, recall, and F_1 for different combinations of μ and γ_1 on the Ohsumed dataset. Maximum per metric marked orange. Parameter recommendation by Zha & C. Li (2019) marked green.

Figure 5.5 shows parameter evaluation results for the Ohsumed dataset. Once again, with respect to precision the model is sensitive to changes in γ_0, but less so for changes in μ. Similarly to the 20NG data, in the mixed-membership setting, both γ_0 and μ are positively correlated with precision, but negatively correlated with recall. For F_1 the experiment on Ohsumed data, too, reproduces the parameter recommendation by Zha & C. Li (2019) with the maximum F_1 score at $\gamma_0 = 3.0, \mu = 0.2$. Overall, three important insights are apparent from the parameter evaluation: (1) Parameter choice shows similar impact on evaluation results in both single-membership and mixed-membership classification settings. (2) The link between category count and the γ_0 parameter ($\gamma_0 = \frac{50}{|C|}$) as established by Zha & C. Li (ibid.) is reproduced for precision and F_1; for recall-oriented settings, the sparsity should be decreased by increasing the γ_0 parameter. (3) The recommendation of $\mu = 0.3$ holds true for both recall and F_1; for precision-oriented settings, increasing μ seems to be beneficial.

5.1.4 Comparison with Supervised Methods

This section focuses on comparing SMTM with standard supervised classification methods. Three different widely used supervised approaches and two naïve approaches were chosen to contextualize the performance of the semi-supervised method. Each of them is described below, with details provided on parameter choices and training procedure. Alongside the scores of these 'baseline' methods, for each dataset, a selection of current state-of-the-art (SOTA) approaches is reported, with scores taken from the respective papers. Previous studies (Jin et al., 2021; 2020; Zha & C. Li, 2019) have shown that SMTM outperforms other multi-class, semi-supervised, seed-based classifiers, and it is considered SOTA for this classification paradigm.

SMTM. The SMTM method is initialized with the optimal parameters as found in Section 5.1.1 and Section 5.1.3: $\mu = 0.2$ and $\gamma_0 = \frac{50}{|C|}$. Seed terms are extracted from the training data using the PMI-based method, with the top 20 seed terms being used as prior to condition the model. The model was then fitted with 100 sampling iterations. An average over 10 runs is reported.

Random Assignment. As first naïve baseline approach, a random assignment of labels to categories, with class probabilities derived from the respective distribution in the training data, is given. The average performance over 10 runs is reported.

Seed Term Overlap. As second naïve baseline approach, a classificator based on seed term occurrence is derived. It follows a simple decision rule given seed words extracted from the training set: if a seed term occurs in a document in the test set, this document is deemed part of the category of the seed term. For single-membership classification, the most occurring category is chosen. For mixed-membership classification, all categories present are used. Its purpose is to provide an estimation of the lower bound of information the seed terms hold without any contextual cooccurrence information from the corpus.

SVC. Also used by Zha & C. Li (2019) as supervised baseline, SVC is a widely used classifier. Parameters where found found using an exhaustive grid search and 5-fold cross validation. Training was conducted on TFIDF vectors. Both linear and RBF kernels were tested, with parameters $C \in [1, 10, 100, 1000]$, and (for RBF only) $\gamma \in [0.0001, 0.001, 0.01, 0.1]$ Best parameters are $C = 1$, and a linear kernel, corroborating the setup used by Zha & C. Li (ibid.).[1]

Multinomial Bayes Classification. As second supervised classification algorithm, MNB was used, since it is a probabilistic approach to classification and thus conceptually more related to the dataless method than SVC. Term occurrence vectors where used as text representation. Parameters where found found using an exhaustive grid search and 5-fold cross validation, with tested values $\alpha \in [0.001, 0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.15, 0.2]$. Best parameters are $\alpha = 0.05$.[1]

[1]Mixed-label output is not naturally supported by SVC and MNB. They were trained in a multi-class, single-label setting accordingly, by reducing the Ohsumed labels to the first label of each sample.

Table 5.1: Accuracy, precision, recall, and F_1 scores for different classification algorithms on the 20NG dataset, with text representation used given. Approaches marked with † are semi-supervised; all unmarked are supervised. All evaluation scores of approaches marked with ‡ are taken from the respective papers.

	Name	Representation	Acc.	Prec.	Rec.	F_1
	SMTM†	Bag-of-Words	0.498	0.672	0.494	0.545
Baselines	Multinomial Bayes	Bag-of-Words	0.687	0.683	0.674	0.671
	SVC	TFIDF	0.663	0.672	0.652	0.655
	Decision Tree	TFIDF	0.402	0.400	0.393	0.394
	Seed Term Overlap†	Bag-of-Words	0.370	0.533	0.365	0.387
	Random Assignment†	Bag-of-Words	0.051	0.050	0.050	0.050
SOTA	Guidotti & Ferrara (2021)‡	Sparse Tensors	0.864	0.863	0.856	0.856
	Gupta et al. (2020)‡	Multi-Sense Emb.	0.862	0.862	0.862	0.862
	Yamada & Shindo (2019)‡	Bag-of-Entities	—	—	—	0.862

Decision Tree Classification. Decision tree classifiers are often used in domains where model explainability is valued highly. As SMTM also offers high interpretability of the process through the probability distributions over words, decision tree classifiers are included in the comparison as representative for highly interpretable models. Parameters were found found using an exhaustive grid search and 5-fold cross validation. Both Gini and Entropy-based fitting criteria where tested, with Gini fitting criterion performing best.

Each of the aforementioned methods was trained and evaluated on both datasets. For the 20NG data, Table 5.1 summarizes the comparative evaluation results, with accuracy, precision, recall, and F_1 score for every method, alongside evaluation scores taken from three recent SOTA publications.

In comparison to the supervised baseline methods, SMTM proves competitive. It places second after MNB for precision, and third for recall and accuracy. Yet, while recall is lower by 0.158 compared to the next best (SVC), it is important to note that in the single-membership setting, the selection process of SMTM yields a precision-oriented choice. Thus, lower recall values are to be expected, at the cost of high precision, which in turn is on par with other methods. Overall, the semi-supervised approach of SMTM performs well on the 20NG data, reaching close to the performance of established supervised classification techniques, and even outperforms the Decision Tree method in every regard while offering similar levels of model explainability. Though, compared to the supervised SOTA, it shows room for improvement.

Table 5.2: Accurracy, precision, recall, and F_1 scores for different classification algorithms on the Ohsumed dataset, with text representation used given. Approaches marked with † are semi-supervised; all unmarked are supervised. All evaluation scores of approaches marked with ‡ are taken from the respective papers.

	Name	Representation	Acc.	Prec.	Rec.	F_1
	SMTM†	Bag-of-Words	0.254	0.498	0.497	0.478
Baselines	SVC	TFIDF	0.338	0.721	0.363	0.452
	Multinomial Bayes	Bag-of-Words	0.306	0.667	0.317	0.391
	Decision Tree	TFIDF	0.150	0.594	0.094	0.135
	Seed Term Overlap†	Bag-of-Words	0.102	0.337	0.697	0.415
	Random Assignment†	Bag-of-Words	0.000	0.038	0.000	0.000
SOTA	Lin et al. (2021)‡	GCN Embeddings	0.728	—	—	—
	Wu et al. (2019)‡	GCN Embeddings	0.685	—	—	—
	Camacho-Collados & Pilehvar (2018)‡	CNN+LSTM	0.375	—	—	—

Furthermore, the SMTM approach shows a clear gain in performance compared to the seed term overlap baseline, with a net increase of 0.158 for the F_1 score. Thus, it can be concluded that the generative approach is successful in utilizing the seed term co-occurrence information in the training data to derive a more robust classifier. Yet, the good result of the simple overlap approach provides further validity to the seed term selection method, as based on the terms only, a sufficient classification can be achieved in many cases.

Table 5.2 shows the results of the comparative evaluation on the Ohsumed dataset. For each method, accuracy, precision, recall, and F_1 score is given, alongside evaluation scores taken from three recent SOTA publications. In this mixed-membership setting, SMTM fares even better than on the 20NG data: while accurracy and precision are comparatively lower than for other approaches, a much higher recall leads to SMTM placing first by F_1 score. While the SOTA approaches only provide accuracy, they once again suggest that the semi-supervised method could be improved upon. For Ohsumed, too, the seed term overlap approach scores fairly high. While precision is sub-par, recall is the highest among all tested systems, achieving a high F_1 score. While the evaluation results are slightly lower than the scores given by Zha & C. Li (2019), this is likely to be attributed to the more stringent approach to data cleaning, stripping metadata that could favor overfitting when training. All things considered, SMTM continues to be a highly competitive semi-supervised classification approach, offering classification performance nearly on-par with related supervised methods while offering high degrees of model flexibility, explainability, and usability.

5.2 Evaluation of SMTM Scalability

The second experiment addresses questions regarding the scalability of the SMTM model. While it has shown favorable performance on established baseline datasets, both 20NG and Ohsumed are comparatively small. Therefore, the much larger Wikipedia dataset is used to train and evaluate the model in-domain, i.e. seeds are mined from the same dataset the model is then fitted and evaluated on. While the Wikipedia dataset is similar to Ohsumed in structure, featuring multi-class mixed-membership categorization, this experiment differs from the Ohsumed one in three regards (see also Table 4.1): (1) larger corpus size, with almost a sevenfold increase in document count (2) increased document size, with average document length being about 2.5 times as long (3) fewer, but more imbalanced document categories. Differences in evaluation scores between the two experiments are therefore likely to be due to one of these three factors.

The previously established optimal parameter choice was used: $\mu = 0.2$ and $\gamma_0 = \frac{50}{|C|}$. Seed terms are extracted using the PMI-based method, with the top 20 seed terms being used as prior to condition the model (Table C.1). The model was then fitted with 200 sampling iterations, an increase by 100 to account for the larger document count and vocabulary size. This choice was validated by fitting secondary models at 50 and 1000 iterations, with the former producing lower overall scores, and the latter being near-identical. The other two parameters seem well-chosen, too: the average number of predicted labels (influenced by γ_0) is near-equal to the actual average labels in the ground truth (1.8 vs. 1.67), while a model fitted with $\mu = 0.5$ produces worse overall scores. Table 5.3 lists evaluation scores per class label, alongside number of label instances in the training data and average document size per class. Evaluation scores are further reported as micro-average, macro-average, and weighted average.

