FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Text-Mining Based Incident Identification

in the Domain of Sustainability

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Florian Wieser
Matrikelnummer 0825449

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung: o.Univ.Prof. Dipl.-Ing. Dr.techn. A Min Tjoa
Mitwirkung: Dipl.-Ing. Dr. Alexander Schatten, Projekt Ass.

Wien, 31.01.2015

(Unterschrift Florian Wieser) (Unterschrift Betreuung)

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Text-Mining Based Incident Identification

in the Domain of Sustainability

MASTER'S THESIS

submitted in partial fulfilment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Florian Wieser
Enrolment number 0825449

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: o.Univ.Prof. Dipl.-Ing. Dr.techn. A Min Tjoa
Assistance: Dipl.-Ing. Dr. Alexander Schatten, Projekt Ass.

Vienna, 31.01.2015

(Signature of Author) (Signature of Advisor)

Erklärung zur Verfassung der Arbeit

Florian Wieser
Im Vogelsang 29, 3340 Waidhofen/Ybbs

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

_____ _____

(Ort, Datum) (Unterschrift Florian Wieser)

Abstract

Sustainability topics have become increasingly popular in recent years due to a growing awareness in society regarding sustainability as well because of a stronger awareness about the long-term consequences of business activities. Companies seem to care more about their corporate social responsibility, possibly due to a redefinition of sustainability, which may have induced more awareness. Nowadays, sustainability has a broader definition and concerns social and economic matters in addition to environmental ones. Sustainability issues can pose a risk for corporate success and, within companies, risk management departments are responsible for the identification and assessment of potential sustainability risks, which is not trivial a task.

This thesis focuses on solving the problem of the identification of environmental sustainability incidents within text documents. The widespread availability of the World Wide Web has led to enormous growth in accessible and reported information all around the world and such possible relevant information could be published on online news portals, blogs, and social media streams. The challenge is to find and extract the important information from this enormous amount of data that increases every day. In order to automate the detection of sustainability incidents, a data mining approach was formulated. To enable the detection of environmental sustainability incidents, it was necessary to develop a formal definition of such incidents and map this definition to a natural language processing approach, which is suitable for an event identification within text.

The system is developed by using the state of the art in natural language processing technologies. It gathers text sources from blog sites that publish in the environmental domain. The system works on a rule-based approach that identifies the presence of an environmental context and its possible relation to a company on a sentence level. Further, dependencies between words are examined and deep learning solutions are applied for sentiment classifications. Several set-ups are evaluated and the results are presented in this work. Finally, the thesis also concludes with a critical review of the ethical concerns in data mining and addresses credibility issues for online sources.

Sustainability wurde ein immer populäreres Thema in den letzten Jahren, sicherlich bedingt dadurch, dass die Gesellschaft ein stärkeres Bewusstsein für Problematiken aus diesem Bereich entwickelt hat und auch ein Umdenken über Langzeitkonsequenzen, von unternehmerischen Tätigkeiten, sich in den Köpfen von den Entscheidungsträger in den Unternehmen festgesetzt hat. Sustainability hat auch eine Neudefinition erfahren, heute behandelt Sustainability auch soziale und ökonomische Anliegen und fokussiert sich nicht nur mehr auf ökologische Thematiken. Das Vernachlässigen von Nachhaltigkeitsthemen kann auch ein potentielles Risiko für Unternehmen darstellen. Deswegen befassen sich innerhalb von Unternehmen das Risikomanagement, unter anderem, mit der Identifikation und Beurteilung von potentiellen Umweltrisikofaktoren, welches nicht als eine triviale Aufgabe aufgefasst werden kann.

In dieser Diplomarbeit liegt der Fokus darauf, ob es möglich ist, Environmental Sustainability Incidents innerhalb von Texten zu identifizieren. Die immer weite Verbreitung des World Wide Web hat dazu geführt, dass eine immer größere Anzahl an berichterstattenden Information verfügbar ist. Mögliche Quellen, in jenen relevante Informationen veröffentlicht werden können, sind Online News, Blogs oder Social Media Streams. Die Herausforderung hierbei ist es, die relevanten Informationen zu erkennen und zu extrahieren. Für eine Automatisierung der Identifikation von Sustainability Incidents, wurde eine Lösung mit der Hilfe von Data Mining Methoden umgesetzt. Dafür war es notwendig eine formale Definition von Environmental Incidents zu finden und diese in eine Natural Language Processing (NLP) Lösung zu überführen.

Die entwickelte Lösung verwendet den State of the Art in NLP und verwendet als Textquellen den Content von Blogseiten welche im Umweltbereich publizieren. Das System funktioniert regelbasierend und identifiziert ob ein Umweltkontext vorhanden ist und erkennt möglichen grammatikalischen Beziehungen zu einem im Satz vorkommenden Unternehmen. Für die Erreichung einer besseren Performance, wurden fortgeschrittene Methoden wie Dependency Detection und Deep Learning Algorithmen verwendet. Nach einem iterativen Knowledge Engineering Ansatz, wurden verschiedene Setups für mögliche Lösungswege formuliert, diese wurden in Folge evaluiert und die Ergebnisse dokumentiert. Letztendlich schließt die Arbeit mit einer kritischen Betrachtung und ethnischen Bedenken in Bezug Data Mining. Des Weiteren wird auch noch die Problematik der Glaubwürdigkeit von online Medien adressiert.

Contents

CHAPTER 1

Introduction

1.1 Motivation and Problem Statement

Incidents are instances of something happening—an event or an occurrence[1]. From a corporate perspective, incidents are mostly regarded as events that trigger consequences that could influence the operative business. Risk management departments in companies are usually concerned about these kinds of events and a fundamental part of risk management is the detection and assessment of risk factors, according to corporate risk management plans, like for example in NASA's[2] risk management management plan 'Zeus'(Jones, 2008). Blog posts represent a possible source for the detection of incidents. News Blogs can be created by everyone who is able to access the internet and the assumption is, that they may not just focus on the most popular events, like ordinary online newspaper do. Micro blogging services (like Twitter) may have that in common with blog sites, but the reported stories are much larger in text size, due to possible restriction in text length, twitter for example is limited to 140 characters [3]. "More than just text, blogs provide significant structural information about the author, such as precise timestamps, geographical location, age, gender, and explicit friendship links"(Lloyd et al., 2006). Additionally, people blog to post their opinions about topics in order to influence others from a personal point of view, thus blog sites can be used to determine public opinions on current events (Lloyd et al., 2006). Thus, blog posts are a different to approach from a text engineering point of view. Especially in the context of sustainability, it is possible that events are reported that are related to companies. If an incident affects a company or some subcontractor in the company's supply chain, sustainability-related incidents become a significant potential risk. Changes in a dy-

[1] http://www.oxforddictionaries.com/definition/english/incident

[2] http://www.nasa.gov/

[3] https://media.twitter.com/de/best-practice/anatomy-of-a-tweet

namic global business environment, requires firms to be more flexible and to adapt and respond quickly to potential risks which could induce market changes.

Among the forces that drive change, the requirements for corporate responsibility and sustainability are becoming increasingly urgent. During such difficult times as the recent economic downturn, companies are faced with hard choices to survive (Dao et al., 2011). Research has acknowledged that addressing sustainability issues is critical to the long-term existence and success of companies (Porter and Kramer, 2006). In current research, sustainability is considered in a broader perspective and the focus is on topics beyond the reduction of energy consumption and of pollutants and a perspective has emerged that defines sustainability as including three components: the natural environment, society, and economic performance. In scientific literature, this perspective is generally referred to as the triple bottom line (TBL) (Elkington, 1994). Even though this thesis will address only the environmental part of sustainability and examine only web blogs that deal with environmental issues, it is important to also mention the broader perspective of sustainability.

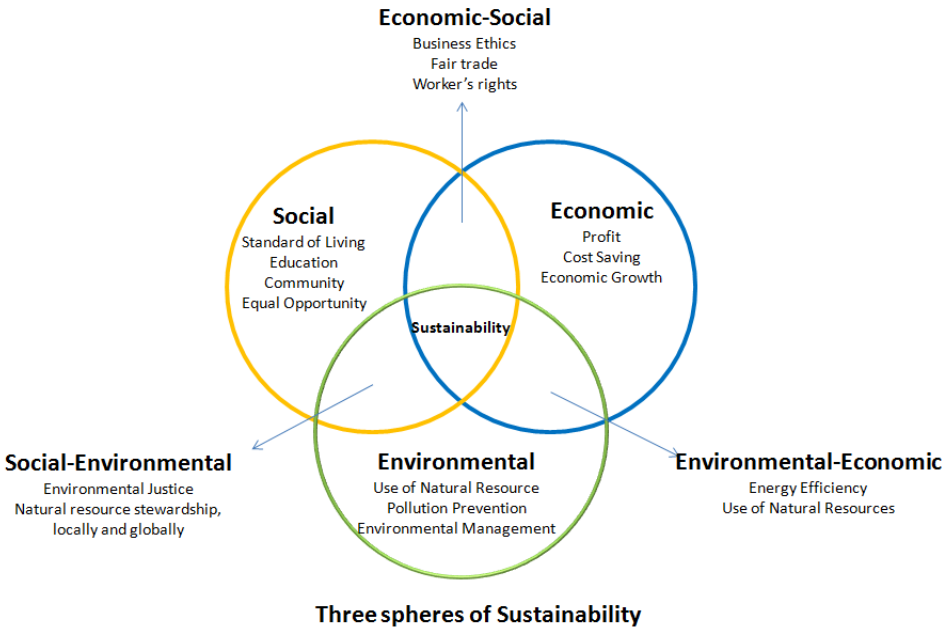


Figure 1.1: The three spheres of sustainability (Envecologic, 2014)

Until now, the major contribution of information technology to sustainability has been in reducing its energy consumption, but beyond that, information technology should also enable companies to develop sustainability capabilities (Dao et al., 2011).

Especially in the field of sustainability incident identification, several different dimensions can be considered and, in this case, IT should be able to process data from information sources that could be relevant for incident identification and to determine their relevance further in order to enable an assessment of their possible impacts. Within and outside companies, a huge amount of information exists that could also be used for analysis. A large proportion of this information is also available as text that can be processed with text mining and natural language processing (NLP) (Dao et al., 2011), as used in this thesis.

As mentioned already, this thesis will focus on the environmental sphere of sustainability. Hence, a definition of a sustainability incident is important. According to the (University of East Anglia, 2014), an environmental incident can be defined as follows: "An environmental incident or emergency is a sudden onset accident or disaster resulting from natural, technological or human-induced factors, or a combination of these, that causes or threatens to cause environmental damage as well as impacts on human lives and/or property. Emergency preparedness is a term increasingly used to describe the interface between accidents, human life, health, and the environment."Real-world examples for such a definition would be spillages of chemicals, oil or fuels, floods or events like the collapse of a building or storm damages. The next section will describe the methodology and the approach for detecting environmental incidents.

Sustainability and Risk Management

This section discusses the importance of sustainability in risk management and the importance of an awareness for sustainability issues within companies. Sustainability can be seen as "a development that meets the needs of the present without compromising the ability of future generations to meet their own needs"(Commission et al., 1987). The main motivation for companies to address sustainability should be, to take into account the communities in which they operate, maintain the highest standards of governance and ethics, and mitigate its overall impact on the environment.

As stated previously, sustainability incidents can be considered a possible risk for companies; thus, they also affect risk management in companies. Environmental risk can have effects on the supply and demand of a company's supply chain. For example, a fire that was caused by lightning in a factory of a supplier could trigger a supply risk for all companies in the supply chain. This would also affect all three aspects of the already-mentioned TBL sustainability baseline. Customer demand could also be affected if a company is unable to ensure a stable supply chain and, additionally, if products are produced under safe environmental and working conditions, the costs which emerge from possible accidents reduce. Sustainability risk has a very broad scope and some examples can demonstrate this. The oil spills of the Exxon Valdez catastrophe in Alaska, which occurred in 1989, caused huge environmental damage and disturbed local business. Exxon had to pay a $5 billion punitive damage award (Anderson, 2006).

To present a different example, a reason for a lower customer demand could be a boycott against the firm's products caused by bad reputation, which has been triggered by sustainability risk events. The consequences of such a boycott would be a loss of revenue. A boycott by Greenpeace of an offshore oil company that planned to sink a platform in the sea caused the company's retail sales to drop by approximately 30% in some European countries. Protests against Nike by students groups for their sweatshop practices resulted in a drop in its stock price and revenues (Anderson, 2006). A further significant sustainability risk is global warming and climate change; studies and climate change models show that the weather had become more extreme in recent decades and its impacts and consequences more severe (Frich et al., 2002),(Easterling et al., 2000). Climate change and global warming are just a part of environmental conditions; many other environmental conditions are undoubtedly produced by mankind, like the increasing pollution of the environment through waste produced by the growth of the world economy and industrialization (Mani and Wheeler, 1998), or urban air pollution in huge cities (Mage et al., 1996).

Sustainability covers a wide area of influences that can be considered if companies want to manage risks within this domain. After all these examples of possible events that trigger environmental risk, the question arises as to how companies should establish a strategy for sustainability. A strategy that should be able to "develop capabilities that enable radical clean technologies and processes that help solve social and environmental issues"within the company and, from an external point of view, to "include the core sustainability capabilities in all products, processes and supply chains"(Dao et al., 2011, p. 69). Companies must be able to measure their use of hazardous substances, emission of pollutants, employee health and safety, and integrate such metrics within key business processes (Kleindorfer et al., 2005). By collaborating and utilizing up-to-date information and standards, firms can improve sustainability, while also increasing operational efficiency and performance (Vachon and Klassen, 2008). One example is a sustainability benchmark like the Dow Jones Sustainability Index (Knoepfel, 2001). This is a global benchmark and it covers the sustainability performance of the 2,500 biggest companies that are listed on the Dow Jones Global Stock Market. The index focuses on economic, environmental and social developments, like the TBL approach. The companies are monitored and rated on the basis of a variety of sources, like the cross-checking of information, company questionnaires, company documents, publicly available information, stakeholder relations, media screening and company interviews. Considering such benchmarks for the decision process would lead to a firm's selecting suppliers based not just on the price and would also concern various TBL metrics (Angell and Klassen, 1999).

The sharing of information and the creation of a web of interactions that enable a network effect and knowledge exchange among firms, their supply chain partners and stakeholders will also be necessary. The closed collaborations between firms, their supply chain partners and stakeholders enabled by such networks create capabilities that

are difficult for competitors to imitate (Hart, 1995).

1.2 Research Question

The aim of this work is to use publicly available information for a media screening in the domain of environmental sustainability. In order to achieve this goal, it is necessary to develop a suitable solution that is able to analyse the information. This thesis attempts to solve this issue by using the capabilities of text mining and NLP. For the development of a text mining system, it is necessary that an approach is worked out which maps the definition of an environmental incident to an NLP solution. The text sources that will be processed by the solution will be blog articles from different authors who publish in the domain of environmental sustainability.

A further important part of this thesis is to formulate and apply an evaluation approach in order to determine and ensure the quality of the developed solution. Without any comprehensible evaluation and the usage of state-of-the-art performance measures, it would not be possible to judge the performance and enable a comparison with other possible solutions.

The following research questions are considered in this work:

- *RQ1: How can the interpretation of environmental sustainability incidents, according to a formal definition, be mapped to a natural language solution approach?*
- *RQ2: What is a suitable information extraction approach and what are possible evaluation methods in order to ensure the quality of the solution?*
- *RQ3: At what performance level is it possible to identify sustainability incidents, that are within an article from a sustainability blog, through the usage of NLP?*
- *RQ4: An evaluation of sustainability incident-related articles, which are published within a certain time frame on popular sustainability blogs.*

1.3 Methodological Approach

- **Literature Review**
  * Determination of incidents in the context of sustainability
  * Determination of the coherence between incidents and sustainability risk
  * Review of text mining solutions
- **Methodology**

    * Determination of a potential text source, like sustainability blogs
    * Usage of suitable data mining/NLP tools
    * Development of a data mining solution and extraction of relevant data

- **Proof of concept**

    * Implementation of the incident detection solution for an exemplary field

- **Evaluation**

    * Evaluation of the context extraction and the incident detection from text sources, through test scenarios

1.4 Structure of the Work

The work is structured in the following way:

Chapter 2—State of the Art: In this chapter, a literature review will be carried out in order to determine the state of the art in NLP. The chapter will also outline the similarities and differences of the solution developed in this thesis compared to those that already exist.

Chapter 3—Solution Approach: The methodology that is used for the system is described along with how to solve the problem that is addressed in this thesis. In this section, it is necessary to define how a sustainability incident could be identified using an NLP approach.

Chapter 4—Proof of Concept: In this section, the system will be described from a technical point of view. The chapter will focus on which technical approach is chosen and which technologies and frameworks are in use for the implementation.

Chapter 5—Evaluation: To ensure the quality and to enable a quantification of the quality, an evaluation approach has to be defined. This section will describe how evaluation can be carried out in NLP and which concepts exist.

Chapter 6—Results: In this chapter, the results of the evaluation will be presented.

Chapter 7—Critical Reflection: This chapter provides a critical reflection of the pros and cons of data mining techniques and the validity of information that is acquired through such a system.

Chapter 8—Summary and Future Work: A summary of the work will be provided here, along with an outlook for possible future work.

CHAPTER 2

Related Work

This section will discuss the research field of NLP. The aim of the thesis is to extract information from online text sources. State-of-the-art approaches and techniques that are suitable for such a task will also be described here.

2.1 Information Extraction

Information extraction (IE) defines any process that selectively structures and combines data that is found, explicitly stated or implied, in one or more texts (Cowie and Lehnert, 1996a). Its goal is to extract facts about systems and to process machine-readable, semi- or unstructured text documents in order to detect information about pre-specified types of events, entities or relationships.

IE should not be mistaken for information retrieval. Information retrieval concerns retrieving a set of documents that are relevant to a query-based keyword search (Singhal, 2001). In contrast, IE is about extracting facts from an input source like, for example, text documents. From the viewpoint of NLP, IE is attractive for many reasons, including the following:

- Extraction tasks are well-defined
- IE uses real-world examples
- IE poses difficult and interesting NLP problems
- IE performance can be compared to human performance on the same task

(Cowie and Lehnert, 1996b)

Principles for the Design of an Information Extraction System

Before discussing the various parts of IE and working on an implementation, it could be beneficial to study the design principles and approaches that exist in this domain. According to (Appelt, 1999), there are two basic approaches for the development of an IE system. The *learning/training* approach and the *knowledge engineering* approach.

Knowledge Engineering

Knowledge engineering requires system engineers that are familiar with the concept of IE systems. These engineers are responsible for the definition of rules in order to extract relevant information from the text input. For this task, a corpus of relevant text sources is usually available and the engineer is responsible for applying general knowledge or intuition in designing the rules. Here, the skill of the knowledge engineer plays an important role in determining the performance level that will be achieved by the system. Besides the detailed knowledge that is needed, knowledge engineering also requires considerable labour and can be repetitive in nature. The process of building a system with a good performance is an iterative one where a set of rules need to be defined. These rules are applied on a text training corpus and the output is examined in order to assess if and where the rules under- or over-generate. If a deviation from the performance level has to be achieved, the knowledge engineer modifies the rules and iterates the process again.

