Universität Zürich[UZH]

Bachelorarbeit
zur Erlangung des akademischen Grades
Bachelor of Science
der Wirtschaftswissenschaftlichen Fakultät der Universität Zürich

Neural Approaches to Sentiment Inference

Verfasser: Dylan Massey
Matrikel-Nr: 17-734-583

Referent: Prof. Dr. Martin Volk

Betreuer: Dr. Manfred Klenner

Institut für Computerlinguistik

Abgabedatum: 01.05.2023

Abstract

In recent years, increased attention in research has been devoted to the Sentiment Analysis (SA) of texts that express positive and negative attitudes more subtly, such as news articles related to politics. The field of research dedicated to inferring such subtle attitudes from text is known as Sentiment Inference (SI). The precise goal of SI is to find out who is opposed to / in favour of whom or what in a given text or who / what is good for / bad for what / whom in a given text. Until now, only a rule-based system has been available for performing SI in the German language. The aim of the present thesis is to investigate and assess the viability of two different neural approaches for German SI, and to compare the two. One approach relies on a text-to-graph Semantic Parser, while the other relies on two separately trained models for entity recognition and relation classification. Since the neural approaches in this thesis rely on training data, and because such data is not readily available for German, the rule-based system is used to generate a silver standard dataset on which the neural approaches are trained and assessed. This thesis provides a first baseline for neural German SI and aims to point out potential directions for further research in this field.

Zusammenfassung

Die Sentimentanalyse von Texten mit subtil ausgedrückten positiven sowie negativen Einstellungen und Haltungen wie sie beispielsweise in Nachrichtentexten mit Politikbezug zu finden sind hat in den vergangenen Jahren in der Forschung an Bedeutung gewonnen. Das Forschungsgebiet, solche subtilen Einstellungen in Texten zu erkennen, wird als Sentimeninferenz (SI) bezeichnet. Das Ziel von SI ist es also herauszufinden, wer in einem gegebenen Text gegen wen oder gegen was ist oder wer / was in einem Text gut oder schlecht für etwas ist. Für die deutsche Sprache gibt es bisher nur ein regelbasiertes System zur SI. Das Ziel dieser Arbeit ist es, die Viabilität von zwei verschiedenen neuronalen Ansätzen für eine deutsche SI zu untersuchen und zu bewerten sowie die beiden gewählten Ansätze miteinander zu vergleichen. Der eine Ansatz basiert auf einem text-to-graph Semantic Parser, während der andere auf zwei separat trainierten Modellen für Entitätserkennung respektive Relationsklassifizierung basiert. Da die in dieser Arbeit gewählten neuronalen Ansätze auf Trainingsdaten angewiesen sind und solche Daten für die deutsche Sprache nicht verfügbar sind, wird das regelbasierte System verwendet, um einen Silver Standard Datensatz zu generieren, auf dem die neuronalen Ansätze trainiert und bewertet werden. Diese Arbeit stellt eine erste Basis für eine neuronale deutsche SI dar und zielt darauf ab, weitere mögliche Forschungsrichtungen aufzuzeigen.

Acknowledgement

First and foremost I would like to acknowledge the support and encouragement I received from my advisor, Manfred Klenner in the form of discussions and pointers to relevant literature. Furthermore, I would like to thank Anne Göhring for supporting me in discussions, with literature pointers, by proof-reading my manuscript and by giving me a place to work. I would also like to thank Martin Volk and the other members of the Institute for Computational Linguistics for making it possible to explore the fascinating discipline of Computational Linguistics here at the University of Zurich. I would also like to thank Francesco Tinner for proof-reading my manuscript. Last but not least, I want to thank my family and my friends for enduring me during the highs and lows of my writing this thesis.

Contents

List of Figures

List of Tables

List of Acronyms

DG	Dependency Grammar
DL	Deep Learning
DS	Dataset
ER	Entity Recognition
ML	Machine Learning
NLP	Natural Language Processing
OIE	Open Information Extraction
ORE	Opinion Role Extraction
PAS	Predicate Argument Structure
RE	Relation Extraction
RQ	Research Question
SI	Sentiment Inference
SMD	Swiss Media Database
SRL	Semantic Role Labelling
SSA	Structured Sentiment Analysis
XML	eXtensible Markup Language

1 Introduction

1.1 Motivation

Recent research in the field of Sentiment Analysis (SA) has increasingly paid attention to the extraction of more subtly expressed sentiment in text. This stands in contrast to texts expressing sentiment more explicitly, such as text which contains an explicitly charged expression like "hate" or "terrible" [Hamborg and Donnay, 2021, p. 1663]. Inferring implied sentiments from texts, in which no explicit lexical items related to sentiments can be found, is especially interesting for text types in which attitudes are expressed or *attributed* in more subtle ways, such as newspaper articles for example [Jacobs and Hoste, 2021; Hamborg and Donnay, 2021]. A guiding question for such a SA of newspaper articles can be the identification of "who or what is portrayed how (by whom)?", or "Who is negative/positive toward X?" [Stoyanov et al., 2005]. Clearly, finding answers to such questions in a wide array of text types can prove of immense value in domains such as intelligence, policy analysis or finance.

Implied sentiment and the means by which such sentiment is elicited, Sentiment Inference (SI), have so far mainly been implemented as rule-based systems. For the German language, a rule-based system for the analysis of implied sentiment has been developed by Klenner et al. [2017]. The approach of Klenner et al. [2017] is verb-centered, which means that the verb in a statement plays an important part in deciding who / what is positive / negative towards what / whom. A statement such as *X loves Y* would thus lead to the conclusion that X is positive towards Y.

Rule-based SI systems that make use of lexicons however do not generalize well on unseen or unspecified input data. This means that, for example, if a negatively connoted word is out-of-vocabulary (OOV)[1], it will not be recognized by the system, which leads to a recall performance problem. A neural system could reduce the OOV problem and potentially offer better generalizability. For example, if the neural system is trained on the relation *X loves Y*, but not on *X adores Y*, the model might

[1]This means the connoted word is not present in the lexicon of a rule-based system.

still be able to infer that X is positive towards Y for the latter statement at inference time due to the distributional semantics usually underlying neural approaches. However, a neural approach to capture implied sentiment, or more specifically, the polar relation between two entities in a German text does not yet exist to the present author's best knowledge.

This thesis therefore aims to offer an initial perspective on neural approaches for performing sentiment inference on German texts. It is centred on a study designed to apply data generated by a rule-based SI system in a semi-supervised fashion to neural models initially developed for other tasks in NLP, namely Entity Recognition, Relation Extraction and Semantic Parsing.

1.2 Research Questions

The study presented here is guided by two primary research questions:

- **RQ1:** Can the task of Sentiment Inference (as defined in 2) on German newspaper articles be adequately replicated through a neural system? This question can be further subdivided into:
  - How well does the neural system work on a random train-test split?
  - How well does the neural system generalize on verbs[2] unseen in the training data?
- **RQ2:** Given two approaches for performing neural Sentiment Inference (as defined in 2), which architecture proves superior? In other words,
  - Does an end-to-end sequence-to-graph based approach or a pipeline-based approach perform better in relation to **RQ1**?

Furthermore, this thesis implicitly addresses the question of how well models originally designed for multilingual or purely English settings work for the German language. It also quantifies the quality of the generated silver standard dataset produced by the rule-based component.

1.3 Thesis Structure

In this first chapter we look at the motivation behind this study and state our research questions.
Chapter 2 introduces the reader to the notions of SI & SA, defines our task of SI, and then discusses some related tasks from which two candidate neural approaches for SI are then derived.
In Chapter 3 we discuss the methodology for answering our research questions and how we evaluate the two neural approaches introduced in Chapter 2.
Finally, Chapter 4 discusses the obtained results before ending with the conclusion in Chapter 5.

[2]We have discussed in the introduction that the rule-based system by Klenner et al. [2017] is verb-centered and cannot deal with OOV words (verbs). Strictly separating the verbs in the training set from verbs in the test/validation set can show us how well the neural approaches to SI can handle OOV words (verbs) and thus prove as a viable complement to a rule-based SI system.

1.4 Experimental Section

In SI the goal is to determine the polar relation between sources and targets of attitudes, emotions, opinions and actions. The present thesis is concerned with the viability of neural approaches to SI. Neural approaches have proven to be successful in a variety of tasks in SA. For German, there has not been any implementation of a neural SI system to the present author's best knowledge. Since we strive for supervised approaches, data is needed. For German, however, no sufficiently large, manually labelled, SI dataset is available. A potential remedy to such a problem is the generation of a silver standard (a dataset that has been generated automatically). The silver standard gives insight into how effective the neural system "could" be on high-quality gold standard data. Further, we can also evaluate how good the silver standard is. Luckily, there is already a rule-based system for SI for German, which greatly eases the generation of a silver standard. We will use the rule-based system to generate 4 silver standard datasets that vary in two properties that attempt to answer our research questions.

2 Background & Related Work

In the following section we will attempt to explain and provide a working definition for what SA and SI are, by referring to key examples from the wide-ranging literature covering the field. We will show that SA and SI (which can be seen as a subfield of SA) are established fields and that tasks within these fields are extremely diverse. We will then focus on deriving the task of SI relevant for this study, namely determining any polar relation between a textual source[1] and a target within a sentence, which, we argue, is quasi-analogous to the goal of attitude relation prediction presented in Klenner et al. [2017]. Finally, some more linguistic (rather than conceptual or technical) aspects are discussed, which are necessitated by the later introduction of a rule-based system (section 2.3.1) covering the SI task which serves as the basis for the empirical part of this thesis. The overall goal of this section is to familiarize the reader with SI and offer a coherent conceptual (human-centered) as well as technical (NLP) perspective on the topic.

2.1 Definitions

2.1.1 Sentiment Analysis (SA)

SA can be considered a field of research that is concerned with the analysis of attitudes and emotions expressed in text [Liu, 2020, p. 1]. Liu [2020] further notes that SA often times acts as an umbrella term for a variety of connected, but subtly different tasks that are to be solved computationally. These tasks can vary in text type (e.g., tweets vs. newspaper articles) and granularity (i.e., determining whether a tweet is overall negative/positive vs. which entities mentioned in the tweet have positive/negative sentiment directed towards them), among other aspects.

[1] In the literature other terms used to describe "source" can occur, such as "holder". Holder, however, can be seen as more intentional than source. We will use the term source over the span of this thesis.

From a human-centered perspective, sentiment is a private state [Rambow and Wiebe, 2015, p. 7]. Private states are mental and emotional states with a source and an intensity [Wilson, 2008]. Sentiment can be further defined as negative and positive judgements, thoughts, attitudes and emotions prompted through feelings [Liu, 2020]. Private states and subsequently sentiment can be expressed in language [Wilson, 2008, p. 1]. Expressions of sentiment in language give rise to what are called subjective statements [Taboada, 2016, p. 3].

In its most basal form, SA is concerned with assigning some polarity, positive, negative, or neutral to a segment of text (e.g., sentence or phrase). For example, determining that the statement *This stock has plummeted* has a negative polarity[2] can be seen as a classical SA example. In some further cases, the goal might also be to determine the polarity a text expresses towards specific aspects, which has been used in product review analysis [Ravi and Ravi, 2015]. In such a case of finer-grained SA, where it must be determined whether the text is positive, negative, or maybe neutral towards some target, the term used to describe it is "aspect-based" (ABSA). A variety of methods have been used to technically realize SA systems, such as lexicon-based approaches, ML (Machine Learning)-based approaches or DL (Deep Learning)-based approaches, on which we will not further elaborate at this stage.

2.1.2 Sentiment Inference (SI)

SA can become more sophisticated when an attempt is made to additionally identify the **source** of the sentiment expressed. In such a case we want to elicit who is against/for what/whom in a given segment of text. In other words, we want to determine the **polar relation**[3] that exists between a source and a target. In figure 1 below, we see that the sentence is "the minister hates the terrorist", hassen (to hate) is a directly exposed private state that tells the receiver (reader of the statement) that the minister has a negative attitude towards the terrorist. It is directly exposed because there is an explicit expression (the verb), which gives us a cue about polarity of the relation from the minister towards the terrorist. Further, *hassen* can be seen as an opinion verb [Wiegand and Ruppenhofer, 2015], by which the source expresses

[2] Another similar term for polarity used in the literature is sometimes valence [Neviarouskaya et al., 2009]. We will use polarity in this thesis and not further differentiate between the two terms. Sometimes also charge is used.

[3] As we will see below, further similar terms used for describing a charged relation between entities could be attitude relation [Klenner et al., 2017] and good for/ bad for (gfbf) relation [Deng et al., 2013]. We use the most generic form, polar relation, which we frame as an overarching term in this thesis.

a position on a particular entity, action, or state underlain by the attitude. As we will see below, not all authors would call the identification of such a relation SI (for example Wiebe and Deng [2014] define the task more narrowly). We, however, already call such identification of a relation SI, since it is implied by an author that the minister is against the terrorist. What is implied can be inferred. It follows that SI is a subfield of SA, where the goal is to determine the polar relation between entities.