5.2.1 Impact of Corpus Size

So far, no studies assess the performance of the SMTM model beyond the size of the previously used 20NG and Ohsumed datasets. Yet, in real-world classification tasks, such as the proposed social system detection, much larger amounts of data are to be expected. On the first look, the overall performance seems to decrease for the Wikipedia dataset compared to the previous Ohsumed evaluation: a drop in (macro-averaged) F_1 score to 0.274, compared to the previous of 0.498 is observable. This is due to a significant reduction in precision. While the recall of the model actually increases (0.566 vs. 0.498), the precision drops to 0.297 compared to a previous 0.497.

Class	Prec.	Rec.	F_1	#Label.		Avg. Len.
				Gold	Pred.	
Kommunikation & Medien	0.378	0.508	0.434	15 563	20 888	228
Politik	0.056	0.412	0.099	1858	13 583	350
Wirtschaft	0.338	0.551	0.419	14 995	24 402	235
Gesundheit	0.505	0.780	0.613	13 996	21 610	246
Wissenschaft	0.986	0.198	0.329	79 398	15 905	237
Kunst & Kultur	0.177	0.122	0.144	27 833	19 279	244
Bildung	0.087	0.799	0.157	1638	15 021	293
Religion	0.079	0.896	0.145	1806	20 507	308
Recht	0.066	0.831	0.123	1606	20 080	302
Micro Average	0.299	0.323	0.310			
Macro Average	0.297	0.566	0.274			
Weighted Average	0.641	0.323	0.332			

Table 5.3: Evaluation scores, label count in training data, and average document length for SMTM fitted on Wikipedia data in-domain; per class (upper part) and averaged (lower part).

However, attributing this decrease purely to the increased corpus size is misleading: when evaluating per-category (Table 5.3), extreme variation in performance per category is observable. This variation however is not indicative of scaling problems: when computing the overall performance only on classes with at least 10 000 instances, it increases to 0.388, making up almost half of the difference between original Wikipedia and Ohsumed performance, while still representing over 95% of the training data. Thus, the problem does not seem to be inability to scale up to larger datasets, but rather the Wikipedia data in question. This warrants a closer investigation of the other two experimental factors, document scale and label imbalance, with the latter a likely being a highly contributing factor to the observed differences.

5.2.2 Impact of Document Size

Not only is the Wikipedia corpus larger in overall size compared to previously evaluated datasets, its documents are also longer on average. From a theoretical standpoint, the model is likely to be sensitive to the average document length, as its initialization depends on document-level co-occurrence (Section 3.2.2)—the more often a term co-occurs with a seed term of a category, the more indicative it is of that category. Increasing the document length effectively

increases the number of categories a term is likely to be associated with: the more words in its neighborhood (i.e. document), the more likely a seed term is to appear. Moreover, the longer a document is, the more multi-faceted it possibly is; in the case of Wikipedia, longer articles usually consists of many sections, each featuring a focused sub-part of the overall content. This in turn yields a higher chance of more categories appearing near a term.

Investigating the evaluation data per category, there is a moderate negative trend observable between document length and precision ($\rho = -0.66$), and a slight positive correlation between document length and recall ($\rho = 0.37$). Given the impact theorized before, these trends are expected: if a term is likely to be indicative of multiple categories, and the higher document length tends to favor such an effect, its predictive power for a single class decreases. This yields a higher amount of multiple class assignments or wrong assignments, thus decreasing precision. On the other hand, multiple class assignments increase the overall count of predictions made in a multi-class setting, increasing recall.

Given that the average number of predicted categories per document is comparatively low ($\bar{c} = 1.8$) due to the models' parameters, wrong assignments seem to be the majority of cases here over multiple assignments, explaining the less pronounced dependency between recall and document length opposed to precision and document length. This apparent connection between document length and model performance warrants a revisiting the model initialization in future studies, as a sliding-window approach with shorter lengths, defining co-occurrence with seed terms by maximum distance, as opposed to on the document-level could yield improvements to model accuracy.

5.2.3 Impact of Class Imbalance

The final difference between the baseline evaluation experiment on Ohsumed data and the scalability evaluation on Wikipedia data is the number, size, and characteristics of classes in the label data. While there exist less classes overall, they are highly imbalanced, with number of training instances per class ranging from 1606 to 79 398 samples, nearly 50 times as much.

It is important here to consider the inherent difference of the SMTM model to comparable supervised training methods. A supervised training model is likely to show degraded performance on imbalanced training data, as less information is available for less-represented classes, in turn not allowing the supervised model to generalize well about that class, resulting in diminished prediction accuracy and imbalanced output. SMTM on the other hand faces the inverse problem: since the seed term mining process yields the same amount of terms (i.e. information) about each class, any imbalanced training set is effectively made balanced from the perspective of the model. Any imbalances

might only have influence on the respective quality of seed terms for each class. The class predictions of the model are thus more likely to follow a uniform distribution rather than the skewed distribution in the training data.

This effect is observable on the Wikipedia data (Table 5.3): while the label distribution in the ground truth gold labels is highly skewed, the label distribution of the model predictions is much more uniform. Further, underrepresented classes generally exhibit a high recall (i.e. a significant portion of the to-be-identified instances is correctly assigned by the model), yet a dramatically worse precision – likely the result of the model "over-assigning" the class label to other training samples. In essence, this means that the model is successful in capturing the general characteristics of each class from the provided seed terms (high recall), yet the imbalanced gold label data penalizes against the balanced output of the model (low precision).

Whether this is shortcoming of the model or the label data depends on the classification scenario and application the model is used in. Generally, SMTM is advisable for use in settings where a balanced output is desirable. However, it can be argued that the classification of social systems is such a balanced setting: as the union of social systems constitute society, we expect all systems to occur more or less equally in large text corpora, and we want the model to be equally sensitive to each system. This further substantiates the suitability of the model for the task at hand.

5.3 Social System Classification (In-Domain)

The third experiment, for the first time, applies the SMTM model to classify social systems in text data. As a first step, this classification is conducted in-domain: seed terms are mined on the Luhmann corpus, which is also the data collection the classification is performed upon. This allows to (1) quantitatively assess the models performance, since gold labels for the Luhmann data are present, which is not possible in a cross-domain setting; and (2) investigate the models properties to establish a baseline to compare future cross-domain experiments to. This allows to assess the general feasibility of classifying social systems using the SMTM model.

5.3.1 Seed Term Selection

To obtain descriptive seed terms for each of the social systems, PMI-based seed term selection (see Section 3.3.2) is applied to the Luhmann corpus. Each monograph is mapped to one 8 social systems: education, art, moral, politics, law, religion, economy, and science. A minimum term frequency of 3 is used

for the frequency-adjusted PMI calculations. Given the model evaluation in Section 3.3, a value of $k = 20$ terms yields the best results. Table C.2 shows the top 30 terms per category, with an indicator drawn at the cutoff.

The majority of seed terms are semantically coherent with the classes they are supposed to characterize. While some are less specific (e.g. 'neigen' for *Religion*, or 'physische' for *Politics*), no ill-fitting terms are extracted. Some of the chosen terms however are very specific to Luhmanns' writing, e.g., 'reflexionstheorien' for *Education*, 'kommunikationsfähigkeit' for *Religion*, or 'autopoiesis' for *Science*. Also, domain-specific words such as the names of other sociologists ('durkheim', 'weber') appear, however at a very low rate. Some terms appear in multiple forms ('kopplung' and 'kopplungen', both for *Politics*), suggesting that stemming could yield further improvements in the future. Overall, the set of seed terms is deemed qualitatively appropriate, lending further validity to the PMI-based approach. An overall trend from very specific and descriptive to less specific and descriptive can be observed with increasing rank, yet even beyond the quantitatively determined cutoff value of 20 terms, appropriate seed terms are included.

Table 5.4 provides quantitative insight into the quality of the seed term sets by calculating seed coverage statistics: the mean covered document ratio, i.e., the number of documents that contain at least one seed term of the respective class divided by the total count of considered documents, and the mean number of seed terms per document, i.e. the occurrence count of unique seed terms of a class. These value are calculated once in a within-class setting, considering only the documents labeled as belonging to the respective class, and using the complete set. The within-class coverage ratio is balanced across all classes and fairly high, hinting at a good descriptiveness of the extracted terms. The overall cover ratios are less balanced, approximately mirroring the imbalance of document count per class. Still, virtually all documents contain at least one seed term of any class, and the mean seed term ratio across all classes and documents is fairly high at 3.62 terms. This further validates the PMI-based term extraction, as both the overall coverage, yet also the class-specific coverage of documents is high, prompting a high classification performance.

5.3.2 Model Evaluation

To evaluate the classification performance of the SMTM model, it was fitted using the previously mined seed terms in single-class prediction mode, meaning that at the end of the training process, the predictions of each document are reduced to the class with the highest probability. The parameter choice established in Section 5.1 was used: $\mu = 0.2$ and $\gamma_0 = \frac{50}{|C|}$.The model was fitted to the data for 200 iterations, mirroring the setup on the Wikipedia data.

Class	Covered Document Ratio		Mean #Seeds p. Doc.	
	Within Class	All Docs.	Within Class	All Docs.
Politics	0.78	0.25	1.63	0.39
Moral	0.78	0.36	1.71	0.66
Art	0.89	0.48	2.30	0.91
Economy	0.78	0.14	1.51	0.23
Education	0.72	0.09	1.25	0.13
Law	0.81	0.23	1.62	0.34
Religion	0.95	0.05	1.95	0.06
Science	0.93	0.40	2.43	0.69
Overall	—	0.95	—	3.23

Table 5.4: Ratio of documents containing at least one seed term, mean number of seed terms per document, and F_1 score per class, w.r.t only within-class / all documents.