The advantage of this approach is that, for a skilled and experienced engineer, a well-performing system is not hard to develop. The downside is that the process is labour-intensive and identifying some of the changes and modifications that are required for improving the process could be difficult.

Learning/Training

In contrast to the knowledge engineering approach, automated learning/training does not require someone with detailed knowledge of the IE system or the ability to write rules for the extraction. He must simply have enough domain knowledge about the addressed topic in order to be able to generate a text corpus and to annotate the text appropriately. This annotated text is the input to the system that runs a training algorithm on the text. The system gathers knowledge from the annotated text and uses this knowledge on new (not annotated) text sources in order to extract relevant information. The advantages and disadvantages of the automatic learning/training approach are complementary to those of knowledge engineering. System expertise is not required, the data-driven rule acquisition ensures full coverage of the examples and it is more portable to a different domain.

The disadvantages are that, sometimes, training data may not exist or is not easy to obtain—for example, if one wants to develop an IE system for a domain in which there

are just a few relevant examples in the training corpus. Moreover, if the relations that are searched for are very complex, then it may be difficult to produce sufficient annotations for a training corpus that would meet the requirements of a proper evaluation.

2.2 Natural Language Processing

NNatural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications“(Liddy, 2001, p. 1).Research in the field of NLP has been active since the beginning of the modern computational age, which started in the early 1950s, but the field has become increasingly popular in recent years (Gil et al., 2013, 150).

In the domain of NLP, different methods exist to extract information from text sources. The main distinction is (again) between the **rule-based** and **statistical/machine learning** approaches. Both approaches have their advantages and disadvantages but, in recent years, the statistical and machine learning techniques have become increasingly dominant (Chiticariu et al., 2013).

According to (Chiticariu et al., 2013), in actual scientific research, statistical or machine learning approaches for NLP are more popular. Between the years 2003 and 2012, 75% of the published scientific papers in the field were related to machine learning NLP approaches, 21% to hybrid solutions and the rest to rule-based ones. The situation is different in the commercial world: According to the industry survey from (Chiticariu et al., 2013), the industrial landscape does not reflect the research efforts of recent years. Considering the large number of vendors, the fraction of rule-based solutions is around 67%.

Both approaches have their advantages and disadvantages, and even if rule-based solutions are not popular in scientific research, they seem to work well for business applications.

Comparison of Approaches

Statistical/machine learning classifiers are highly robust, can easily be trained and necessitate little supervision during learning, but require labelled training data, retraining for domain adaptation and often suffer from poor generalization when data is insufficient. Statistical methods also do not perform well when something with a high sparsity has to be detected, as will be the case in this thesis. Grammar-based robust parsers are expressive and portable, can model the language in terms of granularity, and are easy to modify by hand in order to adapt to new language usages (Wang et al., 2002).

Although grammars are learnable, they often require more supervision, are more difficult to maintain and demand a lot of manual labour. While they can yield an accurate and detailed analysis when a spoken utterance is covered by the grammar, they are less robust for those sentences that are not covered, even with robust understanding techniques. Due to these reasons, statistical classifiers are often used for broad and shallow understanding, and robust parsers are frequently used for narrow and deep understanding in a specific domain where grammars can be crafted carefully to cover as many usages in the domain as possible (Chiticariu et al., 2013). The approach that was chosen for this thesis will be described in detail in the *Solution Approach* chapter.

2.3 Tasks and Approaches in NLP

This section focuses on common tasks in NLP in order to explain how knowledge is extracted from text sources.

Named Entity Recognition

Two fundamental tasks in IE are named entity recognition (NER) and relation extraction (Jiang, 2012). The purpose of NER is to detect single words, or sequences of words that represent a real-world entity like 'Apple Inc.' or 'Steve Jobs' and determine if they are a *person* or an *organization*. Usually, this non-trivial task cannot be solved using simple string matching algorithms, because a named entity could also be context dependent and, for solving the problem of NER two different approaches exist, a rule-based and a statistical learning solution (Jiang, 2012).

Rule-based NLP

The general idea of rule-based solutions is that a set of rules is manually defined (by a knowledge engineer) or automatically learned. Each token in the text possesses a set of features, that are compared to a rule. The structure of such a rule is quite simple: It consists of a pattern and an action that must be taken if the input matches the rule pattern. The following example, taken from (Jiang, 2012, p. 17), illustrates how to label any sequence of tokens of the form "Mr. X "here X is a capitalized word as a person entity. The following rule can be defined:

```
(token = "Mr." orthography type = FirstCap) –> person name
```

On the left-hand side of the rule is a regular expression that fires if a sequence of two tokens occurs, where the first one is equal to the string "Mr"and the second token has the orthography type *FirstCap* (first letter of a word has to be a capital letter). The right-hand side of the rule indicates that such a token sequence should be labelled as *person name*, which suggests that the name of a person has been found.

The manual creation of such rules is very labour-intensive and demands human expertise and domain knowledge. Besides the manual development of rules, an automatic approach is also possible. In the case of automatic learning, two different methods exist: top-down and bottom-up. The similarity between these approaches is that both require a set of training documents with manually labelled entities (Soderland, 1999). The downside of automatically generated rules are that they tend to have a lower precision than manually created ones.

Statistical Learning NLP

NER based on statistical learning is a sequence labelling problem. Sequence labelling is a general machine learning problem and has been used for many natural language tasks, like parts of speech tagging (which will be discussed in the following section) or NER.

A sentence can be seen as a sequence of observations and each observation is represented as a feature vector. The aim is to assign a label to each observation and to map NER to a sequence labelling problem. Each word in a sentence is treated as an observation. The class labels have to clearly indicate the boundaries and the types of named entities within the analysed sequence. More recent work on NER uses statistical machine learning methods. Name finders can be based on hidden Markov models (Rabiner and Juang, 1986), entropy models (Chieu and Ng, 2002) and maximum entropy Markov models (Bender et al., 2003), and support vector machines (Isozaki and Kazawa, 2002) have been applied to NER.

Parts of Speech Tagging

Parts of speech (POS) tagging describes a process in which each word in a sentence is assigned to a contextually appropriate grammatical descriptor Voutilainen (2003). These classes are content words (like nouns, verbs, adjectives, adverbs) and function words (like pronouns, determiners, qualifiers, etc.). POS tagging is a crucial task in IE and important for word sense disambiguation. For the task of POS tagging, a set of tags needs to be chosen. One of the most common tag sets that is used is the PEN Treebank dataset (Marcus et al., 1993), which contains 36 POS tags and 12 other tags (for punctuation and currency symbols) [1]. POS tagging is not a trivial task and has to face some difficulties and challenges. The example, taken from (Nof, 2009, p. 262), considers the following sentence:

Flies like a flower.

The words in this sentence have several disambiguate traits:

- Flies: noun or verb?

[1] http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html

- like: preposition, adverb, conjunction, noun, or verb?

- a: article, noun, or preposition?

- flower: noun or verb?

Different approaches exist to solve the difficulties of POS tagging, and the most notable ones are rule-based, stochastic and transformation-based learning approaches. Rule-based taggers assign a tag to each word using a set of handwritten rules. These rules could specify, for instance, that a word which is following a determiner and an adjective must be a noun. The consequence of such an approach is that the set of rules must be properly written and checked by human experts. The stochastic (probabilistic) approach uses a training corpus in order to pick the most probable tag for a word. Many probabilistic methods are based on first-order or second-order Markov models. There are a few other techniques that use probabilistic approaches for POS tagging, such as the Tree Tagger (Schmid, 1994).

Finally, the transformation-based approach combines the rule-based approach and the statistical approach. It picks the most likely tag, based on a training corpus, and then applies a certain set of rules to see whether the tag should be changed to anything else or not. It saves every new rule that it has learned in the process for future use (Hasan et al., 2007). An example for an effective tagger from this category is the Brill tagger (Brill, 1992), which will be discussed later.

Rule-based POS tagging is the oldest approach and uses handwritten rules to identify the correct tag when a word has more than one possible tag. Disambiguation is done by analysing the linguistic features of the word, its preceding word, its following word and other aspects. For example, if the preceding word is an article, then the word in question must be a noun (World Of Computing).

Statistical/stochastic taggers work on the basis of a simple algorithm: For each word, the tag with the highest likelihood is assigned. For example, if a word is more often used as a noun than as a verb, it will be assigned with the tag noun. Before such a tagger is able to fulfil this task, it needs to be trained on a training corpus, which is used to determine which tags are most common for a word. Hidden Markov models try to maximize the probability of a suitable tag sequence for a given sequence of words. They also take into account the surrounding context of a word—for example, a word is more likely to be tagged as a noun if it comes directly after an article because the probability is much higher in this context. Such models maximize the following formula:

$$P(word|tag) * P(tag|previousntags) \tag{2.1}$$

Brill tagging is transformation-based learning. The general idea is to guess the tag of each word, then go back and rectify previous errors. In this way, a Brill tagger transforms a bad tagging of a text into a better one. It is a hybrid approach, as it starts by using statistical techniques to extract information from the training corpus and then uses a program to automatically learn rules which reduce the faults that would be introduced by statistical mistakes (Brill, 1992). The process behind this method is called transformation-based error-driven learning (TEL) (Brill, 1995).

Figure 2.1: Transformation-based error-driven learning (TEL) (Brill, 1995)

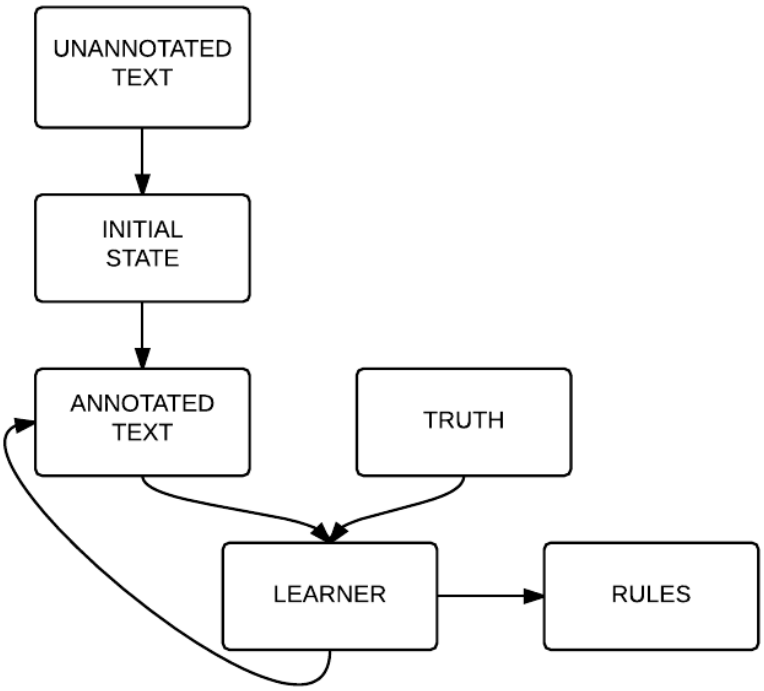


Figure 2.3 illustrates how the process of Brill tagging works. Unannotated text is passed trough the initial-state annotator, which assigns tags to the input. The output of the initial-state annotator is a temporary corpus that is then compared to a manually tagged goal corpus. Each time, the temporary corpus is passed through the learner. The learner produces one new rule, and the single rule that improves the annotation the most (compared with the goal corpus) is chosen and the result replaces the temporary corpus. Through this repetitive process, the learner produces an ordered list of rules.

Performance of POS Taggers

The state of the art in POS tagging methods and its implementations reaches very high accuracy measures. The performance is usually in the area of 97% accuracy (Association for Computational Linguistics, 2014) and even the ANNIE POS tagger, which is used in this work, has an accuracy of the same level (Hepple, 2000).

2.4 Relation Extraction

Relation extraction describes the task of detecting semantic relations between entities in a text. For example, the following sentence

```
Bill Gates works at Microsoft Inc.
```

would lead to the relation

```
Person-Affiliation\add{ }(Bill Gates, Microsoft Inc\add{.})
```

Much of the work on relation extraction is based on the task definition from the Automatic Content Extraction (ACE) program (Doddington et al., 2004). ACE focuses on binary relations—i.e. relations between two entities—which are also referred to as arguments. A set of major relation types and their subtypes are defined by ACE. Examples of ACE major relation types include physical (e.g. an entity is physically near another entity), personal/social (e.g. a person is a family member of another person), and employment/affiliation (e.g. a person is employed by an organization). A different domain of application is the detection of dependency trees in sentences, which is used in the Stanford Parser (De Marneffe et al., 2006).

According to (Cowie and Lehnert, 1996b), common tasks on different levels in the field of IE are:

- Word Level
  Mark words with their part of speech (POS tagging) (Voutilainen, 2003); usually carried out by using statistical methods that are trained from pre-tagged text.

- Noun Phrase Level
  Recognizes major phrasal units in the domain and marks them with a semantic information.

- Inter-sentence Level
  Recognizes and unifies referring expressions; detection of links between previously extracted named entities (co-reference).

Dependency Parsing

Despite a long and venerable tradition in descriptive linguistics, dependency grammar had a fairly marginal role in both theoretical linguistics and NLP.

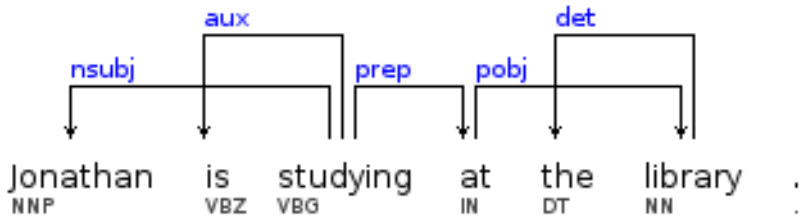
"The increasing interest in dependency-based representations in natural language parsing in recent years appears to be motivated both by the potential usefulness of bilexical relations in disambiguation and by the gains in efficiency that result from the more constrained parsing problem for these representations."(Nivre, 2005, p. 1).

The basic idea in this field is that a syntactic structure consists of lexical items, which are linked by binary asymmetric relations called *dependencies*. (Arrivé, 1969) described the grammatical relations between words and their neighbours. The mind perceives connections; these in totality form the structure of the sentence. The structural connections establish dependency relations between the words and each connection in principle unites a superior term and an inferior term. The superior term receives the name *governor* or *head* and the inferior term is called *subordinate* or *dependent*.

Dependency Graph

The already-mentioned relations between words in sentences (dependency structure), can be expressed and visualized with the help of dependency graphs. Such a graph can be formalized as directed graph *G=(V,E)* with a set of nodes V and a set of arcs E. The nodes are labelled with the word form and the arcs are labelled with the dependency types. To provide a better understanding, in Figure 2.2, a dependency graph of an example sentence is illustrated. The example sentence contains the following relationships:

Figure 2.2: Dependency graph; source: author's illustration



- nsubj(studying-3, Jonathan-1)
  nominal subject; this is a noun phrase which defines the syntactic subject of a clause.

- aux(studying-3, is-2)
  auxiliary; is a non-main verb of the clause.

- prep(studying-3, at-4)
  prepositional modifier; a prepositional modifier of a verb, adjective, or noun is any prepositional phrase that serves to modify the meaning of the verb, adjective, noun, or even another preposition.

- det(library-6, the-5)
  a determiner is the relation between the head of a noun phrase (NP) and its determiner.

- pobj(at-4, library-6)
  object of a preposition; the object of a preposition is the head of a noun phrase following the preposition, or the adverbs “here“and “there“.

Stanford Dependencies

In this work, the Stanford Parser (De Marneffe et al., 2006) and the Stanford typed dependencies representation (De Marneffe and Manning, 2008) are used, which were designed to provide simple representations of the grammatical relationships in a sentence. These representations can be understood and effectively used by people who lack linguistic expertise, but want to extract textual relations. In particular, instead of the phrase structure representations that have long dominated the computational linguistic community, it represents all sentence relationships uniformly as typed dependency relations and, currently, there are approximately 50 different grammatical relationship types.

In comparison to other parsers like the MiniPar Parser (Lin, 1999) and the Link Grammar Parser (Sleator and Temperley, 1995), the Stanford Parser achieves better results on the same evaluation set (De Marneffe and Manning, 2008).

2.5 Sentiment Classification and Analysis

Sentiment classification is a branch of NLP and a new research field that has gained a lot of attention in the scientific community in recent years (Trindade et al., 2013). The term ‘sentiment’ is related to feelings, attitudes, emotions and opinions [2]. There was always a huge interest from scientific research, pollsters and marketing departments regarding the emotional attitudes that people express their opinion on—for example, political developments or how satisfied they are with a product of a company. The growing volumes of opinion-rich resources (like blog sites, social media, micro blogging, etc.) represent new opportunities and challenges that arise since people now can, and do, actively use information technologies to seek out and understand the opinions of

[2] http://dictionary.reference.com/browse/sentiment

others (Pang and Lee, 2008). The main task in this field is to analyse textual documents and to detect subjective expression within these documents. According to (Bhuiyan et al., 2009), the research field of sentiment classification can be divided into two main directions—*sentiment classification* and *feature based opinion mining*. *Sentiment classification* investigates possibilities in order to classify documents as positive, neutral or negative; the level of analysis can also be carried out on the sentence level, where each sentence is classified into the categories mentioned earlier. *Feature-based opinion mining* focuses on the phrase level of a text and aims to extract the meaning of the opinion holder about certain features of an object. This is done by setting the features in relation to opinion-bearing words that make it possible to determine a sentiment orientation for every feature.

2.6 Lexical Databases

Lexical databases are useful tools in the field of NLP and sentiment analysis. They provide semantic information about words and their relation to other words in a certain language. Two common lexical databases are WordNet (Miller, 1995) and SentiWordNet (Esuli and Sebastiani, 2006b).

WordNet

WordNet is a lexical database for the English language. It contains nouns, verbs, adjectives and adverbs that are organized into sets of synonyms (synsets) and semantic relations that link the synonyms. WordNet contains more than 166,000 word forms and more than 90,000 different word senses. The relations of words within WordNet are:

- Synonymy
  words with the same or similar meaning; is a symmetric relation between word forms.
- Antonymy
  opposing name; is used to organize the meaning of adjectives and adverbs
- Hyponymy
  sub name; are transitive relations between synsets
- Meronymy
  part name and its inverse (holonymy and whole name) are semantic relations.
- Troponymy
  manner name; is to verbs what hyponymy is to nouns
- Entailment
  relations between verbs

SentiWordNet

Like WordNet, SentiWordNet is also a lexical resource, but SentiWordNet was developed for the domain of opinion mining (Pang and Lee, 2008). Opinion mining (or sentiment analysis) aims to detect the mood in text documents that express opinion statements, like product reviews, Twitter posts and blog articles. For this task, word relationships and grammatical dependencies are used, like in NLP, but in order to extract further semantic information, resources like SentiWordNet are needed. SentiWordNet is based on the synsets of WordNet by defining the PN polarity (positive-negative polarity) (Esuli and Sebastiani, 2006a) and the SO polarity (semantic orientation polarity) (Esuli and Sebastiani, 2006a). SO polarity describes if a text has a factual nature—in other words, it describes a given situation or event, without expressing a positive or negative opinion. In SO polarity, text is categorized into the subjective and objective categories. PN polarity means to decide if a given subjective text expresses a positive or negative opinion in its subject matter. In particular, each synset is assigned three sentiment scores: positivity, negativity, objectivity.