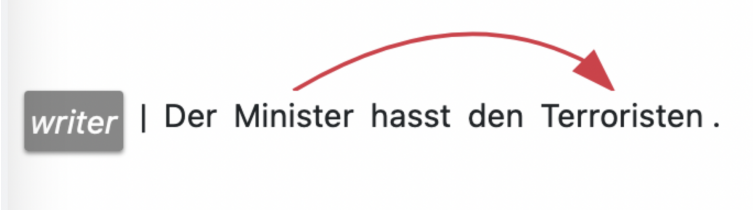

Figure 1: A German example of explicit sentiment between two entities within a sentence. (Authors own diagram, 2023)

Beyond the wider definition of SI above, another form stems from the realisation that some private states (as introduced in section 2.1.1) can be inferred from other private states and must not be explicitly expressed in text [Wiebe and Deng, 2014]. Inference rules may be applied in order to obtain implicit sentiments held by the writer. For example, in the sentence: *The president is going to fight against the* ***terrible*** *market conditions.* A term that explicitly shows the writer's opinion on the market conditions is ***terrible***, which would be elicited by a conventional (aspect-based) SA system. ***Terrible*** would commonly also be referred to as an (explicit) subjective expression (SE) within the domain of SI, which is "[...]any word or phrase used to express an opinion, emotion, evaluation, stance, speculation[...]" [Wilson et al., 2005, p. 2]. However, the sentence also suggests that the writer may be in favour of the president since they (the president) are fighting against something that the writer portrays as terrible. Colloquially speaking, we could say that the rule the "enemy (president) of my enemy (market conditions) is my friend (president)" applies, which is none other than an inference rule.

On a further note, polar relations can be established between two entities from seemingly objective statements[3]. Objective statements can describe events. From the objective statement that a *lorry hits a car* we can infer that the lorry is bad for the car, or in the statement *the policeman saves the child* we can infer that the policeman is good for the child. SI therefore further is the field of study of the

[3]It is difficult to say whether *X hates Y* can be regarded as an objective statement. But since we count hate as a subjective expression, such a statement is subjective for us. Objective statements are statements that are absent of expressions explicitly exposing private states.

means by which private states such as attitudes, opinions, judgements, thoughts and emotions and more generally polar relations can be *inferred* from seemingly objective statements. What is important is that, if a lorry hits the car, no a priori **intentionality** is given as opposed to the above forms of SI. Terms such as Opinion Implicature (Wiebe and Deng [2014]), an implicature being understood as the "the act of meaning or implying one thing by saying something else" [Davis, 2019] and Sentiment Propagation (Deng and Wiebe [2014]) have also been used instead of SI to describe similar phenomena, as Klenner and Göhring [2022, p. 147] point out. In the present thesis, however, SI will be used throughout.

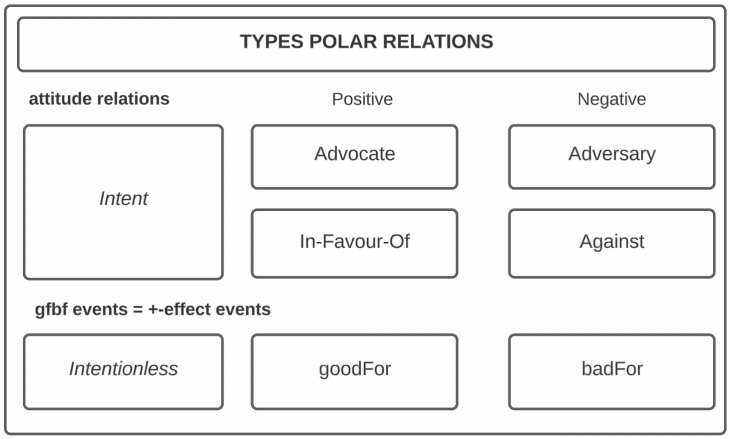

Figure 2: Types of polar relations between entities present in the literature. Broadly polar relations discussed in the literature appear to be categorizable by their level of intent. (Author's own diagram, 2023)

A central notion present in a sizeable portion of surveyed literature on SI appears to be the notion of the **event** [Deng et al., 2013; Ding and Riloff, 2016]. Events can, for example, be actions (a terrorist kills) and changes of state (a dam breaks). It can be argued that events are a semantic primitive, in the sense that they cannot be described in simpler terms but are nevertheless understood universally. Events happen and can be reported on in objective statements, as shown in the example with the "market conditions". Events have participants, such as for example entities that cause events or objects affected by them. Events can be benefactive or malefactive (see Deng et al. [2013], and below) towards entities (including participants), in which case we call such events **affective events**, in line with [Ding and Riloff, 2018, p. 1919], but argue that affective events can not only positively or negatively affect people but also inanimate things. The target of an event (which is an entity), can be affected positively or negatively. Affective events induce a polar relation between its participants [Klenner et al., 2017]. The type of charge[4] of the polar relation induced

[4]By charge here simply positive or negative is meant. Other descriptors have also been used in

by an event is not only dependent on the target.

On a related note, Deng et al. [2013] introduce an annotation scheme that aims to capture how certain events negatively or positively affect their participants. They distinguish between benefactive "goodFor" and malefactive "badFor" events. An event that falls into either category is abbreviated as a *gfbf event*. More concretely, their annotation scheme stipulates that each annotated event is representable as an (agent, gfbf event, object)-triple. Each element of the triple is to be a contiguous span of text. The triple can be agentless, but ought to contain an object. For example the sentence *he is killed*, does not contain an agent but is a gfbf event with the triple (-, kill, he). The two aforementioned characteristics are introduced in order to highlight the parallels from the field of SA to related NLP tasks discussed in 2.2. Gfbf events can also be seen as included in our notion of the polar relation. The subject of the event has, however, as is displayed in figure 2, not necessarily any intent. If the *snow* ***blocks*** *the car drivers* this might imply that the described event is badFor the car drivers, but it does not mean that the snow is adversary of the car drivers, since it (snow) does not have any attitude. If the *demonstrators block the drivers* we also cannot necessarily ascribe a negative attitude of the demonstrators towards the driver, but we might at least assume that the event (and the demonstrators) are badFor the drivers. Such relations between animate entites are included within our task of SI.

Verbs (as a syntactical category) are a type of event predicate that can be used to denote events in statements. Verbs that denote affective events are polar verbs. For instance, as we saw in the example in the prior paragraph, it is not only implied that the writer may be in favour of the president, but there is also a necessary presupposition that the president may be against the (current) market conditions, since they are ***fighting*** against them. Therefore, the verb serves as the cue of a negatively charged polar relation between its two fillers[5] (from the president towards the market conditions). On a more conceptual level, we could state that the event of FIGHTING implies a negative polar relation from the aggressor towards the entity aggressed upon. The president is an adversary of the market conditions, or more naïvely put, they (the president) are against them.

However, it would be wrong to assume that the event predicate - which will be called the verb from now on - alone is necessarily sufficient to mediate (and trigger) polarity between its participating entities. There is also a dependency on the type

the literature such as (against, in-favour-of), (adversary, advocate), (benefactive, malefactive) and (badFor, goodFor). We note that there are subtle differences between these notions.

[5]Fillers are the roles or arguments associated with a verb. More on this concept is introduced in section 2.2.1.

of event participants, and the polar connotation[6] applied to them. Klenner et al. [2017] defines a polar relation — which he conceptualizes more narrowly as attitude relation — of a verb as being entirely dependent on the polar assignment of the target role. Further, a verb can cast positive (peff) and negative effects (neff) on its fillers. Such **effects** a verb exerts on its participants (and which reside behind or beneath the actual lexical items used) are outside the scope of this thesis and emphasis is put exclusively on the relation between a source and its target. The reader may consult Klenner [2015]. These effects are, however, visible in the various examples, as the reader might notice. The cast effects are not dependent on the verb in its entirety, but also on the its fillers, such as when someone ***kills*** *a terrorist*, that someone might get assigned the role of being a positive actor (pac) in the event. In the more general case however, the actor killing someone makes them negative (nac).

We have seen that approaches (and concrete tasks) to SI differ from each other in terms of how inference is defined and rules are expressed, which is in line with [Deng et al., 2014, p. 108]. Some approaches aim to determine how a writer may be positive or negative towards entities in a text where no explicitly voiced opinion can be found, while other approaches base their inferences on the notion of events (shown in figure 3). However, a common characteristic between all the consulted work is the goal of establishing some *type* of polar relation between entities in discourse [Klenner and Göhring, 2022], which serves as our working definition of what SI is for this thesis. Similarly, [Choi et al., 2016, p. 333] define their SI as the goal of "detecting all directed opinions for all entities in a given text". Their definition of the task is however different since they focus solely on opinions, while we include any type of polar relation (attitudes, opinions, events).

In this thesis SI is operationalised as the verb-mediated extraction of polar relations[7] between real[8] textual participants on the sub-sentential level. We restrict our investigations to two different entity types, a textual[9] opinion source and a textual opinion target.

To give the reader a further understanding of what the SI task in this study exactly entails, we will now discuss a few examples, as shown in Figure 4. Example **(I)** illus-

[6] In our case, we understand the polar connotation of a word to be the conventional polar association of it (see Löbner [2013, p. 36]). The verb "rescue", for example, is conventionally associated with positivity.

[7] We use the umbrella term polar relation, since our notion of SI is more widely defined compared to other authors.

[8] As real we mean common and proper nouns, rather than pronouns.

[9] Textual is explicated here, since, as discussed above, an opinion source could either be the writer of a sentence or could indeed be in a different sentence of the text (anaphora)

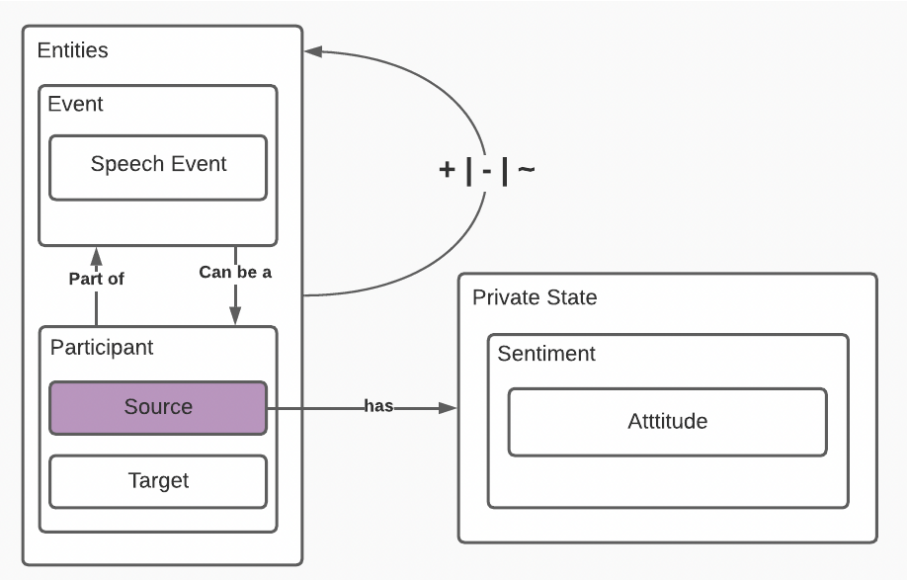

Figure 3: An entity-relationship diagram of concepts around SA / SI. + / - / ∼ (neutral) stands for an attitude relation (more narrowly defined than the polar relation) between entities, or a neutral relation. Events can contain relations to other events, such as in a sentence like *She criticized the attack.* (Author's own diagram, 2023)

trates a reciprocal polar relation between the protesters and the car drivers. Given the event described by the verb ("to impede"), we assume they are both adverse to each other, a plausible assumption if one considers recent protests staged by the group "Die letzte Generation"[10]. Example **(II)** signifies a positive polar relation between a doctor helping children in need, and so we infer a positive attitude from the doctor towards the children. **(III)** was already discussed above, the implication being that the president is negative towards something they are fighting against (in a figurative sense). **(IV)** is a rather special case but shows that polar relations can theoretically be reflexive. Here the event of suicide can clearly be seen as an act of violence against oneself, which the informal German term "Selbstmord" accentuates. We shall not consider such examples further, firstly since the word "Ich" is not a real entity as mandated in our task of SI above, and secondly because we only consider textual entities, which the writer is not a part of. The sentence is nonetheless useful to illustrate the concept of polar relations as they have been introduced here.

Our task also deals with more complex cases, otherwise known as nested events, as illustrated in **(V)** and **(VI)**. In **(V)** the journalist is judged to be an adversary of the APPRAISAL event since they criticize it (which is itself an action event). It can

[10]https://letztegeneration.de/

therefore be further inferred that the journalist is an adversary of the minister. An interesting point to highlight is that APPRAISAL is realised not as a verb, but as a noun (nominalisation). In such cases, the positive attitude of the senate towards the appraisal is therefore not accounted for, since we focus on verb-mediated polar relations and not nominalisations of verbs reporting events. At the same time, the senate is the agent in the APPRAISAL, with the implication that the senate is pro (for) the minister. Similarly, the population in **(VI)** is seen to be in-favour-of the Syrians since they REGRET the event of the Syrians being LET-DOWN. Being LET-DOWN in such a context, may very well also imply that a negative sentiment towards the "down-letter" is bred, that is, that (some) Syrians are "against" the USA.