Table 5.5 shows precision, recall, and F_1 for each class and aggregated using micro, macro, and weighted average. These scores can be directly compared to the results on the 20NG corpus, which, too, is a single-class task.

On average, the model performs even better than in the baseline setting. A (macro) F_1 of 0.604 (Luhmann) vs. 0.545 (20NG) can be observed, with an increased recall (up to 0.645 from 0.494) at a similar precision (0.645, slightly decreased from 0.672). The documents in the Luhmann corpus are all of a fixed, equal length of 90 terms (including stop terms, deviations might occur for the last document in a book). This effectively removes the influence of document length on classification performance, and substantiates the claim that a label imbalance is the most contributing factor, as the effect is also observable in this experiment: a strong positive correlation between label count and precision ($\rho = 0.76$), a medium negative correlation between label count and recall ($\rho = -0.51$) is present, replicating the results observed on the Wikipedia data. However, since the class imbalance is less severe, the absolute effect on scores is diminished. When relating the evaluation results with the seed coverage statistics in given prior in Table 5.4, a positive correlation between overall coverage and F_1 score ($\rho = 0.50$) is apparent. This corroborates the findings of X. Li et al. (2018), who note that a high document coverage of seed terms positively impacts the classification performance.

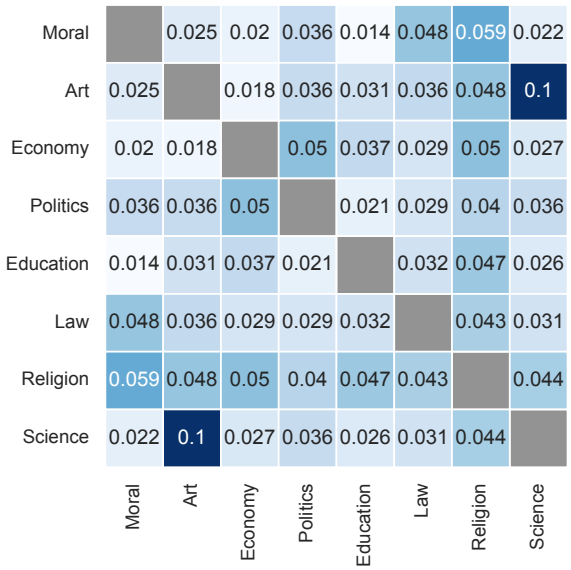
Class	Precision	Recall	F_1	Label Instances	
				Gold	Pred.
Moral	0.905	0.584	0.710	731	472
Art	0.639	0.751	0.691	449	527
Economy	0.695	0.619	0.655	236	210
Politics	0.514	0.601	0.554	278	325
Education	0.735	0.592	0.656	169	136
Law	0.623	0.675	0.648	252	273
Religion	0.221	0.895	0.354	19	77
Science	0.475	0.697	0.565	244	358
Micro Average	0.645	0.645	0.645		
Macro Average	0.601	0.676	0.604		
Weighted Average	0.697	0.645	0.654		

Table 5.5: Precision, recall, and F_1 per class and averaged for the classification model trained on Luhmann data without domain transfer, with number of ground truth and predicted label instances.

Overall, the model can therefore deemed sufficiently equipped to categorize text, even in special domains such as Luhmanns' books. The classification performance is on par with the baseline setting, even exceeding in recall, and the per-class observations replicate previous results as well. The parameter choice is further validated, proving successful in different text domains, scope, and scale. The PMI-based seed term selection method is both qualitatively and quantitatively appropriate.

5.3.3 Investigation of Model Properties

The evaluation of the model has shown very good classification performance. This makes it an interesting candidate for downstream tasks beyond classification of documents. This section illustrated two possible use cases: (1) measuring the strength of association between categories, providing insight into how social systems are interrelated in the Luhmann data; (2) analyzing the most relevant words for each topic for possible downstream tasks in document highlighting and qualitative analysis of the model itself. Note that throughout this section, the multi-class mode of the model was used, instead of the single-class mode in the previous subsection on classification performance, in order to be able to extract inter-class relations.



	Count	Ratio
Moral	530	0.20
Art	678	0.25
Economy	235	0.09
Politics	377	0.14
Education	153	0.06
Law	321	0.12
Religion	107	0.04
Science	473	0.18

Document Count	2659
Mean C. per Doc.	1.08

Table 5.6: Cosine similarity between all system pairs (left side) and predicted label count and label ratio (right side).

Class Relation Analysis. The output of the model in form of the α-matrix allows to interpret the relation of social systems to each other. To quantify the relation of social systems, the cosine similarity between the columns corresponding to each social system in the α-matrix is calculated. Two systems that appear together in higher frequency yield a comparatively higher similarity than two systems that do not appear together. Table 5.6 shows the cosine similarity of each pair of social systems (left side) and the per-system prediction count, as well as overall number of documents and mean number of predicted systems per document.

Since the count of predicted categories per document is very low, it roughly mirrors the label distribution of the single-class setting. Furthermore, since rarely more than one label is assigned, the cosine similarity between systems is extremely low. Only between *Science* and *Art* some degree of similarity is observable, albeit also very low at just 0.1. However, this is consistent with social system theory from a qualitative point of view: as set forth in Section 2.1, Luhmann postulates his systems to be distinct and complementary to each other. Therefore, given the very short document lengths encountered here, it is to no surprise that only very little overlap occurs. This however gives rise to the necessity of performing the experiment in another text domain with much more and much larger documents, such as Wikipedia, to properly investigate the inter-system relationships.

Term Relevance Analysis. A second property of interest, complementary to the document-level analysis, is the investigation of terms, i.e. what terms are highly relevant for each category, as this allows to better assess the semantic scope of a category, and therefore establish or reject correspondence with the desired social system classification. To rank terms with respect to each category, the relevance metric proposed by Sievert & Shirley (2014) is used:

$$\rho(t \in T, c \in C) = \lambda \log\left(\phi_{t,c}\right) + (1-\lambda)\log\left(\frac{\phi_{t,c}}{p_t}\right), \tag{5.1}$$

where $\phi_{t,c}$ is the probability the model predicts for term t and category c, and p_t is the marginal probability for term t. A value of $\lambda = 0.5$ was chosen for a balanced mix between predicted probability and marginal probability. Table D.1 shows the Top 30 relevant non-seed terms per social system. Note that seed terms were deliberately omitted as they still occupy almost all of the top ranks for each system, showing only terms that are learned by the model.

The learned relevant terms are coherent with the social system they are supposed to represent in all cases. While Luhmann-specific terminology is present throughout, the model seems to generalize well beyond the information given by seed terms. Not only are the terms descriptive of their respective class, almost no overlap between classes is observable. A very little amount noise is contained, for example numbers ('39', '40', '14', '1700') or non-descriptive terms like 'off'. The stopword filtering should therefore be improved further. As noted for seed term extraction as well, stemming could yield better results, as multiple forms of a single term are frequently included.

Overall, the descriptiveness and semantic coherence of the single topics give validity to the results, and the experiment can be regarded as successful. In the in-domain setting, both classification performance as well as model properties show promising signs for future downstream tasks.

5.4 Social System Classification (Cross-Domain)

In the final experiment, the cross-domain setting for classification is tested: a seed term set mined from the Luhmann data is applied to categorize documents in the Wikipedia corpus. Since the text domains of the two data sources are different, a covariate shift is likely and has to be accounted for. Besides evaluation of the classification performance, an exemplary analysis of the models properties is included, illustrating possible use cases of the model for social science experiments.

5.4.1 Seed Term Selection

The seed term extraction process is repeated from the previous experiment (Section 5.3.1). However, in the proposed domain-adaptive setting, this process only adheres to two of the three criteria for a high-quality seed term set (Jin et al., 2020): that seed terms should be representative for their associated categories, and that categories should have little to no overlap in their seeds. The third, that seed terms should not be rare words, is dependent on the term frequencies in the target data due to covariate shift. As X. Li et al. (2018) note, the document coverage of seed words correlates positively with the models classification performance. Thus, "rare" can be operationalized as "has low document coverage". Consequently, the candidate set is further refined using document frequency information from the target Wikipedia data. First, a candidate set of 50 terms for each category is extracted from the Luhmann data using the previously established appraoch. Then, to account for covariate shift, this candidate set is ordered descending by the document frequency of terms in the Wikipedia data. The top k seed terms are chosen, where k corresponds to the desired number of seed terms for model training, with $k = 20$ in this case.

Table C.3 shows the top 30 mined seed terms per category, with an indicator drawn at the cutoff value. Once more, all seed terms are semantically coherent. In comparison to Table C.2, slight changes in the seed term set of each category are observable due to the document-frequency-based reordering. The seed terms are less specific and more common in everyday language. No 'Luhmann-specific' terms, i.e. terms that relate to general properties of system theory, like 'selbstorganisation', 'reflexionstheorie', or 'kommunikationsfähigkeit' are present, increasing the widespread usability of the seed term set. Once again, usable seed terms occur beyond the twenty term cutoff.

Table 5.7 repeats the analysis of seed term coverage from Section 5.3.1 using the adjusted seed term set. However, since no gold labels are present, the within-class analysis cannot be conducted. Instead, listed in the table for comparison are the coverage statistics on in-domain data over all documents for comparison purposes. The cross-domain coverage is similar to the in-domain coverage. Yet, while overall coverage is down by seven percentage points, all but one class have an increased coverage value. Especially for underrepresented classes in the in-domain data (like *Religion*), the cross-domain setup improves. Not only the coverage, but also the mean number of seed terms per document increases in the per-class evaluation, and once again does so the most for underrepresented classes. In half the cases, the mean almost doubles. A significant increase is also present in overall number of seeds. In total, the increase of seed coverage is a promising sign that the domain transfer is indeed possible.