Deep Learning Sentiment Analysis

Most sentence-level and even document-level classification methods are based on the identification of opinion words or phrases. There are basically two types of approaches—*corpus-based approaches* and *dictionary-based approaches*. Corpus-based approaches find co-occurrence patterns of words to determine the sentiments of words or phrases (Hatzivassiloglou and Wiebe, 2000). Dictionary-based approaches use synonyms and antonyms from WordNet to determine word sentiments based on a set of seed opinion words (Bhuiyan et al., 2009). A new approach is to use deep learning methods for sentiment classification.

Deep Learning

Deep learning refers to a relatively recently developed set of generative machine learning techniques that autonomously generate high-level representations from raw data sources and, using these representations, can perform typical machine learning tasks such as classification, regression and clustering.

The main purpose of deep learning is to move machine learning closer to artificial intelligence and to mimic the efficiency and robustness by which the human brain represents information (Arel et al., 2010). The phrase ‘deep learning’ describes a set of techniques for the learning in neural networks and outperforms, at the moment, other solutions in the fields of image/speech recognition/classification and NLP.A major problem is that most classification applications demand pre-processed data, which leads to the problem that the intelligence behind many pattern recognition solutions have an human-engineered feature extraction process (Duda et al., 2012). In order to develop more intelligent classification systems, research from the field of neuroscience has to

be considered, such as how the brain processes and governs information. A key finding from the field of neuroscience is that the neocortex, the part of the brain which is responsible for cognitive abilities, does not pre-process signals. In the neocortex, the signals are processed through a complex hierarchy, which learns over the time how to represent observations based on the regularities that are exhibited (Lee and Mumford, 2003).

How can these theoretical hierarchies and abstractions be realized? (Bengio, 2009) wrote in their paper that theoretical results suggest that, in order to learn the kind of complicated functions that can represent high-level abstractions, one may need deep architectures. Deep architectures are composed of multiple levels of non-linear operations, such as in neural nets with many hidden layers, or in complicated propositional formulate re-using many sub-formulae. Searching the parameter space of deep architectures is a difficult task, but learning algorithms such as those for deep belief networks (Hinton, 2009) have been proposed to tackle this problem of beating the state-of-the-art in certain areas. Besides deep belief networks, there are also convolutional neural networks that are well-established in the deep learning field and show great promise for future research (Arel et al., 2010).

Stanford Sentiment Analysis

Most approaches in sentiment analysis use bag-of-words representation (Pang and Lee, 2008). Bag-of-words classifiers can work well on longer documents by focusing on words with strong sentiment like “awesome“or “exhilarating“. However, sentiment accuracy even for the binary positive or negative classification of single sentences has not exceeded 80% for several years. For the more difficult multi-class case, which includes also a neutral class, accuracy is often below 60%, when short messages on Twitter are analysed (Wang et al., 2012). (Socher et al., 2013) proposed a new recursive neural tensor network for a fine-grained classification into five sentiment classes.

- very negative
- negative
- neutral
- positive
- very positive

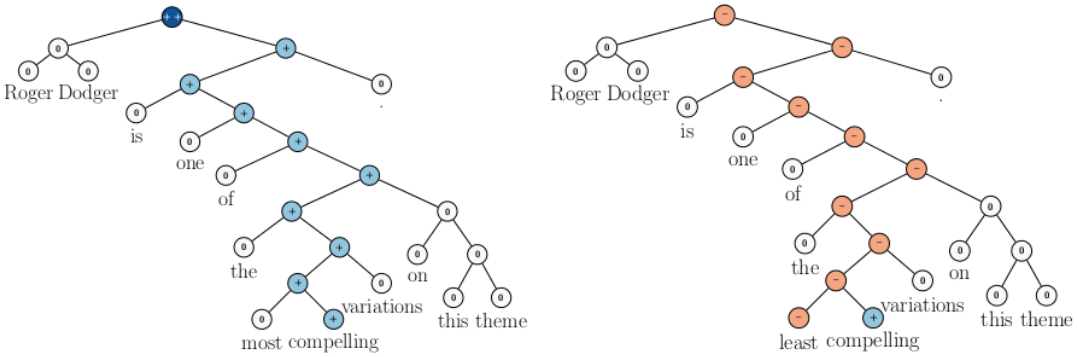
This approach pushes the state of the art in sentiment detection on the sentence level from 80% accuracy up to 85.4% for positive/negative classification. Finally, according to their own paper, their model is the only one which can *accurately capture the effects*

of negation and its scope at various tree levels for both positive and negative features. For example, the sentence

Rodger Doger is one of the most compelling variations on this theme.

has a very positive sentiment orientation. If the word 'most' was changed to 'least', the entire meaning of the sentence would change. This change is also detected by the Stanford sentiment solution, which can be seen in Figure 2.3.

Figure 2.3: Prediction of positive and negative sentences; source: author's illustration



2.7 Semantic Tagging

Semantic tagging, or semantic annotation, is a method for retrieving and attaching additional information to entities within a text. This additional information could be attributes, comments or descriptions that provide context about the entity, which is one of the major advantages of the Semantic Web. The Semantic Web is all about adding formal structure and semantics (metadata and knowledge) to web content for the purpose of more efficient management and access. Since the realization of this vision depends on the presence of a critical mass of metadata, the acquisition of this metadata is a major challenge for the Semantic Web community (Kiryakov et al., 2004).

2.8 Incident and Event Detection with NLP

This section presents the current state of research in the detection of incidents from news sources. The focus is on the kinds of solutions that exist and how similar challenges, that might also occur in this thesis, have been solved in other solutions and domains.

NLP Analysis of news sources and for incident detection

The detection and identification of sustainability-related issues and, besides sustainability, the detection of incidents, events and risks is an emergent topic in academic research. At the moment, there is no published work in the domain of text mining that concerns the available sustainability but, in general, the field of data mining and extracting knowledge from text sources is quite popular. Some publications with similar aspects and challenges, as in this work, will be presented here.

The text mining of news has already been used to analyse the relation between the content of business news and long-term market trends by classifying news into positive and negative categories and how their ratio behaves in the context of long-term market trends (Kroha et al., 2006).

A further application is the detection of weak signals for long-term business opportunities. These weak signals are defined as imprecise and early indicators of impending important events or trends, which are considered key to formulating new potential business items. (Yoon, 2012) presented a quantitative method that identifies weak signal topics by exploiting keyword-based text mining. Different approaches are already available, even in the field of risk management, although many of them are applied to the domain of financial news. (Cheng, 2010) used the text mining of news for forecasting the change in intraday stock prices by determining characteristic words through their high frequency of appearance in a news article. A more advanced approach that does not rely merely on very simple textual representations, such as the bag-of-words model (Zhang et al., 2010), has been used by (Hagenau et al., 2013). The list of words used for text representation is created either on the basis of dictionaries or retrieved from the message corpus based on actual occurrences of the words. They examined financial news using context-capturing features for stock price prediction. Their approach allows selecting semantically relevant features of the text.

Due to the increasing amount of information all around the world, text mining is a useful method for keeping pace with newly published information. (Atkinson and Van der Goot, 2009) presented a solution for *near real-time information mining in multilingual news*; the purpose of their work was to enable media impact analysis—for example, how often one topic is reported in different newspapers worldwide. In order to determine the entities (persons, companies, ...), topic, location and date of the reported news, they used NER for the detection of companies and also detected homonyms by implementing a multilingual disambiguation module. Unfortunately, a precise explanation about how they implemented this part of their work was not provided. Due to the usage of concepts like conceptual-semantic and lexical relations, which are also important in this thesis, it would have been of interest. Even if the addressed problem is different than in this thesis, the methodology is familiar because of used methods like named entity detection and the usage of word relations. In the domain of news classification, microblogging streams are very popular. Due to their restriction in the length of a message,

they are very clearly formulated and hashtags could be an indication for the topic of the tweet. (Weng and Lee, 2011) formulated an approach for event detection on Twitter[3]. They presented a sophisticated statistical approach that, in its baseline, works on the frequent occurrence of words and has achieved good performance levels by doing so.

Another example for event detection has been presented by (Toyabe, 2012). They were using NLP on electronic media to detect in-patient falls in order to support incident reporting in hospitals. The aim of their study was to determine if it is possible to promptly detect serious injuries after in-patient falls and to determine which data source is more suitable for this task. In addition to electronic sources, they used incident reports, discharge summaries, progress notes and image-order entries. Image-order entries contain brief information about a possible diagnosis. They have chosen a rule-based approach (as in this thesis) and used a set of incident reports as a training set to generate approximately 170 decision rules. The F-measure results for detecting falls was 0.12 when using progress notes, 0.24 for discharge summaries, 1.00 for incident reports and the result of image-order entries was 0.91. Since the decision rules were trained on the incident reports, the high result in not a big surprise and results can differ depending on which input source is chosen, how precise it is and the domain on which they focus. The last two cited papers both concerned event identification. (Weng and Lee, 2011) used a statistical and (Toyabe, 2012) a rule-based approach. The latter uses text sources that are extensive and not restricted like the posts on Twitter, which makes it more similar to the problem addressed in this thesis.

Research in Semantic Interpretations

Especially the field of semantic interpretations of words is of interest for this thesis and the state-of-the-art solutions used for carrying out a semantic classification are WordNet (Miller, 1995) and SentiwordNet (Baccianella et al., 2010).

SentiWordNet is useful for evaluating positive or negative statements in texts. (Rill et al., 2012) has used SentiWordNet for building an opinion value list with both, opinion bearing adjectives and adjective-based phrases. (Wogenstein et al., 2013) has also used the same approach in a test scenario. Their results point out that the detection of positive phrases is satisfying but, in the case of negative phrases, just 14% of the phrases has been detected correctly in the evaluation. The reasons for these results are manifold, like indirect expressions in the text or incorrect opinion values in the SentiWordNet lexicon.

[3] https://twitter.com/

CHAPTER 3

Solution Approach

In this section, the methodology for the implementation is described and the formal solution approach for the first research question will be given.

3.1 Methodology

The purpose of the used methodology is the indentification of environmental incidents, for doing so it is necessary to identify the following information within the text sources.

- Words and/or phrases that can be considered as being from environmental nature.
- Names of companies and organisations.
- Classification if a sentence has a positive or negative Sentiment.
- Detection of a relation between the environmental word and the detected organisation.

The focus of this thesis is the analysis of blog articles that are published on environmental blog sites. The reason for using these articles is the assumption that blog posts may report events which are not covered in the headlines of popular news websites. The formal methodology used here involved the processing of tasks in the following order:

- Content extraction
- Information extraction
    - Determination of an environmental context

- Named Entity Recognition
- Sentiment classification on the sentence level
- Dependency parsing

The methodology has evolved and changed step by step due to the knowledge engineering approach. The first attempt was the determination of the environmental context and the identification of companies in the text content. Dependency parsing and sentiment classification are attempts to improve the performance. A step-by-step evaluation of the results will be given in Section 7.

3.2 Text Extraction

Before the analysis of text is possible, the text has to be collected from different sources. The system collects blog articles from various blog websites via RSS feeds (Winer, 2002). Some of these feeds already provide the text content directly via the feed, but most of the websites do not provide the content through the RSS channel; hence, the content needs to be processed directly from the news website. Extracting the main content from a news website is not a trivial task because websites have different style schemes. A parser for this task is the library boilerpipe (Kohlschütter et al., 2010), which has been used in this thesis. This Java library provides methods to extract the text content from news websites in order to define a corpus for the NLP task.

3.3 Incident Detection

For the incident detection, methods from the domain of NLP are used. The core element for this task is carried out by using the GATE (Cunningham et al., 2011) suite, which provides the possibility to build processing pipes or applications that consist of different processing resources. A processing resource is an element in an application that is responsible for a certain task, like a POS tagger.

The application that is built in GATE is based on an ANNIE (Cunningham et al., 2002) default application, extended with custom processing resources (PRs) and JAPE (Cunningham et al., 2000) rules.

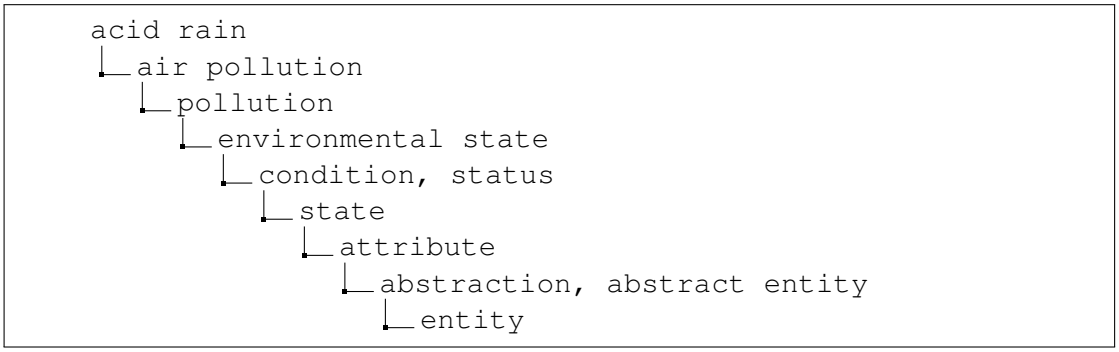
Detection of Words in an Environmental Context

The extraction of a sustainability context from text is a difficult task. A lexical approach that focuses on the usage of domain-specific dictionaries would be a possible solution. For example, if a word from a domain-specific dictionary matches a word from the analysed input text, an indication for a domain-specific context is given. The problematic aspect is that such a specific dictionary is necessary; in the absence of such a list, it is necessary to have access to the knowledge of an expert who would be able to specify

such a word list, even if it meant that this would be a highly labour-intensive task. The approach in this thesis is similar to a lexical approach but needs a generic definition for the domain of interest by using the capabilities of the word taxonomy and lexical database WordNet (Miller, 1995).

How this approach works will be shown through an example. WordNet offers the concept of hyponymy; a hyponymy word shares a type of relationship with its hypernym. For example, *owls*, *flamingos*, *parrots* and *storks* are hyponyms of *bird* (which is their hypernym) and *bird* is a hyponym of *animal*.
This approach also works for terms within the environmental domain. For example, the phrase *acid rain* has a hypernym hierarchy[1] that is shown in Figure 4.2.

Figure 3.1: Hypernym hierarchy of acid rain



The hierarchy starts from a finer granularity and leads to a coarser one, and each level includes more words from different domains. The last level can be seen as a root node of a tree and each level down the tree is a node with multiple edges leading to new nodes. The advantage of this approach is that it is not necessary to find every word or synonym for *acid rain* or words that describe air pollution or pollution. It is necessary only to search for words with the hypernym 'air pollution' in order to find words or expressions that are related to this concept. The only problem is that a user must pay attention to which level to choose, because the more general the term becomes, the more words of different domains could be included. In the hierarchy from the example of *acid rain*, the level with *condition, status* would be too coarse a definition and would open paths in the hierarchy tree to terms that are related to very different domains like, for example, human health or relationships.

So far, just the hierarchy of nouns has been considered, but adjectives or verbs with an environmental context are also of interest. In WordNet, adjectives and verbs are related to nouns via the *derivationally related form* relation. For example, the adjective *radioactive* is related to the noun *radioactivity*, and this noun can then be checked to see if it is a part of the hierarchy, as described earlier.

[1] http://wordnetweb.princeton.edu/perl/webwn

Detection of an Organization in Context

After the identification of words with an environmental context, the focus is on the detection of companies. The companies need to be detected and checked to see if they are linked to an environmental word or phrase. This is necessary in order to address the environmental sustainability risk aspect of this work and, for example, to provide the information a certain company is linked to an environmental event. The identification of a company in a text can be done by using NER techniques, which have been addressed in the second chapter of this work. Possible approaches for the detection of links between an organization and an environmental incident are the following:

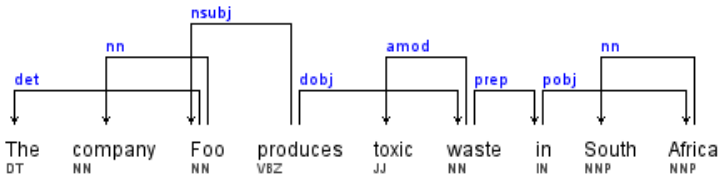
Distance-based Approach

In the first approach, the connection is determined by using a distance-based method. For example, if a negative phrase occurs in a distance of zero to five words, it can be assumed they belong to each other and a sustainability incident is defined. (Hagenau et al., 2013) have used the same distance-based feature extraction approach for determining stock price predictions based on financial news.

Dependency Parsing

In this case, the detection of an environmental incident is done by dependency parsing. For this task, the Stanford Dependency Parser (De Marneffe et al., 2006) is used. The task of the parsers is to work out the grammatical structure of a sentence and the grammatical relations between the words in the sentence. To provide a better understanding, here is a small example:

Figure 3.2: Dependency parsing Source:author's illustration



As we can see in Figure 3.2, the environmental term is *toxic waste*, which was discovered in the previous step. Between *toxic* and *waste*, there is an *amod* relationship which states that it has an adjectival phrase modifying the meaning of *waste*.If we take a look at the company *Foo*, it is possible to extract a tree stating *Foo produces waste*. Starting from *produces*, there are two relations: the *nsubj* pointing at *Foo* and *dobj* which is pointing at waste. This relation indicates that *Foo* is the subject and *waste* the object. This makes it possible to link to the company that produces the *toxic waste*.

How about the dependency in very long sentences, where a large distance is between both words? In Figure 3.4 one can see the dependency relations in a very long sentence. With an ordinary approach, the sentence would be considered an environmental incident because it contains a company and an environmental term at the end of the sentence. If we examine the path of relations from the company *Foo* to the environmental term *pollute*, we can assume that there is no connection if the other word cannot be reached or a large number of steps is needed to reach the other term.

For the extraction of the dependencies and in order to determine the distance between both words, several steps are necessary:

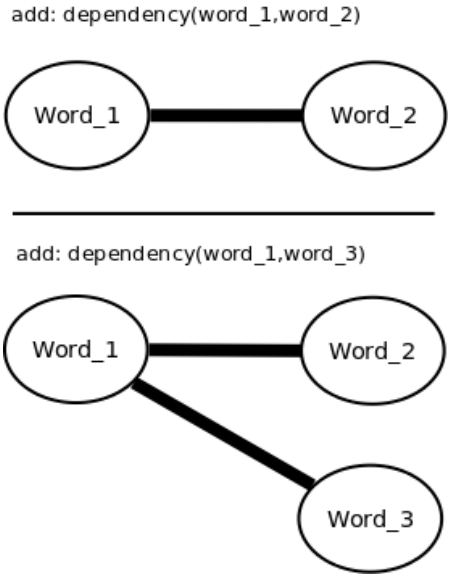
- Extraction of the dependency relations by using the Stanford Dependency Parser.
- Generation of the graph, as illustrated in Figure 3.3.
- Path detection using the Djikstra algorithm (Skiena, 1990). This is a graph search algorithm for the detection of a (shortest) path between two vertexes.

In the second step, the output of the dependency parser has the form *dependency(word1, word2)*. *word1* and *word2* are the nodes and the dependency relation between these two words indicates that a link between them does exist. If a word is already present in the graph, like by adding *dependency(word1,word3)*, *word1* is not added to a graph, and only a new link is established to the new node *word3*.