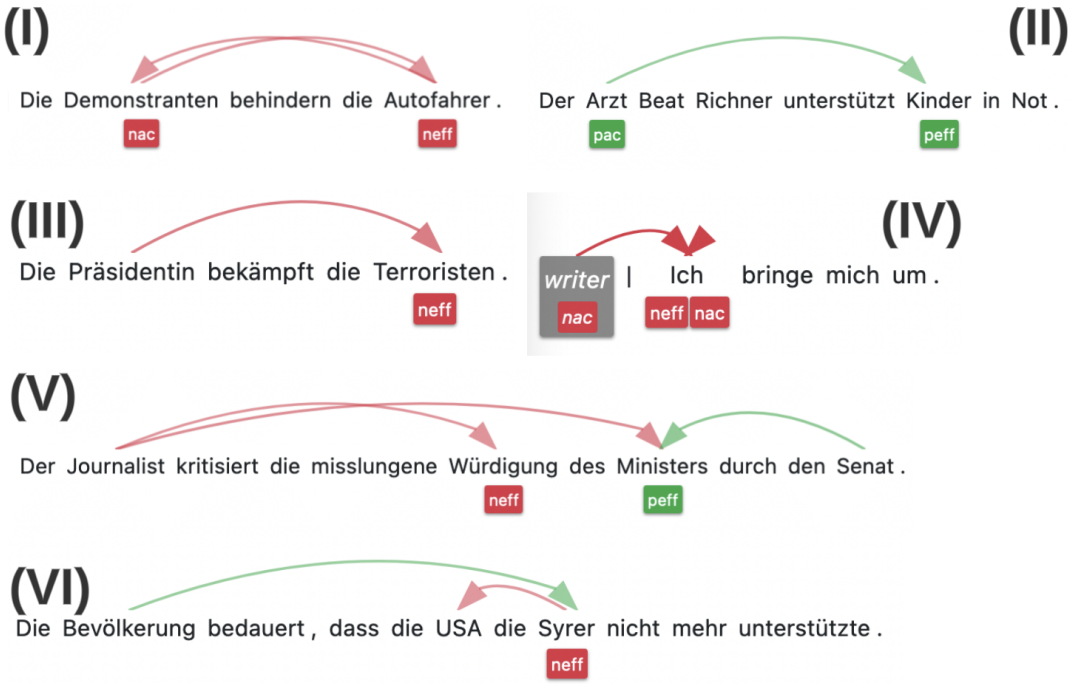

Figure 4: 6 example sentences where implied sentiments are extracted. **(I)**: The protesters impede the car drivers. - **(II)**: The doctor Beat Richner supports children in need. - **(III)**: The president fights the terrorists. - **(IV)**: I kill myself. - **(V)**: The journalist criticizes the failed appraisal of the minister by the senate. - **(VI)**: The population regrets, that the USA didn't support the Syrians more. Generated with the help of the stancer by Klenner [2015] further discussed in section 2.3.1. (Authors own diagram, 2023)

We have looked at what SA and SI are and defined our task of SI. We have also offered a broad conceptual overview of "where" our study is situated in. We further explained our approach to SI by providing six salient examples. Conceptual notions have prevailed so far over linguistic aspects in this discussion. As a next step, the

linguistic aspects of compositionality and negation are introduced.

2.1.3 Aspects

2.1.3.1 Compositionality

Frege's principle of compositionality claims that the "[...] meaning of the whole is a function of the meaning of the parts and their mode of combination." [Dowty et al., 2012, p. 8]. It has also been claimed that this principle can be harnessed for a more accurate SA [Moilanen and Pulman, 2007].

Moilanen and Pulman [2007] use the compositionality principle for sentiment classification on the sentence level and argue that compositionality can overcome problems caused by "naïve" SA (counting how many words are positive vs. negative and classifying by the majority vote). As a proof of concept, they present an example in which such a naïve approach would result in deadlock (equal amount of + and - words). The solution for the deadlock is based on the idea that the polarity of a sentence is defined bottom-up (over a syntax tree), from lexical items (e.g., words) with a polarity (e.g., their connotation) via the constituents to the sentence, a process they call *sentiment propagation*[11]. A series of rules based on syntactic patterns govern which constituent dominates over the other: given a composition of two constituents, where each constituent can be + / - / $\sim$ (neutral), what will be the polarity of the resulting constituent?

2.1.3.2 Negation

Negation changes the meaning of an event ("not to accept" and "to accept" are different events) that is described in a statement. However, due to the asymmetric nature of negation [Taboada, 2016, p. 16-22] it cannot be assumed that the negated form of an event also automatically leads to an inversion of a polar relation it induces. Negation can be regarded as distinct from compositionality [Tron, 2013, p. 15-19]. For example, if someone states that the "food was not bad" we cannot simply invert the polarity of the statement and assume it is equivalent to the statement that the "food was good". As we can see in figure 6, to NOT-ACCEPT someone implies an adversary relation between *him* and *terrorist*, whereas ACCEPT implies (at least a weak) positive attitude from *him* towards the *terrorist*. It is evident that to NOT-ACCEPT something may induce a stronger attitude (in negative intensity) than

[11]Different from the sentiment propagation by Deng and Wiebe [2014] discussed in Section 2.1.2

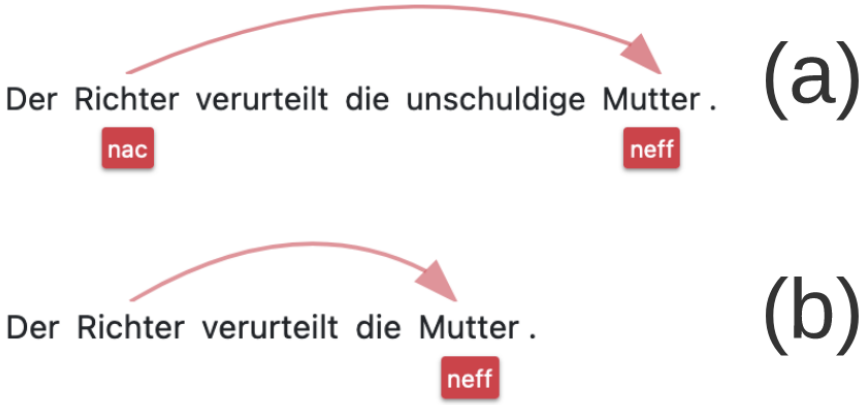

Figure 5: The sentence "the judge sentences the ***innocent*** mother" Example of how the modifier "unschuldig" (innocent) has an effect on the judge, who becomes a negative actor (nac) when the mother is innocent (a). The mother, in either case, is the bearer of a negative effect (neff).

to ACCEPT something does in positive attitude intensity. This is indicative of an asymmetry.

Because intensity lies outside the scope of the present study and thesis, asymmetries in negation are only relevant if a negation word within a statement renders a polar relation void or if it induces a polar relation. In other words, negation is relevant insofar that a negation word within a statement could bring a neutral attitude between entities into a charged state and vice versa. An example of how negation can impact (invert) the charge of the polar relation is shown in figure 6.

2.2 Related Tasks

Deng and Wiebe [2016] also discuss how topics and systems in SA and NLP can mutually profit from each other. More pertinently, Liu [2020, p. 15] notes that "[...] every subproblem of NLP is also a subproblem of sentiment analysis, and vice versa". They propose that rather than SA (and thus also SI) being a subfield of NLP, it is its own mini-NLP world. We agree and argue that advancements in NLP topics such as information extraction (IE) or semantic analysis do indeed offer promising perspectives, including in our case. This stance will form the foundation for the empirical part of this thesis. As Wilson [2008, p. 3] describes in her thesis, SA has

Figure 6: How the polar charge can switch with a negation indicator such as “nicht”. The subject (he) who accepts the terrorist is an advocate of the terrorist (and also, from an exterior perspective, a negative agent). The subject (he) who does not accept the terrorist is an adversary of the terrorist.

helped NLP progress on tasks such as question answering (expanding answerable questions, “who is opposed to who”, as noted in the motivation of this thesis) and IE (filtering out subjective statements). In the following section, we will discuss how advances in IE can help our task.

2.2.1 Open Information Extraction

A part of the IE field of study is the task of the extraction of - so called - SRA (Subject-Relation-Argument) triples [Del Corro and Gemulla, 2013, p. 2]. According to [Del Corro and Gemulla, 2013, p. 1], Open Information Extraction (OIE) ”aims to obtain a shallow semantic representation of large amounts of natural-language text in the form of verbs (or verbal phrases) and their arguments”. In this context, we would like to stress that OIE deals with the surface forms of text, as Soares et al. [2019] notes. A logical consequence is that many of the triples will have similar, but not identical, relations, although they are essentially the same. In other words, relations can “essentially” be the same, but vary in their surface forms, such as when a verb is in past tense, rather than present tense. For example, the triple (Peter, condemns, attack) and (Peter, has condemned, attacks) are not the same but are similar.

As we have already mentioned in section 2.1.2, verbs act as event predicates. Therefore Event Argument Extraction is another similar task that is a subtask of Event Extraction (EE), which is part of the larger field of IE Ellis et al. [2014] and also has the potential to be concerned with the extraction of triples from text. Similar tasks to OIE can be considered Semantic Role Labelling (SRL) and Predicate Argument Structure (PAS) extraction. The "roles" targeted in SRL are mostly the same as the arguments in OIE [Christensen et al., 2011], while Klenner [2005] develops a rule-based approach for PAS extraction in German.

A variety of published literature represents sentiments expressed in (or inferred from) text - on the sub-sentential entity/expression/event level - as triples over which a polarity exists [Kim and Hovy, 2006; Deng et al., 2013; Wiebe and Deng, 2014; Ding and Riloff, 2016; Barnes et al., 2021]. A core component of the modelled triples appears to be a given word that "mediates" sentiment from a source towards a target. An example of such a triple was already given in section 2.1.2, with the *gfbf* events posited by Deng et al. [2013]. A more recent example, which is not primarily concerned with eliciting implicit sentiment, is Barnes et al. [2021]. Here, an opinion is defined as a tuple $O = (h, t, e, p)$, where h is defined as holder (source), t is the target, e is a (not necessarily contiguous) text indicating sentiment and p is the polarity (+/-/$\sim$ (neutral)). The elements of the sub-tuple (h, t, e) appear similar to the aforementioned *gfbf* triples. Whatever the model, triple extraction is obviously vital for the performance of the subsequent downstream task of classifying any possible sentiment in a sentence.

Finally, a task within SA that appears to be a subtask of triple extraction and therefore also appears to be related to IE is Opinion Role Extraction (ORE), which can be defined as the "assignment of opinion source and target given some opinion verb" [Wiegand and Ruppenhofer, 2015, p. 215]. By extension, Bamberg et al. [2022] take a non-verb-centered approach and mention the similarity of ORE to the task of SRL. They set out to exploit the link between the two tasks by using SRL data to improve ORE and achieve state-of-the-art performance on the IGGSA-STEPS dataset from Ruppenhofer and Struss [2016]. The promising results generated by using such a transformers-based model for token classification in ORE encouraged us to train a similar entity labelled on the entire triple extraction task and use transformers (as Bamberg et al. [2022] did) as our base model. This is described in section 2.3.2.1.

2.2.2 Relation Classification

Relation Classification has the goal of predicting what kind of relation two entities $e1$ or $e2$ have between each other (e.g. lives-in or born-on) [Lyu and Chen, 2021]. More formally, given some statement x, which contains within it, two entities $e1$ and $e2$, we are interested in learning a representation of the relation expressed in statement x between the two entities [Soares et al., 2019]. It is clear that Relation Classification presupposes some cues as to where the entities in a statement are to be found.

Soares et al. [2019] experiment with different possibilities of transformers-based representations (from BERT) for relation classification. First, they just feed the standard input into the encoder without providing any further information about the entities for which a relation is to be predicted. To explicitly mark the entities in question, they then use a) positional embeddings and b) entity marker tokens. The outputs of the model are then combined into fixed-length representations in different ways (see figure 7) such as using the CLS-token ((a) & (d)), by performing entity mention pooling ((b), (c) & (e)) or by using the entity start state (f). They report that method (f) is the most performant.

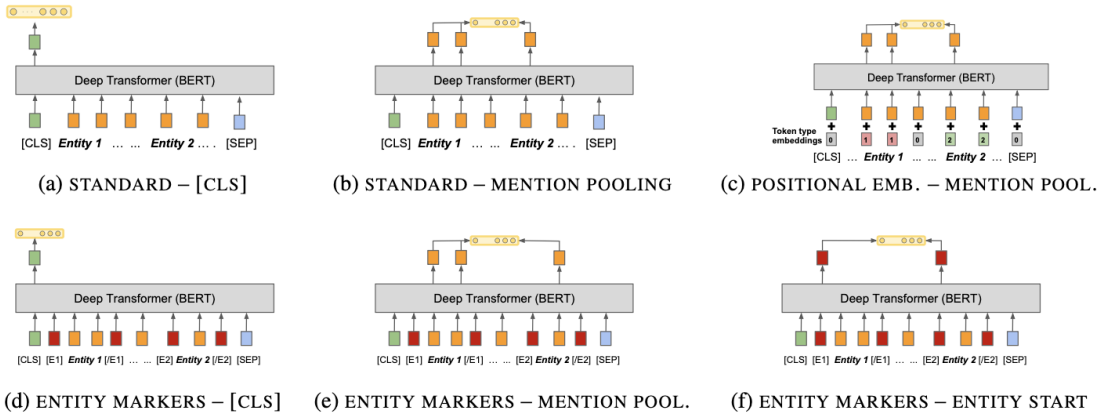

Figure 7: Different forms of relation representations. (Source: [Soares et al., 2019, p. 3]).

Based on the work of Soares et al. [2019], we will adopt a similar approach to (e). More precisely, we use entity mention pooling because it was already provided out-of-the-box in a web tutorial. We will, however, re-conceive the classical binary relation classification task as a ternary classification task since our SI task includes not only the identification of two entities and subsequent classification but also includes the identification of the relation mediating verb.

2.2.3 Semantic Parsing

The task of Semantic Parsing has its goal in bringing a sentence into a type of formal, machine-readable representation pertaining meaning [Samuel and Straka, 2020]. The resulting graph is a deep rather than shallow structure. Naturally therefore Semantic Parsers have seen applications in concept-level sentiment analysis [Agarwal et al., 2015; Cambria et al., 2022]. Since we also discussed SRL in section 2.2.1, we describe the relation between the two tasks as SRL[12] being the shallow version of Semantic Parsing, since it does not take compositionality into account. For example, in the sentence "the president criticizes the attack on the mosque by the terrorists", the attack is an argument of the criticizing event, with which SRL cannot deal in a straightforward way.