Class	Covered Document Ratio		Mean #Seeds p. Doc.	
	In-Domain	Cross-Domain	In-Domain	Cross-Domain
Politics	0.29	0.31	0.44	0.49
Moral	0.57	0.53	0.86	1.16
Art	0.57	0.64	0.93	1.46
Economy	0.17	0.30	0.23	0.50
Education	0.18	0.36	0.22	0.62
Law	0.26	0.28	0.34	0.45
Religion	0.05	0.18	0.06	0.24
Science	0.42	0.49	0.67	0.90
Overall	0.97	0.90	3.67	5.60

Table 5.7: Ratio of documents containing at least one seed term and mean number of seed terms per document per class, within-domain (Luhmann data) and cross-domain (Wikipedia data).

Further, a higher document count and vocabulary size seem to be the improving factor, heightening the chance of encountering high-frequent seed terms. Adjusting the seed term set using the target domain document frequencies of terms has proven beneficial as well.

5.4.2 Evaluation of Classification Performance

Evaluating the models' classification performance in a cross-domain setting presents a problem: no gold labels are present in the target domain. However, one option is to train on the target domain, but evaluate on the (known) labels of the source domain. In this case, the model is trained on Wikipedia data using seed terms extracted from the Luhmann corpus, and its performance is evaluated on the Luhmann label data. Two options to derive predictions for this test set are possible: (1) concatenate both corpora, as to train on both and have the Luhmann documents included in the models' α-matrix; and (2) use the NB-like prediction process introduced in Section 3.2.4, circumventing the need to include Luhmann documents when training the model.

The second strategy is chosen in order to not introduce a train/test leakage in the models training process, which could potentially impact classification results and undermines the cross-domain setup. However, as noted in Section 3.2.4, the NB-like approach is expected to perform sub-par. Therefore, only the relative change in performance between the in-domain and cross-domain setting, not the absolute values should be interpreted. A label probability dis-

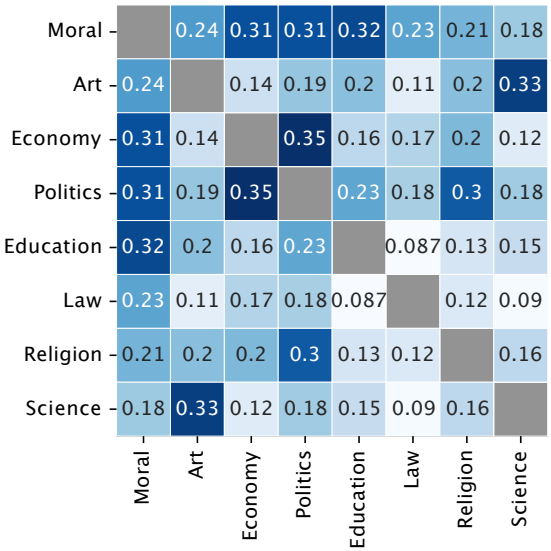
	Precision		Recall		F_1	
	In-D.	Cross-D.	In-D.	Cross-D.	In-D.	Cross-D.
Politics	0.14	0.09	0.00	0.33	0.01	0.15
Moral	0.28	0.31	0.83	0.62	0.41	0.41
Art	0.14	0.19	0.68	0.94	0.24	0.31
Economy	0.00	0.09	0.00	0.44	0.00	0.15
Education	0.00	0.06	0.00	0.24	0.00	0.09
Law	0.05	0.03	0.13	0.01	0.08	0.02
Religion	0.00	0.00	0.00	0.00	0.00	0.00
Science	0.07	0.12	0.62	0.22	0.13	0.15
Micro Avg.	0.15	0.16	0.46	0.49	0.23	0.25
Macro Avg.	0.09	0.11	0.28	0.35	0.11	0.16

Table 5.8: Precision, recall, and F_1 score for NB-like predicitions of the In-Domain and Cross-Domain model, per class and averaged.

tribution is calculated for each document in the Luhmann corpus, once using the cross-domain model trained on Wikipedia, and once using the in-domain model trained on the Luhmann corpus. The two top-scoring classes of each document are assigned as predictions to emulate the mixed-membership setting. Both models are evaluated against the Luhmann gold labels, with precision, recall, and F_1 score given in Table 5.8, per class and averaged.

As expected, the NB-like approach performs much worse than the original predictions made by the in-domain SMTM model. Suprisingly, the cross-domain model performs slightly better than the in-domain model. The micro average increases less than the macro average for all three measures from in-domain to cross-domain setting, due to the reduced variation in per-class scores in the cross-domain setting. The highest increase in observable for macro-averaged recall, which is 0.07 higher from 0.28 to 0.35.

When comparing the in-domain results to the cross-domain results, three classes (economy, education, religion) are never predicted by the in-domain model, as both precision and recall are zero, yet recover a bit under the cross-domain model. This is likely due to the much expanded vocabulary size: while every term of the test set occurs in the vocabulary of both models, the latent probability distributions of the cross-domain model are apparently more useful, as more context information derived from word cooccurrences is encoded in the cross-domain model. Given the higher term counts and vocabulary size of the Wikipedia corpus, terms can be observed in more training samples, and in turn, more information about them can be encoded, improving predictions.



	Count	Ratio
Moral	39 046	0.41
Art	27 781	0.29
Economy	24 023	0.25
Politics	18 920	0.20
Education	18 977	0.20
Law	12 297	0.13
Religion	11 481	0.12
Science	16 377	0.17

Document Count	95 222
Mean C. per Doc.	1.77

Table 5.9: Cosine similarity between all system pairs (left side) and predicted label count and label ratio (right side).

However, *Religion* remains a problematic class, as even by the cross-domain model, the label is never assigned. This is presumably due to an 'overshadowing' effect: since for each document, only the two highest-scoring categories are selected, and since *Religion* is the least frequently occurring class in the gold label set, more 'powerful' categories are selected by default, even when terms relating to the *Religion* class are indeed present in a document.

Extrapolating from these results, and given the promising classification performance in both the previous experiment on Wikipedia data as well as the in-domain Luhmann classification, the models' ability to correctly identify social systems in large text corpora is deemed sufficiently high.

5.4.3 Investigation of Model Properties

To further validate the model given its favorable classification results, the analysis of model properties is repeated, to compare the cross-domain model to the original in-domain one. Once again, class relations and term relevance are analyzed using the same methodology as before.

Class Relation Analysis. Table 5.9 shows the cosine similarity of each pair of social systems (left side) and the per-system prediction count, as well as overall number of documents and mean number of predicted systems per document. Compared to the in-domain setting, an increase in mean detected

categories per document leads to increased cosine similarity scores, as more document-level class cooccurrence can be observed. For example, *Moral* is assigned to about 40% of articles, and thus shows an increase in similarity to other high-frequent classes. Beyond the pairings with *Moral*, the other three systems pairs with high similarity are *Economy* and *Politics*, *Politics* and *Religion*, and *Art* and *Science*.

While these results appear to be reasonable at first glance—politics and economy are intertwined, moral judgment occurs in all contexts, and both art and science assign meaning to the world—the interpretation should be done with caution. Since this is a cross-domain setting, the one-to-one correspondence between assigned classes and social systems given before cannot be accepted without reservation. In fact, as the term relevance analysis will show, there is a significant semantic drift between both corpora, altering the interpretation of categories.

Term Relevance Analysis. Table D.2 shows the Top 30 relevant non-seed terms per social system. Note that seed terms were deliberately omitted as they still occupy about half of the top terms for each system. All other terms were learned by the model. The results are of mixed quality overall. Compared to the in-domain setting, the strong coherence and correspondence between terms and system is lost. Instead of a focused and well-delineated selection of appropriate terms in each system, a common theme across all systems in the cross-domain setting are terms that semantically relate to scientific descriptions of plants, animals, or chemical compounds. This hints at an inherent flaw of the Wikipedia data underlying the classification: even the filtered subset of Wikipedia is highly skewed towards articles describing species or substances, which in turn influences the topic detection. This skew also influences the semantics of the seed terms, introducing a topical drift.

One prominent example of this is the *Art* system: its seed term set includes terms like 'ordnung', 'form', or 'einheit' — terms that are semantically ambiguous, and occur in different contexts in the Luhmann data (where e.g. 'ordnung' refers to 'Gesellschaftsordnung') and Wikipedia (were 'ordnung' overwhelmingly refers to 'order' in the biological sense). This prompts to formulate a fourth requirement for seed term selection: seeds should be semantically unambiguous and refer to the same concept in source and target data. This effect is also very noticeable in the system of *Education*. A strong topical drift into terms relating to *Healthcare* is observable. This is likely due to seed terms like 'mensch/en', 'geburt', or 'kind'. Their semantic meaning in Wikipedia is overshadowed by articles with a medical context, and thus the intended educational interpretation of the terms in the Luhmann data is lost.

The system with the best semantic correspondence is, unsurprisingly, *Science*. Second to that are *Economy*, featuring a topical focus on concepts like automatization, technology, and processes, and *Politics*, with a focus on germany, people, interpretation of meaning ('sachbegriff', 'wortherkunft', 'sinne'), and conditional frameworks ('praxis', 'rahmen', 'zusammenhang'). The system *Religion*, which has the lowest seed term coverage both in-domain and cross-domain, yet features one of the most specific (in the sense of unambiguous) seed term sets is also capturing concepts like history, meaning of words, and translations, which are core tasks for exegesis, and can thus be related to *Religion* in a broader sense. Yet, the overall correspondence of detected term distributions in the Wikipedia to the desired Luhmannian systems is scarce.

Some noise words are also apparent ('displaystyle', 'gnd', 'frac', 'com'), suggesting that more strict text preprocessing is needed. While the same trend of observing multiple forms a the same word ('person', 'personen') formerly noted for the seed term sets is also noticeable here, in light of the ambiguity problems, aggressive stemming might not yield further model improvements. The tradeoff between the reduced vocabulary size yielding higher predictive power for single words and the increase in ambiguity introduced by stemming is to be investigated closely in future work.