Regarding the path detection using Djikstra, the main criterion in this step is to detect if a path exists in the graph between the company mentioned in the sentence and the environmental term. If a path exists, it is also of interest how many nodes are between the starting point and the destination point.

Usually, the typed dependencies from Stanford are mapped in a straightforward manner onto a directed graph representation, but the graph for the detection of the path between both points needs to be undirected because, most times, the company and the environmental word are the end nodes of the directed edge, and a path detection starting from an end point with no outgoing directed graph is not possible.

Figure 3.3: Dependency detection graph; source: author's illustration



3.4 Sentiment Extraction

Sentiment classification is used in order to extract the subjective polarity of a sentence. The reason for using sentiment classification is to determine if the previously extracted information (environmental term and company) is mentioned in a negative context and if performance can be pushed by using this additional information. For evaluation, sentiment values other than the negative ones will also be tested.

Sentences in the documents will be tagged with the following tags if the described criteria match:

Negative Sentence
Each sentence that has a *negative* and/or *very negative* sentiment score is tagged as a negative sentence. Again, the reason for this rule is that it is easier to change this one parameter in the evaluation when just *negative* or *very negative* sentiment scores are considered.

Negative Sentence with a Company
A negative sentence from the previous step is marked if it contains a company in addition to the already-mentioned criteria.

Negative Sentence with a Company and an Environmental Term
Again, the result from the previous step is marked if it contains, additionally to the company, an environmental word or phrase.

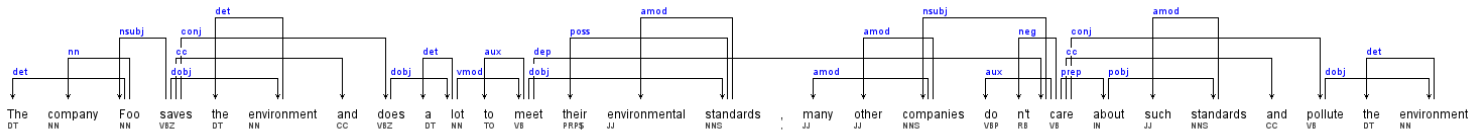


Figure 3.4: Dependency graph of a long sentence; source: author's illustration

CHAPTER 4

Proof of Concept

This chapter will explain the practical part of the master thesis and how the result of the implementation can be used. First, the approach of the implementation will be described and then the technology used and a detailed description of the system will be provided. In the last part, problems that have occurred during the development and also in the analysis of language resources will be presented.

4.1 Approach

Based on the theoretical background, which has been worked out in the section 2, a solution has been implemented that identifies environmental incidents within text documents from news blog sources.

In the first phase, the blog sites are gathered from RSS (Winer, 2002) feeds and the main content is extracted. In the next step, the extracted content is loaded into the GATE (Cunningham et al., 2011) application, where the text analysis is done. The detailed description of this GATE application will be given in Section 4.2. The application consists of a modular structure of different standard processing resources that are available in GATE as well as three custom PRs that have been developed for the purpose of environmental incident detection. Those custom PRs analyse the text for the mentioned companies and environmental words, and determine the sentiment of each sentence. The output of the system will be the sentences of the documents in which an environmental incident occurs.

The advantage of this module structure is that the application can be changed or extended without much effort. Due to this structure, different setups can be executed in the text documents for evaluation purposes. The different results of these setups will be shown in the section 5. In this thesis, the GATE application is used in a Java pro-

gram that forwards the input from the RSS feeds to the GATE application and processes the output, but it would also be possible to open and run the application in the GATE Developer application.

4.2 Implementation

In this section, the technology and frameworks that have been used for the development of the incident detection solution will be described.

Technology Stack

The developed solution is a Java application that uses several third-party frameworks, libraries and packages.

GATE

The general architecture for text engineering (GATE) is in a Java-developed open source software that is capable of solving most existing text engineering problems. It is a widely used system with many active users in academic and industrial contexts. GATE offers various solutions. In this thesis, the GATE Developer has been used, which is an IDE for language processing, and the GATE Embedded framework, which is an object library for the development of language engineering applications in Java.

The central elements in GATE are language resources (LRs) and processing resources (PRs). LRs refer to data-only resources like lexicons, thesauri or corpora. PRs refer to resources whose character is programmatic or algorithmic, like lemmatizers, generators, translators, parsers or speech recognizers. For example, a PR could annotate the names of companies in a text or determine if a word is a noun or a verb (POS tagging). PRs are available as plugins and can be added to a GATE application—i.e. an application that consists of a set of PRs that analyses and manipulates the LRs. Each PR has interfaces through which it receives the output from previous PRs. It processes the input and can propagate its own results to the next PR. The output of such a process is an XML document in which the text content is stored and the annotations have been added by the PR.

GATE is used in this work because it offers several built-in standard functionalities, its expandability through the custom processing resources and the evaluation module of GATE which is used for the work presented in Chapter 5.

OpenCalais

OpenCalais [1] is a web service that automatically creates semantic metadata to the content which is submitted to it. OpenCalais uses machine learning and NLP techniques to

[1] http://www.opencalais.com/documentation/opencalais-documentation

identify entities like persons, cities, companies, etc. The web service is free for commercial and non-commercial use.

The advantage of OpenCalais is that there is already a processing resource available in GATE and it provides more detailed information about the annotations, which allows better filtering of the results than with the standard NER and gazetteers, which are available in GATE. For example, consider if a newspaper like *The Guardian* is mentioned in a text like this:

The Guardian reports the spillage of toxic fluids.

In this work, the entity of publishing mediums like *The Guardian* are not of interest because the goal is to identify the companies that have caused the environmental incident. With the limited capabilities from GATE's standard NER, it would not be possible to determine what kind of organization *The Guardian* is, because it tags newspapers, TV stations, companies and manufacturers with the same tag; hence, making a distinction is not possible. OpenCalais offers a more granular distinction between organizations—for example, *The Guardian* would be tagged as a company, like the standard Gate PR would do, and additionally provide a tag called published media, thereby enabling a distinction. According to (Rizzo and Troncy, 2011), OpenCalais has solid performance measures. Two solutions—AlchemiApi[2] and Zemanta[3]—would outperform OpenCalais, but due to their restricted access for non-commercial users, OpenCalais is more suitable for this work.

Boilerpipe
In this work, text from new blog articles is extracted. The extraction of the main content, especially from very detailed web pages that contain a large amount of elements, can be a non-trivial task. In this work, the Java library boilerpipe has been used, which is based on the work from (Kohlschütter et al., 2010). It provides algorithms to detect and extract the main textual content of a web page. The focus of boilerpipe is the extraction of information from news websites, which is, besides the good performance measures of boilerpipe, the reason for choosing this implementation. The performance measures of boilerpipe indicate an F-measure value of 95% (Tomaz Kovaziz, 2014), and the recall and precision results are at the same value. Moreover, in comparison to other solutions, boilerpipe is one of the leading implementations (Tomaz Kovaziz, 2014).

WordNet (Miller, 1995), Stanford Dependencies (De Marneffe and Manning, 2008) and Stanford Sentiment (Socher et al., 2013) have already been described in detail in the previous section.

[2]http://www.alchemyapi.com/
[3]http://www.zemanta.com/

System Description

In this section, the developed solution will be described and the responsibility of every part of the implementation will be identified.

RSS Reader

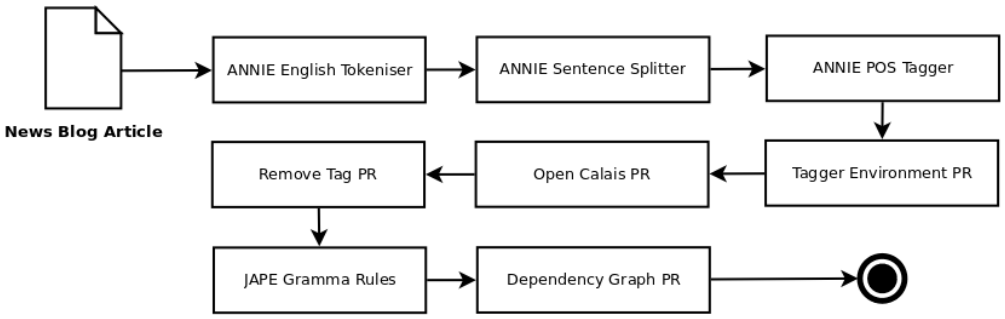
In the first step of the solution, it is necessary to gather input data. In order to achieve continuous up-to-date data, the news blog sites are accessed via RSS. RSS is a structured data format for publishing content (like news articles) from websites. An RSS feed provides a set of articles from a website with a short description and (most important) a link to the original website.

The RSS reader has been implemented in Java and uses a list of RSS feeds as an input source; it parses every RSS link, invokes the news article website and extracts the main content of the article by using boilerpipe. For evaluation purposes, the articles have been stored in a MySql (Widenius and Axmark, 2002) database, but it would also be possible to generate a corpus directly from the gathered articles and forward them to the GATE application.

GATE Application

For detecting the environmental incidents, a Gate Application, or so-called pipeline, has been developed. The elements of the pipeline will be described briefly in this section and an overview is given in Figure 4.1. The input of such a pipeline is a document in the

Figure 4.1: GATE pipeline for incident detection; source: author's illustration



text format. Each element of the pipeline processes the document and adds annotations to the text. Annotations are tags that provide meta information about elements in the text. This information is stored in an XML format. From the application in Figure 4.1, the following elements have been developed for this work: Remove Tag PR, Tagger Environment PR, JAPE Grammar Rules and the Dependency Graph PR. The pipeline can be constructed in the Gate Developer user interface. Each processing resource can be added to a pipeline and parameters can be set. Two parameters are mandatory for every

plugin; these are the names of the input annotation set and the output annotation set, and the parameters need to have the same value in all processing resources in the pipeline. The constructed pipeline is stored in an XML representation as a GAPP (GATE saved application state) file.

Listing 4.1: XML Sentence Annotation

```
<Annotation Id="11478" Type="Sentence" StartNode="703"
EndNode="832">
<Feature>
  <Name className="java.lang.String">Sentiment</Name>
  <Value className="java.lang.String">negative</Value>
</Feature>
</Annotation>
```

Listing 4.1 shows a code snipped from a processed document that is stored in an XML file. The snipped portion represents how a sentence is tagged by the system. Each annotation has the attributes ID and type. The attributes *StartNode* and *EndNode* describe the position in the document. The sentence starts at character 703 and ends at character 832 in the document. In addition, features can be stored. Usually, a feature has a *Name* and a *Value*. In this example, the name is *sentiment* and the value is *negative*, which denotes that this sentence has a negative sentiment. The set of output annotations from a processing resource in the pipeline can also be used as an input annotation set by the subsequent elements of the pipeline and the provided metadata about the documents can be used for further processing to generate more information. Each processing resource generates metadata about the text and adds its information to the XML output. Step by step more and more information is gathered. The role of the JAPE rules is to enable conclusions from the gathered information—for example, a rule which checks if a sentence contains an environmental word and a company name.

Listing 4.2: JAPE Rule Example

```
Input: Sentence EnvironmentWord OpenCalais

Rule: rule1
(
   {Sentence contains {OpenCalais._type=="Company"}}
   {Sentence contains {EnvironmentWord}}



):rl
-->:rl.SentencewithEnvCompany
     ={rule\add{ }= "Sentence with Env Word and Company" }
```

For each rule, it has to be defined which kind of tags are used as input. The rule part defines that a sentence has to contain an environmental word and a tag from OpenCalais that has the type company. If the condition holds true, then the sentence is marked with the tag *SentencewithEnvCompany*.

ANNIE Components
The first three steps in the pipeline are carried out by components from the ANNIE (Cunningham et al., 2002) IE system. The first three steps are elementary in NLP and are necessary to enable further processing of the text and provide information about sentences, words, word types, etc.

ANNIE English Tokenizer
The ANNIE English Tokenizer splits the input text into tokens and categorizes them into numbers, punctuation, space tokens, symbols and words of different types. For this purpose, the tokenizer uses grammar rules in which the LHS of the rule indicates a regular expression that has to be matched and the RHS describes which annotation has to be added to the annotation set. For example:

Listing 4.3: ANNIE Tokeniser Rule Example

```
UPPERCASE_LETTER LOWERCASE_LETTER *
> Token;orth=upperInitial;kind=word;
```

The expression in Listing 4.3 states that a sequence of letters has to begin with an upper case letter followed by zero or more lower case letters. This sequence will then be annotated as type *Token*. The attribute *orth* (orthography) has the value *upperInitial*; the attribute *kind* has the value *word*.

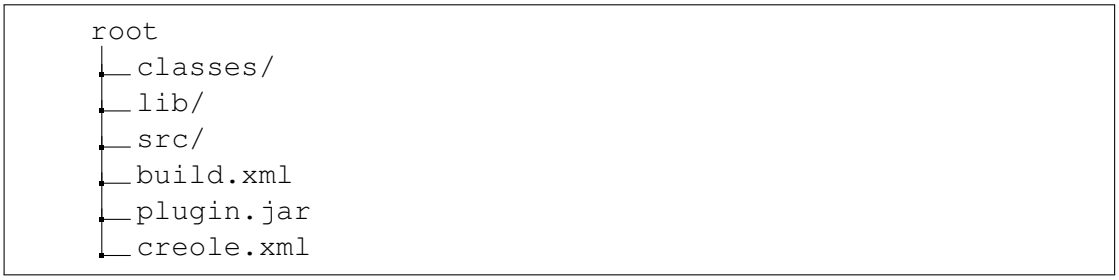
ANNIE Sentence Splitter
The sentence splitter segments text into sentences. Like the tokenizer, the sentence splitter is based on grammar rules and adds the sentence annotation to the annotation set which will be used by the following components of the processing pipeline, because this focuses on incident detection on the sentence level.

ANNIE POS Tagger
The ANNIE POS tagger, or Hepple POS tagger (Hepple, 2000), is a modified implementation of the Brill POS tagger (Brill, 1992), which uses a default lexicon and a rule set that has been acquired by training on a large corpus from the *Wall Street Journal*. The concept of POS tagging has already been addressed in this work.

Custom Processing Resources
GATE provides a large set of built-in components and functionality, but some applications require additional functionality that is not provided by these components. In this case, GATE offers the possibility to add new custom-developed processing resources to the suite. The resources that are integrated into GATE are called CREOLE (Collection of Reusable Objects for Language Engineering).

In principle, the implementation of a such a processing resource is a Java class plus an XML metadata file in the same path. It is also possible to do the resource annotation in the Java source file by using Java annotations, but in this work the configuration has been done with the help of the XML file. Furthermore, the custom PRs are Java projects that are built and defined with the help of Apache Ant [4]. The directory structure of such a processing resource is the following:

Figure 4.2: Directory structure of a processing resource



The folder *classes* contains the in the build process-compiled Java sources from the *src* folder. The *lib* folder contains the used external .jar libraries, build.xml is the ant build file and plugin.jar the builded Java container file. The creole.xml is the previously mentioned file that contains metadata about the plugin, describes information about the application (like the name or a description) and, most importantly, which resources to

[4] http://ant.apache.org/

use (e.g. libraries from the *lib* folder) or which initial parameters—e.g. the name of the input and output annotation set—have to be set before the processing resource can be executed. A detailed documentation of the CREOLE GATE component model can be found in the GATE manual [5].

Tagger Environment PR

The Tagger Environment PR is responsible for the following two tasks:

- Detection of environmental terms and phrases
- Sentiment classification on sentence level

As outlined earlier, the solution attempts to detect environmental incidents by searching for environmental terms. For some domains like technical terms, there are word lists which state that a word which is in the list is of a technical nature. For the domain of environmental words it was not possible to find a suitable list. The creation of an environmental terms list is not an easy task, because there is always the threat of missing out on important words, terms or synonyms. In this work, an approach was chosen that uses WordNet and its hierarchies of words. An additional problem are compound words because it is difficult to detect them properly. For example, *acid rain* is such a word. A possible solution is to analyse every combination of words in the sentence, which would lead to a huge increase in processing time and a very bad performance. A closer look at the structure and the grammar relations between the words in the compound word may be beneficial. It is possible that a compound word consists of two nouns, where one noun serves to modify the head noun, while a further possible relation is an adjective that modifies the meaning of the noun. These relations have been extracted with the help of the Stanford Dependencies Parser. Passing a sentence through the parser will return all grammatical relations between the words in the sentence. The relationship types *noun compound modifier (nn)* and *adjectival modifier (amod)* describe the aforementioned relations. If a compound word has one of these two grammar relations, it will be analysed whether it has a hypernym from an environmental domain.

If the detection of an environmental word or compound word is positive, it will be tagged as an *EnvironmentalWord* or *EnvironmentalPhrase*, and the tags will be added to the output annotation set, which can be used in the subsequent steps of the application pipeline.

Additionally to the detection of environmental words, the processing resource is also responsible for the sentiment classification. Since the plugin is working on the sentence level, like the sentiment classification, it is convenient to complete this operation in the

[5] http://gate.ac.uk/sale/tao/splitch4.html

same procedure. For the classification of the sentiment, the Stanford sentiment solution, which is explained in the section 2.6, has been used. The detected sentiment level is added as a feature to the *Sentence* tag. The respective tag in the XML format has been shown in Listing 4.1. The PR needs two additional parameters to be set in GATE Developer when it is added to the pipeline. The first parameter is a filepath to the local root folder of the local WordNet installation and the second one is the file path to the file that contains the defined hypernyms for the environmental domain. For the development of the sentiment classification, the Stanford Core NLP library was used.

Open Calais PR
GATE already offers a plugin for OpenCalais. For configuration purposes, just a free API key has to be requested from the OpenCalais website. This processing resource is used to retrieve and identify companies in the input text. OpenCalais tags all identified entities with a tag called *OpenCalais*. This tag also provides a property called *type*, which gives the information about what kind of named entity the tag is. The advantages and reasons for using OpenCalais have already been discussed.

Remove Tag PR
As mentioned in the description of OpenCalais, a distinction between organizational types is possible. The Remove Tag PR is necessary because, for example, an organization like *The New York Times* is also a company but is not of interest because it is a publishing medium that reports the news and does not cause the negative environmental event or situation which is of interest. For a human being, it is quite easy to make this distinction.

The advantage of OpenCalais compared to other NER solutions or gazetteers is that OpenCalais provides two tags for an organization like the *The New York Times*: a tag with the value *company* and a tag with the value *publishing medium*. In principle, the Remove Tag PR removes every *company* tag from an element that has also been tagged as a *publishing medium*. This was the only case for which a removal was necessary, because OpenCalais was precise in determining other organizational types.

Dependency PR
Additionally to the detection of environmental words/phrases, companies in sentences and the determination of sentiment levels, it is also of interest if a semantic dependency exists between the environmental word and the company. For the implementation of the dependency methodology the Stanford Dependencies are used. These provide information about the dependency types between two words. The JGraph[6] library is used for the creation of the graph structure in order to enable the path detection by using the Dijkstra algorithm. For the implementation of semantic relation parsing, the Stanford Parser 3.5 library has been used (which contains the Stanford Dependency functionality).

[6] http://www.jgraph.com/

Before these steps are carried out, each environmental incident is tagged as *IncidentSentence*, which at this point of time gives no information regarding whether a dependency relation is present. If such a relation has been discovered, a feature called *hasDependencyRelation* is added to the *IncidentSentence* annotation.