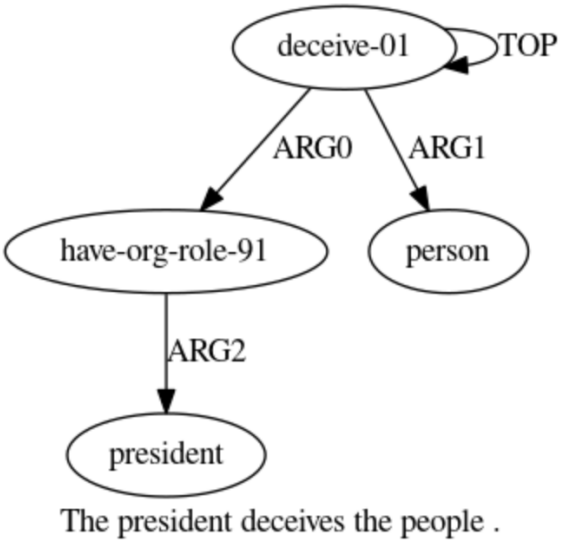

Figure 8: An example of a semantic graph in Abstract Meaning Representation (AMR), where the surface forms are brought into relation with "deeper" concept level forms. The sentence is "*The president* ***deceives*** *the people.*". Generated with https://bollin.inf.ed.ac.uk/amreager.html.

Samuel et al. [2022] apply their semantic parser PERIN (closer discussed in 2.3.2.2) to parsing text directly to a structured representation denoting a polar relation between entities in a sentence. Similar to our task of SI, they are not just interested in extracting source, target and polarity but also extract expressions mediating polarity (similar to how our verb mediates polarity). Since PERIN is a general text-to-graph parser, it can predict ternary relations. As mentioned, their approach

[12]SRL is sometimes also known as frame-semantic parsing in the literature. The theoretical background of SRL is frame-semantics.

is dedicated to explicit sentiment analysis, or more concretely is concerned with structured sentiment analysis[13] as defined by Barnes et al. [2022] and discussed in section 2.2.1. Their motivation is based on how Barnes et al. [2021] encode the graph for the task of structured sentiment analysis (example in figure 9), which is lossy. For us, the conversion performed in Barnes et al. [2021] is not lossy, since we have no multiword expressions in our prediction and are only interested in the heads (e.g., similar to the notion of the eTarget in [Deng et al., 2014, p. 180]). They experiment with different encodings that are non-lossy and find that the best way for encoding the information for their semantic parser is using what they call the opinion tuple representation. With their approach, they outperform a considerable amount of submissions in the SemEval 2022 Task 10. Since their model is available on GitHub publicly and shows a promising performance, we attempt to use this model for our SI task.

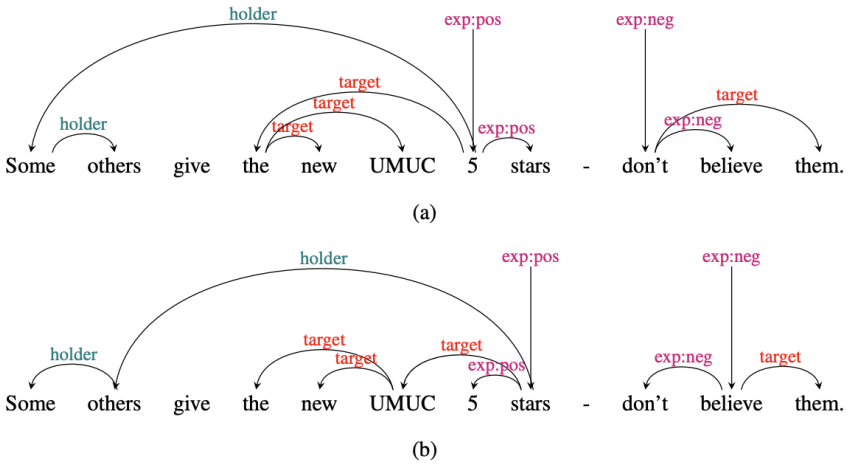

Figure 9: An example of a sentiment graph, head-first (a) and head-final (b). (Source: Barnes et al. [2021, p. 5]).

[13] Structure Sentiment Analysis can be seen as analogous to fine-grained SA as described in [Katiyar and Cardie, 2016].

We have now discussed how a variety of tasks or fields of study in NLP, including ***structured prediction tasks*** such as Semantic Parsing and Open Information Extraction relate and have been applied in the field of SA. We further justified the means by which we attempt to solve our task of SI in the empirical part of this section, both of the **approaches discussed in this thesis were chosen under the presupposition that if they work for explicit sentiment analysis they might also work for SI as we define it**. In the coming section we will describe the reference systems used in the empirical part of this thesis in more technical detail.

2.3 Technical Realisation

Our goal in this section is to describe the technical details of the rule-based system that solves our task of SI. We will then set out to explain the neural systems that attempt to solve the SI task.

2.3.1 Rule-based Approach

The rule-based approach described here [Klenner et al., 2017] attributes and detects polar relations from event statements. Informally we call it the **stancer**. It is based on a lexicon comprised of syntactical (subcategorization) frames around verbs. Depending on the triggered subcategorization frame for a parsed sentence and the connotation of words, for example, the presence of the word ***innocent***, a frame can trigger and induce a polar relation between a source and a target (or not). Additionally, there can be effects (negative effects (neff) and positive effects (peff)) cast on the sources and targets (which we will not follow up more closely here; see Klenner [2015]). Examples of outputs the stancer generates include figures 1, 4 and 6.

The stancer takes as input a sentence (or multiple sentences) and relies on a pass through a dependency parser [Sennrich et al., 2009] for named entity recognition, animacy detection (`bel`; detecting whether an NP-head is animate or not) and a coreference resolver [Tuggener, 2016] as well as the extraction of predicate-argument-structures (`pas`) for each clause (as discussed briefly in 2.2.1). Then the tokens (`words`) and lemmas (`sent`) of each sentence along with its PAS, grammatical functions, coreference information (`mycoref`) and dependency structure (`dep`) is fed into a Prolog logic program (the actual stancer), which performs the SI.

The outputs are all polar relations between sources and targets, as well as effects they have. The stancer is not publicly available.[14]

For example, in the sentence: "Er hat die Menschen, die ihre Kinder geschützt haben, diffamiert". *He **defamed** the people who **protected** their children.*, returned facts of the Prolog program would be:

```
[0-nac(1),0-neff(4),0-peff(8),0-c(1,4),0-p(4,8)]
```

As we can see, we get `c(1,4)` and `p(4,8)`, which are the polar relations between textual entities. The numbers between the brackets indicate the word positions of the original German sentence. The number preceding the predicates are related to the sentence number. Since only one sentence is fed into the system it is 0. The outputs make sense because "people" who protect their children are in-favour-of their children and the one who defames such an act of protection of them is against the people (and most likely also against the children). The `neff` and `nac` predicates are not in the scope of this thesis but are displayed here for completeness of the output.

2.3.2 Neural Approaches

Neural approaches appear to have been successfully employed on tasks in the field of SA. Katiyar and Cardie [2016, p. 920] claim to be the first ones to use a (neural) deep learning approach (LSTMs; see Hochreiter and Schmidhuber [1997]) for the full entity- and relation-extraction problem in SA, which means predicting sources, targets and subjective expressions and then linking them together. They frame the goal of fine-grained SA as predicting which entities are sources/targets/expressions, and as linking these by classifying from whom the expression is (target-source linking; assigning IS-FROM) and finding out about what or whom the opinion is (assigning IS-ABOUT). They are not concerned with determining the polarity.

We use transformers-based approaches, based on Devlin et al. [2019]; Vaswani et al. [2017] as they are an integral part of many of the state-of-the-art NLP systems and have appeared promising on a variety of related tasks as we have shown in Section 2.2. We will now describe in more detail the architectures of our chosen approaches.

[14] A demo can be found at: https://pub.cl.uzh.ch/demo/stancer/index.py/

2.3.2.1 Entity Recognition & Relation Extraction (ERRE)

Similarly to Katiyar and Cardie [2016], Barnes et al. [2022] provide a baseline for their competitive SemEval task of Structured Sentiment Analysis[15] (SSA). As one of their baselines they provide a model that contains three separate BiLSTM models to extract source, target and sentiment expression. The resultant representations are pooled and fed into a BiLSTM-based relation predictor. This approach can be seen as pipeline-based (two-step) because separate models are trained. None of the winning teams of the task uses the pipeline-based baseline as their basis, but instead most rely on the dependency-graph-based baseline further explicated in Barnes et al. [2021]. Admittedly, the pipeline-based approach has its limitations since it is prone to error propagation and no interactions are able to be modelled between the two subtasks [Zhang et al., 2019a, p. 56]. For example, an unrecognized or falsely recognized entity in the first step (entity recognition) can directly affect the performance of the downstream task (relation extractor). Since the training process of the two models is not connected, there is no possibility for the models to learn together in the intended setting.

The first model we are going to use for our task of SI will follow the pipeline-based route. As a justification for utilizing such a joint approach - even though, as discussed, it may suffer from error propagation - we rely on Zhong and Chen [2020] who, by using a joint approach on datasets not related to SA, achieved state-of-the-art performance for their transformers-based system **PURE**. Therefore, we will implement two separate models, an Entity Recognizer and a Relation Extractor[16]. The two models are trained independently of each other. One model is trained to recognize entities within text. In our specific case, this would mean deciding whether a given word is either source, target or verb. The second model is then trained for predicting the type of relationship that exists between these three extracted terms. This is analogous to a classification task. No multitask learning will be used, which is the idea of sharing parameters between the two tasks.

We shall abbreviate this model as **ERRE** (Entity Recognition & Relation Extraction) from here on in. As of current, our implementation of this model only supports the extraction of a single polar relation from a sentence, since we have not implemented a linker that determines which verb belongs to which source/target in the case of multiple opinions.

[15] Structured-sentiment-analysis SSA is concerned with extracting all opinions from a sentence in the form $O = (h, t, e, p)$ as described in section 2.2.1

[16] Some code of the relation extractor is adapted from https://github.com/sujitpal/ner-re-with-transformers-odsc2022, and is adapted to support three entities.

2.3.2.2 Permutation-invariant Parsing (PERIN)

Contrary to ERRE, the Permutation-invariant parser (PERIN) directly parses text into a structured representation and therefore follows a joint approach to structured prediction[17]. As discussed in section 2.2, the system was initially developed for the Semantic Parsing task but can be regarded as a general text-to-graph parser.

Samuel and Straka [2020] follow a transformers- and graph-based approach to Semantic Parsing and, as discussed in section 2.2.3, apply their parser to the structured sentiment analysis task [Samuel et al., 2022]. A novelty in the approach is that a permutation-invariant loss function and model are presented.

The (simplified) idea behind permutation-invariance in the context of a text-to-graph parser like PERIN, is that in prior text-to-graph approaches for Semantic Parsing, the graphs that should be predicted had to be linearized for the training process in neural networks to take place [Zhang et al., 2019b]. Most approaches rely on a fixed linearization strategy. As a linearization strategy for tree-like graphs, pre-order traversal can be chosen for example [Zhang et al., 2019b]. If, during training, the network would generate the wrong linearization, but with a near-perfect predicted graph, the entire output would be counted as wrong. PERIN alleviates this problem by using dynamic matching, instead of relying on fixed-order representations to potentially increase the performance of the model. In dynamic matching, only the "true errors" of the predicted graph are accounted for through the loss function. This is illustrated in figure 10.

As mentioned in section 2.2.3, PERIN can be trained on different forms of representations and each of these representations may yield a different performance. In their paper evaluating the performance of PERIN on the SSA task, Samuel et al. [2022] found that the ordered-tuple representation achieved the highest performance among several other representations. This form of representation we shall also use for our empirical part. Figure 11 shows different representations that PERIN can take.

[17] Our task of SI, that is predicting source, target and verb as well as the polar relation (if existent) that exists between entities can recast as a structured prediction problem, where the goal is to extract a tuple (source, verb, target, polarity), which we will call the sentiment tuple.

We have now explained, in technical detail, the three systems that will play a role in our empirical study. First, we looked at the system performing our task of SI in a rule-based manner. This system generates our silver standard and is the reference point when it comes to our task of SI. The two other models, which are from other domains, but, as we have argued, may prove successful for our task, are our neural approaches (transformers-based).

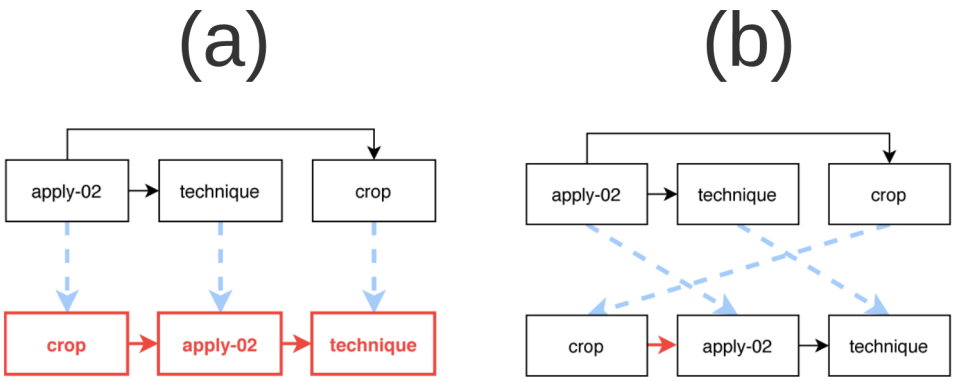

Figure 10: Permutation-invariant loss: At the top is the gold standard and at the bottom is the predicted graph. In (a), we have a node-by-node comparison and this makes the loss function evaluate the entire linearized graph as wrong. However, in (b) only the "true error" is counted, independent of whether the prediction is linearized the same way as the gold standard.(Adaptation from: Samuel [2021, p. 26-27]).