Chapter 6

Conclusion

In this thesis, the possibility of performing automatic detection of social systems in text data was explored. Starting out with a description of social system theory as notably developed by sociologist Niklas Luhmann, it was then operationalized within the framework of seed-guided classification. The motivation behind this is that a specific realization of the seed-guided classification method, called SMTM, provides all the benefits of topic modeling, which has found widespread application in the fields of digital and computational humanities due to its properties, while offering the possibility of influencing which latent topics are detected in a given text corpus through a set of so-called seed terms. This allows to condition the model to detect latent categories in text that correspond to each of the to-be-detected social systems without the expense of creating a large amount of domain-specific training data.

To achieve this goal, seed-guided classification in general and the SMTM model in particular were described in detail, reflecting on its advantages, drawbacks, and possibilities of use. Two properties make it especially useful for the task at hand: first, it takes a set of representative seed terms for each category as input, circumventing the expensive and tedious process of creating labeled training data in a large enough quantity for supervised methods. It further allows to mine seed terms directly from relevant primary social science literature for a consistent operationalization of otherwise notoriously difficult concepts. Secondly, it provides as output probability distributions on two levels of granularity, namely a term-category and a document-category probability matrix. This allows for a high degree of model explainability and interpretability, as well as a large variety of downstream tasks.

Yet, the proposed method was never applied for social science research prior to this thesis. Therefore, the major contribution of this thesis is an in-depth evaluation of the model's properties and different methods of automated seed term extraction, both with respect to established benchmark corpora for clas-

sification in computer science, as well two novel datasets created to investigate the models' suitability in two settings: in-domain, detecting social systems within texts written by Niklas Luhmann, mining seed terms and applying the model on a dataset of nine monographs, each corresponding to one social system; and cross-domain, which mines seed terms on the aforementioned Luhmann corpus, and applies the model on a large-scale Wikipedia corpus, encompassing nearly 100 000 articles covering a wide variety of content.

The baseline evaluation on the 20NG and Ohsumed datasets has shown the SMTM model to be a competitive semi-supervised classification method, performing on par to other supervised baseline methods. Previous results of other studies have been successfully replicated, enhancing the validity of the method and establishing a universally well-performing parameter choice, as well as suggestions for adapting the model to the task at hand. The three influential parameters μ, γ_0, and iteration count allow the model to adapt to various settings, with different operating points regarding seed influence, prediction sparsity, and precision/recall tradeoff.

The comparison of different seed term extraction methods yielded a clear recommendation to use the PMI-based selection, which produces seed term sets that are both quantitatively optimal as well as qualitatively sensible. A HITL-based procedure for integrating human-curated seed terms has continued to prove well-performing as per previous studies and this evaluation. This is a promising sign for employing seed-guided modeling for social science research, as labeled data to mine seed terms using automatic approaches may not be available for all tasks. Integrating domain experts into a hybrid process seems to a be a fruitful area of future research.

Adding insight to existing work, the scalability evaluation proved SMTM to be a highly flexible and scalable model that can cope well with imbalanced data, generating largely balanced output. In addition to class imbalance, document length is identified as an influencing factor, which suggests the need for further research in model initialization techniques, for example using a sliding-window over a document-based seed occurrence approach. Overall, smaller document lengths together with a large vocabulary size improves results.

In an in-domain setting, SMTM shows outstanding performance, being able to produce classification results and to exhibit model properties that are not only quantitatively impressive, but also align well with theoretical expectations. The term relevance analysis in particular shows the model to be able to generalize well beyond the given seed terms, and to identify well-delineated categories, in close semantic correspondence to the social system theory underlying the experiment. This is a successful first foray into utilizing seed-guided methods to assists in quantifying phenomena in digital humanities research that have so far been addressed in a qualitative way exclusively.

However, the cross-domain setting performs less favorable. The model has difficulties transferring the semantics encoded in the seed term set to the target domain, and shows to be sensitive to the latent structures of the underlying data. The Wikipedia dataset seems ill-suited to further pursue social system classification on, as it is too skewed to technical articles and the natural sciences. Repeating the analysis of a dataset with properties closer to the domain seed terms are sourced from, for example newspaper articles, seems necessary.

The SMTM model itself could be improved further: as NB-like inference for unseen documents produces largely unreliable results, a specialized post-training method for inferring classes would be a welcome addition. Also, more research into model initialization methods is warranted. Methods that rely on word embeddings instead of cooccurrence to incorporate supervision into the model possibly mitigate the impact of term ambiguity, and improve the cross-domain performance by making it more robust to covariate shift. Using an externally pre-trained embedding model allows to incorporate seed terms that are not part of the target vocabulary, or occur only very infrequently. Embedding-based methods could also be used to quantify the semantic shift of terms between source and target corpora, combating the noted problems.

Given the use cases for topic modeling in the social sciences exemplified in Section **??**, employing SMTM beyond the analysis of social systems is suggested for all cases where influence on the resulting topics is wanted. This could include analysis of free-text surveys, where answer categories of interest are formulated by the investigating researchers and need to be mapped to the responses at hand. By working in a bi-directional process, the validity of these pre-defined categories could be further interpreted by comparing the model properties, i.e. the relevant terms and class associations to the intended descriptions of topics, or their seed term sets, respectively.

SMTM further helps to overcome the problems noted in Section **??**: as the categories are pre-defined by their seed sets, the problem of characterizing and assigning meaning to topics is alleviated. Also, the traditional initialization problems of defining a topic count, and the possibility to derive different outcomes on the same data are diminished: the topic count is an inherent part of the classification process, and since the initialization process relies on word cooccurrence patterns instead of chance, the model variation is (while still present) reduced to a minimum. In total, SMTM offers all the advantages of a topic-model-based data analysis process, such as the intuitive interpretation, and possibilities for building visual interfaces to the data, while reducing the impact of classic problems associated with the method. The possibility of training the model without extensive label data, but instead on seed terms either curated by a domain expert or mined directly from relevant literature opens up exciting possibilities for future applications of the SMTM method in social science and digital humanities research.

Bibliography

Albert, Mathias (2019). *Luhmann and Systems Theory*. DOI: 10.1093/acrefore/9780190228637.013.7 (cit. on pp. 2, 5–8).

Blei, David M., Ng, Andrew Y., & Jordan, Michael I. (2001). "Latent Dirichlet Allocation". In: *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*. Ed. by Thomas G. Dietterich, Suzanna Becker, & Zoubin Ghahramani. MIT Press, pp. 601–608 (cit. on pp. vi, 9).

Blitzer, John, Dredze, Mark, & Pereira, Fernando (2007). "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, pp. 440–447 (cit. on p. 15).

Camacho-Collados, Jose & Pilehvar, Mohammad Taher (2018). "On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 40–46. DOI: 10.18653/v1/W18-5406 (cit. on p. 41).

Cer, Daniel et al. (2018). "Universal Sentence Encoder". In: *CoRR* abs/1803.11175. arXiv: 1803.11175 (cit. on p. 13).

Chang, Ming-Wei, Ratinov, Lev-Arie, Roth, Dan, & Srikumar, Vivek (2008). "Importance of Semantic Representation: Dataless Classification". In: *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*. Ed. by Dieter Fox & Carla P. Gomes. AAAI Press, pp. 830–835 (cit. on pp. vi, 2, 11, 12).

Chen, Xingyuan, Xia, Yunqing, Jin, Peng, & Carroll, John A. (2015). "Dataless Text Classification with Descriptive LDA". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015,*

Austin, Texas, USA. Ed. by Blai Bonet & Sven Koenig. AAAI Press, pp. 2224–2231 (cit. on pp. 25, 35).

Chuang, Jason, Manning, Christopher D., & Heer, Jeffrey (2012). "Termite: visualization techniques for assessing textual topic models". In: *International Working Conference on Advanced Visual Interfaces, AVI 2012, Capri Island, Naples, Italy, May 22-25, 2012, Proceedings*. Ed. by Genny Tortora, Stefano Levialdi, & Maurizio Tucci. ACM, pp. 74–77. DOI: 10.1145/2254556.2254572 (cit. on pp. 15, 25).

Church, Kenneth Ward & Hanks, Patrick (1990). "Word Association Norms, Mutual Information, and Lexicography". In: *Comput. Linguistics* 16.1, pp. 22–29 (cit. on p. 24).

Grimmer, Justin & Stewart, Brandon M (2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts". In: *Political Analysis* 21, pp. 267–297 (cit. on p. 2).

Guidotti, Emanuele & Ferrara, Alfio (2021). "An Explainable Probabilistic Classifier for Categorical Data Inspired to Quantum Physics". In: *CoRR* abs/2105.13988. arXiv: 2105.13988 (cit. on p. 40).

Gupta, Vivek, Kumar, Ankit, Nokhiz, Pegah, Gupta, Harshit, & Talukdar, Partha P. (2020). "Improving Document Classification with Multi-Sense Embeddings". In: *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*. Ed. by Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, & Jérôme Lang. Vol. 325. Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 2030–2037. DOI: 10.3233/FAIA200324 (cit. on p. 40).

Hai, Zhen, Chang, Kuiyu, & Cong, Gao (2012). "One seed to find them all: mining opinion features via association". In: *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*. Ed. by Xue-wen Chen, Guy Lebanon, Haixun Wang, & Mohammed J. Zaki. ACM, pp. 255–264. DOI: 10.1145/2396761.2396797 (cit. on p. 13).

Heydebrand, W.V. (2001). "Structuralism, Theories of". In: *International Encyclopedia of the Social & Behavioral Sciences*. Ed. by Neil J. Smelser & Paul B. Baltes. Oxford: Pergamon, pp. 15230–15233. ISBN: 978-0-08-043076-8. DOI: https://doi.org/10.1016/B0-08-043076-7/01976-8 (cit. on pp. 2, 5).

Hingmire, Swapnil & Chakraborti, Sutanu (2014). "Topic labeled text classification: a weakly supervised approach". In: *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*. Ed. by Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, & Kalervo Järvelin. ACM, pp. 385–394. DOI: 10.1145/2600428.2609565 (cit. on p. 13).

Hingmire, Swapnil, Chougule, Sandeep, Palshikar, Girish K., & Chakraborti, Sutanu (2013). "Document classification by topic labeling". In: *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013*. Ed. by Gareth J. F. Jones, Paraic Sheridan, Diane Kelly, Maarten de Rijke, & Tetsuya Sakai. ACM, pp. 877–880. DOI: 10.1145/2484028.2484140 (cit. on p. 13).

