BRIAN VAN DER BIJL

TWEET, REACTION, ACTION

Abstract

This document details the intended direction and steps necessary for my master thesis. It discusses the relevant background, research methodology and planing. Risks and plans for their mitigation are discussed as well.

Board of Examination

This proposal was submitted on April 22, 2024.
Student: Brian van der Bijl
Programme: Master of Artifical Intelligence (GSNS)
Student Nr: 3098176

Adviser: Dr. Huib ALDEWERELD, Hogeschool Utrecht
First examiner: Dr. Rick NOUWEN
Second examiner: Vacancy

Latest update

53deada (Mon, 22 Apr 2024 12:36:38 +0200)
    Post mortem

Contents

1 Introduction

Present society is more than ever connected, but simultaneously exceedingly polarised[Wor]. Public discourse is dominated by complex and divisive issues: the handling of COVID-19[Onr], the climate emergency, Black Lives Matter and Brexit to name a few. The rise of modern technology has facilitated that everybody can have a voice in such discussions, and thus everybody can be an influencer, regardless of motivation, accountability or ability to envision consequences. This has had the unintended side-effect that these discussions can quickly escalate from being innocent online chatter to becoming part of large societal issues[Lyn]. This online debate is often publicly visible, occurring on websites such as Twitter and Facebook, but it appears to be difficult for authorities to ascertain whether a topic is going to cause a major stir or not. This is unfortunate, as it is in this online parlance that the seeds of societal unrest are often instigated[Lyn]. For the purpose of this chapter, we will refer to the concept of communication designed to incite a violent reaction as *incitement*:

Definition (Incitement, provisionary)**.** A series of statements, e.g. a tweet, designed to provoke societal violence in a target audience.

A more thorough definition is introduced in Chapter 2.

To ACCURATELY PREDICT and prevent mobilisation and societal violence, authorities employ newsrooms to monitor online conversations and intervene if necessary. Due to the large amount of daily tweets, some form of filtering is required. This can be done based on known keywords or hashtags, but this mainly relies on pre-existing knowledge of then-current divisive subjects. That is to say, a human operator will generally know the societal context required to differentiate trending from divisive, whereas software will not. Conversely, by broadening the view large amounts of noise are introduced along with any potential novel relevant discussions. In the current situation, the tradeoff therefore comes down to opting to miss relevant information because it is filtered out, or opting to miss relevant information because it is lost within large amounts of irrelevant data.

THE SUBJECT OF THIS RESEARCH is whether and how AI can help to solve this issue. Computers are very efficient at rapidly analysing

vast quantities of data, provided one can teach the computer what to look for. If an AI-based application can be trained to get a good estimate on whether a specific tweet contains incitement, or is likely to lead to incitement further down the road, it can then relegate the tweet to a human professional who can judge whether and how to act. The role of AI here would be to speed up the process by working through large amounts of information, reducing this to only the more relevant threads, and prioritising the selection before presenting it to human newsroom operators.

Unfortunately, current text-analysis algorithms are ill-equipped to recognise subtext and sentiment in natural language text. Previous attempts focussing on Twitter data generally appear to have been made to classify approval or disapproval with the intent of classifying public reaction to products or cultural media such as films or music. Outside the domain of Twitter data, the applications of sentiment analysis are more broadly applied but still appear skewed towards interpreting user reviews. Previous work to detect intent such as incitement seems to be more limited, and not to consider the limitations of shorter text messages such as tweets.

Organisational Context

This research is performed in cooperation with the Artificial Intelligence Research Group of the HU University of Applied Sciences[Res]. The primary connection is the research group for Artificial Intelligence headed by Stefan Leijnen, which sponsors the research project described in this document. The research group is committed to human-centred AI and is a subdivision of the Knowledge Centre for Digital Business and Media. The mission of this knowledge centre is to research human centred and data driven solutions for digital transformation.

The research project proposed here is an exploratory component of a larger collaboration between various instances including four research groups of the HU University of Applied Sciences, a number of Dutch municipalities (including Utrecht, Amsterdam, Rotterdam en Den Haag), and the Association of Netherlands Municipalities (VNG). This project, “Goed Gereageerd”, endeavours to find novel data driven solutions to the issues descibed above. The project described here explores a solution based on semantical and syntactical information combined with the structure of Twitter conversations, the results of which will inform future direction for the larger project.

Problem Statement

ToDo: Huib: terugwijzen op wat je eerder geschreven hebt in introductie: bedoel je hier met een letterlijke terugverwijzing of door voorbeelden te herhalen? Is het OK om vanuit deze sectie

naar elders te verwijzen of moet dit deel opzichzelf staan (dacht ik namelijk)? Social media such as Twitter have come to play an important role in the public debate in recent years. Particularly on divisive issues, the unmoderated nature of these platforms lends itself well for malignant actors to incite societal violence without proper accountability[Tse]. Due to the growing trend in the number of monthly active Twitter users[Twi], the scale of communication on these platforms is likely to increase as well. This in turn complicates adequate supervision intended to mitigate the unwanted effects described above, by increasing the manpower required to continue manual newsroom analysis. The expected result is that, if this trend continues and no improvements to current workflow are implement, this might result in providing criminals an increasing means to effectively instigate and organise violence whilst denying authorities the ability to react accordingly.

Addressing this problem without outlawing a platform[1] based on the behaviour of a small subset of users and thus sacrificing freedom of a majority of users requires a more effective way for authorities to observe and moderate the discourse.

[1] Not only would this be an undesireable direction, it is also unlikely to effectively tackle the problem as previous experience has shown that this will only move the problem to outside the perception of authorities and public. Furthermore, the legal basis for such a move would probably be very thin indeed.

Project Goals

ToDo: Huib: nu zonder exploration, maar gaat het zo niet meer lijken op een project opdracht dan een academische exercitie? Ik zoek een beetje naar de sweet spot tussen beide uitersten. The desired outcome of this project is the application of artificial intelligence to detect incitement and proto-incitement from tweets. The solution should be able to combine techniques from discourse analysis, sentiment analysis and computational semantics to consider conversations as they evolve over time, and use the context this provides to estimate a priority rating for newsroom analysts to direct their attention to the most relevant tweets and threads. The challenges associated with this goal include the following:

- The limited lenght of tweets;
- the amount of noise present in tweets;
- the large amount of daily tweets; and
- the lack of previous work and existing tooling to deal with this issue.

This list of challenges is unlikely to be exhaustive; during the design of a prototype solution, more challenges are likely to be encountered, both within and without the scope of this project.

Research Question

In order to structure this project, a main research question has been formulated to translate the desired outcomes — a prototype AI able

to determine sentiment — into a research oriented project. This project is aimed at providing not only the aforementioned prototype, but also describing reproducable strategies on how such a system should be designed to optimally interpret available information in the given context of conversations of tweets or similar short messages.

The main research question additionally specifies the starting point of which indicators to consider in this process, based on implicit hypotheses on what factors best indicate sentiment within the tweets. Given this, the main research question is posed as follows:

What is the most accurate way to automatically determine intent and sentiment from conversations of short informal text messages, such that the level of incitement can be estimated, focussing on syntactical and semantical aspects, and the conversational structure of the text messages?

In the context of this question, accuracy should be measured in the context of recall[2] over precision[3]. The main purpose of an AI in this setting would be to reduce the size of the search space for further human involvement, and as such should ideally not produce any false negatives[4] as a first priority. A secondary objective would be to reduce the number of false positives[5]. As such, it stands to reason to label incoming tweets not on a binary scale (incitement or no incitement), but instead provide a more detailed degree of truth based on the certainty of the classification, and present the results as a sorted queue.

[2] Recall represents the fraction of relevant instances recognised; a recall of 100% corresponds to not missing any relevant tweets.

[3] Precision represents the fraction of relevant instances over the total number of returned instances; a relevance of 100% corresponds to not returning any irrelevant tweets.

[4] A false negative in this context means discarding information that would have been relevant in preventing incitement and escalation.

[5] A false positive in this context is when a harmless message is identified as potentially problematic, which is a lesser problem but would still waste valuable human time.

Next Chapters

This research combines approaches from various different fields of textual analysis. Chapter 2 describes important background information about the context provided by these different fields. Firstly, the term *incitement*, which has hitherto been kept vague, is given a canonical definition for the remainder of this research. Following this, an overview is given for the subjects of *discourse analysis*, *sentiment analysis* and *natural language processing*. This chapter is concluded with a description of the moral frameworks used in the ethical recommendations.

For the execution of this research project, the Design Science approach as described by Hevner[HMPR] is utilised. Chapter 3 describes the research approach, question and methods.