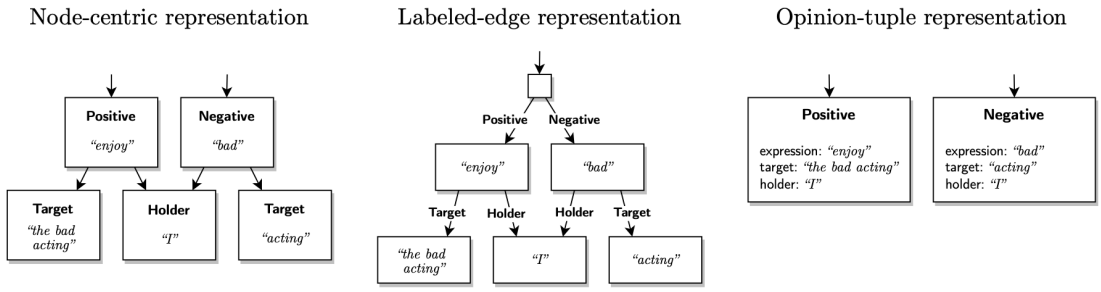

Figure 11: Different forms of representing attitude triples for the PERIN parser. Each representation can achieve a different performance. We will use the opinion-tuple representation as it appeared promising in the paper of Samuel et al. [2022]. (Source: Samuel et al. [2022, p.4]).

3 Data & Method

While the previous chapter has provided the necessary background on the conceptual origin of SI and offered a glimpse of the basic features of the systems involved, in this chapter we will explain in more detail how we shall proceed in order to answer our two research questions. The origin of the data is explained further in section 3.2. Preprocessing is detailed in section 3.3. The silver standard is generated on the basis of the system detailed in section 2.3.1. Then 4 different train-test splits are generated with different properties detailed in section 3.4. We round off the chapter by briefly giving the metrics used for evaluation in section 3.5.

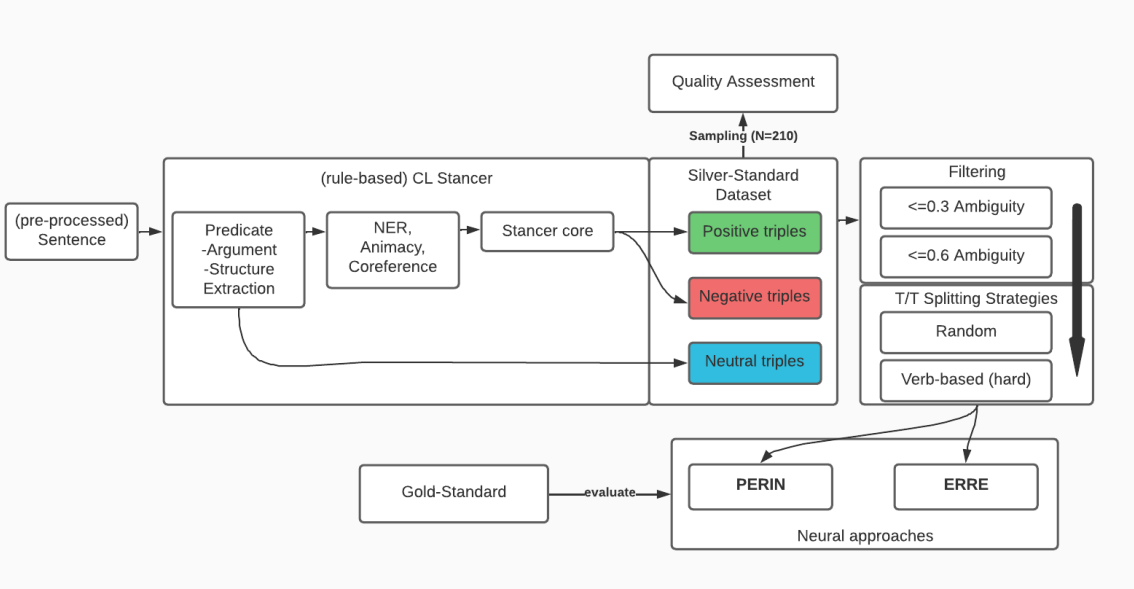

Figure 12: A high-level overview of the methodology we follow in order to answer our two research questions. The silver-standard data is generated by the stancer, then filters and two train-/test splitting strategies are applied to assess the performance of our neural approaches. (Source: own creation)

Figure 12 is a visual abstract of our method. We start by passing each sentence in our dataset into the rule-based system. As a first step, PAS extraction is performed as outlined in section 2.2.1. Triples that are identified can then either be neutral or, if detected as attitudes further on in the stancer pipeline (in the **stancer core**), be positive or negative (analogous to our notion of the ***polar relation***). We then assess

the quality of our generated silver standard by manually annotating 210 randomly drawn samples.

3.1 Implementations

We have already introduced the two approaches used in this thesis in chapter 2 and have elaborated on the approaches from an architectural and technical (high-level) perspective. In the following section, we shall briefly discuss the implementation details of the two approaches.

3.1.1 ERRE

The ERRE consists of two components implemented as Jupyter Notebooks using the Python programming language. The Entity Recognizer uses the Hugging Face[1] [Wolf et al., 2020] trainer classes with a pre-implemented head for token classification, while the Relation Classifier uses a custom head on top of the transformers encoder XLM-RoBERTa, which is inspired by Zhong and Chen [2020]. The head architecture itself is similar to option e) in Soares et al. [2019].

3.1.2 PERIN

PERIN, as implemented in the Github repository by Samuel et al. [2022], can be used for our experiment without extensive modification. We forked the repository and the code was minimally modified to accommodate external datasets. We also wrote dedicated instructions on how to handle external data. We also created a converter for bringing the data from a dataframe (as is outputted by the handler scripts enabling the generation of the silver standard) into the format accessible for PERIN.

As we have already discussed in section 2.3.2.2, we opt for the opinion-tuple representation with PERIN, based on its good performance in the paper by Samuel et al. [2022]. For each representation form, the parser has its own head. The variant of the PERIN model capable of learning and predicting the opinion-tuple representation utilizes three anchor[2] classifiers in conjunction with a multi-class node head. An edge classifier is not needed in this particular case.

[1] See https://huggingface.co/docs/transformers/main_classes/trainer

[2] Anchoring is the process of connecting a semantic node to surface-form textual features.

3.2 Datasets

3.2.1 Silver Standard

The dataset subject to our evaluation originates from the Swiss Media Database (SMD/Swissdox).[3] The dataset consists of news articles from the Swiss German-language newspapers Blick, Tagesanzeiger and Neue Zürcher Zeitung (NZZ). It contains all articles from these in the SMD from 2018-01-01 to 2022-11-01. In total that amounts to 266'647 news articles. No filtering was undertaken so as to not restrict potential domains of application.

The news articles were then passed into the stancer system as described in section 2.3.1 to generate the silver standard datasets. Source, target, verb and the polar relation between them in a sentence were extracted from the stancer. We also extracted "neutral" relations between two heads of noun-phrases. We are only interested in the heads of the noun phrases (NPs) and the relation among them, but not in extracting the entire NP in our task. This differentiation between choosing the entire phrase vs. only the head is present in the literature and has effects on evaluation [Deng et al., 2014, p. 180].

We sampled and manually annotated 210 sentences of the silver standard and determined the source, target and polar relation. The annotation was performed using the Universal Data Tool [Ibarluzea, 2021]. The sentences were randomly sampled in a disproportionately stratified way.

Through the evaluation of the silver standard based on these 210 manual annotations, we obtain an F1-score of $\mathbf{87.3}\%$ for the task of identifying the source of sentiment and $\mathbf{85.4}\%$ for identifying the target of the sentiment. We did not conduct an evaluation based on the performance of identifying the verbs and supplied these to the annotator as cue. The F1-Score of jointly correctly identifying the source and the target of a sentiment (with the verb given) is $\mathbf{79.88}\%$. The F1-Score of correctly predicting the sentiment as either + / - / $\sim$ (neutral), after having correctly identified the sources and targets, leads to a $\mathbf{59.45}\%$ F1-score. This last score is the full evaluation measure for our task of SI. The metrics are further detailed in section 3.5 along with an example in section 3.5.4.

A brief manual qualitative assessment of the results reveals that ambiguity of a verb (further discussed in section 3.4.1.1) can be seen as a root cause of the problem concerning the correct polarity assignment. For example, the verb ***ausnutzen***

[3] https://www.liri.uzh.ch/en/services/swissdox.html

(exploit) can either be neutral or negative (from source towards target). If *Mary* ***exploits*** *the situation*, then she is not necessarily an adversary of the situation. If, however, *Mary* ***exploits*** *her employees*, she is definitely bad for the employees and an adversary of them. Furthermore, it might not always be clear who the source and target of an implied sentiment might be, which can be exemplified in the verb ***irritieren***. If *The tenant* ***irritates*** *the landlord.*, there might be a negative polar relation from the tenant towards the landlord. For example, if the tenant refuses to pay. But if we would know that the landlord is simply irritated by the tenant for being a member of a different political party the situation changes and we could interpret the polar relation as being negative from the landlord towards the tenant. Further, the negative attitude could be implied as mutual, such as shown in example (I) of figure 4.

3.2.2 Gold Standard

The gold standard[4] contains data from three annotators which classified the relations between entities as either positive (pro) or negative (con). No neutral annotations are included in the dataset. This dataset will be applied to our neural approaches in order to check how well the neural approaches would perform in a "real-life" analytical setting.

To obtain our gold standard data, we only include annotations that all three annotators marked as valid. No verbs are marked in the gold standard and pro/con annotations are only undertaken between two entities. We evaluate the gold standard against both of our systems. It is not used as training material but rather we evaluate how well the system performs given a supplied training dataset from section 3.4.1. The total size of the gold standard amounts to 500 sentences, each of which may contain several polar relations.

3.3 Preprocessing

The original documents of the SMD corpus contained XML tags to delimit and wrap around the lead texts (`<ld>`), paragraphs (`<p>`) and sections like legends (`<lg>`). Since problems in terms of quality appeared in text enclosed by, e.g. legend tags, emphasis was put on the text within the paragraph tags (p-tags). After extraction

[4]Available at https://www.cl.uzh.ch/en/texttechnologies/research/opinionmining/sentiment-inference.html

of all text within p-tags, as a next step, the concatenated paragraphs of news documents were split by their sentences using the sent-tokenize function from NLTK [Bird et al., 2009]. In some instances, the sentence tokenization failed due to a missing delimiting symbol between sentences. It could be observed that such sentences often contained a token that with an upper-case character midword, which is why such sentences were ignored.

Sentences with a length shorter than 4 words were also ignored since it appeared that a considerable amount of such sentences consisted of terminal text indicating e.g., simply the author's name or the signature of the Swiss News Agency (sda) and would not have contributed in a productive way to the training and evaluation process.

The sentences were concatenated together again for each news document and passed into the dependency parser as a preprocessing step for the stancer in order to generate the silver standard. Then the texts were passed into the rule-based stancer and the data was extracted into a Pandas[5] dataframe. Since in German, verbs can occur separated, as in ***ankündigen*** in *Er* ***kündigt*** *eine neue Geschäftsidee* ***an.***, such discontiguous cases were only included if the lemmatizer component of ParZu recognized the lemma correctly, and if subsequently the verb matched with the lexicon of the stancer. Only the first main part of the verb was marked for entity prediction.

As mentioned while defining our task of SI (in chapter 2), we do not consider pronouns as either argument of a verb. We only include proper nouns and common nouns. As a consequence, reflexive uses of a verb (such as *X hates themselves*) are also discarded.

3.4 Configurations

As is visible in the figure 12, our method consists of generating different datasets with different properties, or as we shall call them, **configurations** in order to determine the viability of our approaches to SI. These datasets (varying in the properties described below in 3.4.1) will serve as means to answering our **RQs**. In other words, in this section, we will discuss the "variables" that are altered in order to generate our different datasets as well as the rationale behind modifying these variables. Further, we will detail the parameters used during training for the systems of each approach.

[5]https://pandas.pydata.org/

3.4.1 Subsets

3.4.1.1 Ambiguity

Some verbs are ambiguous with regard to our task of SI. The sentence *Er* ***stellt*** *sich auf den Tisch.* (he steps on the table) and *Er* ***stellt*** *sich hinter den Präsidenten.* (he backs the president) have the same subcategorization frame but the latter implies a positive attitude from "him" towards the president, while the former doesn't. The verb ***stellen*** therefore may have a lot of ambiguity and occur multiple times in our dataset indicating different polarities between entities. A further example is the word ***bedauern***, where *Sie* ***bedauert*** *den Vorfall.* (she regrets the incident) and *Sie* ***bedauert*** *ihre Schwester.* (she feels sorry towards her sister) are inverse to each other regarding attitude. The latter indicates a positive attitude and the former a negative attitude from "her" towards the respective entities. The verb ***bedauern*** therefore too can be understood as ambiguous in our dataset.