Holmström, Susanne (2007). "Niklas Luhmann: Contingency, risk, trust and reflection". In: *Public Relations Review* 33.3. Special Issue on Social Theory, pp. 255–262. ISSN: 0363-8111. DOI: https://doi.org/10.1016/j.pubrev.2007.05.003 (cit. on p. 6).

Hu, Minqing & Liu, Bing (2004). "Mining Opinion Features in Customer Reviews". In: *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, July 25-29, 2004, San Jose, California, USA*. Ed. by Deborah L. McGuinness & George Ferguson. AAAI Press / The MIT Press, pp. 755–760 (cit. on p. 13).

Huang, Chung-Chi, Chen, Mei-Hua, & Yang, Ping-Che (2015). "Bilingual Keyword Extraction and its Educational Application". In: *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*. Beijing, China: Association for Computational Linguistics, pp. 43–48. DOI: 10.18653/v1/W15-4407 (cit. on p. 14).

Huang, Ronggui (2019). "Network fields, cultural identities and labor rights communities: Big data analytics with topic model and community detection". In: *Chinese Journal of Sociology* 5, pp. 1–28 (cit. on p. 10).

Jacobs, Thomas & Tschötschel, Robin (2019). "Topic models meet discourse analysis: a quantitative tool for a qualitative approach". In: *International Journal of Social Research Methodology* 22.5, pp. 469–485 (cit. on pp. 2, 8, 10, 11).

Jagarlamudi, Jagadeesh, Daumé III, Hal, & Udupa, Raghavendra (2012). "Incorporating Lexical Priors into Topic Models". In: *Proceedings of the 13th*

Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: Association for Computational Linguistics, pp. 204–213 (cit. on p. 15).

Jin, Yiping, Bhatia, Akshay, & Wanvarie, Dittaya (2021). "Seed Word Selection for Weakly-Supervised Text Classification with Unsupervised Error Estimation". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Online: Association for Computational Linguistics, pp. 112–118. DOI: 10.18653/v1/2021.naacl-srw.14 (cit. on pp. 14, 38).

Jin, Yiping, Wanvarie, Dittaya, & Le, Phu T. V. (2020). "Learning from noisy out-of-domain corpus using dataless classification". In: *Natural Language Engineering*, pp. 1–31 (cit. on pp. 13, 15–17, 24, 25, 30, 32, 33, 38, 51).

Joachims, Thorsten (1998). "Text Categorization with Support Vector Machines: Learning with Many Relevant Features". In: *Machine Learning: ECML-98, 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Proceedings*. Ed. by Claire Nedellec & Céline Rouveirol. Vol. 1398. Lecture Notes in Computer Science. Springer, pp. 137–142. DOI: 10.1007/BFb0026683 (cit. on p. 27).

Ko, Youngjoong & Seo, Jungyun (2004). "Learning with Unlabeled Data for Text Categorization Using a Bootstrapping and a Feature Projection Technique". In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. Barcelona, Spain, pp. 255–262. DOI: 10.3115/1218955.1218988 (cit. on pp. 12, 14).

Kouw, Wouter M. & Loog, Marco (2019). "An introduction to domain adaptation and transfer learning". In: *CoRR* abs/1812.11806. arXiv: 1812.11806 (cit. on p. 15).

Li, Chenliang, Xing, Jian, Sun, Aixin, & Ma, Zongyang (2016). "Effective Document Labeling with Very Few Seed Words: A Topic Model Approach". In: *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*. Ed. by Snehasis Mukhopadhyay et al. ACM, pp. 85–94. DOI: 10.1145/2983323.2983721 (cit. on pp. vi, 13, 16, 18, 21).

Li, Ximing, Li, Changchun, Chi, Jinjin, Ouyang, Jihong, & Li, Chenliang (2018). "Dataless Text Classification: A Topic Modeling Approach with Document Manifold". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*. Ed. by Alfredo Cuzzocrea et al. ACM, pp. 973–982. DOI: 10.1145/3269206.3271671 (cit. on pp. 13, 25, 33, 47, 51).

Lin, Yuxiao, Meng, Yuxian, Sun, Xiaofei, Han, Qinghong, Kuang, Kun, Li, Jiwei, & Wu, Fei (2021). "BertGCN: Transductive Text Classification by Combining GCN and BERT". In: *CoRR* abs/2105.05727. arXiv: 2105.05727 (cit. on p. 41).

Liu, Bing, Li, Xiaoli, Lee, Wee Sun, & Yu, Philip S. (2004). "Text Classification by Labeling Words". In: *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, July 25-29, 2004, San Jose, California, USA*. Ed. by Deborah L. McGuinness & George Ferguson. AAAI Press / The MIT Press, pp. 425–430 (cit. on pp. 12, 14).

Lopez, Patrice & Romary, Laurent (2015). "GROBID - Information Extraction from Scientific Publications". In: *ERCIM News* 2015.100 (cit. on p. 29).

Luhmann, Niklas (1976). "Generalized Media and the Problem of Contingency". In: ed. by Jan J. Loubser, Rainer C. Baum, Andrew Effrat, & Victor M. Lidz, pp. 507–532 (cit. on p. 6).

— (1977). "Differentiation of society". In: *Canadian Journal of Sociology* 2, pp. 29–54 (cit. on p. 5).

— (1987). "Tautologie und Paradoxie in den Selbstbeschreibungen der modernen Gesellschaft". In: *Zeitschrift für Soziologie* 16.3, pp. 161–174. DOI: doi:10.1515/zfsoz-1987-0301 (cit. on p. 5).

— (1995). *Social systems*. Stanford: Stanford University Press (cit. on p. 7).

Maas, Andrew L., Daly, Raymond E., Pham, Peter T., Huang, Dan, Ng, Andrew Y., & Potts, Christopher (2011). "Learning Word Vectors for Sentiment Analysis". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 142–150 (cit. on p. 15).

Maier, Daniel et al. (2018). "Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology". In: *Communication Methods and Measures* 12, pp. 118–93 (cit. on p. 10).

Michel, Jean-Baptiste et al. (2011). "Quantitative analysis of culture using millions of digitized books". In: *Science* 331.6014, pp. 176–182 (cit. on p. 8).

Müller-Hansen, Finn, Callaghan, Max W., & Minx, Jan C. (2020). "Text as big data: Develop codes of practice for rigorous computational text analysis in energy social science". In: *Energy research and social science* 70, p. 101691 (cit. on p. 1).

Ramage, Daniel, Rosen, Evan, Chuang, Jason, Manning, Christopher D., & McFarland, Daniel A. (2009). "Topic modeling for the Social Sciences". In: *NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond*. Vol. 5, pp. 1–4 (cit. on p. 11).

Reimers, Nils & Gurevych, Iryna (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3982–3992. DOI: 10.18653/v1/D19-1410 (cit. on p. 13).

Remus, Robert, Quasthoff, Uwe, & Heyer, Gerhard (2010). "SentiWS - A Publicly Available German-language Resource for Sentiment Analysis". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA) (cit. on p. 13).

Ritzer, George (2010). *Sociological Theory*. English. 8. Ed. New York: McGraw-Hill (cit. on pp. 6, 7).

Roberts, Margaret E., Stewart, Brandon M, & Airoldi, Edoardo M. (2016). "A Model of Text for Experimentation in the Social Sciences". In: *Journal of the American Statistical Association* 111, pp. 1003–988 (cit. on pp. 1, 2).

Roberts, Margaret E. et al. (2014). "Structural topic models for open ended survey responses". In: *American Journal of Political Science* 58, pp. 1064–1082. DOI: 10.1111/AJPS.12103 (cit. on pp. 9, 10).

Roth, Steffen (2015). "Free economy! On 3628800 alternatives of and to capitalism". In: *Journal of interdisciplinary Economics* 27.2, pp. 107–128 (cit. on p. 5).

Roth, Steffen, Clark, Carlton, & Berkel, Jan (2017). "The Fashionable Functions Reloaded: An Updated Google Ngram View of Trends in Functional Differentiation (1800-2000)". In: *Writing Technologies eJournal* (cit. on pp. 2, 5, 6, 8).

Roth, Steffen & Schütz, Anton (2015). "Ten Systems: Toward a Canon of Function Systems". In: *Cybernetics and Human Knowing* 22, pp. 11–31 (cit. on pp. 6–8, 29).

Saalmann, Gernot (2016). *Soziologische Theorie: Grundformen im Überblick*. Vol. 27. Wiesbaden: Springer VS (cit. on pp. 4, 5).

Sievert, Carson & Shirley, Kenneth (2014). "LDAvis: A method for visualizing and interpreting topics". In: *Proceedings of the Workshop on Interactive*

Language Learning, Visualization, and Interfaces. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 63–70. DOI: 10.3115/v1/W14-3110 (cit. on pp. 50, 76, 77).

Song, Yangqiu, Upadhyay, Shyam, Peng, Haoruo, & Roth, Dan (2016). "Cross-Lingual Dataless Classification for Many Languages". In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*. Ed. by Subbarao Kambhampati. IJCAI/AAAI Press, pp. 2901–2907 (cit. on p. 13).

Spärck-Jones, Karen (1972). "A statistical interpretation of term specificity and its application in retrieval". In: *J. Documentation* 28.1, pp. 11–21 (cit. on pp. vi, 15, 24).

Vanderstraeten, Raf (2002). "Parsons, Luhmann and the Theorem of Double Contingency". In: *Journal of Classical Sociology* 2, pp. 77–92 (cit. on p. 7).

Weber, Max & Winckelmann, Johannes (1972). *Wirtschaft und Gesellschaft Grundriß der verstehenden Soziologie*. Studienausg., 5., rev. Aufl. Tübingen: Mohr. ISBN: 3165336318 (cit. on p. 4).

Wesslen, Ryan (2018). "Computer-Assisted Text Analysis for Social Science: Topic Models and Beyond". In: *ArXiv* abs/1803.11045 (cit. on p. 11).