Chapter 4 describes the risks associated with the execution of this projects, and additionally suggests measures taking to mitigate these risks. This chapter will also provide a more thorough overview of the ethical questions that are considered.

2 Background

This chapter contains a broad overview of the theoretical basis of this research, starting with the definitions of some central subject: Utterances and incitement. The rest of the chapter explores the fields of natural language processing (NLP) that this research takes as its starting points, including the relevance and shortcomings of each. Within semantics we consider two broad research areas, *compositional semantics* and *distributive semantics*. The latter will be discussed first, as the central concept of *word embeddings* will be referenced in various later sections. Following this, *discourse pragmatics* and *sentiment analysis* are discussed. This chapter closes by an exploration of how these research areas have previously been combined, and where there still exists room for novel exploration.

Definitions

This section provides some definitions for terms that will be used extensively in the remainder of this text.

Tweets, Utterances, and Discourse

For the purpose of analysing text, the first thing that must be agreed upon is which unit of language to consider central. In this research, the most obvious choice for this would be a single tweet. A tweet can be seen as a unit of language which can be subdivided into smaller units such as sentences, hashtags, etc. These subdivisions can be analysed for sentiment and intent, but the overall intent of the tweet is central to consideration. On the other hand, multiple tweets together can form a conversation, which can also be attributed sentiment and intent. This level will be considered as a central part in this research, but the intent of a conversation is determined by the intent of its constituent tweets. The reason to define the tweet and not the conversation as central is because the tweet is the largest unit in which we can be reasonably certain that the content reflects the sentiment and intent of a single author. By combining multiple tweets into a conversation, the resulting entity *can* have a dominant intent behind it, but this is unlikely to be necessarily the case.

The remainder of this chapter explores existing fields of

knowledge dealing with analysing text on different levels and with different goals. Though some prior work has been done on tweets in particular, the vast majority of scientific writing predates Twitter and as such considers other units of text. The unit most similar to a tweet appears to be that of a *discourse*, in that it can be seen as a building block of conversation. The concept of discource originates in the field of pragmatics, which together with discourse analysis is explored in Section 2.3. Fetzer[Fet] describes discourse as being built up from sentences or *utterances*. Despite the prevalence of the latter term, most authors refrain from providing an exact definition and presume reader familiarity. Levinson[Lev] attemps to do so by juxtaposing the concept of utterance with the concept of sentence. He notes that the sentence is defined based on grammatical considerations, whereas the concept of utterance resides in the uttering of a sentence within a context. Multiple utterances together can form the aforementioned discourse, which in turn forms the building blocks of conversation. Framing the concept of discourse in utterances in lieu of sentences is more relevant in the context of this research, due to the nature of tweets as a digital equivalent to spoken text and will thus be preferred. The term *sentence* will be used purely in a grammatical sense when discussing from a semantic of syntactic perspective.

Incitement

In order to be able to detect incitement in tweets, or any medium in general, the first requirement would be a usable definition of the concept. Timmermann[Tim] spends the first chapter of *Incitement in International Law* on providing a definition, which he summarises as generally including five elements:

> *(i) Negative stereotyping of the target group.*
>
> *(ii) Characterization of the target group as an extreme threat.*
>
> *(iii) Advocacy for an "eliminationist" or discriminatory solution to the perceived threat in the sense of excluding the target group members from society or the human community.*
>
> *(iv) The incitement is carried out in public.*
>
> *(v) The incitement is part of a particular context which dramatically increases the effectiveness of the inciting words, usually through the involvement of the State or another powerful organization.*
>
> — Timmermann, 2014

It is not stated outright whether all of these should be present to constitute incitement, or whether certain combinations are valid on their own. The indicators quoted above are listed as "general components" of incitement to hatred. Similarly, the author does not put any clear requirements on the medium used for the delivery of incitement, and can thus be taken to refer to spoken or written text, images, etc. For the purpose of describing incitement within

the context of international law it stands to reason not to limit the definition to a specific medium as new forms of communication can arise and should be automatically included if the relevant indicators are present. As an example, laws on content and (the limits of) freedom of speech have been around for longer than the medium of video-games, but the same rules should generally apply as for other forms of expression, insofar as this makes sense for the new medium. Conversely, for the purpose of this research, a definition on incitement should focus on the more narrow scope of the project: Twitter and similar forms of communication. This difference is reflected in the definition provided below by considering discourse and utterances.

Returning to the indicators provided by Timmermann, the third item appears to be the most relevant to this study as it captures a *call to action* which suggests imminent violence against a targeted group — a situation that a newsroom should be able to react to in short order. It therefore stands to reason to view the presence of this indicator, even in isolation, as a priority in detecting incitement.

Of the other indicators, the first and second describe what could be considered hate-mongering, but these indicators alone lack the call to action that warrants immediate response by authorities. This missing aspect could follow from the context: if a user is known to advocate violence against group A and in a later tweet compares group B to group A, this can be viewed as a call to violence against group B as well. The first two indicators, therefore, should not be dismissed entirely, but in isolation do not warrant the level of scrutiny as discourse including a direct call to action.

The fourth, and to a lesser extent the fifth of Timmermann's indicators are more of a given in the context, as tweets are by definition[1] public and the amplifying context is provided by the platform and the people on it reading the tweets. Still, the amount of followers a user has, or more generally the projected reach the tweet has can be considered a relevant aspect within the spirit of this element of the definition.

[1] It is possible to put a Twitter profile on private, thus removing the public aspect of its tweets. Tweets on a private profile cannot be seen by the general public, which will also result in those tweets not being visible to newsroom analysts or any potential AI-based solutions. This is per design of Twitter and therefore private tweets are left outside the scope of this research.

Definition (Incitement)**.** Discourse or utterance implying or advocating hostile action against a demonised person, people or status quo.

Computational Semantics

In order to make claims about whether an utterance contains incitement, we need be able to reason about the meaning or content of natural language. The study of meaning in general is known as *semantics*[Par], and various theories exist within this domain using different formalisms to capture semantic content. The study of *computational semantics*[BB] specifically works with approaches

applicable to automated processing, considering representations of meaning usable for computers. As the goals of this project lie in an AI based solution, computational semantics are considered as a starting point. Within this field, two broad approaches are explored for application in this study: *Distributional semantics* and *compositional semantics*. Both are explored in the following sections. Following this, a synthesis of the two approaches, *distributional compositional semantics* is discussed as a means to combine the strengths and mitigate the weaknesses of both.

Distributional Semantics

Translating the concept of semantics to computers and AI is challenging: The language used in general purpose computers is one of numbers[2], where no real equivalents exists for most real word concepts. The field of *distributional semantics*[Sch] aims to derive meaning from the statistics of word cooccurrence and encode this information in terms of *word embeddings*[JMb]: vectors, which are essentially ordered lists of numbers.

[2] In essence, these numbers are binary integers. Using clever encodings, floating point numbers can be worked with as well.

Using a single number, values on a single scale can be given meaning; adding a second number provides a two-dimensional space in which meanings can be assigned to words. For example, Figure 2.1 shows one possible way to encode a set of animals in a 2-dimensional semantic space. Whilst there are infinitely many possible ways to assign or *embed* these points in the space, the chosen embedding is not arbitrary[JMb]. Given the positions of the words and our knowledge of the animals, we could interpret the x axis as representing average size, and the y axis as a subjective measure of pettability.

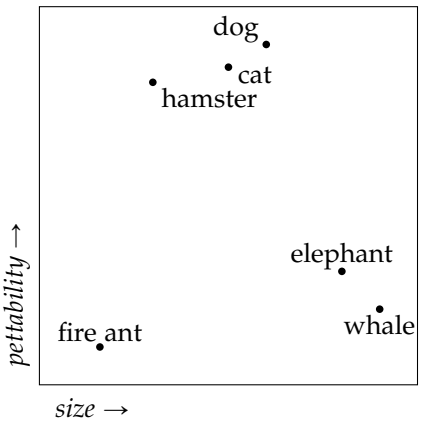

Figure 2.1: An example 2D word embedding space

The result of the informed embedding given in the example is that words assigned closely together represent words with some similarity in semantic context. Pets, in the example in Figure 2.1, are clustered towards the top (highly pettable animals) and mainly centred on the x axis (not too large or too small to keep around). The example here is limited, but serves to illustrate a central point in distributional semantics: linguistic items with similar embed-

dings have similar meanings. By adding additional axes, more semantic context can be encoded.

To encapsulate the complexity of human language semantics, the number of dimensions required is generally in the order of magnitude of a few hundred axes[JMb]. In general, greater dimensionality allows for greater granularity in expressing the meaning of terms and the space available for relational connections. At the same time, high dimensionality incurs a computational cost which gives rise to a tradeoff.