The reason we filter by ambiguity is that our system could potentially get "too confused" if it encounters too many verbs with high ambiguity. We can use a modification of the **shannon entropy** as a measure of ambiguity of certain verbs[6]. Our version of the shannon entropy is given by the formula:

$$H(v) = -\sum_{i=1}^{k} \frac{c_i}{n_v} \log \frac{c_i}{n_v}$$

$H(v)$ is the entropy of verb v. In our case, n_v would amount to the total number of observations of verb v in our dataset. k are the number of classes in the dataset. In our case $k = 3$ (+ / - / $\sim$ (neutral)). c_i denotes the number of instances of verb v that are assigned the label of class i. If the dataset is very unbalanced then the entropy tends towards 0. Otherwise, the entropy tends towards $\log k$. Naturally, therefore we can divide by $\log k$ to bound the value of ambiguity between 0 and 1. This leads to:

$$\text{ambiguity(v)} = \frac{H(v)}{\log k} = \frac{-\sum_{i=1}^{k} \frac{c_i}{n_v} \log \frac{c_i}{n_v}}{\log k}$$

If we now assume that the above verb ***stellen*** appears 50 times in its neutral context and 50 times in a charged context (con), we get a higher ambiguity score than the verb ***hassen*** which may occur 100 times and is associated only with a negative

[6]Idea adapted from https://stats.stackexchange.com/questions/239973/a-general-measure-of-data-set-imbalance

charge (polar relation).

In order to verify whether ambiguity potentially has an impact on the performance of our approach we filter our dataset. We generate one dataset which only contains verbs whose ambiguity score is $\leq .3$ and a second one whose ambiguity score is $\leq .6$. These values appear reasonable to us since the resulting datasets still appear to be sufficiently large so the model is capable of performing (and improving) over the training time.

3.4.1.2 Generalization

As part of **RQ1**, we want to find out how well our system generalizes on verbs unseen in the training data to validate our neural approaches. The general advantage of neural approaches (based on distributional semantics) in contrast to a lexicon-based (rule-based) approach, is that neural approaches would expectedly work even in out-of-vocabulary (OOV) settings. For example, the sentences *Der Bäcker* ***verabscheut*** *seine Konkurrenz* (the baker despises their competition) and *Der Bäcker* ***hasst*** *seine Konkurrenz* (the baker hates their competition) both induce a negative attitude of the baker towards their competition. If now ***verabscheuen*** (despise) were not in the lexicon of the rule-based system. We would be curious to find out if our neural system were to abstract from a set of examples it was trained on with ***hassen*** to the related word ***verabscheuen*** and whether ***verabscheuen*** would be still marked as inducing a negative attitude from baker to competition, without ever having been part of the train set.

To answer that part of **RQ1**, we will generate a specific split for our dataset in which verbs from the train set are strictly separated from verbs in the test/development set. We call this approach hard verb splitting. Soft splitting refers to a random split in this thesis.

After performing the filtering from section 3.4.1.1 with the two thresholds 0.6 and 0.3, as well as one splitting of train and development/test data (hard) by the verbs and at random (soft) we receive 4 datasets of which the sizes are reported of in table 1. From hereon we abbreviate the dataset in the first quadrant of table 1 as **DS1** (low-ambiguity, randomly split verbs), the one in the fourth quadrant as **DS2** (high-ambiguity, randomly split verbs), the one in the third quadrant as **DS3** (high-ambiguity, strict separation of verbs;hard-split) and in the one in the second quadrant as **DS4** (low-ambiguity, strict separation of verbs; hard-split). The reason that the datasets are unbalanced is obvious for **DS3** and **DS4** since we stipulate that the lemma of one verb cannot be in another dataset. If a verb lemma is thus

associated with, for example, 20 datapoints, and another verb is associated with 40 datapoints, it will not be possible to achieve a perfect balance between the train set and development/test set. The random splits can also be slightly unbalanced, since we stipulate that a sentence of the same document cannot be shared between the train and the development/test set.

We have already pointed out that the ERRE system is only capable of extracting a single relation per sentence. As it turns out, this is not a major problem for our task of SI (more specifically in the context of our generated datasets), since in our datasets there are only very few sentences (less than **0.35**%) that contain multiple relations. Therefore even if the system is only capable of extracting a single relation it may be competitive with the more feature-rich PERIN system.

		$\text{split}_{\text{hard}}$			$\text{split}_{\text{random}}$		
		train	**dev**	**test**	**train**	**dev**	**text**
$\text{amb}_{\leq 0.3}$	**Negative**	1276	480	479	1564	343	328
	Neutral	1407	414	414	1553	333	349
	Positive	1238	498	499	1584	325	326
$\text{amb}_{\leq 0.6}$	**Negative**	10139	1624	1623	9385	2034	1967
	Neutral	10128	1629	1629	9407	2013	1966
	Positive	9765	1811	1810	9280	2035	2071

Table 1: Comparison of dataset sizes and how many labels are in each category. The numbers correspond to the polar relation counts of each dataset and split.

3.4.2 Parameters

As already mentioned we use XLM-roberta-base by Conneau et al. [2019] as our base language model for the entity recognition component of the ERRE system and for PERIN, which is multilingual and therefore also suited for text in German. Bert-base-german-cased is used for the Relation Classification part of the ERRE system[7]. Additionally training is carried out on a single NVIDIA GeForce GTX TITAN X for each system.

For the Entity Recognizer component of the **ERRE** system, we use a Hugging Face `AutoModelForTokenClassification` head, which refers to a given model's out-of-the-box default for entity recognition tasks. We rely on sensible defaults in the settings of the parameters and orientate ourselves at the parameters given on the

[7]Available at: https://huggingface.co/bert-base-german-cased

Hugging Face page[8]. This amounts to a batch size of **16** for both training and development/test sets, 3 epochs, a learning rate of $\mathbf{2 \times 10^{-5}}$ and a weight decay of **.01**. Cross-entropy is used as a loss function and Adam as the optimizer[9].

For the Relation Classification component, we use a custom head with AdamW as the optimizer and a linear scheduler (without warmup). A weight decay of $\mathbf{1 \times 10^{-2}}$ is used, while the learning rate is set to $\mathbf{2 \times 10^{-5}}$. The model is trained again for 3 epochs. Cross-entropy is the loss function again.

For the PERIN system we rely (in part) on the defaults provided in the paper of Samuel et al. [2022]. We use a weight decay of **0.1**, a learning rate of $\mathbf{6 \times 10^{-6}}$ and train the model for 20 epochs. Initially, the default provided amounted to 100 epochs in the scenario where the transformer encoder layers were unfrozen[10], but the performance gain over the further 80 epochs was minimal, which led us to only train for 20 epochs. The scheduler is set to use 1000 warmup steps. As already mentioned in section 2.3.2.2, the loss function is non-generic. The optimizer is AdamW. The batch size used is again 16.

We will not perform hyperparameter tuning in the scope of this thesis and shall be rather concerned with evaluating our different datasets as detailed in the above section 3.4.1, as these provide possible grounds to validate our research questions.

3.5 Metrics

In this section, we describe the means by which we assess the performance of our approaches to neural SI. We will first give the formulas for the three primary metrics: Precision, Recall and F1-score. Then we will briefly show by an example how these are calculated in the "complex case" where an attitude triple has to be extracted from text and then classified.

3.5.1 Precision

Precision is a common metric used in a variety of NLP tasks. It gives us the proportion of correctly classified elements over the total elements classified as belonging to a given class.

[8] https://huggingface.co/course/chapter7/2

[9] Further details can be found at https://bit.ly/3MfHa6M

[10] Freezing in this context refers to the decision of whether to only train the weights of the head (frozen) or whether to also include training of the underlying language model layers.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

where TP stands for true positives and FP stands for false positives.

3.5.2 Recall

Recall gives us the proportion of correctly classified elements over all true elements belonging to a given class.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where TP stands for true positives and FN stands for false negatives.

3.5.3 F1-score

The F1-score combines the precision and recall measures to provide a single overall evaluation of a system's performance. It is the harmonic mean of precision and recall. It is calculated as follows:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.5.4 Composite Measures

The notion of a sentiment tuple (ST) can be helpful for understanding how the results are reflected in chapter 3. A labelled sentiment tuple can formally be defined as a 4-tuple comprising of the elements (source, target, polar_expression, polarity). The unlabelled sentiment tuple can be considered the 3-tuple with the elements (source, target, polar_expression). In our case, the polar expression always refers to the verb. Let us now assume that in the evaluation of a dataset with $n = 4$ (where n denotes the number of sentences) on a given system we get $F1_{source} = .75$, $F1_{target} = 1$, $F1_{polar\_expression} = 1$ for our extracted entities. Then, under the assumption that for each of our sentences, we only have a single possible tuple to extract, the precision, recall and F1-score ($\mathbf{ST\text{-}F1_{unl}}$) for our unlabelled sentiment tuple will be 0.75. This means that already a single mistake can violate and affect the performance of the entire unlabelled sentiment tuple score. Now also assume that the tuple that has a source extraction mistake and in a further tuple, the polarity

has been assigned improperly. This leads to 0.50 $\mathbf{ST\text{-}Pr_{lab}}$ (precision of the labelled sentiment tuple), 0.50 $\mathbf{ST\text{-}Rec_{lab}}$ (recall of the labelled tuple) and 0.50 $\mathbf{ST\text{-}F1_{lab}}$ (F1-score of the labelled tuple). As we will often see in the results chapter 4, Precision and Recall are closely aligned, this has to do with the fact that the models are trained to always extract three expressions in all cases. Intuitively therefore, the approaches will always attempt to find "something", but if that something is wrong, the precision will be negatively impacted and what is to be found usually is then not found (since at least the ERRE model is trained only to extract a single triple). This will lead to very similar precision and recall.

4 Results

As discussed previously, we generated a total of 4 datasets with different properties. In this chapter, we are concerned with detailing the performance of both of our systems. Each system is evaluated on a dedicated test set that was generated for the dataset. In addition, we report on the independent performances of the two components of the ERRE approach. We also evaluate how well the approaches perform on the gold standard with a given training set from one of the generated datasets.

4.1 ERRE

4.1.1 Entity Recognizer

Table 2 shows the performances we get for each of our datasets. The counts listed in the support column represent tokens, since in the standard entity recognition task, the goal is to assign a label to each token. The seqeval[1] package by Nakayama [2018] was used to obtain these results. The evaluations are based on the test set. The reason that the support size is the same for each of the categories (apart from None) is that, as discussed above, the ERRE system is only capable of extracting a single potential polar relation per sentence.

The smallest dataset size in terms of tokens is the random split with low ambiguity (**DS1**, quadrant I, in table 2), where roughly 6643 tokens are classified. On the whole, the weighted F1-score[2] for **DS1** amounts to **70**%. This is 10 points more than **DS4**, which only yields **60**% on the same measure. We note that, as was discussed in section 3.4.1, **DS4** contains marginally more data, but still underperformed.

[1] https://github.com/chakki-works/seqeval

[2] Weighted Precision/Recall/F1-score are the per-class scores weighted by the relative strength of support (against the support of the other classes).

		$split_{hard}$				$split_{random}$			
		Precision	**Recall**	**F1-score**	**Support**	**Precision**	**Recall**	**F1-score**	**Support**
$amb_{\leq 0.3}$	**None**	0.59	0.54	0.57	5086	0.70	0.69	0.69	3663
	Source	0.67	0.68	0.68	1383	0.65	0.68	0.66	994
	Target	0.67	0.50	0.57	1383	0.70	0.70	0.70	993
	Verb	0.72	0.64	0.68	1383	0.72	0.73	0.73	994
$amb_{\leq 0.6}$	**None**	0.52	0.49	0.51	18343	0.55	0.52	0.54	21749
	Source	0.50	0.50	0.50	4998	0.42	0.43	0.42	5915
	Target	0.47	0.43	0.45	4998	0.42	0.43	0.42	5915
	Verb	0.48	0.46	0.47	4998	0.51	0.51	0.51	5915

Table 2: Performance matrix for the entity recognizer of the ERRE approach. The performance depends on the different splits. The support number corresponds to each token in the dataset. amb stands for ambiguity.

4.1.2 Relation Classifier

The performances of the relation classification component are detailed in table 3. For **DS1**, we get a weighted average F1-score of **88**%, with an overall accuracy of **58.25**%, while we achieve an average F1-score of **58**% for **DS4**, with an overall accuracy **58.2**%. For high-ambiguity datasets **DS2** and **DS3**, we get a weighted average F1-score of **76**%, with an accuracy of **75**% for **DS3**.

		$split_{hard}$				$split_{random}$			
		Precision	**Recall**	**F1-score**	**Support**	**Precision**	**Recall**	**F1-score**	**Support**
$amb_{\leq 0.3}$	**Negative**	0.77	0.39	0.52	473	0.84	0.92	0.88	322
	Neutral	0.44	0.90	0.59	412	0.89	0.84	0.86	345
	Positive	0.85	0.49	0.63	495	0.92	0.89	0.90	318
$amb_{\leq 0.6}$	**Negative**	0.89	0.64	0.74	1593	0.93	0.98	0.96	1926
	Neutral	0.59	0.95	0.73	1598	0.97	0.85	0.91	1914
	Positive	0.96	0.68	0.80	1779	0.92	0.99	0.95	2040

Table 3: Performance matrix for the relation classification. The performance depends on the different splits.