Wu, Felix, Jr., Amauri H. Souza, Zhang, Tianyi, Fifty, Christopher, Yu, Tao, & Weinberger, Kilian Q. (2019). "Simplifying Graph Convolutional Networks". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Kamalika Chaudhuri & Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 6861–6871 (cit. on p. 41).

Yamada, Ikuya & Shindo, Hiroyuki (2019). "Neural Attentive Bag-of-Entities Model for Text Classification". In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, pp. 563–573. DOI: 10.18653/v1/K19-1052 (cit. on p. 40).

Yao, Limin, Mimno, David M., & McCallum, Andrew (2009). "Efficient methods for topic model inference on streaming document collections". In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*. Ed. by John F. Elder IV, Françoise Fogelman-Soulié, Peter A. Flach, & Mohammed Javeed Zaki. ACM, pp. 937–946. DOI: 10.1145/1557019.1557121 (cit. on pp. 22, 23).

Zha, Daochen & Li, Chenliang (2019). "Multi-label dataless text classification with topic modeling". In: *Knowl. Inf. Syst.* 61.1, pp. 137–160. DOI: 10.1007/s10115-018-1280-0 (cit. on pp. i, vi, 12, 13, 17, 18, 20, 21, 25, 27, 30–33, 35–39, 41).

Zhang, Mingxi, Li, Xuemin, Yue, Shuibo, & Yang, Liuqian (2020). "An Empirical Study of TextRank for Keyword Extraction". In: *IEEE Access* 8, pp. 178849–178858. DOI: 10.1109/ACCESS.2020.3027567 (cit. on p. 14).

Appendix A

Wikipedia Categorization

Table A.1: Overview of German Wikipedia root categories. Articles have to belong to at least one required (Req.) category, and cannot belong to a reject (Rej.) category.

Category	Rej.	Req.	Related Social System
Bildung		✓	Education
Digitale Welt			
Energiewesen			
Erde	✓		
Ereignisse	✓		
Feuer			
Fiktion			
Geographie	✓		
Geschichte			
Geschlecht			
Gesellschaft			
Gesundheit		✓	Health
Internationalität			
Kommunikation und Medien		✓	Mass Media
Kunst und Kultur		✓	Art
Lebensstadien			
Lebewesen als Thema			
Methoden, Techniken und Verfahren			
Militärwesen			
Naturwissenschaft und Technik			
Organisationen	✓		
Personen	✓		
Planen und Bauen			
Politik		✓	Politics
Raum			
Recht		✓	Jurisprudence
Rekorde			
Religion		✓	Religion
Sicherheit			
Spiele			
Sport			
Umwelt und Natur			
Verkehrswesen			
Wasser			
Weltraum			
Werke			
Wirtschaft		✓	Economy
Wissen			
Wissenschaft		✓	Science
Zeit	✓		

Appendix B

Stop Word Lists

Wikipedia Stopwords. fur, displaystyle, the, isbn, wurde, dass, uber, and, konnen, etwa, wurden, zwei, sowie, art, doi, jedoch, beim, auflage, hrsg, kommt, mehr, gibt, meist, dabei, englisch, bzw, pdf, wahrend, oft

Luhmann Stopwords. unmarked, la, niklas, luhmann, suhrkamp, taschenbuch, bibliothek, he, wem, entry, re, vgl, law, ibid

Appendix C

Seed Terms

	Recht	Kunst & Kultur	Bildung	Wissenschaft	Gesundheit	Wirtschaft	Kommunikation & Medien	Politik	Religion
0	deutschland	bezeichnet	deutschen	band	band	verwendet	verwendet	deutschland	menschen
1	recht	verwendet	bezeichnet	bildern	bezeichnet	bezeichnet	bezeichnet	bezeichnet	bezeichnet
2	deutschen	band	deutschland	bezeichnet	menschen	bildern	beispiel	begriff	jahrhundert
3	bezeichnet	form	begriff	sammlung	springer	eingesetzt	beispielsweise	deutschen	zeit
4	begriff	bildern	zeit	arten	form	sammlung	seit	seit	begriff
5	seit	beispiel	beispiel	verwendet	berlin	heute	form	beispiel	geschichte
6	insbesondere	zeit	seit	drei	häufig	besteht	deutschen	zeit	band
7	beispiel	seit	jahren	besteht	besteht	form	zeit	insbesondere	sei
8	gemäß	heute	bereits	new	for	deutschland	internet	berlin	welt
9	gilt	sammlung	universität	form	beispielsweise	seit	weitere	deren	münchen
10	internet	deutschen	entwicklung	familie	new	zeit	immer	jahre	gott
11	österreich	häufig	form	teil	eingesetzt	genannt	band	beispielsweise	heute
12	abs	beispielsweise	berlin	häufig	verwendet	beispiel	ersten	deutsche	berlin
13	gnd	bereits	denen	berlin	kommen	beispielsweise	bereits	münchen	seit
14	bereits	weitere	menschen	weitere	insbesondere	band	neben	sei	form
15	normdaten	eingesetzt	rahmen	innerhalb	erfolgt	häufig	viele	jahren	leben
16	ognd	berlin	deren	denen	führen	jahren	heute	immer	jahrhunderts
17	aks	denen	beispielsweise	beispielsweise	stuttgart	teil	eingesetzt	gilt	bedeutung
18	sachbegriff	teil	insbesondere	gattung	weitere	daher	denen	bereits	buch
19	januar	neben	münchen	deren	liegt	denen	deren	neue	kirche
20	beispielsweise	ersten	neben	beispiel	with	neben	berlin	gesellschaft	wort
21	deutsche	drei	heute	liegt	aufgrund	weitere	möglich	siehe	vgl
22	zeit	besteht	neue	ebenfalls	behandlung	bereits	häufig	denen	deutschen
23	fall	genannt	übersetzungen	bekannt	teil	deren	begriff	daher	denen
24	deren	immer	besonders	stuttgart	daher	jahre	teil	theorie	stuttgart
25	personen	deutschland	jahre	wobei	führt	regel	erste	teil	normdaten
26	berlin	besonders	müssen	genannt	bereits	möglich	allerdings	entwicklung	genannt
27	denen	jahren	bereich	daher	deren	drei	ebenfalls	menschen	aks
28	allerdings	deren	bedeutungserklärungen	heute	verschiedene	aufgrund	besteht	jahr	ognd
29	sinne	ebenfalls	wortherkunft	bereits	möglich	deutschen	genannt	internet	gnd

Table C.1: Top-30 seed terms per social system for the cross-domain model as extracted from Luhmanns' monographies. Sorted descending by document frequency in Wikipedia data. Top-20 (above line) are used for model training.

	Education	Art	Moral	Politics	Legal	Religion	Economy	Science
1	erziehung	kunst	moral	macht	rechts	glauben	geld	bewußtsein
2	bildung	kunstwerk	ethik	politik	argumentation	religionen	wirtschaft	kommunikation
3	unterricht	wahrnehmung	moralisch	politischen	recht	weber	preise	beobachtungen
4	lebenslauf	welt	achtung	politische	rechtstheorie	religion	geldes	wissen
5	schüler	kunstwerke	moralische	wirkungen	argumentieren	neigen	knappheit	kopplung
6	kinder	zweiter	moralischen	ursachen	begriffe	heilige	arbeit	irritationen
7	schulen	ordnung	gut	gewalt	interessen	heiligen	zahlungen	erkenntnis
8	ungewißheit	beobachter	erwartungen	kausalität	varietät	organisatorischen	zahlung	gleichzeitigkeit
9	schule	kunstwerken	moralischer	machthaber	redundanz	kommunikationsf...	steuerung	beobachten
10	lehrer	kunstwerks	verhalten	sanktionen	gründe	religiöser	konkurrenz	moment
11	selbstorganisation	werk	schlecht	kopplungen	regeln	glaubt	preis	erkenntnistheorie
12	reflexionstheorie	formen	normen	herrschaft	gerechtigkeit	vorgegebene	selbststeuerung	sprache
13	kindes	imitation	mißachtung	mediums	geltung	holen	markt	ereignis
14	kind	form	risiken	einfluß	juristen	verstummen	instabilität	beobachter
15	reflexionstheorien	beobachten	risiko	staat	juristischen	kollektive	tausch	operation
16	erziehen	space	durkheim	politischer	redundanzen	freigegeben	programme	beobachtet
17	redescription	wahrnehmen	ego	kopplung	prinzipien	starke	differenzen	autopoiesis
18	unterrichts	beobachtung	werte	legitimation	rechtsquelle	irrtümer	wirtschaftlichen	ereignisse
19	unbestimmtheit	wahrnehmungen	theorie	physische	rechtsgeltung	speziellen	symbolische	gleichzeitig
20	profession	unterscheidung	solidarität	demokratie	logik	brauchbaren	nichtzahlung	gedanken
21	veränderungen	innen	arbeitsteilung	unsicherheitsab...	juristische	gesellschaftsb...	knappe	strukturelle
22	lebenslaufs	fremdreferenz	alter	absichten	konsistenz	verbindliche	steuern	mitteilung
23	mikrodiversität	geschmack	interaktion	freiheitsgrade	begründungen	rettung	zahlen	nachher
24	selbsterzeugte	erster	gefahr	politisch	argumentations...	hierarchische	inputs	bewußtseins
25	muster	unterscheidungen	supertheorien	stadt	common	selektionskriterien	medien	beobachters
26	kultur	realität	erwarten	politisches	richter	arme	de	psychischer
27	stützen	kunstarten	gesellschaften	entscheidungen	rechte	intendiert	unternehmen	vollzug
28	mensch	information	verständigung	selbstbeschreib...	of	identifizierbar	zugriff	ursache
29	autonomie	betrachter	heuchelei	gemeinschaften	rechtsdogmatik	verringern	orientiert	beobachtens
30	nichtwissen	schönen	erwartet	machthabers	recognition	bekämpfen	ressourcen	selbstbeobachtung

Table C.2: Top-30 seed terms per social system for the in-domain model as extracted from Luhmann data, sorted descending by PMI score. Top-20 (above line) are used for model training.