The vector space model for semantics is powerful in that it allows computers to reason about the semantics of words using vector arithmetic operations[BZ], and as such will be referenced in other research areas going forward.

As each word encountered in a text needs an associated vector, and each vector requires a large amount of numeric values, assigning these embeddings manually is unfeasible, nor will randomly assigned vectors work in providing the desired semantic context. The field of distributive semantics therefore also deals with methods to automatically generate such vector spaces[AX]. The ability to do so depends on the statistical nature of language: certain combinations of words occur more or less frequently together, for example "green grass" vs. "green ideas". In order to capture these observations, distributional models are trained on *corpora*, large datasets of text, based on the assumption that words that occur together, or are used in similar sentences, are more closely related than words that do not. During this process, a set of *stop words*[WS] is often ignored. Stop words are generally taken to be short, common words without any real semantic context. Examples are articles, prepositions and similar words, all of which have signifance in syntax rather than semantics. There are multiple ways[DDF+][ST] to extract the distributional information to produce the desired word vectors, which can be subdivided into two categories:

Count based models[DDF+] work by counting, for each word in a corpus, how often it occurs near each other word and storing this information in a *cooccurrence matrix*. The definition of near can vary according to the specific strategy used in calculating the word embeddings, but usually means something like "next to eachother" or "both occur within a shared 3 word window". Generally, stop words are removed when building this cooccurrence matrix. The resulting matrix is typically *sparse*[Duf], i.e. containing lots of zeroes. This is undesirable both from a computational point of view, and because this cause for example similarity measures to tend to 0.

The answer to this issue is to apply some form of *dimensionality reduction*, using linear algebra to effectively find the most relevant axes for a lower dimensional space and embedding the sparse count-based vectors (rows or columns of the cooccurrence matrix)

via a change of basis into this new space.

Predict based models[ST] begin by assigning each word in the provided corpus to a *one hot encoded* vector or basis vector: A vector with only 0s and a single 1 the position of which uniquely identifies the word. Again, stop words have generally been removed from the corpus at this point. Then, training examples are generated based on the chosen model and the cooccurrence of words in the corpus. For example, we consider a *Continuous Bag of Words* (CBOW) model with windows size 1: each word is only associated with the words directly before and after it. For each word in the corpus, a positive training example is generated as follows: Given a sentence such as "Jackdaws love my big sphynx of quartz", the combined input of the vectors associated to the pair ("jackdaws", "my") should be associated to "love"[3], the input ("love", "big") should be associated to "my", etc. These training examples are supplemented by negative training examples (signifying word combinations that should not occur together) which can be generated by randomly combining words and removing randomly generated examples corresponding to actually occurring training examples. This training data will then be fed to a single layer neural network. The final resulting weight matrix can then be used to transform a one hot encoded input vector to a word embedding.

[3] This example could be read as a fill-in-the-blanks for "Jackdaws _ my ...", the answer to which should be "love".

The field of distributional semantics and the concept of word embeddings deriving from it form the basis of many forms of textual analysis, which warrants their consideration for this project. Distributional semantics capture one aspect of natural language: statistics. There is, however, another aspect to language of similar importance, which is compositionality[Bra].

Compositional Semantics

Compositionality refers to the effect of combining words in a language. The "green grass" mentioned above combines two words to describe a more specific idea than just "green" or "grass" separately. The way that these words can be combined relies on the *syntax* or grammar of a language, which in English for example dictates that "green grass" and "green ideas" are valid, but "green else" or "exist grass" are not. The syntax of a language and the validity of sentences emerging from this syntax can be studied using formal theories, such as logics[Lamb] with their associated[Cha] type theories or pregroups[Lama].

The field of *compositional* or *formal* semantics takes the syntactical information as its starting point, attempting to tag each word using grammatical type and constructing a *parse tree* (iff a sentence is valid) or failing to do so (which happens for invalid sentences). An example of a parse tree is shown in Figure 2.2.

In addition to determining whether a given sentence is grammatically valid, parse trees can also be used to determine how various

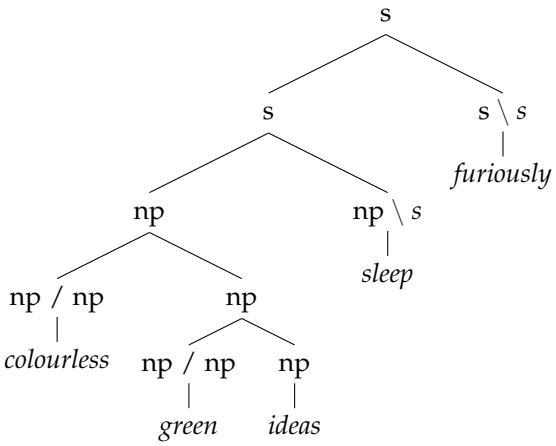

Figure 2.2: An example parse tree. The type $np \setminus s$, as seen for "sleep" means that provided a noun-phrase (np) on the left, this token will produce a sentence (s).

parts of a sentence compose and to differentiate between multiple distinct sentences that can be formed using the same words. For example, the two sentences "The man bit the dog" and "The dog bit the man" are built up from the exact same set of words, which each have the same semantic value regardless of the sentence. Despite this, the two sentences themselves clearly have different meanings, a fact which is represented in their different parse trees.
ToDo: Huib: ik heb hier nog wat meer context toegevoegd mbt formal semantics; hiermee wordt de voornaamste tekortkoming van het systeem ook beter uitgelicht

THE SECOND STEP in formal semantics is to assign semantics to the syntactic structure recovered. This is done in terms of set theory and higher-order logic, generally using *lambda abstractions*[DWP], also known as *anonymous functions*. An anonymous function is formed using a lambda symbol λ to capture a single variable. The function $\lambda x[\mathbf{exist}(x)]$ captures an variable x, which can be bound by calling the anonymous function on an argument. Every occurence of x within the function body, i.e. between the square brackets, gets replaced by the value bound by it. For example, to state "I exist", the function can be applied to the token **me** as such: $\lambda x[\mathbf{exist}(x)](\mathbf{me})$ will evaluate to $\mathbf{exist}(\mathbf{me})$.

As a simple example of using formal semantics, a sentence "Alice sees Bob" would be translated to $\mathbf{see}(\mathbf{Alice}, \mathbf{Bob})$. This semantic valuation can be built up from the following items:

"Alice"	noun phrase / proper name	**Alice**
"Bob"	noun phrase / proper name	**Bob** [4]
"sleeps"	transitive verb	$\lambda y[\lambda x[\mathbf{see}(x, y)]]$

Given the example phrase, the first step is to combine the object with the transitive verb to get the verb phrase "sees Bob": $\lambda x[\mathbf{see}(x, \mathbf{Bob})]$. This verb phrase can then be combined with the subject to get the logical expression for the complete sentence. Finally, the truthiness of the logical phrase can be analysed based on a given model of reality. The identity of the two individuals, as well as the meaning of the concept of seeing are represented within the

[4] Other conventions exists, such as using the lambda for nouns. In this convention, a noun phrase such as "Alice" gets mapped to $\lambda P[P(\mathbf{Alice})]$ with the lambda capturing a predicate, and a verb such as "sleeps" becomes plain **sleep**, referring the set of sleepers. In this example, $\lambda P[P(\mathbf{Alice})](\mathbf{sleep})$ evaluates to $\mathbf{sleep}(\mathbf{Alice})$.

system only as far as they have been defined within the model.

THE SHORTCOMINGS OF THIS APPROACH for the context of this research are apparent. The meanings of words have to be inserted into the system on a per-word basis, and similarity between words is not captured unless specifically provided. The result of comparing the sentences "Socrates sees Plato" and "Socrates sees a man", for example, will depend on whether the system knows that Plato is a man. Though this form of semantics can be powerful in capturing how words combine to form sentences, the meaning of the words themselves is left as an exercise to any programmer attempting to automate text analysis.

Distributional Compositional Semantics

The two approaches to computational semantics mentioned in the previous two sections are each based on different starting points and at first glance do not appear to overlap. This section discusses a new framework to combine the two approaches to solve the most pressing problems in each:

- Distributional semantics' difficulty in dealing with the grammatical aspect of a sentence[SKB], exemplified in the distincion between subject and object.

- Compositional semantics' dependency on formalised knowledge to deaal with meaning on a word-based level.

This composite approach, often referred to as *distributional compositional semantics*[CSC], seeks to use parse trees and their associated types to make grammar-informed compositions of distributional word vectors. This done using tensors[KSP], higher order equivalents of vectors representing (multi)linear functions[Lan]. This approach allows distributional semantics to provide meaning at the word level, and assign meaning to larger sentences based on both the meaning of the words and the way that these words interact.