JAPE Grammar Rules
To add some additional tags for processing and especially for the evaluation and making conclusions, several JAPE rules are necessary. In order to get results at different levels for the evaluation, changes in the JAPE rules are carried out and maintained more easily than in the custom PRs.

Noun Marker
This marks different kinds of nouns with one tag; the tag is needed for the processing using Tagger Environment PR. The noun types that are marked are:

- nouns (POS tag: NN)
- proper noun singular (NNP)
- proper noun plural (NNPS)
- noun plural (NNS)
- proper noun singular (NP)
- proper noun plural (NPS)

Sentiment Sentence
Tags a whole sentence with the tag *SentimentSentence* if the *Sentence* tag property has in the rule a specified value. This seems redundant, but is necessary because the built-in evaluation in GATE works only on the tag level and cannot consider the property values from a tag.

Sentiment Sentence with Company
If sentence is tagged as *SentimentSentence* and a detected company is mentioned as well, the tag *SentimentSentenceCompany* is assigned.

Sentiment Sentence with Company and Environmental Word
If the tag *SentimentSentenceCompany* also contains an *EnvironmentalWord* tag, it will be marked as an *IncidentSentence*.

4.3 Implementation Issues

The implementation phase of the text mining system had to face some technical challenges and several factors that influenced the performance and the quality of the output. Some of these factors have been addressed in the implementation phase. Most of these problems were minor, like word-sense disambiguation or problematic text input.

Word-sense Disambiguation
One problem was the term GE, which is commonly used to describe *genetically engineered* food and, sometimes, is mentioned very frequently in a news article. The problem is that GE is also related to the company General Electric; hence, in many articles on genetically engineered food, the OpenCalais PR tags the term GE as company and as an environmental term. Thus, the term GE had to be ignored by an exception in the PR.

Boilerpipe
The extraction of the news content using boilerpipe was occasionally not perfect. On some websites, it did not extract the main content well enough. One problem was, that sometimes the description text from images was included and, in this description text, in many cases, a company or publishing company that had the copyright on this image, was mentioned. This text was not separated by a punctuation from the other content on the website, which had the consequence that the text from the image description is always attached to the next sentence. This problem sometimes created a false positive output from the system and lowered the performance. It was necessary to remove the company tag from these image descriptions.Another problem was a social media plugin that was not properly filtered by boilerpipe. Besides the buttons for sharing on several social media platforms, it also contained some tags. This combination of tags, which could be of an environmental nature, and the company names of the social media vendors sometimes led to a false positive output from the system.

CHAPTER 5

Evaluation

In this section, the evaluation of the system will be presented. First, an insight into evaluation in NLP and some concepts will be given. Second, the evaluation approach that is used in the thesis will be discussed. Finally, the evaluation results of the developed solution will be presented.

5.1 Evaluation Methods in NLP

"I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be."(Thomson, 1894, p 73)

In order to be able to judge the quality of the system, it is crucial to quantify the performance by measuring the difference between an expected result and the final result. For evaluation in the area of information extraction, different approaches exist as do different terms to describe them. Like in software engineering, there are two major types of evaluation approaches: black box and white box (Illingworth, 1997). Black box considers the relation between the input and output of a system without considering the the internal structure of the software while white box evaluation takes into account the structure of the system (Popescu-Belis, 1999).

In this thesis, the black box evaluation approach is used. The following sections will describe this approach and its methods.

Gold Standard

As stated previously, in automated black box testing, the results of a system are compared to the expected results, which (in the field of NLP and information retrieval) is referred to by the term gold standard. The gold standard is a set of documents (corpus) that was previously annotated by one (or more) human reader(s). Although the cost of producing the gold standard can be quite high, automatic evaluation can be repeated as often as needed without incurring much additional cost (on the same input data). However, for many NLP problems, the definition of a gold standard is a complex task, and can prove impossible when inter-annotator agreement is insufficient (Kumar, 2011). It is not always easy or obvious what this gold standard should be, as different people may have different opinions about what is correct. Typically, this problem can be solved by using more than one human annotator, and by comparing their annotations. This can be done by calculating inter-annotator agreement (IAA), which is also known as inter-rater reliability (Cunningham et al., 2011). In this work, only one human annotator was available, namely the creator of this work, which made the application of the before mentioned IAA not possible and definitely leads to a subjective point of view tagging of the environmental incidents in the evaluation corpus.

Performance Measures

Most of the research in information retrieval (Rijsbergen, 1979) over recent decades has been connected to the MUC (Chinchor, 1992) competitions. Hence, it is not unsurprising that the MUC (Message Understanding Conference) evaluation metrics for precision, recall, and F-measure also tend to be used, along with slight variations. These metrics have a very long tradition in the field of information retrieval. For these measures, certain terms are relevant—such as *true positive* which expresses the number of terms correctly classified by the system, while *true negative* describes the results that have been rejected correctly, *false negative* , is the number of items that has been assigned to a certain class, but that does not belong to this class, and *false negative* represents those items that have been rejected from a certain class but should belong to it.

Table 5.1: Classification

	Relevant	Non-relevant
Retrieved	true positive	false positive
Not Retrieved	false negative	true negative

Precision

Precision measures the number of correctly identified items as a percentage of the items identified. The higher the precision, the better the system is at ensuring that what is identified is correct.

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \tag{5.1}$$

Recall

Recall measures the number of correctly identified items as a percentage of the total number of correct items. The higher the recall rate, the better the system is at not missing correct items.

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \tag{5.2}$$

Measuring recall can be problematic because it is often difficult to know how many relevant records exist in a database. Often, recall is estimated by identifying a pool of relevant records and then determining what proportion of the pool the search retrieved (Jizba, 2007).

F-Measure and the Relation of Precision and Recall

The two quantities, precision and recall, trade off against each other. It is always possible to get a recall of one by retrieving all documents for all queries, but it is very likely thatthe precision is very low in such a scenario. Recall is a non-decreasing function of the number of documents retrieved. On the other hand, in a good system, precision usually decreases as the number of documents increases. In general, it is good to achieve a strong recall value while tolerating only a certain percentage of false positives.

Recall and precision are inversely related. As recall increases, the precision decreases; otherwise, as the recall decreases, the precision increases. The reason for this trade-off lies in the nature of languages. For example, if the goal is a comprehensive retrieval, it is obvious to include synonyms, related terms, and broad and general terms for each concept. A consequence of this approach is that precision will decrease because synonyms may not always be exact synonyms; hence, the likelihood of retrieving irrelevant results will increase and, in addition, broader terms may result in the retrieval of results that do not discuss the narrower search topic (Jizba, 2007).

A single measure that takes into account the trade-off between precision and recall is the F-measure, which is the weighted harmonic mean of precision and recall (Manning et al., 2008).

$$F = 2 * \frac{precision\ *\ recall}{precision + recall} \tag{5.3}$$

5.2 Evaluation Approach

In this section, it will be described how evaluation will be carried out practically. For the evaluation of the output from the proof of concept implementation, an evaluation corpus, or a so-called gold standard, will be generated.

Evaluation Corpus

The corpus for the evaluation will comprise 200 human annotated text documents. These text documents are a subset of articles from different new environmental blog sites that have been published since November 2013. The following blog sites have been used as input sources:

- The Pollution Blog
- The Guardian Environment Blog
- Grist.org
- Environmental Working Group Blog
- Friend of the Earth
- Greenpeace
- WWF
- Natural Resource Defence Council (NRDC) Blog
- Huffington Post Green Blog
- Oxfam

These blogs were chosen because they focus on environmental topics and some are maintained by well-known and prestigious environmental NGOs like Greenpeace, NRDC and the WWF. The generation of the evaluation corpus has been carried out in two steps. In the first step, a corpus was created that consisted of 100 documents. In this corpus, the likelihood of true positive and false negative environmental incidents is higher. In the second step, a true random corpus from the database has been chosen and manually annotated. The reason for this approach is that environmental incidents do not occur very often in blog articles and this sparsity would lead to a variation in the results. The first half of the corpus should measure how the system performs on articles that contain a larger set of true positives and false negatives. The second half of the corpus should give an estimation on how the results will be if a random, and not biased by pre-selection, sample is chosen.

Summarizing: The total corpus for the evaluation consists of 200 documents that have been tagged on the sentence level. Due to the sparsity of environmental incidents, only 30 documents contain such incidents. In total, the documents in the corpus have approximately 5,200 sentences, of which 88 sentences contain incidents.
The fraction of the 100 random corpus documents is approximately 2,200 sentences that contain 12 incidents.

If, according to the understanding of the human annotator, a sustainability incident occurs in one sentence, then the sentence is marked with the label *IncidentSentence*. The evaluation corpus is manually generated by a human reader and stored in the XML schema that GATE uses for documents. These documents are compared in the next step with output from the incident detection system.

Evaluation with GATE

The GATE suite offers a set of tools for the evaluation of IE results. In this work, the corpus benchmark tool (Cunningham et al., 2010) will be used. This enables evaluation to be carried out over a whole corpus rather than on a single document. And it provides detailed information regarding annotations that differ between the different versions of the corpus. The corpus benchmark tool is used to evaluate an application with respect to a gold standard. For this evaluation, it is necessary to provide the following data setup stored in one main folder:

- clean
  directory of the non-annotated documents in XML form

- marked
  directory containing the human-annotated (gold standard) document in XML form

- processed
  directory containing the annotated document output from the application in XML form

Furthermore, it is important to define the corpustool.properties file that holds the annotation set containing human-marked and system-marked annotations, which annotation types (in this case, it is the annotation type *IncidentSentence*) and which feature values should be considered.

The results of this process is the HTML document file report with statistics like how many annotations are correct, partially correct or are missing; furthermore, precision, recall and F1-score metrics are reported as well.

Evaluation Settings

For the evaluation, different settings for the incident detection solution have been chosen in order to provide an overview of how the results change when the approach for the detection of incidents moves from a general and very low detailed search approach to a more formalized one. There are three parameters which can be changed:

Semantic Dependency
Defines if a semantic relationship between the environmental word/phrase and the company is considered for the detection of environmental incidents.

Sentiment Value
Level of the sentiment value of a sentence. For the application of this thesis, the sentence sentiments *negative* and *very negative* are of interest. Besides, tests have been conducted to determine the performance level of the sentiment levels *positive*, *very positive*, *neutral* and no sentiment at all.

Company Name
For the detection of an incident, the occurrence of the name of a company is of interest. In one test case, it is evaluated how good the performance is if just the presence of a company name in a text from an environmental blog site could be an indication for an incident, due to the assumption that a blog article from an environmental blog site already has an environmental bias and no further analysis regarding an environmental context is done.

These parameters lead to the following different settings shown in Table 5.2, which were applied on the evaluation corpus. Not all variations of parameter settings are documented or tested in this work. First of all, the parameter for an environmental context is missing. This context is considered as a baseline for all other setups. A further reason is that most sentiment values were negative, and results involving different classes led to a bad performance; hence, further variations were not necessary.

Table 5.2: Evaluation: Settings Settings

	Semantic Dependency	Sentiment Value	Company Name
Setup 1	no	no	no
Setup 2	no	negative and very negative	no
Setup 3	no	very negative	no
Setup 4	no	positive and very positive	no
Setup 5	no	neutral	no
Setup 6	yes	negative an very negative	yes
Setup 7	no	negative an very negative	yes
Setup 8	yes	no sentiment	yes
Setup 9	no	no sentiment	yes

CHAPTER 6

Results

In this section, a comparison of the results is given. The results are compared based on the values for true positive, false negative and false positive as well as on those for precision, recall and F-measure. Furthermore, a distinction between sentence level and document level is made. Although the developed solution works on the sentence level, it is also interesting to see how good the performance is if the document level is considered. For example, as soon as one true positive sentence occurs in a document, the document as a whole is considered as relative.

6.1 Performance Baseline

In order to provide an estimation on the performance level, it is important to do a comparison with a primitive baseline approach. Let us consider a case on the document level and a random approach that categorizes documents into two distinct classes, *relative* and *not relative*. As a baseline comparison for results, a random approach is considered. For example, in a test set of 200 documents that includes 30 true positive documents, a random approach that categorizes documents with a probability of 0.5 as *relative* and *not relative* would, on average, detect 15 documents as true positive and 15 documents as false negative, from the 30 positive documents in the evaluation corpus. From the remaining 170 documents, 85 would be considered as false positive. By using these values, the recall, precision and F-measure can be calculated.

$$Recall = \frac{15}{15+15} = 0.5 \tag{6.1}$$

$$Precision = \frac{15}{15+85} = 0.15 \tag{6.2}$$

$$F - Measure = 2 * \frac{0.5 * 0.15}{0.5 + 0.15} = 0.23 \tag{6.3}$$

6.2 Results on the Sentence Level

Tables 6.1 and 6.2 show the results for the reduced evaluation corpus. In this reduced set, there are 100 selected documents that contain a higher amount of human-annotated true positives. Tables 6.3 and 6.4 contain the results for the full evaluation set, to which 100 documents are added that have been selected randomly.

Table 6.1: Evaluation: Results of reduced evaluation set

	true positive	false negative	false positive	Semantic Dependency	Sentiment Value	Company Name
Setup 1	60	13	288	no	no	no
Setup 2	48	14	247	no	negative and very negative	no
Setup 3	5	67	13	no	very negative	no
Setup 4	2	70	16	no	positive and very positive	no
Setup 5	0	72	0	no	neutral	no
Setup 6	41	31	39	yes	negative and very negative	yes
Setup 7	47	25	61	no	negative and very negative	yes
Setup 8	42	30	44	yes	no sentiment	yes
Setup 9	49	23	68	no	no sentiment	yes

Table 6.2: Evaluation: Performance measures of a reduced evaluation set

	Recall	Precision	F-measure	Semantic Dependency	Sentiment Value	Company Name
Setup 1	0.833	0.172	0.286	no	no	no
Setup 2	0.774	0.163	0.269	no	negative and very negative	no
Setup 3	0.069	0.278	0.111	no	very negative	no
Setup 4	0.028	0.111	0.044	no	positive and very positive	no
Setup 5	0	-	-	no	neutral	no
Setup 6	0.569	0.513	0.539	yes	negative and very negative	yes
Setup 7	0.653	0.435	0.522	no	negative and very negative	yes
Setup 8	0.583	0.488	0.532	yes	no sentiment	yes
Setup 9	0.681	0.419	0.519	no	no sentiment	yes

Table 6.3: Evaluation: Results

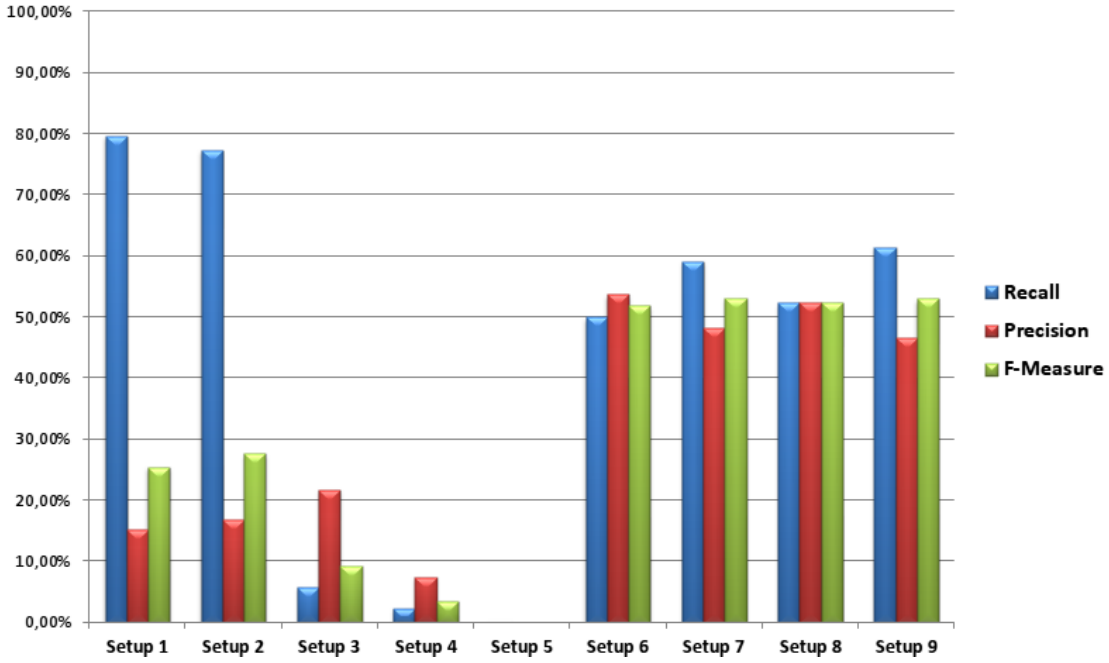
	true positive	false negative	false positive	Semantic Dependency	Sentiment Value	Company Name
Setup 1	70	18	394	no	no	no
Setup 2	68	20	398	no	negative and very negative	no
Setup 3	5	81	18	no	very negative	no
Setup 4	2	84	25	no	positive and very positive	no
Setup 5	0	86	0	no	neutral	no
Setup 6	46	40	47	yes	negative and very negative	yes
Setup 7	56	30	77	no	negative and very negative	yes
Setup 8	48	38	53	yes	no sentiment	yes
Setup 9	58	28	79	no	no sentiment	yes

The results clearly indicate that as the search approach becomes more restrictive, the precision increases and the recall decreases. Concerning the results from the full evaluation corpus, Setups 1 and 2 are more or less primitive approaches that are searching for companies in sentences. Those have achieved the highest recall values and find the highest number of true positive incidents. But the downside of such an approach is that it also leads to approximately 400 false positives, which lowers the recall to a value of 0.27, which puts it in the same area as a random approach.

Table 6.4: Evaluation: Performance measures

	Recall	Precision	F-measure	Semantic Dependency	Sentiment Value	Company Name
Setup 1	0.795	0.151	0.254	no	no	no
Setup 2	0.773	0.164	0.275	no	negative and very negative	no
Setup 3	0.058	0.217	0.092	no	very negative	no
Setup 4	0.023	0.075	0.035	no	positive and very positive	no
Setup 5	0	-	-	no	neutral	no
Setup 6	0.5	0.537	0.518	yes	negative and very negative	yes
Setup 7	0.591	0.481	0.531	no	negative and very negative	yes
Setup 8	0.523	0.523	0.523	yes	no sentiment	yes
Setup 9	0.614	0.466	0.529	no	no sentiment	yes

Figure 6.1: Comparison: Sentence-level results



Setups 3, 4 and 5 are developed to test the impact of the different sentiment levels. In Setup 2, the sentiment levels negative and very negative have been used. In Setup 3, only the very negative sentiment has been considered, which resulted in a poor performance: only five true positive incidents have been detected and the performance measures are all close to zero. Setup 6, which has been explained in the Section 3, and Setup 7 change this approach by not considering the semantic dependency. Setups 8 and 9 additionally do not consider any sentiment level at all. Regarding the results of the last four setups, the F-measure values for all four are similar at about 52%. The only difference which can be observed is the impact of the semantic dependency on the recall and precision measures. The setups with the semantic dependency show higher precision values and

lower recall, and vice versa without the semantic dependency.