4.1.3 Joint Performance

The joint performances of each dataset are reported in the comparison table 4. The evaluation metrics used are further explained by example in section 3.5.4. We can observe that the **ST-F1$_{\text{lab}}$** amounts to **71.7**% for **DS1**, whereas it only achieves roughly **20.5**% on the **DS4**, where verbs are strictly separated. Recall and precision appear to be more or less equal over most of the systems for the reasons explained in section 3.5.4. In the unlabelled case (where we only consider how well the approaches extract triples), **DS4** achieves **44.5**%, whereas **DS1** manages **77.8**%. So the performance drop when considering labels additional to the triple extraction alone is larger for **DS4** than for **DS1**.

dataset	amb	split	system	F1$_{\text{source}}$	F1$_{\text{target}}$	F1$_{\text{verb}}$	ST-F1$_{\text{unl}}$	ST-Pr$_{\text{lab}}$	ST-Rec$_{\text{lab}}$	ST-F1$_{\text{lab}}$
DS4	≤ 0.3	hard	ERRE	0.833	0.605	0.768	0.445	0.205	0.206	0.205
DS4	≤ 0.3	hard	PERIN	0.098	0.032	0.029	0.018	0.056	0.005	0.009
DS1	≤ 0.3	random	ERRE	**0.907**	0.861	0.945	0.778	0.718	0.716	0.717
DS1	≤ 0.3	random	PERIN	0.851	0.799	0.934	0.704	0.602	0.620	0.611
DS2	≤ 0.6	hard	ERRE	0.828	0.734	0.817	0.650	0.488	0.489	0.488
DS2	≤ 0.6	hard	PERIN	0.473	0.421	0.470	0.418	0.514	0.206	0.294
DS3	≤ 0.6	random	ERRE	**0.907**	**0.887**	**0.968**	**0.852**	**0.820**	**0.819**	**0.820**
DS3	≤ 0.6	random	PERIN	0.895	0.858	0.959	0.839	0.787	0.792	0.790

Table 4: Overall performance, suitable for comparison with the results from the PERIN system below. Highlighted in boldface are the highest achieved performances under the different metrics. Amb stands for ambiguity.

4.2 PERIN

Since PERIN is trained and designed as an end-to-end system, we only can discuss the end performance. As indicated in the table 4, the PERIN approach dramatically underperforms in **DS4** and achieves **1.8**% for **ST-F1$_{\text{lab}}$** under the less ambivalent dataset where the verb lemmas are separated in the train and test/dev splits from each other. More performance is achieved in the settings of **DS1**, with the random split. There, PERIN manages **61.1**% for **ST-F1$_{\text{lab}}$**. In the unlabelled case (**ST-F1$_{\text{unl}}$**), PERIN achieves **70.4**%. So roughly 10 points are lost if we consider labels additionally to pure triple extraction performance. The system performs best for **DS3**, where PERIN achieves **79**% **ST-F1$_{\text{lab}}$**. Further, in **DS3**, the performance difference from the unlabelled F1 to the labelled F1 for the tuples only consists of

4.9%, which is the smallest drop apart from **DS4**. Source, target and verb identification also appear to work the best on **DS3**. Even though the ambiguity is higher, this appears to have been compensated by the increase in samples.

4.3 Gold Standard

In the results discussed above, the test dataset had the same characteristics as the training dataset, insofar as these results only allow insights into how well the system performs (can learn) under the different treatments (properties). The gold standard dataset presents a picture of how the approaches would perform in a real-life setting. As discussed in section 3.2.2, the verb is not considered in this case. The sentiment tuple only consists of prediction (source, target, polarity).

The ERRE approach performs best independently of the dataset it was trained on. The most accurate performance in terms of $\mathbf{ST\text{-}F1_{lab}}$ was when the pretraining took place on **DS2**, which is the dataset with the high ambiguity and a random splitting of the verbs. Both approaches performed best for **DS2** with regard to such a manner. For PERIN, there appears to be only a minor performance difference of **.2**% in the case of **DS2** and **DS3**. It is noticeable that the recall in the case of PERIN is consistently lower compared to the recall of the ERRE system.

dataset	system	$\mathbf{F1_{source}}$	$\mathbf{F1_{target}}$	$\mathbf{ST\text{-}F1_{unl}}$	$\mathbf{ST\text{-}Pr_{lab}}$	$\mathbf{ST\text{-}Rec_{lab}}$	$\mathbf{ST\text{-}F1_{lab}}$
DS1	ERRE	**0.763**	0.634	0.515	0.395	0.316	0.350
DS1	PERIN	0.303	0.237	0.219	0.342	0.068	0.114
DS2	ERRE	0.409	0.597	**0.570**	**0.474**	**0.389**	**0.428**
DS2	PERIN	0.565	0.490	0.454	0.420	0.174	0.246
DS3	ERRE	0.604	**0.656**	0.536	0.405	0.326	0.361
DS3	PERIN	0.566	0.479	0.407	0.4125	0.174	0.244
DS4	ERRE	0.756	0.635	0.474	0.329	0.263	0.292
DS4	PERIN	0.247	0.198	0.150	0.218	0.042	0.071

Table 5: Performance of the gold standard on the different systems trained on the train splits of the different datasets. In bold are the best performances for a measure.

5 Discussion

Returning to the aims set out in the introduction, we shall now attempt to discuss our results concisely in the context of the research questions posed. **RQ1** was concerned with how well the two compared neural systems perform and also how well they generalize on verbs unseen in the training data. We subdivided **RQ1** into two questions that are directly visible from the results in chapter 4 (table 4 resp. 5). The best performance was achieved under the configuration of **DS2** (high-ambiguity, random split of verbs) for both systems and also for the gold standard evaluation. It is clear that the achieved labelled ST-F1 score on the gold standard of 42.8% and 24.6% is not necessarily a convincing performance for all application domains (e.g., policy analysis) outlined in the introduction. However, it could serve as an initial starting point for further steps in the direction of a SI as defined in this thesis. It also must be noted that the gold standard does not contain any annotations of neutral relations which may skew the results.

It is also important to point out that the task of SI that has been defined here is by no means trivial. For instance, Samuel et al. [2022] achieved $34.1^{\pm 1.1}\%$ $\mathbf{ST\text{-}F1_{lab}}$ on their task of structured sentiment analysis for the MPQA[1] dataset. It lies outside the scope of the present thesis and remains a matter of debate as to how similar the MPQA dataset is to our datasets. However, in the MPQA dataset, just as in our dataset, triple extraction and polarity assignment have been used as means to learn and predict annotations, as is evident from Katiyar and Cardie [2016].

For PERIN, the hard (verb) vs. random (soft) splitting factor appears to have hardly impacted the performance in the high-ambiguity datasets, since there is only a **.2**% difference between the two labelled ST-F1 scores for the gold standard. On the training of low-ambiguity datasets, PERIN performs poorly in labelled ST-F1 concerning the gold standard, which appears to be mostly attributable to bad recall.

Another observation worth mentioning may be the fact that gold standard performance is better than test-split performance in the case of **DS4** for both systems.

[1] The MPQA (multi-perspective question answering) dataset by Wiebe et al. [2005] is known to be a fine-grained dataset for (more explicit) sentiment analysis.

Especially in the case of PERIN, the performance of the test-split of **.09**% $\mathbf{ST\text{-}F1_{lab}}$ to **7.1**% on the gold standard is surprising. It is observable from table 4 that the problem appears not necessarily to reside in the labelling (assigning a polarity), but to lie in the identification of the source, the target and the verb. The gold standard does not include the identification of the verb, which may be an explanation for its better performance compared to the test split. This, however, cannot explain the bad performance on the test split of **DS4** in the identification of the other surface forms beyond verb identification. The reason for the problem may be found in the model design (anchoring). Interestingly, this problem appears to dissipate in the performance of test splits as well as the gold standard of the high-ambiguity datasets (DS2 & DS3), since we cannot see such a performance drop. In the low ambiguity random splitting (DS1) (which has almost an equal train dataset size), we might conclude that the poor performance is attributable to several factors, such as ambiguity, the verb splitting and the resulting low sample size. Based on a brief qualitative examination the bad performance attained in **DS4** appears to be a recall problem since few opinions appear to have been successfully extracted at all.

As for **RQ2**, where the question was to establish which of the two approaches is better, it appears from our results that the joint ERRE approach outperforms PERIN on all our configurations. Even though the capabilities of the ERRE system are reduced, since multiple polar relations cannot be extracted from a sentence, it outperforms PERIN. It must be noted, however, that further investigations in terms of the settings of hyperparameters could yield different results. Additionally, we cannot call our results statistically significant since we have not used cross-validation (multiple different splits with different random seeds) and did not perform multiple runs. Further research is necessary to conclude with a more certain answer for **RQ2**.

We have conceptualized our task as a structured prediction problem of extracting (source, verb, target, polarity) from a segment of text and investigated two neural network based approaches to tackle the task. In the first system, two models are separately trained, one for identifying the first three elements of the "sentiment tuple" from the text and then a model for subsequently identifying whether the triple is charged or not (and if so, positively or negatively). The second model was initially developed for Semantic Parsing and makes its predictions by essentially learning to parse the sentence into a graph. In the context of our task, we have only focused on the noun phrase heads for the source and the target. Future work could also be concerned with longer spans of source extraction. For example, in the sentence *Präsident Müller* ***attackiert*** *seine Bediensteten* (President Müller **attacks** their officials) the task could also include the extraction of more than just *President*, but the entire phrase *President Müller* as the source.

The first neural approach (ERRE) was not capable of extracting multiple relations from text, as we found out that few of these "multiple relations" sentences existed in our dataset. Further research, however, could focus specifically on such sentences and on adequate approaches to dealing with such multiple relations. Although PERIN would have been able to deal with "multiple relations" sentences, we have not investigated the matter in more detail. In cases of multiple relations in a text segment, one could further investigate inference and the extraction of so-called "higher order" relations, which the rule-based system (from which the silver standard stems) is capable of. An example of such higher-order inference can be seen in the statement *X criticizes that Y attacks Z*. The sentiment tuples (X, criticize, Y, -) and (Y, attack, Z, -) can be extracted. However, the knowledge that X must be *in favour* of Z can be inferred. Specific investigations of problems of such a nature could be the subject of future work.

Finally, there have been important new developments in AI with likely implications for SI. In the months during which this thesis was being written, a new large language model (LLM) called ChatGPT[2] came onto the market. Investigating its capabilities concerning SI could offer promising new directions in SA research. Potential avenues could be researching how to make the LLM reliably generate structured output that could be used for further analysis by a domain expert. An informal investigation also shows that some limits are to be expected since (at least the commercial) LLMs tend to remain vague when it comes to inferring polar relations in statements. In the case of badFor / goodFor relations as shown in figure 13, ChatGPTs inferences appear more precise compared to against / in-favour-of relations. It further might be a matter of prompt engineering for making ChatGPT work for the task of SI as defined in this thesis.

[2]chat.openai.com

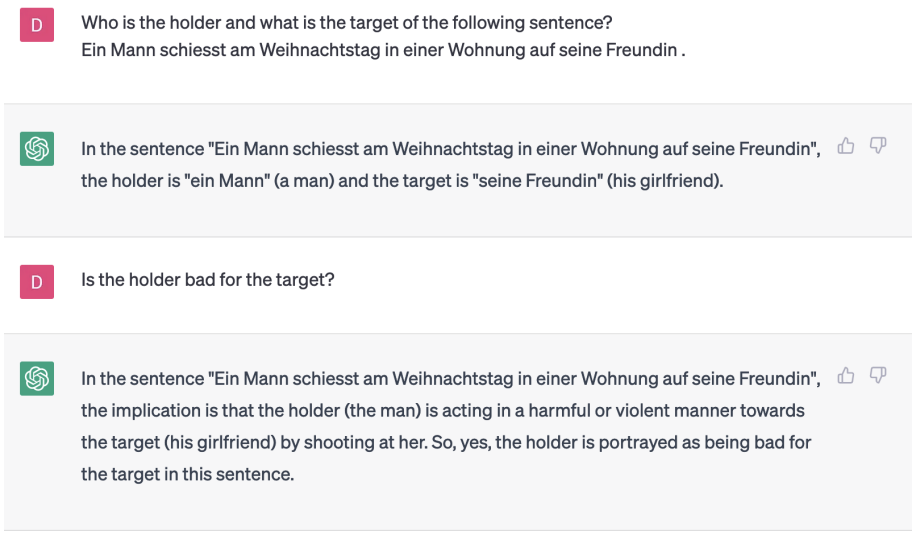

Figure 13: ChatGPT performing reasoning in line with our task of SI based on the sentence *A man* ***shoots*** *his girlfriend in his flat on Christmas day.* (Source: Own screenshot, 2023).

6 Conclusion

This thesis has set out to assess the viability of neural approaches to SI. The first approach investigated is pipeline-based, where two separate neural models are trained, one for recognition of the source, the target and the relation mediating verb, and one for classifying the polar relation potentially induced by the recognized entities (ERRE). The second approach exploits a model initially designed for Semantic Parsing and performs recognition of entities and subsequent classification of the potential polar relation that holds between them in a single step (PERIN). In response to the specific research questions addressed, the first subquestion of **RQ1** (i.e., how well do the two neural approaches investigated work on a random train-test split of the silver standard?) can be answered by stating that the ERRE approach achieves its best performance on our task of SI with 82% **ST-F1$_{\text{lab}}$**[1] on the test set, while the other approach (PERIN) achieves 79% **ST-F1$_{\text{lab}}$** as its best performance on the test set. For the second subquestion of **RQ1** (i.e., how well do the two neural approaches investigated work on data unseen during training [generalizability]?), ERRE achieves 48.8% **ST-F1$_{\text{lab}}$** and PERIN 29.4% **ST-F1$_{\text{lab}}$**. The ERRE approach appears to outperform the PERIN system on the majority of datasets. The highest performance for the gold standard was attained under ERRE with 42.8% **ST-F1$_{\text{lab}}$**.