ONE AREA OF INTEREST within the study of distributional compositional semantics is the handling of ambiguity in sentences[MW]: A single sentence can sometimes lead to multiple structurally different parse trees corresponding to multiple possible readings of the same sentence. Specifically in Dutch, which is the target language for this study, the way that relative clauses are handled can easily lead to sentences with more than one meaning, for example "de man die de vrouw haat" can be translated either as "the man that hates the woman" or "the man that is hated by the woman". These forms of ambiguity are of potential interest to this research, as this provides a way to formulate a tweet with multiple meanings, one of which is the intended reading whilst another provides a form of plausible deniability.

Semantics, summary

Given the above exploration, the usage of distributional compositional semantics is preferable as a basis for this research project. The applicability, however, largely depends on the availability of higher-order tensors, corresponding to words with special roles such as verbs or adjectives[KSP], for the Dutch language. As most work on this subject is relatively new, it cannot be guaranteed that high quality higher order word embeddings are readily available for a relatively small language. As a fallback, regular distributional semantics appears to be preferable, due to the requirement for expansive set theoretic models in compositional semantics. Even without tensor availability, some of the lessons provided by distributed compositional semantics may be used to inform this research.

Discourse Pragmatics

The mentioned existing work on semantics provides a foundation to determine the sentiment of a tweet. However, it should be taken into account that text rarely exist in a vacuum; additional context is required to make sense of an utterance. In the case of tweets, part of this context is determined by the conversational structure provided by Twitter. Tweets can act as a reply to another tweet, and can themselves be replied to. In other words, they form part of a conversation, the content of which is built from the semantics of the individual tweets, but also their interdependence.

The domains concerned with this contextual aspect to meaning in text are the closely related fields of *pragmatics*[Lev] and *discourse analysis*[JMa]. The focus of the latter appears to be on the linguistic components of the context, and that which can be inferred from the surrounding discourse[AhS]. The former discipline tends to shift this notion of context to focus more on external or physical context and to analysing speaker intention. In this regard, discourse analysis appears to be more applicable to the platform at hand, whereas pragmatics more closely aligns to the stated goals of this project as incitement is one potential form of speaker intention. Most ideas referenced from both of these fields exist in the intersection between the two, or arose in one field but find application in the other. The term *discourse pragmatics*[AhS] is used to describe the hybrid field arising from the collaboration between the two subjects and as such will be preferred as the general term for the combination of these fields in this writing; in most instances, more specific terms will be used to refer to concepts used within discourse pragmatics, as described in the following subsections.

For the purpose of this research, two main ideas appear to be of interest: *Speech acts* and *conversational implicature*. Both are briely considered in the following sections. We futhermore investigate the *dialogic principle* and *pragma-dialogue* due to its relevance to

the conversational aspect of tweets.

Speech Acts

A central concept to the field of discourse pragmatics is the concept of *speech acts*, also referred to as an *illocution*, as posited by Austin[Aus] and expanded upon by Searle[Sea]. The central thesis of this concept is that an utterance can be more than merely a statement of information, but can itself be seen as an act with real-world consequences. As human reality is shaped by the power of words, we allow words to impact that reality just as physical actions would. For example, a head of state has the power to enact law or declare or end wars by words — spoken or written — alone. Similarly, a parent naming their child will in essence do so by stating a new fact about the world and thereby creating a world in which a newborn child is named.

IT SHOULD BE NOTED THAT the result of any speech act depends on context and speaker. The sentence "The match has begun!" will have the intention and effect of starting the match when uttered by an umpire, on a pitch where two teams have assembled for a match of cricket. The same sentence, uttered by another person in the exact same situation, or by the same person in a different situation, might have the same intention but will not accomplish the same effect. In the former case, the speech act is considered to have been *felicitous*; in the latter case, the speech act fails to be performed in what is referred to as a *misfire* — the person making the declaration has no authority to start a match in the given circumstances, and as such nothing happens.

A second way in which a speech act can fail to be felicitous is in the case of *abuse*. An example would be when Alice promises Bob to peform an action without intention to follow up on it. In this scenario, the speech act *can* be considered to have been performed, but the act is not felicitous. The rules which dictate whether a speech act can be considered felicitous are called *felicity conditions*.

AN UTTERANCE cannot be seen separately from intention. For example, the intention behind a question such as "Do you think it's cold in here?" will in many contexts be to get a person to close a window or turn up the heat, not to start a debate on that person's perception on the temperature. Austin[Aus] describes three levels on which a speech act can be analysed:

- The *locutionary act* is the actual act of speaking or writing the sentence.
- The *illocutionary act* signifies the implied request or demand and represents the intention or purpose of the speech act: What is the speaker trying to accomplish by making a statement?
- The *perlocutionary act* is the actual effect of the speech act.

In the example of "Do you think it's cold in here?", the locutionary act consists of a speaker uttering the question, the illocutionary act is to request the addressee closes the window, and the perlocutionary act is at the very least conveying the speaker's discomfort with the temperature, and potentially persuading the addressee to help solve source of the problem.

Searle[Sea] goes on to categorise illocutionary acts into five categories:

- Assertive or representative, which state a fact one believes to be true, committing to the validity of the preposition, and attempting to convince the receiver thereof;
- directives, where one wishes to persuade the receiver to do something, including but not limited to ordering, requesting or suggesting;
- commissives, where one makes a promise or threat, or enters a verbal contract;
- expressives, which reflect emotions or attitudes such as apologies and expressions of gratitude or (dis)approval;
- declarations, which by their utterance (attempt to) change the world by representing its new state, such as christening a child or declaring war.

It should be noted that utterances can fall in an overlap between some of these categories: The sentence "I promise to obey" both commits the speaker to obediance, and declares the promise. The difference between these types of speech acts is relevant to interpreting the intent of the speaker. In the context of incitement, assertives and expressives can for example serve to convince the receiver of a perceived threat or injustice, thereby providing a context for commissives (in this case, threats), directives (suggesting violence) and declarations (of a state war[5]). This flow could play a part in determining the intended effect of a series of statements.

[5] In this case, war is used in its broader definition; it is meant to include wars on peoples, groups or concepts within a nation state, not just war between separate political entities.

Conversational Implicature

In order to gauge speaker intention, which has been established can differ from the literal meaning of an utterance it is important to separate what is said from what is implied. The latter is called the *conversational implicature*, and can be detected by how a speaker deliberately fails to obey certain unwritten rules of conversation.

These unwritten rules or *maxims* have been postulated by Grice[Gri] in the theory of the *cooperative principle*. In it, the assumption is that both speaker and listener are trying to communicate effectively. The listeners should be confident that in any case of ambiguity, the most likely intended meaning is the correct one. In order to achieve this, the speaker will generally obey 4 maxims:

The maxim of quantity states that the speaker should communicate the right amount of information; they should not leave relevant information out or include unnecessary details. For example, when discussing what main course to order in a restaurant, one could list the available options on a menu. By excluding a dish without good reason[6] or including a dessert, one breaks this maxim.

[6] An example of a good reason to exclude a dish would be to conform to a listeners dietary preferences, thereby favouring the maxim of relation above the maxim of quantity.

The maxim of quality states that the speaker should not communicate information believed to be false, or for which there is insufficient evidence. In the restaurant example, one could break this maxim by suggesting dishes not on the menu. Both deliberate lies and overstatement of confidence in a fact are included in counterexamples to this maxim.

The maxim of relation states that the speaker should only communicate relevant information to the context at hand. Continuing the example of the restaurant, starting a discussion on the weather or the state of politics whilst deciding what to eat would in general violate this maxim.

The maxim of manner states that the speaker should communicate in a clear manner, avoiding obscure or ambiguous terms, being succint and not leaving out crucial steps in reasoning. In most cases, listing the original Chinese names of dishes to an English speaker, or adding irrelevant information about the origins of each dish, one could be in violation of this maxim.

GRICE STATES THAT in general conversation people implicitly and unconsciously try to obey these rules. Any overt deviance, then, could be interpreted as a deliberate *flouting* of a maxim, which in turn signals to the receiver that the information conveyed is not or not merely the information semantically contained within the sentence.

For example, a tweet containing a turn of phrase such as "It would be a shame if someone were to X" can be understood as, depending on context, being either an honest expression of desire *not to see X happen*, or an covert suggestion to a target audience to perform X. If the concept of X has not been hitherto mentioned and can generally be perceived to be a negative thing, this utterance would at the same time flout the maxim of quantity by introduce more information than needed — as the negativity of X is common knowledge, flout the maxim of relation — as prior to this noone was openly considering X to happen, and flout the maxim of manner — by being more verbose and indirect than appropriate. Incidentally, the remaining maxim of quality is also flouted by the author, who themselves do not accept the generally agreed upon truth of X being negative, which further signals to an informed audience the intended meaning. In this example, the tweet can reasonably be flagged to potentially inciting.