The trade-off between precision and recall leads to a similar F-measure performance, although the impact of the semantic dependency cannot be considered as very strong and, so far, the most positive impact on the performance involved the combination of an environmental and company context within a sentence. Besides, the consideration of the sentiment did not have a high impact on the performance, which is shown by Setups 8 and 9. In order to provide an explanation for this effect, a closer look at the sentiment values of all sentences in the evaluation corpus could be of interest.

Table 6.5: Sentiment in evaluation corpus

	Sentiment Level	Percentage
very positive	23	0.4%
positive	704	12.3%
neutral 3	778	13.6%
negative 4	4052	70.9%
very negative	158	2.7%

Table 6.5 shows the distribution of the sentence sentiment levels in the evaluation corpus. The fraction of negative and very negative sentiment levels is very high and sums up to approximately 74%. So, if most of the sentences in the news blog articles are already negative, a restriction on negative sentences will not lead to a large increase. Possible reasons for the skewed deviation of the sentiment could be the negative nature of news reporting and that these article discuss problems on environmental issues. Hence, such an outcome is not very surprising. The skewed distribution was also a reason for not focusing on additional test cases on the sentiment level.

Regarding the difference between the full set and the reduced one, if the full set is used, then the performance measures drop in small numbers and not significantly. Figure 6.2 presents a comparison of the F-measure vale.

6.3 Results on the Document Level

It is also interesting to see how well the system performs if just a classification of relevant and non-relevant documents is considered. The approach is the following: If a document contains at least one incident, it is considered as relevant, regardless of how many other false positives, true negatives and false negatives are found in this document.

The results in Table 6.6 show that in all setups the performance of the F-measure increases, and especially the recall is much better than on the sentence level. The performance of the first four setups is still worse than in the last four. Especially the usage

Figure 6.2: Comparison of F-Measure: Reduced and full evaluation corpus

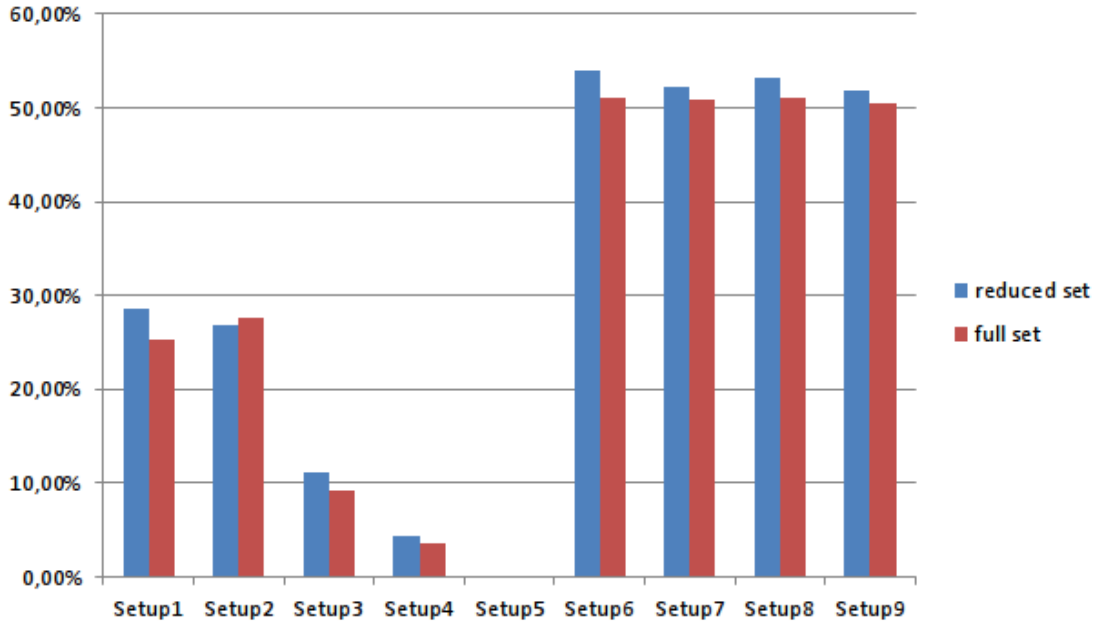


Table 6.6: Evaluation: Performance measures on the document level

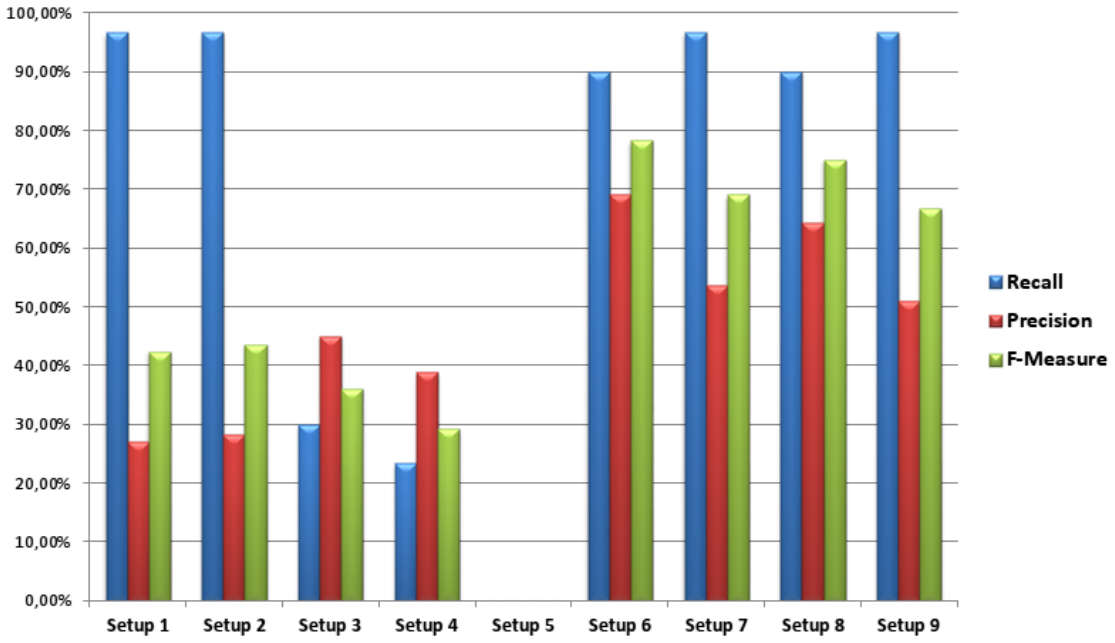
	Recall	Precision	F-measure	Semantic Dependency	Sentiment Value	Company Name
Setup 1	0.967	0.271	0.4234	no	no	no
Setup 2	0.966	0.281	0.436	no	negative and very negative	no
Setup 3	0.300	0.450	0.360	no	very negative	no
Setup 4	0.233	0.388	0.291	no	positive and very positive	no
Setup 5	0	-	-	no	neutral	no
Setup 6	0.900	0.692	0.7826	yes	negative and very negative	yes
Setup 7	0.692	0.537	0.691	no	negative and very negative	yes
Setup 8	0.900	0.642	0.750	yes	no sentiment	yes
Setup 9	0.866	0.508	0.666	no	no sentiment	yes

of the semantic dependency leads to an F-measure that is nearly 10% better. However, sentiment classification does not lead to a significant increase in the performance.

In summary, the focus on the classification of relevant and non-relevant documents leads to a significant increase in performance compared to just considering the output from the detected incident sentences.

To provide an estimation about the quality of the results and not just quantitative numbers, it is also interesting to study who is written about in these blog articles, how often companies are mentioned in the articles and how many different companies are mentioned. The question here is whether only the well-known companies are written about

Figure 6.3: Comparison: Document Level Results



and how broad is the perspective of journalistic publishing.

First, we examine all the documents that have been published between October 2013 and May 2014. Within this time frame, a total of 1,884 documents have been stored and the following companies have been mentioned in the articles:

Table 6.7 shows which companies have been mentioned most often in the published articles. A total of 984 different companies were mentioned 3,450 times within 1,880 articles. Of course, some companies are mentioned more frequently due to current events, as in the case of Tepco and Exxon Mobile. Moreover, the media is more concerned with the big players in the environmental sector, like Monsanto. Nevertheless, a large variety of different companies is mentioned in the articles.
Examining the companies mentioned within environmental incidents could provide some additional insights.

Table 6.8 shows the results for the number of times that a company was mentioned in the incidents, that have been tagged manually from the evaluation corpus. 49 different companies were mentioned which is a good distribution on a number of 86 tagged incidents. Certainly, some well-known companies, like Monsanto, Dow Chemical and

Table 6.7: Companies mentioned in Articles

Mentions	Company	Mentions	Company
108	TEPCO	22	Fukushima Daiichi
61	International Energy Agency	22	Loblaw Companies Limited
60	Exxon	21	Canon
57	Coca-Cola	21	Unilever
52	Google	20	Bayer
51	ExxonMobil	20	Earth Hour Capital
50	Bloomberg	19	Enviva
49	BP	19	Exxon Mobil
49	Monsanto	18	TransCanada
46	Chevron	17	Deutsche Bank
45	Enbridge	17	Galapagos
42	PepsiCo	16	Ahmedabad Municipal Corporation
39	Dan River	16	Bumitama Agri
38	Duke Energy	16	Ford
29	Bristol Bay Native Corporation	16	Island Conservation
29	Kellogg	15	GM
26	Repsol	15	TELUS
25	BT	14	Cabo Dorado
25	General Mills	14	Prodigy Network
23	Nestle	13	Associated British Foods

Table 6.8: Companies mentioned in incidents

Mentions	Company	Mentions	Company
9	Monsanto	2	Barrick Gold
5	Dow Chemical	2	Chemtura
5	Koch Industries	2	Chevron
4	Gazprom	2	Chisso Corporation
4	Hooker Chemical	2	Coca-Cola
3	BP	2	Duke Energy
3	Exxon	2	E.ON
3	ExxonMobil	2	Pongola Supergroup
3	Freedom Industries	2	U.S. Steel
3	TEPCO	1	Mammut

Koch Industries, are mentioned more frequently.

6.4 Analysis of Errors

It is important to examine closely the errors that have been produced by the system and the reasons why wrong classifications occur. Setup 7 from the previous section has been evaluated here, along with which steps in this approach led to a wrong outcome.

Table 6.9: Error in evaluation: False positive

Reason	Occurrences
company name as environmental word	1
wrong environmental word	19
wrong sentiment/not relevant	21
dependency error	3
wrong detection of company	13

Table 6.10: Error in evaluation: False negative

Reason	Occurrences
wrong sentiment	6
no dependency	11
environmental word not detected	15
company not detected	15

Most times, in the case of false positives, the wrong detection of the sentiment and the environmental, was one of the main reasons for errors. Moreover, in the case of false positives, a main reason was the false classification of the sentence sentiment, which means that the sentence was not negative or very negative or the sentence was not considered to be relevant at all.

The usage of OpenCalais also led to some wrong results: 15 times a company was not detected at all. One of the reasons was that a synonym for the company was used which is not known by the NER classifier. OpenCalais also sometimes tagged words or names as companies that are not companies at all. Evidently, wrong classifications can appear in every step of the pipeline with more or less the same plausibility.

6.5 Summary of Results

For the IE problem that is addressed in this work, the best approaches are those that considered the appearance of a word with environmental context and the appearance of a company name in a sentence. Using more sophisticated techniques in addition, like sentiment analysis or word dependency detection, had a very small influence on the

performance measures and cannot be considered as significant, regarding the sentence-level evaluation. The results clearly show that the more specific an approach becomes, the more recall will decline and precision values will increase.

For the evaluation on the document level, a document was considered true positive as a whole if it had one appearance of a true positive sentence. The impact of sentiment analysis and word dependencies is more significant. The setups from 6 through to 9 had very similar F-measures on the sentence level, although those with slightly better precision values resulted in significantly better F-measure results on the document level. An explanation for this effect is that environmental incidents are spare with respect to the total amount of sentences in documents, and it is likely that a document contains only one incident. Thus, higher precision on the sentence level leads to a better performance if the results are considered on the document level.

CHAPTER 7

Critical Reflection

In this section, the critical aspects of this work are addressed and, in general, a discussion about the methods, data mining and artificial intelligence techniques like NLP is conducted.

7.1 Credibility

In order to ensure the reliability of the information that is extracted with the solution of this work, the websites that are chosen as input data need to be reliable and trustworthy. Credibility is difficult to measure objectively and is defined in (Iding et al., 2008) as the ability to inspire belief or trust and as information accuracy and veracity (Klemm et al., 2001).

In this work, news blog articles from different sources are used and such news blogs provide important insights into attitudes about global or local events. Furthermore, in studying credibility, an interesting and differentiating feature of blogs is that they are meant to be inherently biased being an alternative media. Readers of blogs admittedly often distrust traditional media and see blogs as a viable alternative, particularly since they believe bloggers do not hide their biases (Johnson and Kaye, 2004). Using these news blogs raises questions about the validity and trustworthiness of these sources. Furthermore, the traditional measures for topic relevance, timeliness, specificity, and credibility are inadequate (Ulicny and Baclawski, 2007).

This problem is important if someone wants to filter additionally which results from this work could be more or less relevant. the purpose for which someone wants to use the gathered information. For example, if a company wants to discover the context in which it is mentioned, the credibility of the news source is not the top priority, because every site that mentions the company, or other affiliated companies, in a negative context

should be considered. However, if the information is used to accuse someone about a certain event, the information needs to be reliable.

The problem with the credibility of news blogs, and other web 2.0 platforms, also had an impact on scientific discussions due to the increasing popularity of such platforms and the amount of information available therein. Web 2.0 also affected the way news is reported. "The aftermath of the Iranian elections in June 2009 provided compelling evidence of the power of user-generated footage, but it also highlighted a battle of wills between a government determined to restrict access to information, and an alliance of newspapers, broadcasters and Iranian citizens equally determined to use new technology to get the story out"(Newman, 2009). Further, during the Arab Spring in northern Africa in 2011, a large fraction of the information and communication was done by using micro blogging services and was adopted by main stream media organisations (Lotan et al., 2011). The challenge in this events has been the vast amount of information, because "organisations like CNN and the BBC, which at one stage was receiving up to five videos a minute"(Newman, 2009). Problematic is the aspect of assessing credibility in this huge amount of information or maybe some people are not aware that some information from such news sources could be invalid. For example, "some people did provide updates from Tehran, but many didn't check out. When someone tweeted that there were 700,000 people demonstrating in front of a mosque, it turned out that only around 7,000 people showed up "(Weaver, 2009).

User studies showed (Fogg et al., 2001) that users judge a website as credible based primarily on structural and author-specific elements. Note that in contrast to standard websites, people can be less controlled and even harder to trace in the blogosphere and other Web 2.0 applications. According to the Stanford Guidelines for web credibility (Fogg, 2002) which provide a set of 10 guidelines in order to improve the credibility of a websites, the first point in this guideline is that it should be easy to verify the accuracy of the information on the website by using references and citations. Several approaches that rely on the references of articles do exist in field of information retrieval, there it is common to rely on the quality of news, for example, by using algorithms to determine the relevance of the document and by using the Google Page Rank (Page et al., 1999) in order to get an additional metric about the quality of outgoing and ingoing web-links from the website.
According to (Ulicny and Baclawski, 2007), just taking into account the quality is not enough for judging credibility. However, retrieval, clustering and indexing techniques that work on ordinary web documents do not work well in the blogosphere, because blog posts are short, of ephemeral importance, highly exophoric, highly quotable, and much less susceptible to PageRank/Kleinberg type analyses because they have relatively few incoming links, especially on a per-author basis (Ulicny and Baclawski, 2007).

A description for measuring credibility was proposed by (Ulicny and Baclawski, 2007). They constructed a measure for credibility that takes into account source, message and the reception by bloggers. In their approach, it is also crucial to measure the credibility

of an author rather than on the blog site itself, because many blog sites are multi-author blogs, and an author could derive credibility that he has not earned. In their work, the researchers have identified 48 source, message and reception features to formulate a measure for credibility. These could include, for example, whether the full name of the author is present (it could also be possible that the author writes under a pseudonym), the affiliation, presence of comments and hyperlink citations. (Rubin and Liddy, 2006) proposed an analytic framework for blog credibility assessment based on four profile factors:

1. the blogger's expertise and the amount of offline identity disclosure
2. the blogger's trustworthiness (or the overtly stated value system including beliefs, goals, and values)
3. information quality
4. appeals of a personal nature

This framework has been used in practice by (Weerkamp and de Rijke, 2012) to construct a re-ranking approach. They suggested two different approaches. The first approach, credibility-inspired re-ranking, simply re-ranks the top n of a baseline based on the credibility-inspired score. The second approach, combined re-ranking, multiplies the credibility-inspired score of the top n results by their retrieval score, and re-ranks based on this score. They used two different groups of indicators for assessing credibility—i.e. post-level and blog-level indicators. Post-level indicators include spelling mistakes, correct capitalization, use of emoticons, punctuation abuse, document length, timeliness (when related to a news event), and how its semantics match formal (news) text. On the blog level, the following indicators are used: average number of comments, average number of pronouns, regularity of posting, coherence of the blog, and the expertise of the blogger.

Both approaches have proved capable of improving an already strong baseline and the best performance is achieved by using a combination of all post-level indicators. In their work, they have identified the most influential indicators and explained why these indicators lead to improvements in retrieval performance.

To assess the credibility of new blog posts, one can rely on existing research in this area. Theoretical approaches and frameworks have already been formalized and practical implementations do exist; hence, all of the cited research papers mention that further work and improvements are required to increase the validity of these solutions.

Credibility has some critical implications for this work and, in general, for the domain of text analysis. However, this work focuses on the detection of environmental incidents

and does not account for credibility issues in its results. In this work, a small sample of blog sources from popular blog sites has been used. The problematic aspect relates to the interpretation of the results and their reliability. Thus, in a real-world scenario, the selected news sources need to be assessed in the beginning. In general, credibility is always an issue regarding the analysis of news and social media content, since it is always possible that, for example, social media marketing agencies try to manipulate the social media stream. They could create fake profiles and write fake posts with positive content in order to influence the overall views on a certain topic. This problem also raises the question of the extent to which the truth can be determined, or, if we apply a constructivist point of view, it may lead to the conclusion that real truth is perhaps always somehow constructed and the real objective reality is difficult to determine.

7.2 Critical Review Data Mining and Artificial Intelligence

New methods from the field of artificial intelligence, like data mining and NLP, combined with the technological abilities of big data have made it possible to gather and process increasing amounts of information. Nowadays, there is little that cannot be tracked in our online and offline lives. People post about private topics on social media platforms [1], write reviews or share their opinions about products, movies, hotels and restaurants. Furthermore, they also use location-based services [2] that are accessible via smartphones and their GPS abilities, use loyalty from grocery stores cards [3] and make it possible for companies to track everything they have bought.

By using this data, companies create customer profiles in order to develop merchandising strategies and customize the shopping experience for the needs of their customers. So far, everyone could be benefiting from this. People and companies receive pure information that leads to a benefit for both of them.

Recent events like, for example, the NSA Leaks (Guardian, 2013), have shown that this new technology is also used for surveillance purposes and that the privacy of many people is threatened by such technologies. Some legitimate questions arise with big data and data mining, such as: Is our privacy in danger? Are peoples being analysed through their behaviour and categorized into groups that could have consequences on our lives in the future? Do we need new ethical standards for using these technologies? These questions are addressed in the following section.