However, there are limitations to the present study. More thorough investigations will be necessary to provide a more definitive answer to **RQ2** (i.e., which of the two investigated approaches to neural SI proves superior?). One possible approach could be to generate the datasets several times using multiple different random seeds and then average out the performances. Furthermore, as no hyperparameters were tuned, actually doing so could provide another fruitful line of inquiry for future research, especially in addressing **RQ2**. A further limitation is our use of a silver standard to train and partly also to assess our models. We find that the silver standard results in an F1-score of **59.45**% for the task of annotating a target, a source and classifying the polar relation that holds between these two entities in a sentence, which addresses the implicit research question about the quality of our

[1] This measure is the joint F1-score of predicting the source, target, the polarity and the mediating verb of an (implicit) attitude, emotion, opinion in a sentence.

silver standard. It should be added, however, that only a single annotator was used. Relying on multiple annotators with harmonization could therefore offer greater clarity.

Despite these caveats, the current thesis makes two main contributions to the field. The first comprises the comprehensive literature survey on SI and its related tasks that it contains. The second, and most important, is the actual study it presents: the evaluation of two neural systems for SI, the one (PERIN) adapted to fit the task of SI, the other implemented by the author based on web tutorials. The resulting models are the first neural SI models for the German language. We hope that the study can stimulate greater interest in this area of research and provide an impulse for other novel approaches that seek to improve the performance of the SI task in a neural setting. We plan to make an online demo version of the best performing model presented in this thesis freely available[2] as of summer 2023.

[2] A link will be published in the GitHub repository for this thesis at https://github.com/mystreamer/ba_thesis.

References

B. Agarwal, S. Poria, N. Mittal, A. Gelbukh, and A. Hussain. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. *Cognitive Computation*, 7:487–499, 2015. Publisher: Springer.

L. Bamberg, I. Rehbein, and S. Ponzetto. Improved Opinion Role Labelling in Parliamentary Debates. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 110–120, Potsdam, Germany, 2022. KONVENS 2022 Organizers. URL `https://aclanthology.org/2022.konvens-1.13`.

J. Barnes, R. Kurtz, S. Oepen, L. Øvrelid, and E. Velldal. Structured Sentiment Analysis as Dependency Graph Parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3387–3402, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.263. URL `https://aclanthology.org/2021.acl-long.263`.

J. Barnes, L. Oberlaender, E. Troiano, A. Kutuzov, J. Buchmann, R. Agerri, L. Øvrelid, and E. Velldal. SemEval 2022 Task 10: Structured Sentiment Analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.semeval-1.180. URL `https://aclanthology.org/2022.semeval-1.180`.

S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

E. Cambria, R. Mao, S. Han, and Q. Liu. Sentic parser: A graph-based approach to concept extraction for sentiment analysis. In *Proceedings of the 2022 International Conference on Data Mining Workshops, Orlando, FL, USA*, volume 30, 2022.

E. Choi, H. Rashkin, L. Zettlemoyer, and Y. Choi. Document-level sentiment inference with social, faction, and discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 333–343, 2016.

J. Christensen, S. Soderland, and O. Etzioni. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the sixth international conference on Knowledge capture*, pages 113–120, 2011.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. *CoRR*, abs/1911.02116, 2019. URL `http://arxiv.org/abs/1911.02116`. arXiv: 1911.02116.

W. Davis. Implicature. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019. URL `https://plato.stanford.edu/archives/fall2019/entries/implicature/`.

L. Del Corro and R. Gemulla. ClausIE: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366, Rio de Janeiro Brazil, May 2013. ACM. ISBN 978-1-4503-2035-1. doi: 10.1145/2488388.2488420. URL `https://dl.acm.org/doi/10.1145/2488388.2488420`.

L. Deng and J. Wiebe. Sentiment Propagation via Implicature Constraints. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 377–385, Gothenburg, Sweden, Apr. 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1040. URL `https://aclanthology.org/E14-1040`.

L. Deng and J. Wiebe. How can NLP Tasks Mutually Benefit Sentiment Analysis? A Holistic Approach to Sentiment Analysis. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 53–59, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-0411. URL `https://aclanthology.org/W16-0411`.

L. Deng, Y. Choi, and J. Wiebe. Benefactive/Malefactive Event and Writer Attitude Annotation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages

120–125, Sofia, Bulgaria, Aug. 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-2022.

L. Deng, J. Wiebe, and Y. Choi. Joint Inference and Disambiguation of Implicit Sentiments via Implicature Constraints. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 79–88, Dublin, Ireland, Aug. 2014. Dublin City University and Association for Computational Linguistics. URL https://aclanthology.org/C14-1009.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

H. Ding and E. Riloff. Acquiring Knowledge of Affective Events from Blogs Using Label Propagation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v30i1.10394. URL https://ojs.aaai.org/index.php/AAAI/article/view/10394.

H. Ding and E. Riloff. Human Needs Categorization of Affective Events Using Labeled and Unlabeled Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1919–1929, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1174. URL https://aclanthology.org/N18-1174.

D. R. Dowty, R. Wall, and S. Peters. *Introduction to Montague semantics*, volume 11. Springer Science & Business Media, 2012.

J. Ellis, J. Getman, and S. Strassel. Overview of Linguistic Resource for the TAC KBP 2014 Evaluations: Planning, Execution, and Results. 2014. URL https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/tackbp-2014-overview.pdf.

F. Hamborg and K. Donnay. NewsMTSC: A Dataset for (Multi-)Target-dependent Sentiment Classification in Political News Articles. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1663–1675, Online, Apr. 2021. Association for

Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.142. URL https://aclanthology.org/2021.eacl-main.142.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. URL https://ieeexplore.ieee.org/abstract/document/6795963. Publisher: MIT press.

S. Ibarluzea. Universal Data Tool, 2021. URL https://github.com/UniversalDataTool/universal-data-tool.

G. Jacobs and V. Hoste. Fine-Grained Implicit Sentiment in Financial News: Uncovering Hidden Bulls and Bears. *Electronics*, 10(20):2554, Oct. 2021. ISSN 2079-9292. doi: 10.3390/electronics10202554. URL https://www.mdpi.com/2079-9292/10/20/2554.

A. Katiyar and C. Cardie. Investigating LSTMs for Joint Extraction of Opinion Entities and Relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 919–929, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1087. URL https://aclanthology.org/P16-1087.

S.-M. Kim and E. Hovy. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Sydney, Australia, July 2006. Association for Computational Linguistics. URL https://aclanthology.org/W06-0301.

M. Klenner. Extracting predicate structures from parse trees. Sept. 2005. doi: 10.5167/UZH-19134. URL https://www.zora.uzh.ch/id/eprint/19134. Publisher: s.n.

M. Klenner. Verb-centered Sentiment Inference with Description Logics. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 134–139, Lisboa, Portugal, Sept. 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-2919. URL https://aclanthology.org/W15-2919.

M. Klenner and A. Göhring. Semantic Role Labeling for Sentiment Inference: A Case Study. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 144–149, Potsdam, Germany, 2022.

KONVENS 2022 Organizers. URL https://aclanthology.org/2022.konvens-1.17.

M. Klenner, D. Tuggener, and S. Clematide. Stance Detection in Facebook Posts of a German Right-wing Party. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 31–40, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-0904. URL https://aclanthology.org/W17-0904.

B. Liu. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2 edition, Oct. 2020. ISBN 978-1-108-63928-6 978-1-108-48637-8. doi: 10.1017/9781108639286. URL https://www.cambridge.org/core/product/identifier/9781108639286/type/book.

S. Lyu and H. Chen. Relation Classification with Entity Type Restriction. *CoRR*, abs/2105.08393, 2021. URL https://arxiv.org/abs/2105.08393. arXiv: 2105.08393.

S. Löbner. *Understanding semantics*. Understanding language. Routledge, New York, NY, second edition edition, 2013. ISBN 978-1-4441-2243-5 978-0-415-82673-0 978-0-203-52833-4.

K. Moilanen and S. Pulman. Sentiment Composition. Sept. 2007. URL http://www.clg.ox.ac.uk/_media/people:karo:sentcompranlp07final.pdf.

H. Nakayama. seqeval: A Python framework for sequence labeling evaluation, 2018. URL https://github.com/chakki-works/seqeval.

A. Neviarouskaya, H. Prendinger, and M. Ishizuka. Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 3, pages 278–281, 2009. URL https://ojs.aaai.org/index.php/ICWSM/article/view/13987. Issue: 1.

O. Rambow and J. Wiebe. Sentiment and belief: How to think about, represent, and annotate private states. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 7–11, 2015. URL https://aclanthology.org/P15-5003.pdf.

K. Ravi and V. Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89:14–46, 2015. Publisher: Elsevier.

J. Ruppenhofer and J. M. Struss. IGGSA-STEPS: Shared Task on Source and Target Extraction from Political Speeches. 29(1):33 – 46, 2016. URL https://nbn-resolving.org/urn:nbn:de:bsz:mh39-49410. Place: Regensburg Publisher: Gesellschaft für Sprachtechnologie und Computerlinguistik.

D. Samuel. Permutation-Invariant Semantic Parsing. Master's thesis, Charles University, Prague, 2021. Publisher: Univerzita Karlova, Matematicko-fyzikální fakulta.

D. Samuel and M. Straka. ÚFAL at MRP 2020: Permutation-invariant Semantic Parsing in PERIN. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 53–64, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-shared.5. URL https://aclanthology.org/2020.conll-shared.5.

D. Samuel, J. Barnes, R. Kurtz, S. Oepen, L. Øvrelid, and E. Velldal. Direct parsing to sentiment graphs, 2022. URL https://arxiv.org/abs/2203.13209.

R. Sennrich, G. Schneider, M. Volk, and M. Warin. A new hybrid dependency parser for German. *Proceedings of the German Society for Computational Linguistics and Language Technology*, pages 115–124, 2009.

L. B. Soares, N. FitzGerald, J. Ling, and T. Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*, 2019. URL https://arxiv.org/abs/1906.03158.

V. Stoyanov, C. Cardie, and J. Wiebe. Multi-Perspective Question Answering Using the OpQA Corpus. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 923–930, Vancouver, British Columbia, Canada, Oct. 2005. Association for Computational Linguistics. URL https://aclanthology.org/H05-1116.

M. Taboada. Sentiment Analysis: An Overview from Linguistics. *Annual Review of Linguistics*, 2(1):325–347, Jan. 2016. ISSN 2333-9683, 2333-9691. doi: 10.1146/annurev-linguistics-011415-040518. URL https://www.annualreviews.org/doi/10.1146/annurev-linguistics-011415-040518.

S. Tron. A verb-centered Sentiment Analysis for French. Master's thesis, Masters Thesis. Zürich University, 2013.

D. Tuggener. Incremental Coreference Resolution for German. Jan. 2016. URL https://www.researchgate.net/publication/318269130_Incremental_Coreference_Resolution_for_German.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

J. Wiebe and L. Deng. An Account of Opinion Implicatures, 2014. URL https://arxiv.org/abs/1404.6491.

J. Wiebe, T. Wilson, and C. Cardie. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3):165–210, May 2005. ISSN 1574-020X, 1572-8412. doi: 10.1007/s10579-005-7880-9. URL http://link.springer.com/10.1007/s10579-005-7880-9.

M. Wiegand and J. Ruppenhofer. Opinion Holder and Target Extraction based on the Induction of Verbal Categories. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 215–225, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/K15-1022. URL https://aclanthology.org/K15-1022.

T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada, Oct. 2005. Association for Computational Linguistics. URL https://aclanthology.org/H05-1044.

T. A. Wilson. Fine-grained Subjectivity and Sentiment Analysis: Recognizing the intensity, polarity, and attitudes of private states. June 2008. URL http://d-scholarship.pitt.edu/7563/.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

M. Zhang, Q. Wang, and G. Fu. End-to-end neural opinion extraction with a transition-based model. *Information Systems*, 80:56–63, Feb. 2019a. ISSN 03064379. doi: 10.1016/j.is.2018.09.006. URL https://linkinghub.elsevier.com/retrieve/pii/S0306437918301182.

S. Zhang, X. Ma, K. Duh, and B. Van Durme. AMR Parsing as Sequence-to-Graph Transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 80–94, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1009. URL https://aclanthology.org/P19-1009.

Z. Zhong and D. Chen. A frustratingly easy approach for entity and relation extraction. *arXiv preprint arXiv:2010.12812*, 2020. URL https://arxiv.org/abs/2010.12812.

A Datasets

The datasets evaluated in this thesis can be obtained under the following link: https://github.com/mystreamer/ba_thesis/tree/main/annex. The most important contents are described below. Further information can be obtained in the README of the repo.

- The `annotations` folder contains the manual annotations evaluating the silver standard.
  - `eval.sh` is the script that downloads the evaluation scripts and then automatically performs the evaluation (prerequisite is a virtual environment called `perin-venv`[1]).
  - `manual_annotation.json` are the sampled sentences of the silver standard.
  - `manual_annotation_ssa_annotated.json` is the file containing the manually annotated sampled sentences of the silver standard.
- The `datasets` folder contains the relevant datasets for this thesis.
  - `eval.sh` downloads and performs the automatic evaluation of the dataset predictions.
  - `tt_03_va_hard` is the folder that contains the dataset with low-ambiguity and verb (hard) splitting (**DS4**).
  - `tt_03_va_soft` is the folder that contains the dataset with low-ambiguity and randomised (soft) splitting (**DS1**).
  - `tt_06_va_hard` is the folder that contains the dataset with high-ambiguity and verb (hard) splitting (**DS2**).
  - `tt_06_va_soft` is the folder that contains the dataset with high-ambiguity and randomised (soft) splitting (**DS3**).

[1] Instructions for setting up perin-venv are found here: https://github.com/mystreamer/direct_parsing_to_sent_graph