ToDo: Huib: is dit waar je op doelde? One challenge in the application of Grice's maxims in this research is that it is not immediately obvious how to automatically determine when a maxim is being flouted, as this process requires a lot of context. Some work[VHLH] has been done on using machine learning, specifically support vector machines, on the automatic detection of irony on Twitter inspired by the Gricean notion of maxim flouting, but it seems this process is mainly informed by Grice instead of directly based on it.

More generally, the concept of implicature relates to the political idea of *dog whistling*[GS], where a speaker uses a specific euphemistic phrase to signify one thing to one part of the audience (the in-group), and another thing to the rest (the out-group). For a historical example, the phrase "state rights" has been used to platform racial segregation in the United States[War]. Instead of providing a direct answer to a question regarding the issue of desegregation, a politician campaigning to maintain the status quo (thereby campaigning against desegregation) would shift the debate to the issue of state rights. By answering a question about A by starting about B, the politican can flout the maxim of relation.

In the case of online communication such as tweets, this process can be used to signify meaning to a target audience — those being aware of a certain context — whilst at the same time being readable by a more general audience without conveying the same message. A modern example can be found in the usage of the okay sign emoji in tweets: in certain alt-right communities the associated hand gesture came to be understood as representing the phrase "white power"[Lea], lending context to the usage of the symbol beyond the more generally understood original meaning of "It's okay" or "I'm okay". Using this context a tweet can signal part of an audience within an in-group a different reading of the same text compared to what an out-group audience would understand. For example, a tweet calling out a succesful person of colour and containing the emoji could be read both as an endorsement by the tweeter, or as a call to action for white nationalists. In this case, the intentions of the author can then be considered harmful, while the tweet itself provides a form of plausible deniability.

The Dialogic Principle

In Pragamatics, Discourse, and Cognition[KH], Horn and Kecskes identify pragma-dialogue as one of three approaches within the field of pragmatics, based on work by Weigand[Wei]. This field shifts focus to the dialogic nature of interaction, where two interactants act and react. The *dialogic principle* states that speech acts are not communicatively autonomous, but that the smallest possible subdivision is the sequence of action and reaction.

In the case of Twitter discourse, the general trend in conversation does not exactly resemble dialogue, i.e. interaction between two constant parties, but rather a multiway conversation where interactants can join and apparently[7] leave without formality. Nevertheless, the focus of this paradigm on action and reaction seems highly relevant to the analysis of incitement and general intent behind tweets. In a general conversation on Twitter, the initial action is a publicly visible tweet or thread[8] of tweets by a single author which also forms an obvious starting point for any automated system considering a conversation. The subsequent reactions can be split into three broad categories: (i) replies and quoted tweets, i.e. publicly interacting with the tweet and adding one own thoughts on the matter; (ii) likes and retweets, i.e. publicly affirming having read the tweet and expressing approval and (iii) having read the tweet and internalised part of the message without visibly interacting. The last form of reaction is clearly the most common — having read and at least understood the content of a tweet can be seen as a requirement for further interaction — and therefore most desirable to use as an indicator. Unfortunately, this type of reaction is also the least visible and thereby hard to consider in any automated capacity. Likes and retweets can be viewed as a rudimentary metric of how often a tweet is read and internalised, but cannot be considered very accurate as the ratio between agreement and publicly expressed agreement is not necessarily the same for different tweets as it may depend on factors such as how vocal the public is and how socially acceptable endorsing the view expressed in a tweet is.

For the purpose of this research, the reactions of replying and quoting are of the most immediate interest, as these reactions are themselves actions which can elicit further reaction. This perspective allows us to consider the tree-like structure formed by tweets and their replies and quotations. Intent can then be analysed on two levels: First on the level of a single tweet, and then on (paths within) a conversation tree. It is part of our hypothesis that the way intent within individual tweets evolves over the course of a discussion can yield patterns which can be used to more accurately judge the intent of individual tweets and the conversation as a whole, allowing extrapolation to locate tweets of interest potentially in advance.

ToDo: Huib: sectie hierboven is uitgebreid, hypothese valt hier logisch maar goed om deze in BG te noemen?

Sentiment Analysis

The problem of detecting emotion from text has been explored to a great degree, combining different fields visited earlier in this chapter. The field of *sentiment analysis (SA)*[Fel] overlaps with distributional semantics as described above, generally treating a piece of text as a series of word vectors. This semantic information is combined with *part-of-speech (POS) tags* to form the input for machine

[7] Due to the nature of the platform, it can be observed that a interactant ceases to contribute to the discourse, but not whether they actually continue to listen in.

[8] It is common practice on Twitter to self-react in order to avoid the 280-character limitation on tweet length.

learning techniques in order to extract information the sentiment expressed in an utterance.

Part of Speech Tagging

Whereas distributional semantics generally disregards syntactic information contained in a text in favour of the semantic content of the individual words, sentiment analysis frequently includes some form of part-of-speech tagging to distinguish homographs[9]. A part of speech in this context describes the grammatical role of a word in a sentence: Verb, noun, determinant, etc. The process of POS tagging is generally based on stochastic methods, frequently employing a *Hidden Markov Model (HMM)*[Mar12][Kup]. The hidden states in this context are the parts of speech to be determined for every word in a sentence.

For example, consider the sentence "Consuming lead will lead to poisoning." Here, all words can be interpreted as at least two parts of speech, and the word "lead" occurs twice in different roles. In order to figure out which POS should be assigned to each word, a Hidden Markov Model is employed. The words in the sentence correspond to the emissions of the model:

consuming → lead → will → lead → to → poisoning

The token "consuming" likely corresponds to a *verb*, but might also be an *adjective* or (rarely) even a *noun*. As this is the first word, the probability for it being a verb does not depend on the previous word, but only on the probability that a sentence starts with a verb[10] multiplied by the probability that any randomly chosen verb would turn out to be "consuming"[11]. The probabilities that the word is an adjective or a noun are calculated similarly. For the second word, there are two possible POS tags: verb or noun. This yields six possible taggings for "Consuming lead": verb → verb, verb → noun, adjective → verb, etc. The probability for the tag verb → verb is the product of three probabilities:

- The probability that "Consuming" is a verb, as calculated before;
- the probability that a verb follows a verb, without considering the specific verbs and
- the probability that a randomly chosen verb will turn out to be "lead".

By repeating this process, the probabilities for each possible tagging of a sentence can be computed, after which the most probable tagging is chosen.

Twitter Data

ToDo: Huib: deze sectie en die daaronder zijn (met grotendeels dezelfde inhoud) geherstructureerd - leest dit wat jou betreft beter

[9] Homographs are akin to homonyms in that they denote a set of words with the same spelling, whilst dropping the requirement of also sharing the same pronunciation. For example, the word "lead" can be interpreted as a verb, meaning to guide, or as a noun, meaning the element. Although these words will rarely be confused in spoken text due to a difference in pronunciation all standard dialects of English, the words are spelled the same and thus in isolation indistinguishable in the context of written text.

[10] This probability can be trained on a corpus and forms part of the trained model.

[11] This probability is similarly part of a trained model

zo? De citaties zijn naar voren gehaald, de voetnoot over sentiment versus intent daaronder. Daarnaast ter overzicht de voorbeelden ook in een tabel gezet

A lot of work has gone into the application of sentiment analysis to Twitter data. This includes procedures to work with noisy data[BF, BS] and emoji[GBH][Rea]. Also of potential interest is the work of González-Ibá, Muresan and Wacholder[GIMW] on identifying sarcasm in Twitter, as this may also be used to code subtext into tweets. Additionally, Nazir, Ghaznafar, Maqsood, Aadil, Rho and Mehmood[NGM+] investigate the combination of tweet volume, hashtags and sentiment analysis to perform signal detection.

Mukherjee and Bhattacharyya[MB] propose a method for polarity detection of tweets using discourse relations. Their work focusses on discourse relations within tweets, considering the tweet as a whole instead of sentences, but not considering conversations consisting of multiple tweets by different authors. The influence of discourse analysis on their work is the consideration of relations between sentences within a tweet, by considering conjunctions signifying coherence relations. This also serves to highlight a different approach to stop words from most semantics oriented work: Conjunctions are generally regarded as carrying no semantic information and thus discarded, but are here considered on a supra-semantic level.