Concerns with Big Data

"What is Big Data? A meme and a marketing term, for sure, but also shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions. There is a lot more data, all the time, growing at 50 percent a year, or more than doubling every two years"(Lohr, 2012). Besides the new technological possibilities, big data might also change our society. Big data changes the objects of knowledge and also provides better insight into how we understand and perceive human networks and communities, because a change in the instruments could lead to a change in the entire social theory that accompanies them (Boyd and Crawford, 2012).

For example, sociology, according to (Latour, 2010), has been obsessed by the goal of becoming a quantitative science, but there was always the struggle of what constitutes quantifiable knowledge and what does not. Big data offers to the humanistic disciplines the possibility to claim that theirs is an objective and quantitative science. But besides

[1] http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users

[2] http://www.statista.com/statistics/294314/share-of-mobile-subscribers-using-location-based-services/

[3] http://www.statista.com/statistics/323588/customer-loyalty-card-schemes-usage-behaviour-preferences-uk/

the hype about big data and all its possible implications, the claimed objectivity of big data may also be questioned.

a large mass of raw information, Big Data is not self-explanatory. And yet the specific methodologies for interpreting the data are open to all sorts of philosophical debate. Can the data represent an 'objective truth' or is any interpretation necessarily biased by some subjective filter or the way that data is 'cleaned?'."(Bollier and Firestone, 2010, p. 13)

In fact, the objectivity of big data is very controversial since researchers always interpret data. Even if mathematical models seem valid and appropriate, as soon as a researcher begins to understand and think about the meaning of the output, the interpretation starts (Boyd and Crawford, 2012) and we are back to subjective influence. The same problem arises in the decision regarding what should be measured, which measures are important and which are not. Additionally, large sets of data are also prone to errors, especially when multiple data sets are used together (Boyd and Crawford, 2012). Twitter [4] is a very popular website for big data because its content is easily accessible through its API. Twitter users don't have to be necessarily representative for society and one account represent might not be equal to one human being. Users could have multiple accounts and many posts are also automated to publish news from websites.

Big data is also modelled to what can fit into mathematical models, which could lead to their being taken out of context, with the result that the data loses meaning. For example, the analysis of connections in social networks is a popular field of study, but these connections are not necessarily equivalent to real connections respectively sociographs and if they really share information with each other and what kind of information (Boyd and Crawford, 2012).

Privacy Issues in Web Data Mining

Data mining provides considerable new insight into data, especially in the field of web data mining because the Word Wide Web can be seen as the largest database available. Based on the definition of (Etzioni, 1996), web data mining is defined as "the whole of data mining and related techniques that are used to automatically discover and extract information from web documents and services". A critical issue here is that a person might not be aware that information and knowledge is being collected about him/her and how it will or could be used. This invisible information gathering happens all the time on the World Wide Web. Another issue is the secondary use of data—for example, data that is given to someone for a certain purpose could be misused for a different one.

Although privacy is an important issue in media and scientific discussion, little is known about how the general public understands privacy and what is important to them. Thus, (McCullough et al., 2014) focused in their study on people's evaluation of society's

[4] https://twitter.com

value of privacy. They found that a large number of people value privacy but do not really try to protect their privacy. This seems paradoxical and questions how much they actually value their privacy. (Acquisti et al., 2013) found evidence that individuals assign very different values to the privacy of their data depending on the order in which they consider different offers for that data, whether they consider the amount of money they would accept to disclose private information, or the amount of money they would pay to protect information. In their findings, they also doubt the abilities of individuals to decide rationally about privacy issues and the question "How much is privacy worth?" does not just depend on the people, it also depends on the kind of private information that is sought. In the domain of information privacy, most people consider it very important to have control over the information and to have the ability to protect this information. Most issues that arise from web data mining fall within the category of informational privacy.

(Van Wel and Royakkers, 2004) divided the privacy risk into three different categories of web data mining, namely

- *content mining*
  To analyse the content in web documents like text, images, audio files, etc.; text documents form the focus of the practical part of this thesis.

- *structure mining*
  Focuses on the analysis of how web documents are linked to each other.

- *usage mining*
  Analyses the transaction data that is logged when users interact with the web.

The information gathered from content and structure mining is publicly available on the websites and has been made available (in the most cases) by the owner of the information itself. It seems that such information does not need protection, but every small kind of information that has been gathered could be used in a new context in order to generate new information and (Nissenbaum, 1997, p. 216) argues that "the assumption, an aggregation of information does not violate privacy if its parts, taken individually, do not, is not correct". Further, experiments with databases that store personal information has shown that the value of information increases by combining the data with other information and a whole industry exists that focuses on assembling and selling data, because "a single fact about someone takes on a new dimension when it is combined with other facts about the individual, or when it is compared with similar facts about other individuals"(Nissenbaum, 1997, p. 216).

Usage mining is an approach where all the interactions of a user on a website are tracked (Srivastava et al., 2000) —for example, if someone visits an online marketplace, every product page that the user visits is logged; if the user is additionally registered and

logged in on the website, he may also provide information about his gender, age, location, etc. With the help of this information, it is possible to generate user profiles. Usually, if a user is not known to a website (not logged in or registered), it is possible to identify him by the usage of cookies, but even if a user turns off the cookies in his browser and uses private browsing settings, it is still not possible to stay anonymous. A recent trend in web analysis is called device fingerprinting or host fingerprinting. Like a human fingerprint, every user that surfs on a website provides some information to the browser, which makes it possible to identify her/him. This includes details like which browser is used and which configurations, screen resolution, operation system, hardware and network. By using these parameters, it is possible to create a profile of the user and to identify the user across different websites (Yen et al., 2012). Detailed information of this can be seen in Figure 7.1.

Figure 7.1: Taxonomy of features which can be used for fingerprinting (Nikiforakis et al., 2013)

Fingerprinting Category	**Panopticlick**	**BlueCava**	**Iovation ReputationManager**	**ThreatMetrix**
Browser customizations	Plugin enumeration$_{(JS)}$ Mime-type enumeration$_{(JS)}$ ActiveX + 8 CLSIDs$_{(JS)}$	Plugin enumeration$_{(JS)}$ ActiveX + 53 CLSIDs$_{(JS)}$ Google Gears Detection$_{(JS)}$		Plugin enumeration$_{(JS)}$ Mime-type enumeration$_{(JS)}$ ActiveX + 6 CLSIDs$_{(JS)}$ Flash Manufacturer$_{(FLASH)}$
Browser-level user configurations	Cookies enabled$_{(HTTP)}$ Timezone$_{(JS)}$ Flash enabled$_{(JS)}$	System/Browser/User Language$_{(JS)}$ Timezone$_{(JS)}$ Flash enabled$_{(JS)}$ Do-Not-Track User Choice$_{(JS)}$ MSIE Security Policy$_{(JS)}$	Browser Language$_{(HTTP, JS)}$ Timezone$_{(JS)}$ Flash enabled$_{(JS)}$ Date & time$_{(JS)}$ Proxy Detection$_{(FLASH)}$	Browser Language$_{(FLASH)}$ Timezone$_{(JS, FLASH)}$ Flash enabled$_{(JS)}$ Proxy Detection$_{(FLASH)}$
Browser family & version	User-agent$_{(HTTP)}$ ACCEPT-Header$_{(HTTP)}$ Partial S.Cookie test$_{(JS)}$	User-agent$_{(JS)}$ Math constants$_{(JS)}$ AJAX Implementation$_{(JS)}$	User-agent$_{(HTTP, JS)}$	User-agent$_{(JS)}$
Operating System & Applications	User-agent$_{(HTTP)}$ Font Detection$_{(FLASH, JAVA)}$	User-agent$_{(JS)}$ Font Detection$_{(JS, FLASH)}$ Windows Registry$_{(SFP)}$	User-agent$_{(HTTP, JS)}$ Windows Registry$_{(SFP)}$ MSIE Product key$_{(SFP)}$	User-agent$_{(JS)}$ Font Detection$_{(FLASH)}$ OS+Kernel version$_{(FLASH)}$
Hardware & Network	Screen Resolution$_{(JS)}$	Screen Resolution$_{(JS)}$ Driver Enumeration$_{(SFP)}$ IP Address$_{(HTTP)}$ TCP/IP Parameters$_{(SFP)}$	Screen Resolution$_{(JS)}$ Device Identifiers$_{(SFP)}$ TCP/IP Parameters$_{(SFP)}$	Screen Resolution$_{(JS, FLASH)}$

(Yen et al., 2012) have shown that it is possible to identify 60--70% of the hosts just by using the http user agent string. They improved these results by generating a user identification from the hardware specifications of a client, which resulted in precision and recall values of about 93%. They also mentioned in their study that using methods like private browsing had no impact. (Nikiforakis et al., 2013) have analysed different commercial fingerprinting solutions and state that browser extensions that help users spoof the user-agent of their browsers can be used as an additional fingerprinting feature by those fingerprinting solutions.

The implications of these results are that absolute privacy for ordinary users is threatened and it is difficult to guarantee it. A possibility which makes it possible to hide oneself on the web are proxy servers, which could protect one from fingerprinting, but it is not popular enough and maybe too complex to use for the ordinary internet user.

The core question that arises from all these new technologies and possibilities for com-

munication, shopping and sharing personal information online is this: Is privacy still possible nowadays and is it possible to protect it from unreasonable government and private sector intrusion? Recent events, like the NSA disclosures about the surveillance by governmental organizations or the leaked private pictures of prominent people, who have used cloud services to store their private pictures (Evershed, 2014), are examples that privacy is not easy to preserve. Besides, people might not be aware of the risks that accompany these technologies. Studies mentioned earlier have shown that people care about privacy but to actually protect their privacy, they would have to change their communication and online habits. Do people care enough to do so?

Besides all the threats that have been mentioned earlier, the internet can also be a medium to empower people, use new possibilities to connect with others and react to perceived threats to privacy, argues (Berman and Bruening, 2001). For example, the company DoubleClick planned to to link personally identifiable information that was collected offline with that collected online. This plan was revealed and led to negative media coverage, thereby forcing the company to stop their plans.

Another example is the proposal of the European Union for the new General Data Protection Regulation (GDPR), which has the objective of establishing a common data protection law in all member states of the EU. The first proposal was very consumer-friendly, with strict regulations for companies and high penalties if they did not comply with the regulations (Hornung, 2012). The proposal became an incentive for companies to lobby for changes to the proposals. Thereafter, several requests for modifications were taken up by Members of the European Parliament (MEPs), which changed the nature of the regulation into a more corporate-friendly one. In order to reveal these changes and the influence of the lobby groups, a journalist and a law student have founded the platform LobbyPlag [5]. Their website provides information about which political party and which MEP favours more or less privacy. They also show if the MEP has made proposals or held discourses in the European Parliament concerning the topic of data privacy before the discussion about the reformation started. Furthermore, they have used text analysis to determine the origin of the change requests that have been handed in by the MEPs. They controlled if the change requests had been taken directly from proposals or papers published by lobbying groups.

Recent history shows that there is an enormous threat to privacy but the public also has the possibility to make significant gains in privacy protection by using new technologies. (Berman and Bruening, 2001) argues, that privacy is a work in progress and that it is still possible in the 21st century. Besides technological evolution, the level of privacy that is achieved or not achieved will always depend on whether society is willing to act in order to ensure privacy and still believes that privacy is possible nowadays. This is because technology inherently lacks a good or a bad attitude.

[5] http://www.lobbyplag.eu/

Ethical Controversy

"Technology is neither good nor bad; nor is it neutral . . . technology's interaction with the social ecology is such that technical developments frequently have environmental, social, and human consequences that go far beyond the immediate purposes of the technical devices and practices themselves."(Kranzberg, 1986, p 545)

The purpose for which technology is used determines how it will be perceived by the society. In addition to the privacy issues in big data and data mining, further ethical issues are of interest. Is it ethical to include someone within large aggregated data and to analyse his details without him being aware of it? How should public posts on blogs or social networks be treated? Is the author always aware that a post is public? What if a public post is taken out of context and used for a purpose that was not intended by the author?

Just because something is publicly accessible does not make it ethical to use it for every purpose. Furthermore, it is also a question of power and control. Companies and researchers are able to access this data and deduct patterns and new information thereby. On the other hand, access is limited to the users of social media platforms that generate this data. Users are not always aware where and what kind of data is gathered and do not know about the the multiple use cases their data has for data analysts. There is also a considerable difference between when someone decides to be public and in which context, or their receiving public attention all the time without any control over it (Boyd and Crawford, 2012).

Access to data storage is also an important issue. Who gets access, to which data sets, for which purpose and under which restrictions? Twitter data, for example, was very restrictive in the past and still is restrictive if ones want to access the full dataset [6].

This produces a serious disparity because only organisations with the financial capabilities can acquire or produce datasets with valuable information. Regarding the already-mentioned capabilities, the analysis of big data demands educated and skilled employees with the necessary computational skills. People without this knowledge would be at a disadvantage. An implication here could also be that big data changes the world into a world of knowledge and may divide people into two categories; Those who are able to understand how the process of data collection works and how to inhibit the data being collected from them, as far as it is possible, and those who do not.

[6] https://dev.twitter.com/streaming/firehose

CHAPTER 8

Summary and Future Work

Finally, this thesis concludes with a summary and an outlook for possible future work on the system that has been developed here.

8.1 Summary

In this master thesis, an approach for the detection of environmental incidents from news blog articles has been developed and presented. From a corporate perspective, such incidents could trigger consequences which influence the oprative business. Awareness of such a topic could be important for companies, if they want to establish sustainability capabilities. Research suggests that companies should address sustainability issues because these are critical for the long-term existence of a company Porter and Kramer (2006). The challenge for companies is to gather information about potential risks, in order to enable a risk assessment. Companies should not only care about their own actions which trigger sustainability risk, they also should care about what their subcontractors in their supply chain do. Especially if a company has a large set of subcontractors, it is more difficult and time-consuming to acquire and asses information about them, to ensure that these subcontractors comply to the sustainability standards or values of the contractee. Sustainability does not only consider environmental issues; social and economic indicators are concerns too. Each of these factors could be considered as a possible risk for companies and could hinder the achievement of corporate goals. Corporate risk management requires data and information,to be able to determine, assess and react to possible risks.

Usually, the role of information technology in terms of environmental concerns was the reduction of energy cost by developing more energy-efficient hardware. In the domain of environmental risk, information technology can play a more important role and should go beyond just green technologies. IT should now be able to provide and con-

tribute abilities that enable the discovery and analysis of data as well as support the risk management processes.

The focus in this thesis is on the analysis of news blog articles that focus on environmental topics. The analysis is done by using the abilities of text mining technologies. In this thesis, the state of the art about information extraction methods has been evaluated. Different approaches, like knowledge engineering or learning/training, have been described. Furthermore, different methods in NLP were discussed, which make it possible to detect parts of speech, grammatical structures or extract the relations, dependencies between words within sentences and determine the sentiments of sentences. A rule based solution approach for the first research question(RQ1), if it is possible to extract relevant information of environmental incidents from text sources, has been formalized in the 3 chapter.

In this thesis, an environmental incident is any event that is reported within a news article and that causes or threatens environmental damage, which could have an impact on the environment, human lives or property. The solution uses a knowledge engineering approach that aims to find sparse patterns within the text sources by formulating rules. The aggregation of different rules formalize different setups that should detect environmental incidents. The performance of these different setups is evaluated with an approach that was formulated in the chapter 5 in order to answer the second research question (RQ2). Those setups are compared by using their performance measures (precision, recall, F-measure) on a manual tagged evaluation corpus out of 200 news documents that contain approximately 5,000 sentences. The sentences have been tagged into the categories *not relevant* and *relevant*.

In the proof of concept section of the thesis, an NLP attempts to develop a system that analyses text input and produces an annotated output on the sentence level. A knowledge engineering approach has been chosen for the development. The system is a GATE CREOLE plugin resource that aggregates functionality from several different state-of-the-art NLP tools like POS tagging, sentiment evaluation, dependency parsing and semantic tagging. The system evaluates each input sentence on the following criteria.

The chapter 6 focuses on the answer of RQ3 and RQ4 and provides results that show a clear pattern: The lower the restrictions on the criteria, the higher is the recall and the lower is the precision. When considering the F-measure, the setup that involved no semantic dependency, negative and very negative sentence sentiment, and the presence of a company in the sentence, reached the best result with a value of 53%. It is also interesting to see how well the system performs if just a classification of relevant and non-relevant documents is considered. An F-measure performance value of 75% was achieved on the evaluation corpus in a setup that considers a semantic dependency between a environmental word and the company in a sentence. The sentiment orientation of the sentences has not been considered, due to the worse performance results. The hugest impact on the results were sentences where companies and environmental words

had been mentioned. Leaving out a semantic dependency or sentiment level changed the recall and precision values accordingly, but the F-measure did not change significantly. Especially the factor of sentiment levels did not have a significant impact. This could be because sentences in news stories are formalized and most of the content that is delivered through news blog posts have a negative attitude. This assumption is supported by the fact that, according to the Stanford Sentiment solution, a very high fraction of the sentences in the evaluation corpus have a negative or very negative sentiment level. Most reasons for errors/wrong classifications are that, in the case of false positives, a wrong sentiment level has been assigned to a sentence or a wrong environmental word has been detected. In the case of false negatives, the most frequent reasons for error have been that a company name was not detected or, again, the environmental word was wrong. Overall in the articles in the evaluation corpus 49 different companies were mentioned and 20 of them were present in detected incidents. The Top 5 mentioned companies were present in approximately in every second detected incident, which leads to the assumption that even reporting in news blogs of environmental issues still focuses more on some well known companies and topics.

Besides the technical challenges in IE, there are different challenges like the credibility of news sources, privacy and ethical concerns. When using online resources for information extraction, the reliability of these sources is of interest in order to assess if the extracted information can be taken seriously. Scientific research has already addressed this issue; the frameworks and practical solutions that exist are introduced in the section 7. The consequences of big data and artificial intelligence technologies, like data mining and IE, on privacy and ethical concerns are also of particular interest. These new technologies offer new possibilities to process huge amounts of data and to detect patterns within these data, thereby enabling people to be categorized into groups with common attributes. But this superficial approach also has its critics, who say that it does not take individuality into account. Moreover, there is the matter of whether it is ethical to use every kind of data that is available and whether the source of this data is aware that this data could be used in a different context than it was intended by the owner.

This implicates a privacy issue too. Data relating to nearly every part of people's lives nowadays is collected and stored, like buying habits, communication patterns and financial situation. Furthermore, people are tracked with or without their consent or knowledge and, for example, staying anonymous on the web is becoming more and more difficult due to increasingly advanced tracking. In conclusion, regarding the subject of ethics and privacy, it is hard to tell how the future will develop, whether we will have to face some kind Orwellian state, or whether the ongoing threat to privacy will elicit a reaction and a demand for more privacy.

8.2 Future Work

In the future, additional improvements and further refinements could be carried out in the following respects:

Credibility Assessment of News Sources
The problem of credibility of new sources has been addressed theoretically in this thesis. For a productive implementation, it would be necessary to do a prior assessment of the credibility of the used news sources as well as an extensive search for sources, especially local ones that are harder to find and more difficult to assess.