	Education	Art	Moral	Politics	Justice	Religion	Economy	Science
1	menschen	form	erst	stadt	recht	starke	unternehmen	zeit
2	bildung	seite	menschen	größere	staaten	speziellen	anhand	unterscheiden
3	kultur	dadurch	gut	macht	rechts	heiligen	arbeit	gleichzeitig
4	kinder	geht	entwicklung	gebildet	gründen	verarbeitung	knapp	information
5	schule	zeigt	gesellschaft	staat	regeln	entstandene	markt	sprache
6	familien	welt	sollen	politischen	begriffe	weber	gebracht	wort
7	vorteil	darauf	alter	beruht	rechte	religion	zahlen	elemente
8	wissen	information	natur	politische	interesse	glauben	paris	beobachtet
9	unterschiede	raum	verhalten	herrschaft	formel	verringern	wirtschaft	wissen
10	veränderungen	formen	bedingungen	entscheidung	fehler	arme	medien	ursache
11	anforderungen	grenze	theorie	positive	gründe	heilige	its	vorher
12	mensch	erster	gute	politik	schrift	ausführen	preis	beobachten
13	festen	beginnt	problem	feste	interpretation	neigen	geld	kommunikation
14	schulen	ordnung	werte	gegebenenfalls	richter	verschiedensten	wirtschaftliche	ereignisse
15	schüler	werk	gefahr	ursachen	interessen	religionen	wirtschaftlichen	notwendige
16	muster	alt	historisch	ständig	texten	freigegeben	zugriff	beobachtung
17	kind	kunst	handeln	politisch	prinzipien	charakterisierung	steuerung	beobachtungen
18	geburt	geschieht	soziale	organisationen	rechtlich	gehabt	programme	operation
19	revolution	tradition	betroffenen	negativen	common	verdeutlicht	steuern	gedanken
20	einstellung	einheit	risiko	entscheidungen	logik	glaubt	orientiert	gehirn
21	lehrer	beobachten	themen	gewalt	juristische	religiöser	konkurrenz	organismus
22	personal	verstehen	schlecht	wirkungen	varietät	bekämpfen	wirtschaftlich	ereignis
23	betont	innen	lernen	variation	abwehr	vorgegebene	ressourcen	moment
24	unterricht	unterscheidung	gefahren	politischer	konsistenz	rettung	eigentum	erkenntnis
25	individuum	kommunikation	gesellschaften	absicht	juristischen	vorlesungen	preise	operationen
26	erzeugten	bezeichnen	normen	demokratie	gesetzgebung	concepts	zahlung	gekoppelt
27	kindes	zweiter	risiken	physische	logische	methodische	ausgegeben	psychischen
28	vorgegeben	objekte	freiheit	strukturelle	textes	kollektive	kreislauf	empirisch
29	stützen	wahrnehmung	erwarten	gemeinschaften	geltung	holen	konkurrenten	sprachliche
30	erziehung	beobachtung	begründung	kopplung	gerechtigkeit	verbindliche	differenzen	wissens

Table C.3: Top-30 seed terms per social system for the cross-domain model as extracted from Luhmanns' monographies. Sorted descending by document frequency in Wikipedia data. Top-20 (above line) are used for model training.

Appendix D

Relevant Terms per System

	Moral	Art	Economy	Politics	Education	Law	Religion	Science
1	böse	künstler	arbeitsteilung	ankommen	14	rechtssystem	variation	kommunikations...
2	ethische	unmarked	wirtschaftssystem	zentrum	erziehungssystems	law	organisationen	wahrnehmungen
3	normativ	kunstsystem	güter	civilis	erziehungssystem	rechtssystems	veränderungen	bewußtseinss...
4	sollens	kunstsystems	wirtschaftssystems	überdies	systemtypen	systemintern	max	kommunikations...
5	bereitet	ästhetik	änderung	bindende	verschoben	zirkulären	motive	kommunikations...
6	wahre	rekursiven	marktwirtschaft	auszeichnet	realisation	spezifik	dürkheim	aktuelle
7	moralisierens	distinction	39	akut	interaktionssy...	kondensieren	reichen	beteiligung
8	sympathie	künstlerischen	motivation	ausmacht	pädagogische	verstoßen	unsicherheitsabs...	real
9	schlechtes	relativismus	bestimmtheit	liberale	pädagogen	ausgezeichnet	vornherein	sätze
10	moralist	beobachtbar	gütern	parasiten	pädagogischen	verträge	gottes	innenseite
11	schematismus	übersetzt	40	wirtschaftssystem	selbstselektion	entscheidungsprä...	lernens	hört
12	gutem	erstmals	gesellschaftssy...	irrtum	vergleichbare	werten	wahrheiten	gehirns
13	moralisieren	künste	verknüpfen	wieweit	verschiedenartigen	römischen	selbstliebe	neurophysiolog...
14	nichtlernen	variation	arbeiter	ausgangspunkte	schwäche	normativer	ausdrücken	gedächtnisleist...
15	menschliche	sinnlichkeit	revidiert	ausgangsbeding...	schönen	coupling	plausibilität	liest
16	bewertet	künstlich	publikationen	regierenden	pädagogik	territorialstaat	religionssystem	füreinander
17	staatliche	auswechseln	knapper	stärkeren	anlässen	sinnvollen	jenseits	bewußtseinssyst...
18	verdienen	angestrebt	geläufige	rechtsstaat	fiktionen	respektieren	integration	absieht
19	versagt	künstlers	schaft	societas	trivialmaschine	ausschließungsef...	stadt	verzögerung
20	ernsthaft	fiktionaler	kauf	intimbeziehungen	übersicht	tautologisch	1700	führung
21	festlegungen	kombinationen	richtlinien	weltweite	ständischen	passenden	heider	kommunikations...
22	lob	festgelegten	heterogene	global	forderungen	unzulänglichkeit	bricht	sozialwissensch...
23	abweichenden	übrig	käufer	auswirkt	soziokulturellen	einstweilen	gesicherten	abbruch
24	utilitarismus	historismus	dominierende	starken	ungleiches	naturrechtliche	jüngste	organismen
25	gefährdung	horizont	beratung	selbstvalidierung	erinnerns	gespeichert	nebeneinander	strukturbildung
26	imperativ	bemerken	leute	demokratisierung	intentionen	sense	regulierungen	organischen
27	deckung	wahrnehmens	konkreter	legitimität	arbeitsplätze	kunstsystems	off	führe
28	positives	crossing	transaktion	zugangs	kleinere	crossing	fritz	draußen
29	personale	bekannte	unbestimmten	science	hinzufügen	kapazität	genießen	wahrnehmens
30	humanistische	optischen	geldsummen	redescription	uberschuß	liefert	schöpfungsthe...	lebenden

Table D.1: Top-30 most relevant terms per category for the in-domain model, ranked using the relevance measure of Sievert & Shirley (2014) with $\lambda = 0.5$. Seed terms were excluded from the ranking.

	Moral	Art	Economy	Politics	Education	Law	Religion	Science
1	verwendet	arten	eingesetzt	begriff	volltext	displaystyle	jahrhundert	arten
2	eingesetzt	gattung	internet	geschichte	springer	entsprechend	jahrhunderts	familie
3	displaystyle	familie	einsatz	deutschen	freier	system	geschichte	gattung
4	springer	bildern	seit	sei	behandlung	definiert	begriff	weibchen
5	verfahren	sammlung	com	jahrhundert	pmc	frac	leipzig	roten
6	anwendung	world	verwendet	bedeutungser...	patienten	bildet	wiktionary	iucn
7	möglich	species	möglich	deutsche	zellen	cdot	übersetzungen	bildern
8	einsatz	pflanzen	deutschland	übersetzungen	therapie	systematik	bedeutungser...	world
9	wasser	gattungen	original	gnd	führen	zusammensetzung	wortherkunft	gefährdeter
10	beispiel	blüten	müssen	wortherkunft	medizin	minerale	lexikon	männchen
11	mittels	verbreitet	automatisch	aks	with	right	mittelalter	eingestellt
12	müssen	zentimeter	memento	personen	thieme	left	antike	kopf
13	bezeichnet	tiere	archive	wiktionary	syndrom	mineral	kunst	tiere
14	eigenschaften	länge	system	ognd	protein	sei	gnd	verbreitungsgebiet
15	verwendung	ulmer	technik	normdaten	auftreten	ergibt	verwendet	sammlung
16	entsteht	volume	software	sachbegriff	erkrankungen	mineralienatlas	chr	nahrung
17	herstellung	verbreitungsgebiet	mittels	sinne	nummer	in	aks	schwanz
18	hergestellt	gefärbt	deutschen	frankfurt	human	eigenschaften	ognd	schwarz
19	chemie	besitzen	verfügung	jahrhunderts	erkrankung	chemischen	johann	gefährdet
20	lässt	flora	begriff	beispiel	begriff	handbook	normdaten	insekten
21	wert	samen	möglichkeit	bedeutung	führt	englischen	friedrich	species
22	daher	unterfamilie	januar	münchen	erfolgt	springer	sachbegriff	gefärbt
23	technik	weibchen	jahr	zusammenhang	krankheiten	klasse	hergestellt	weiß
24	führt	meter	verfahren	rahmen	rahmen	raumgruppe	alten	international
25	erfolgt	roten	rahmen	person	molecular	pro	wörterbuch	eier
26	hohe	evolution	vorlage	ziel	biology	wobei	münchen	schnabel
27	hierbei	org	november	vgl	gen	kristallisiert	sammlung	vögel
28	ergibt	metern	oktober	eigenen	review	schen	wilhelm	birdlife
29	chemische	beträgt	beispiel	praxis	clinical	elementarzelle	buch	beine
30	somit	breit	gemäß	einführung	blut	strunz	synonyme	zentimeter

Table D.2: Top-30 most relevant terms per category for the cross-domain model, ranked using the relevance measure of Sievert & Shirley (2014) with $\lambda = 0.5$. Seed terms were excluded from the ranking.