Finally, Oluoch[Olu] in his masters thesis has studied the application of sentiment analysis for the detection of radicalisation on Twitter. The project follows a fairly standard approach of text classification machine learning approaches and does not consider the syntactic structure of tweets, nor the aspect of conversational flow.

Sentiment versus Intent

Generally, the goal of the work cited above gravitates to determining whether a tweet is positive (happy or excited about a product, person or situation) or negative (sad, angry or less than enthusiastic), which is subtely different from the concept of intent. The intent of any form of communication can be considered beneficious/constructive or malicious (signifying potential incitement) regardless of positivity or negativity. Consider four example utterances corresponding to the four possible combinations[12]:

[12]

	positive	**negative**
beneficial	i	ii
malicious	iii	iv

(i) "I am happy to live in a society were healthcare is accessible!"

(ii) "Utterly dismayed at recent developments, I hope they'll manage to fix this soon!"

(iii) "This politician sucks, and something should be done about him!"

(iv) "Today is a good day to die! We will bathe in the blood of our enemies!"

In this example, (i) is both positive in sentiment and constructive; it should not register as incitement. Example (ii), whilst negative, does not express any incitement. The latter two examples do contain inflamatory language and as such should be flagged, despite (iii) sounding more negative and (iv) being likely to be flagged as positive based on the semantics of the most obvious indicator words. This distinction between sentiment and intent is important when determining whether and how to apply previous solutions to different problems to the subject of this research.

Summary

Based on a thorough search of available literature, the problem central to this research has not been solved in this form but related work is available to inform individual steps in the process of detecting incitement from tweets. The fact that previous research on Twitter data mainly focuses on sentiment than intent does not invalidate this previously work for the context of this research. Whilst the indicators utilised may not be directly applicable, they can provide insight regardless of what information needs to be recovered from messages and how to deal with noise present in the medium. Previous work on computational semantics provides a basis to work with the meaning of the text within a tweet, whereas concepts from discourse pragmatics can be used to inform how to combine intent gathered from individual sentences to the level of tweets and conversations.

3 Research Methodology

This chapter describes and motivates the methodology used in this research project, its objectives, and the central research questions.

The primary methodological framework used for this research project follows the principles of Design Science Research, the application of which to information science was posited by Hevner[HMPR] in 2004. In this approach, the central process consists of of a practical problem, which is then to be progressively solved by the creation of innovative artefacts. This design phase is informed by interpretation on the desires as formulated by the stakeholders and study of existing work in relevant fields. At the same time, the results of this design phase are continuously evaluated, providing further direction for the cycle to repeat with the refinement of existing or creation of new artefacts[JP]. This method matches the objective of this project, where the desired outcome is the creation, study and adoption of a novel solution, rather than the study of an already present phenomenon or framework. The focus, then, is twofold: On one hand, study of existing methods and previous research is an integral part of this research. On the other hand, no batteries-included solution is readily available, so innovative exloration and recombination of existing techniques will also be necessary to solve the issues at hand. Using these two approaches, it is the intention of this project to explore the design and evaluation of a possible conceptual model to tackle the issue. The focus herein is on research, to evaluate whether and why the chosen model is applicable, rather than producing a finished, ready-to-market product.

For the scope of this project, the problem to be addressed is that of identifying sentiment, specifically incitement, from short public social media posts such as tweets. The problem arises from the fact that existing methods of sentiment analysis depend on a certain minimum amount of content in order to correctly predict the general sentiment of a text. Most methods are based on a bag-of-words model, wherein stop words[1] are removed leaving even less text to work with. A tweet, by definition, is short, and thus on itself may fail to provide adequate textual information for the detection of sentiment. This is further complicated by the fact that tweets are generally comprised of informal language, and can include a large amount of information not recognised as text by naive natural lan-

[1] generally short common words with little or no semantic content, instead providing syntactic information on how other words relate.

guage processing: hashtags, accidental or delibarate misspellings, links, emoji, etc. This leaves the amount of parseable text generally even lower than the 280 character limit imposed by Twitter, and causes every rejected word to have a relatively large impact.

To apply the design science approach to this project, the three cycles of Hevner[Hev] are used as a guideline. The central design phase cycles between the construction of artefacts to test out theories, and using the results to refine the assumptions for the next cycle.

Subquestions

This research follows three separate but codependent cycles as described by Hevner: (i) Collecting domain knowledge [rigour cycle], (ii) Model development [design cycle], and (iii) Application context [relevance cycle] The rest of this section will discuss each cycle in more detail and formulate the relevant subquestions. Following this, a summary is provided including a graphical representation of the questions and their dependencies in Figure 3.1.

Collecting domain knowledge (rigour cycle)

The first cycle is designed to collect and evaluate domain-specific knowledge and existing solutions. This phase will be guided by the following questions:

RI1: "WHAT INDICATORS FROM DISCOURSE PRAGMATICS AND SENTIMENT ANALYSIS APPROACHES HAVE THE HIGHEST CORRELATION WITH THE LEVEL OF INCITEMENT?"

This question is intended to get a grounded overview of applicable indicators in the domain of discourse pragmatics pertaining to sentiment, and in the domain of sentiment analysis pertaining to conversations, in order to detect emotion — specifically incitement. This question will be used to fuel the design process required to answer De1 and will be part of the literature review for this project. Relevant papers will include one or more of the following:

- The application of sentiment analysis as regarding to the detection of incitement or negative emotional content.
- The application of discourse pragmatics as regarding the evolution of emotional content within conversations.

RI2: "WHAT ARE THE MOST PROMINENT APPROACHES THAT HAVE BEEN USED TO EXTRACT INFORMATION FROM THE STRUCTURE OF CONVERSATION IN OTHER DOMAINS THAT CAN BE ADAPTED TO THE EXTRACTION OF SENTIMENT?"

This question connects existing NLP techniques to the handling of structured text, as used for distinct yet comparable applications, to

the task at hand. It will be used to fuel the design process required to answer De2 and will be part of the literature review for this project. Relevant papers will include one or more of the following:

- Distributional NLP approaches used to analyse the semantic content of natural language text.
- Compositional NLP approaches used to analyse the syntactic structure of natural language text.
- NLP approaches used for structured conversations consisting of trees of short messages.

Ri3: "What is an adequate and feasible benchmark for the performance of incitement-detection Artificial Intelligences?"

This question explores how a potential solution could be compared to existing or future solutions, which anticipates and facilitates the evaluation step of the design cycle and thus connects to De3. This step will ultimately help in assessing the value of the results of this project and will be the final part of the literature review for this project. Relevant papers will include one or more of the following:

- The existance of benchmarks containing elements of sentiment analysis, specifically incitement detections.
- The formalisation of benchmarks regarding different yet similar problems within the domain of NLP and/or discourse pragmatics.

Model development (design cycle)

In this cycle, a prototype AI will be developed to extract sentiment and incitement information from conversational trees of text fragments.

De1: "How can existing insights on incitement-detection from conversational and sentiment analysis be used to steer the developtment of a semantic, sytactic and structural model of analysis?"

This question serves to connect the grouding developed in the answering of Ri1 to the design of the artefacts utilised to test these ideas in the context of this project.

De2: "How can existing models of semantic, syntactic and structural analysis on conversations be adapted to detect incitement?"

This question is similarly connected to Ri2, aiming to apply the models available within the NLP domain to the design cycle of this research.

De3: "How do the results of the design cycle evaluate using the developed benchmarks?"

This question formalises the evaluation step of the design cycle, applying predefined metrics to the produced artifacts and gauging their effectiveness. The acceptance criteria for this research question are dependent on the results of Ri3.

Application context (relevance cycle)

The final cycle is concerned with the applicability of the prototype artifacts to the goals as formulated by the stakeholders of this project. As this project is primarily connected to a larger research projects that is still in the planning stages, verification by testing the prototype in its intended environment is unfeasible considering the timescale dictated by a master's thesis project. As future adoption of the prototype and/or technologies derived from it are presupposed, the relevance cycle of this project will instead be focussing on guidelines for eventual real-life application.

The primary potential issues with the adoption of the technologies under scrutiny are concerned with the ethical considerations involved in applying such an AI to user-generated data without prior consent or reasonable user-expectation on the usage of their data, and the implications should the technology be adapted beyond the initial scope of incitement detection and into more thorough invasion of an individuals thoughts. To scrutinise these issues, the following two questions have been formulated:

Re1: "Which primary ethical dilemmata should be considered in the design and exploitation of the prototype artefact?"

Re2: "What ethical recommendations should be considered in the design and exploitation of the prototype artefact?"

In order to answer these questions, the first step would be to perform an ethical check by interviewing one or more people with relevant ethical knowledge and to get their view on what dilemmata are involved in the exploitation of the proposed prototype (Re1). A priori, the following considerations will be included in these interviews:

- Ethical considerations on working with user-generated data without prior consent.
- Ethical considerations on automated vetting of private citizens and the prevention of misuse.