Different Input Sources
Besides new blogs, further possible input sources exist. For example, an analysis of Twitter posts could be interesting because these allow easier categorization through hashtags and geographic information about the person who is posting.

Additional Language Support
The limits of my language mean the limits of my world (Wittgenstein, 1963).
One of the major limitations is to find qualified sources, especially sources who write about local events. English is a good starting point, because most of the relevant news around the world is published in English too. Nevertheless, there is a bias towards topics about English-speaking countries. A further point is that even if you know the language that is spoken in a country, you might not be familiar with the media, journalists, NGOs, etc., which publish relevant articles on the subject and it might also prove difficult to find people who are well-versed in this respect. An implementation for the German language would be particularly interesting here.

Improvements in Sentiment Classification
The sentiment classification did not have a huge positive impact on the results, but was responsible for errors. The Stanford sentiment algorithm, which was used in this work, is a classifier based on the deep learning/deep neural network paradigm. It was trained on a set of movie reviews, due to their open accessibility and variety in semantic orientation. Improvements regarding the semantic orientation of environmental words would be beneficial to the performance. At present, it is possible to gather environmental words from WordNet, but so far an assignment of sentiment values is missing. This could also be beneficial to dependency parsing and would make it possible to analyse grammatical relations with other words or phrases with a certain semantic orientation.

Performance Optimization
The performance of IE tools is always an important issue that is influenced by the amount of data chosen, the level of detail in which the data is analysed, and which technology or frameworks are used. One of the main elements that increases the pro-

cessing time is the OpenCalais plugin. Since it is a web service, the invoke, processing and response times of the web service have to be considered.

OpenCalais was used because of its good performance measures and its more granular segmentation regarding the differences between companies, organizations, etc. An extension of an existing gazetteer could be a feasible option if similar performance results can be reached.

Using Linked Data and Ontologies
Linked data could be a possibility to put the extracted information in a different context and provide easy access to additional types of analyses. How many times has a company been mentioned by a certain blogger or journalist? How often have they been connected to positive or negative events for which they are responsible? If a person is mentioned in a news article, it might be possible to determine in which other relations this person has been mentioned (possible connections to companies, political parties), and so on. The idea is analyse the news reporting in order to discover connections and networks.

Focus on Social Sustainability
The solution in this work could be extended to the domain of social sustainability, by using different hypernym hierarchies from WordNet.

Bibliography

A. Acquisti, L. K. John, and G. Loewenstein. What is privacy worth? *The Journal of Legal Studies*, 42(2):249–274, 2013.

D. R. Anderson. The critical importance of sustainability risk management. *RISK MANAGEMENT-NEW YORK-*, 53(4):66, 2006.

L. C. Angell and R. D. Klassen. Integrating environmental issues into the mainstream: an agenda for research in operations management. *Journal of Operations Management*, 17(5):575–598, 1999.

D. E. Appelt. Introduction to information extraction. *Ai Communications*, 12(3):161–172, 1999.

I. Arel, D. C. Rose, and T. P. Karnowski. Deep machine learning-a new frontier in artificial intelligence research [research frontier]. *Computational Intelligence Magazine, IEEE*, 5(4):13–18, 2010.

M. Arrivé. Les éléments de syntaxe structurale de lucien tesnière. *Langue française*, pages 36–40, 1969.

Association for Computational Linguistics. Pos tagging - state of the art. `http://aclweb.org/aclwiki/index.php?title=POS_Tagging_(State_of_the_art)`, 2014. Accessed: 2014-04-21.

M. Atkinson and E. Van der Goot. Near real time information mining in multilingual news. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 1153–1154, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4.

S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

O. Bender, F. J. Och, and H. Ney. Maximum entropy models for named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 148–151. Association for Computational Linguistics, 2003.

Y. Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

J. Berman and P. Bruening. Is privacy still possible in the twenty-first century? *Social Research*, pages 306–318, 2001.

T. Bhuiyan, Y. Xu, and A. Josang. State-of-the-art review on opinion mining from online customers' feedback. In *Proceedings of the 9th Asia-Pacific Complex Systems Conference*, pages 385–390. Chuo University, 2009.

D. Bollier and C. M. Firestone. *The promise and peril of big data*. Aspen Institute, Communications and Society Program Washington, DC, USA, 2010.

D. Boyd and K. Crawford. Critical questions for big data. *Information, Communication amp; Society*, 15(5):662–679, 2012.

E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the workshop on Speech and Natural Language*, pages 112–116. Association for Computational Linguistics, 1992.

E. Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, 21(4):543–565, 1995.

S.-H. Cheng. Forecasting the change of intraday stock price by using text mining news of stock. In *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, volume 5, pages 2605–2609. IEEE, 2010.

H. L. Chieu and H. T. Ng. Named entity recognition: a maximum entropy approach using global information. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.

N. Chinchor. Muc-4 evaluation metrics. In *Proceedings of the 4th Conference on Message Understanding*, MUC4 '92, pages 22–29, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics. ISBN 1-55860-273-9.

L. Chiticariu, Y. Li, and F. Reiss. Rule-based information extraction is dead! long live rule-based information extraction systems! *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 827–832, Ocober 2013.

B. Commission, B. Commission, et al. Our common future, 1987.

J. Cowie and W. Lehnert. Information extraction. *Communications of the ACM*, 39(1): 80–91, 1996a.

J. Cowie and W. Lehnert. Information extraction. *Commun. ACM*, 39(1):80–91, Jan. 1996b. ISSN 0001-0782.

H. Cunningham, H. Cunningham, D. Maynard, D. Maynard, V. Tablan, and V. Tablan. Jape: a java annotation patterns engine. Technical report, University of Sheffield, Department of Computer Science, 2000.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

H. Cunningham, A. Hanbury, and S. Rüger. Scaling up high-value retrieval to medium-volume data. In H. Cunningham, A. Hanbury, and S. Rüger, editors, *Advances in Multidisciplinary Retrieval (the 1st Information Retrieval Facility Conference)*, Lecture Notes in Computer Science, Volume 6107, Vienna, Austria, May 2010. Springer. ISBN 978-3-642-13083-0.

H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters. *Text Processing with GATE (Version 6)*. 2011. ISBN 978-0956599315. URL `http://tinyurl.com/gatebook`. Accessed: 2014-05-20.

V. Dao, I. Langella, and J. Carbo. From green to sustainability: Information technology and an integrated sustainability framework. *The Journal of Strategic Information Systems*, 20(1):63–79, 2011.

M.-C. De Marneffe and C. D. Manning. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics, 2008.

M.-C. De Marneffe, B. MacCartney, C. D. Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, volume 6, pages 449–454, 2006.

G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, and R. M. Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, volume 4, page 837–840, 2004.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

D. R. Easterling, G. A. Meehl, C. Parmesan, S. A. Changnon, T. R. Karl, and L. O. Mearns. Climate extremes: observations, modeling, and impacts. *science*, 289(5487): 2068–2074, 2000.

J. Elkington. Towards the sustainable corporation. *California Management Review*, 36: 90–100, 1994.

Envecologic. Sustainability – the buzzword. `http://envecologic.com/2012/04/26/understanding-sustainability-and-its-need/`, 2014. Accessed: 2014-08-20.

A. Esuli and F. Sebastiani. Determining term subjectivity and term orientation for opinion mining. In *European Chapter of the Association for Computational Linguistics*, volume 6, page 2006, 2006a.

A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, volume 6, pages 417–422, 2006b.

O. Etzioni. The world-wide web: quagmire or gold mine? *Communications of the ACM*, 39(11):65–68, 1996.

N. Evershed. How easy is it to crack into an apple icloud account? we tried to find out, September 2014. URL `http://gu.com/p/4x8tg/sgp`. Accessed: 2015-03-02.

B. Fogg, J. Marshall, T. Kameda, J. Solomon, A. Rangnekar, J. Boyd, and B. Brown. Web credibility research: a method for online experiments and early study results. In *CHI'01 extended abstracts on Human factors in computing systems*, pages 295–296. ACM, 2001.

B. J. Fogg. Stanford Guidelines for Web Credibility. Technical report, Stanford University, 2002. URL `http://credibility.stanford.edu/guidelines/`. Accessed: 2015-03-02.

P. Frich, L. Alexander, P. Della-Marta, B. Gleason, M. Haylock, A. Klein Tank, and T. Peterson. Observed coherent changes in climatic extremes during the second half of the twentieth century. *Climate Research*, 19(3):193–212, 2002.

G. Gil, A. de Jesús, and J. Lopéz. Combining machine learning techniques and natural language processing to infer emotions using spanish twitter corpus. In J. Corchado, J. Bajo, J. Kozlak, P. Pawlewski, J. Molina, V. Julian, R. Silveira, R. Unland, and S. Giroux, editors, *Highlights on Practical Applications of Agents and Multi-Agent Systems*, volume 365 of *Communications in Computer and Information Science*, pages 149–157. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-38060-0.

T. Guardian. The nsa files, 2013. URL `http://www.theguardian.com/us-news/the-nsa-files`. Accessed: 2015-02-12.

M. Hagenau, M. Liebmann, and D. Neumann. Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, 55(3):685–697, 2013.

S. L. Hart. A natural-resource-based view of the firm. *Academy of management review*, 20(4):986–1014, 1995.

F. M. Hasan, N. UzZaman, and M. Khan. Comparison of different pos tagging techniques (n-gram, hmm and brill's tagger) for bangla. In *Advances and Innovations in Systems, Computing Sciences and Software Engineering*, pages 121–126. Springer, 2007.

V. Hatzivassiloglou and J. M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 299–305. Association for Computational Linguistics, 2000.

M. Hepple. Independence and commitment: Assumptions for rapid training and execution of rule-based pos taggers. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 278–277. Association for Computational Linguistics, 2000.

G. E. Hinton. Deep belief networks. *Scholarpedia*, 4(5):5947, 2009. revision 91189.

G. Hornung. Eine datenschutz-grundverordnung für europa. *Licht und Schatten im Kommissionsentwurf*, 25(2012):99–106, 2012.

M. Iding, B. Auernheimer, and M. E. Crosby. Towards a metacognitive approach to credibility. In *Proceedings of the 2nd ACM workshop on Information credibility on the web*, pages 75–80. ACM, 2008.

V. Illingworth. *Dictionary of Computing*. Oxford University Press, Inc., New York, NY, USA, 4th edition, 1997. ISBN 0192800469.

H. Isozaki and H. Kazawa. Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.

J. Jiang. Information extraction from text. In C. C. Aggarwal and C. Zhai, editors, *Mining Text Data*, pages 11–41. Springer US, 2012. ISBN 978-1-4614-3222-7.

R. Jizba. Part 4: Recall and precision: key concepts for database searchers. 2007. URL `https://dspace.creighton.edu/xmlui/handle/10504/7292`. Accessed: 2015-03-02.

T. J. Johnson and B. K. Kaye. Wag the blog: How reliance on traditional media and the internet influence credibility perceptions of weblogs among blog users. *Journalism & Mass Communication Quarterly*, 81(3):622–642, 2004.

D. Jones. Project zeus risk management plan at nasa, 2008. URL `http://sce.uhcl.edu/helm/ZEUS/rmpzeus.pdf`. Accessed: 2015-03-02.

A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff. Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1):49–79, 2004.

P. R. Kleindorfer, K. Singhal, and L. N. Wassenhove. Sustainable operations management. *Production and operations management*, 14(4):482–492, 2005.

E. B. Klemm, M. K. Iding, and T. W. Speitel. Do scientists and teachers agree on the credibility of media information sources?. *International Journal of Instructional Media*, 28(1):83–91, 2001.

I. Knoepfel. Dow jones sustainability group index: a global benchmark for corporate sustainability. *Corporate Environmental Strategy*, 8(1):6–15, 2001.

C. Kohlschütter, P. Fankhauser, and W. Nejdl. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 441–450. ACM, 2010.

M. Kranzberg. Technology and history:"kranzberg's laws". *Technology and Culture*, pages 544–560, 1986.

P. Kroha, R. Baeza-Yates, and B. Krellner. Text mining of business news for forecasting. In *Database and Expert Systems Applications, 2006. DEXA'06. 17th International Workshop on*, pages 171–175. IEEE, 2006.

E. Kumar. *Natural Language Processing*. I K International Publishing House Pvt. Ltd, 2011. ISBN 9380578776.

B. Latour. 10 tarde's idea of quantification. *The social after Gabriel Tarde: debates and assessments*, page 145, 2010.

T. S. Lee and D. Mumford. Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448, 2003.

E. D. Liddy. Natural language processing. In *In Encyclopedia of Library and Information Science, 2nd Ed*. Marcel Decker, Inc, 2001.

D. Lin. Minipar: a minimalist parser. In *Maryland linguistics colloquium*, 1999.

L. Lloyd, P. Kaulgud, and S. Skiena. Newspapers vs. blogs: Who gets the scoop? In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 117–124, 2006.

S. Lohr. The age of big data. *New York Times*, 11, 2012.

G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, et al. The arab spring| the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International journal of communication*, 5:31, 2011.

D. Mage, G. Ozolins, P. Peterson, A. Webster, R. Orthofer, V. Vandeweerd, and M. Gwynne. Urban air pollution in megacities of the world. *Atmospheric Environment*, 30(5):681–686, 1996.

M. Mani and D. Wheeler. In search of pollution havens? dirty industry in the world economy, 1960 to 1995. *The Journal of Environment & Development*, 7(3):215–247, 1998.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

N. McCullough, B. Sims, T. Ballas, and J. Davis. What is privacy? *Epistimi*, 8, 2014.

G. A. Miller. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41, 1995.

N. Newman. The rise of social media and its impact on mainstream journalism. *Reuters Institute for the Study of Journalism*, 8(2):1–5, 2009.

N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *Security and Privacy (SP), 2013 IEEE Symposium on*, pages 541–555. IEEE, 2013.

H. Nissenbaum. Toward an approach to privacy in public: challenges of information technology. *Ethics & Behavior*, 7(3):207–219, 1997.

J. Nivre. Dependency grammar and dependency parsing. *MSI report*, 5133(1959):1–32, 2005.

S. Y. Nof. *Springer handbook of automation*. Springer Science & Business Media, 2009.

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

A. Popescu-Belis. Evaluation of natural language processing systems: a model for coherence verification of quality measures. *A Blueprint for a General Infrastructure for Natural Language Processing Systems Evaluation Using Semi-Automatic Quantitative Black Box Approach in a Multilingual Environment. ELSE Project LE4-8340 (Evaluation in Language and Speech Engineering)*, 1999.

M. E. Porter and M. R. Kramer. Strategy and society. *Harvard business review*, 84(12): 78–92, 2006.

L. Rabiner and B.-H. Juang. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.

C. J. V. Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979. ISBN 0408709294.

S. Rill, J. Scheidt, J. Drescher, O. Schütz, D. Reinel, and F. Wogenstein. A generic approach to generate opinion lists of phrases for opinion mining applications. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 7. ACM, 2012.

G. Rizzo and R. Troncy. NERD: evaluating named entity recognition tools in the web of data. In *ISWC 2011, Workshop on Web Scale Knowledge Extraction (WEKEX'11), October 23-27, 2011, Bonn, Germany*, Bonn, GERMANY, 10 2011. URL `http://www.eurecom.fr/publication/3517`. Accessed: 2015-03-02.

V. L. Rubin and E. D. Liddy. Assessing credibility of weblogs. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 187–190, 2006.

H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK, 1994.

A. Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24 (4):35–43, 2001.

S. Skiena. Dijkstra's algorithm. *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica, Reading, MA: Addison-Wesley*, pages 225–227, 1990.

D. D. Sleator and D. Temperley. Parsing english with a link grammar. *CoRR*, abs/cmp-lg/9508004, 1995.

R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, 2013.

S. Soderland. Learning information extraction rules for semi-structured and free text. *Machine learning*, 34(1-3):233–272, 1999.

J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter*, 1(2):12–23, 2000.

W. Thomson. *Popular lectures and addresses*. MacMillan and Company, London, 1894. ISBN 9781108029780.

Tomaz Kovaziz. Comparison and evaluation of text extraction algorithms. `http://readwrite.com/2011/06/10/head-to-head-comparison-of-tex`, 2014. Accessed: 2014-05-20.

S.-i. Toyabe. Detecting inpatient falls by using natural language processing of electronic medical records. *BMC health services research*, 12(1):448, 2012.

L. Trindade, H. Wang, W. Blackburn, and N. Rooney. Effective sentiment classification based on words and word senses. In *Machine Learning and Cybernetics (ICMLC), 2013 International Conference on*, volume 1, pages 277–284. IEEE, 2013.

B. Ulicny and K. Baclawski. New metrics for newsblog credibility. In *International Conference on Web and Social Media*, 2007. Available at `http://www.icwsm.org/papers/4--Ulicny-Baclawski.pdf`, Accessed: 2015-03-02.

University of East Anglia. Sustainability incidents. `http://www.uea.ac.uk/estates/environmentalpolicy/environmental-incident`, 2014. Accessed: 2014-02-20.

S. Vachon and R. D. Klassen. Environmental management and manufacturing performance: the role of collaboration in the supply chain. *International journal of production economics*, 111(2):299–315, 2008.

L. Van Wel and L. Royakkers. Ethical issues in web data mining. *Ethics and Information Technology*, 6(2):129–140, 2004.

A. Voutilainen. Part-of-speech tagging. *The Oxford handbook of computational linguistics*, pages 219–232, 2003.

H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120. Association for Computational Linguistics, 2012.

Y.-Y. Wang, A. Acero, C. Chelba, B. J. Frey, and L. Wong. Combination of statistical and rule-based approaches for spoken language understanding. In *INTERSPEECH*. International Conference on Spoken Language Processing, 2002.

M. Weaver. Iran's 'twitter revolution' was exaggerated, says editor, June 2009. URL http://www.theguardian.com/world/2010/jun/09/iran-twitter-revolution-protests. Accessed: 2015-03-02.

W. Weerkamp and M. de Rijke. Credibility-inspired ranking for blog post retrieval. *Information retrieval*, 15(3-4):243–277, 2012.

J. Weng and B.-S. Lee. Event detection in twitter. *International Conference on Web and Social Media*, 11:401–408, 2011.

M. Widenius and D. Axmark. *MySQL reference manual: documentation from the source*. Ö'Reilly Media, Inc.", 2002.

D. Winer. Rss 2.0 specification. *Berkman Center for Internet & Society at Harvard Law School.*, 2002.

L. Wittgenstein. *Tractatus logico-philosophicus. Logisch-philosophische Abhandlung*. Suhrkamp, [Frankfurt am Main, 1963. ISBN 3518100122 9783518100127.

F. Wogenstein, J. Drescher, D. Reinel, S. Rill, and J. Scheidt. Evaluation of an algorithm for aspect-based opinion mining using a lexicon-based approach. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, pages 5:1–5:8, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2332-1.

World Of Computing. Rule based pos tagging. http://language.worldofcomputing.net/pos-tagging/rule-based-pos-tagging.html. Accessed: 2014-04-20.

T.-F. Yen, Y. Xie, F. Yu, R. P. Yu, and M. Abadi. *Host Fingerprinting and Tracking on the Web: Privacy and Security Implications*. 2012.

J. Yoon. Detecting weak signals for long-term business opportunities using text mining of web news. *Expert Systems with Applications*, 39(16):12543–12550, 2012.

Y. Zhang, R. Jin, and Z.-H. Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, 2010.