Following this, the identified dilemmata will be subjected to the

following ethical standards (Re2) in order to formulate recommendations:

- Virtue Ethics (Aristotle),
- Deontology (Kant),
- Utilitarianism (Stuart Mill) and
- Justice as Fairness (Rawls)

Section 4.2 contains a more thorough discussion on the expected concerns, which will be used as the basis of the interviewing phase described for Re1.

Summary

In summary, the project has been divided into three relevant sub-problems, each represented in the rigour- and design cycles. These three sub-problems correspond to the two domains of knowledge relevant for the case at hand, supplemented with the need for benchmarking the results. The relevance cycle is less substantive for this project, as the results will be unlikely to be adopted in time for the duration of this research project. Consequently, this cycle will be limited to a consideration of the ethical implications of adopting an AI system as considered in this study. The internal dependencies of the total project are visualised in Figure 3.1.

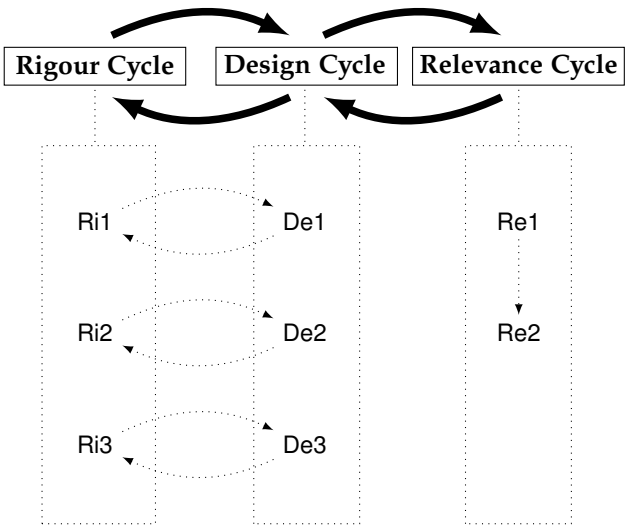

Figure 3.1: Internal Dependency Graph

Sprint Planning

In a practical sense, the work needed for the completion of this project will be organised in sprints. Sprints are organised using vertical slicing, meaning that each sprint works to further all seven research questions, ideally in more or less equal measure. However, the first few sprints are expected to disproportionately favour Ri1

and De1 in order to compensate for the gap in previous knowledge as compared to the subject matter of Ri2 and De2 respectively. The latter two questions are more closely alligned to experience accumulated in the courses and projects of this master's track, which prioritises achieving relative parity early in the project.

As each sprint includes work towards answering seven research questions, the sprint duration chosen should be adequately substantial to ensure progress can be made on all of these within a single sprint. Conversely, the project is of fixed duration, and should contain sufficiently many sprints to facilitate the feedback loop desired in the design science paradigm. Balancing these two desires implies a sprint duration of 2 weeks to be a workable compromise.

Literature Review Methodology

As this project works towards applying existing Natural Language Processing (NLP) technology to a novel problem domain, that of sentiment analysis and discourse pragmatics, it is of vital importance that the literature review part of this endeavour is both expansive as well as grounded in rigour. As the problem domain and the domain of the intended solution are largely disjunct, it is important to include both in this researsch phase, as well as any intersection previously touched upon by research.

In order to ensure a thourough and complete literature review phase, we adopt an approach where we first establish selection criteria on which information would be relevant and need to be included. This would then be followed by a literature search and one or more iterations of refinement. The goal of this cycle is to get a grip on the best working hypotheses on how to detect incitement before proceeding to test the results on the data at hand. The sources included as the starting point of the literature search are the following:

- Web of Science,
- ACM Digital Library,
- Science Direct,
- Springer Link,
- Wiley Online,
- WorldCat.org and
- SAGE

In order to facilitate searching these sources, an aggregate search engine provided by the Hogeschool Utrecht is used. The search terms are based on the criteria sketched for each research question above, and are refined to produce a reasonable number of results per question. If the number of articles is prohibitively high even

after refinement, a select and truncate approach is used based primarily on the number of citations each article has. This ensures that established and relevant literature is considered in this process. After an initial selection has been made, a second round of selection is based on the relevance determined by reading the abstract, introduction and conclusion of the articles.

4 Risks, mitigation and ethics

This chapter addresses the risks involved in this research project, and discusses potential ways these problems can be avoided if possible and mitigated if necessary. Additionally, it includes a short sketch of some ethical consideration which form the basis of the material that will be addressed in Re1.

Risks

As is always the case with any serious project, successful completion is never guaranteed. This section details the risks identified for this project, and the measures planned to mitigate the impact of these risks on the completion of this project.

Quantity and Quality of Available Relevant Literature

The literature review is a significant part of this project, forming the basis for the design cycle and prototype artefact. As such, a large part of this research is directly or indirectly dependend on the succesful acquisition of relevant literature. This potential problem is most serious in the context of Ri1, as this question both addresses a novel domain[1] and the literature study is projected to be the primary means of answering this question[2].

[1] For NLP, previous experience has been provided by the masters course.

[2] Compared to Ri3, where interviewing relevant stakeholders forms a second part of the methodology.

Mitigation

In order to mitigate this risk, should this problem arise, an alternative method of acquiring information is needed. For the specific case of Ri1, we anticipate that the HU Research Group would be able to provide expertise on the subject. As the experiences here are estimated to be relevant, but not entirely specific to the case at hand, this route would provide an adequate, if less satisfactory, means to answer the research question.

This risk is overall not very likely with a probability estitmated at 3/10; the impact to the quality of the end result is small due to adequate possibility of mitigation, estimated at 2/10.

Applicability of Literature Study Results on Artefact

The purpose of the literature review is to inspire the design cycle of this project, and as such must not only yield results, but also

provide enough basis to be translatable to a prototype solution. This is of course, not guaranteed.

Mitigation

If this scenario were to occur, the most prudent course of action would be to enlist help from colleagues via short sparring sessions. It should be noted that this risk also plays into a personal tendency to overthink, which must be avoided by committing to move forward even when stuck. As this part of the process is the most reliant on creativity, it is better to follow a path that turns out to be, in retrospect, suboptimal than to grind to a halt: the former can, at the very least, provide insights into why an approach was fruitless and possibly inspiration on how to move forward. I will depend on my daily supervisor to remind me of this, if necessary.

This risk is overall somewhat more likely but still manageable with a probability estitmated at 4/10; the impact to the quality of the end result is small due to adequate possibility of mitigation, which is also estimated at 2/10.

Absence of Desired Structural Features in the Twitter API

The desired approach to solving the main problem inspiring this research project depends on analysis of the structure of Twitter conversations. It is projected that this information can sufficiently be extracted from Twitter using the APIs. For this, access to the Twitter API is assumed — this could be considered a risk which has already been mitigated. Should there be any problem in extracting the desired information, this will produce delays in the realisation of the prototype artefact.

Mitigation

The situation described above allows for a number of different solutions, some of which can be set in motion concurrently and the choice of which impacts the expected quality of the result of this research: An attempt can be made to manually reconstruct the tree-like structure from tweets, but this approach is quite labour-intensive. An alternative would be to retrieve the structure from Twitter via web-scraping. This will also cost some time, but can more easily be outsourced to an HBO-student as it is mainly a technical challenge that could be relevant enough to warrant a project for undergraduates. The main risk then would be to start and finish such a project on relatively short notice. A third approach, using mock-data, can be viewed as the ultimate fallback: Though it is guaranteed to be possible and the time-investment necessary is relative to the amount of data required[3], the results are likely to be of significantly lower quality than the other two approaches using actual Twitter data. Mock-data would allow the testing of the prototypes performance in a quantative sense, but any qualitative results would be largely dependent on the mock-data and therefore would

[3] Inversely, the ammount of time available can thusly dictate the amount of mock-data that can be feasibly created.

not give an accurate reflection of real performance.

Given the three potential solutions mentioned above, it would be prudent to explore different avenues simultaneously: if the second route can indeed be outsourced, this would create the potential to generate high-quality data (indistinguishable from the data that would be extracted from the API) whilst time could concurrently be invested in manual labelling or generation of adequate mock-data. The latter choice would then depend on the time available.

This risk is probably the most likely with a probability estitmated at 5/10; the impact to the quality of the end result is also larger, and estimated at 6/10.

Ethical Considerations

As with any application of Artifical Intelligence, Machine Learning and related technology, it is paramount to consider the ethical implications of the project and its results in advance. In the case of this project, the prime ethical concerns can be subdivided into two broad classes: those regarding the methods used to gather and interpret data, and the ethical implications of the finished product.

Implications of data usage

In this project, public discourse will be utilised to extract sentiment information. Specifically, the prototype to be developed and evaluated will be targeted at tweets, i.e. short[4] text messages, potentially including images, links, emoji and hashtags. The tweets considered in the development of this prototype are all public: the original author has at minimum agreed to the public visibility and traceability of their input, or has actively intened this.

[4] 140 characters prior to 2017, 280 afterwards

On the other hand, the author might not have expected their input to be subjected to additional scrutiny, be it by human eye or artificial intelligence, and may not agree to their information being used in this manner. This is exacerbated by the fact that the original dataset supplied for this research consists of a series of tweets regarding a specific incident covered in Dutch media in May 2013, a period of time during which public awareness of online privacy was more limited than it is now.

In order to proceed, a few options are on the table:

- Using the data to train and evaluate the prototype, publishing the results without consideration of author consent (ignore this concern);
- using the data to train and evaluate the prototype, refraining from publication of individual tweets;
- using the data to train and evaluate the prototype, and asking for explicit permission before quoting individual tweets in publication;

- preemptively notifying authors that their tweets are used and for what purposes, removing tweets from the dataset on explicit objection (opt-out);
- preemptively asking consent (opt-in).

Of these options, the most favourable options from the researcher's point of view — balancing ethics and practical considerations — appears to be the second: using the data to train and evaluate the prototype and nothing else. In this scenario, if tweets are required in publication to outline (parts of) the process, mock data will be used instead. The main arguments for this approach are twofold: Firstly, the amount of data used per author is small, such that tweets by any individual author will be unlikely to produce identifiable artefacts within the final results. Secondly, the process of contacting people about tweets made eight years ago is not only unpractically intensive, but also likely to be more invasive than the data-usage in itself, specifically considering the first argument. Most people will not have any active memory on what they have posted online almost a decade ago, and bringing this to their attention will likely be met with apprehension both regarding the original subject matter and the persistence of data. Considering these arguments, the latter three options appear to suggest a cure worse than the disease, albeit on varying scales, and as such can be rejected. Of the remaing two options, the first option is simply to ignore the stated concern. Even though this concern is considered minor, the second option can be considered more ethical and thus should be selected.

Implications of the technology

The second, more pervasive concern is that of the product this project envisions and aims to expedite. Though at first the technology to allow authorities to better assess popular sentiment, de-escalate problematic conversations and mitigate results of incitement appears virteous, this notion entirely depends on the benevolence of the user. The same technology used to combat potential domestic terrorism could, with little to no alteration, be utilised by a more repressive government to monitor and control its population in increasingly invasive manners. Similarly, the user of such technology might not constitute a legitimate authority altogether, but a privately owned corporation whose intentions could be orthogonal to the rule of law. Finally, even with good intentions on the part of the user, it has been shown that statistical models trained on data are liable to internalise undesirable societal prejudice, potentially leading to a system prone to cast undue suspicions on vulnerable societal minorities.

Therfore, the consideration on how a finished product could be designed to work in an ethical manner requires further research. As such, this issue has been given its own research question:

As the question on how to guarantee techological progress is used ethically is, in a way, as old as technology itself, this domain has likely been studied intensively. As such, this question will be incorporated into the literature study as part of the relevance cycle.

5 Project Planning

This chapter provides an overview of the planning of the rest of this project, the major milestones and relevant dates. In accordance with the chosen methodology, the main work for this project has been subdivided into three cycles, each structured into a set of subquestions. For the rigour cycle, each subquestion is scoped by a definition of done, the fulfillment of which is considered the associated milestone.

For the design cycle, each subquestion is linked both to a rigour question, as well as to the design cycle as a whole. Milestones for the design cycle are informed by the development of the application as a whole, organised using *vertical slicing* as elaborated upon in Section 5.1. Due the nature of this structuring, milestones for the design cycle will consider the prototype as a whole. The exception to this is De3, where a separate evaluation of the benchmark has been included as a milestone.

For the relevance cycle, two major milestones correspond to the two separate subquestion therein: Firstly, defining the ethical dilemmata associated with this project and the exploitation of its results, and secondly the interpretation of these dilemmata according to the four frameworks presented in the elucidation of Re2.

Vertical Slices

The design cycle corresponds to the development and testing of the prototype, the process of which will be structured according to the Agile[BBvB+] paradigm utilising practices from GTD[All] and Scrum[TN], the latter insofar as these apply to individual projects. One concept adopted for this project from Agile in general and Scrum in particular is the *vertical slice*[RH]. Applied to the context of this research, a vertical slice refers to a cross-sectional slice through each of the component subquestions of this project, furthering both the literature review associated with each subquestion within the rigour cycle and implementing the results in the prototype as part of the design cycle. This process repeats in 2-week sprints, most of which will also contain part of the work required for the relevance cycle. Not every sprint has milestones for every subquestion, as the milestones correspond to more significant steps in the execution of the project.

Milestones

This section lists the proposed milestones in chronological order and provides a definition of done for each. Figure 5.1 illustrates the dependencies between the various milestones.

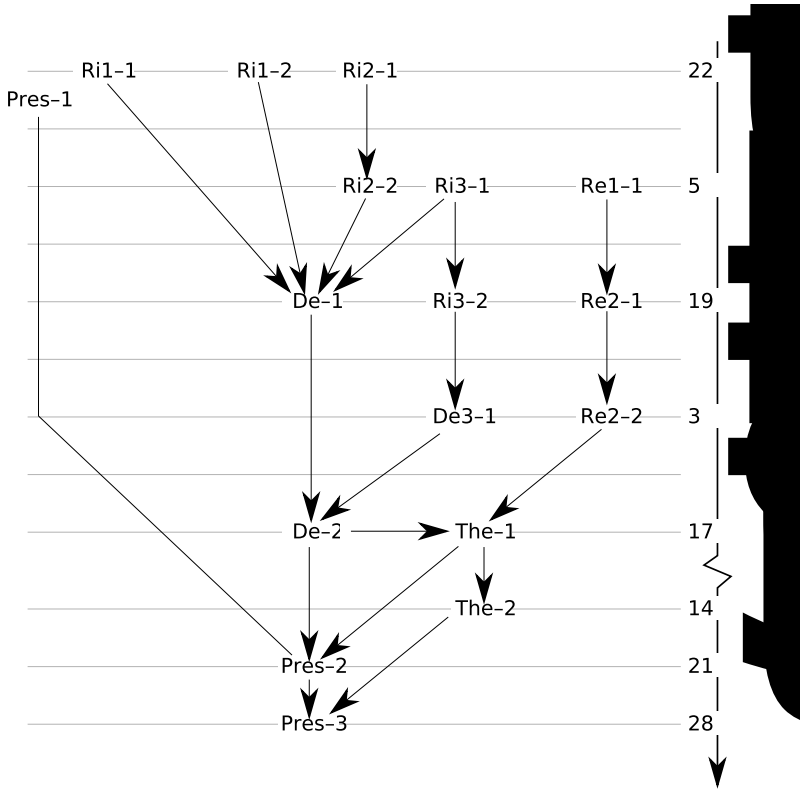

Figure 5.1: Dependency Graph

Milestone Ri1–1: Survey of Sentiment Analysis *Projected Delivery 2021-10-22*

Definition of Done:

- Finalised list of relevant papers on Sentiment Analysis;
- summary of found indicators and
- summary of applicability of indicators in computer code.

Milestone Ri1–2: Survey of Discourse Pragmatics *PD 2021-10-22*

Definition of Done:

- Finalised list of relevant papers on Discourse Pragmatics;
- summary of found indicators and
- summary of applicability of indicators in computer code.

Milestone Ri2–1: Strategies for Semantic Analysis of Tweets *PD 2021-10-22*

Definition of Done:

- Quantative overview of feasiblity of using syntactic information from tweets;
- initial recommendations on utilisation of available syntactic information to inform semantic analysis on individual tweets;
- updated overview of required functionality for Ri2–2.

Milestone Pres–1: Start Research Group Presentation *PD 2021-10-26*

Definition of Done:

- Presentation on research plans and methodology to the AI research group of the host organisation.

Milestone Ri2–2: Strategies for Structural Analysis of Conversations *PD 2021-11-05*

Definition of Done:

- Updated recommendations on utilisation of available syntactic information to inform semantic analysis on individual tweets and
- strategy for combining tweet-level analysis results in conversation trees;

Milestone Ri3–1: Benchmark Exploration *PD 2021-11-05*

Definition of Done:

- Overview of existing benchmarks in related problem spaces;
- suggestion(s) for applicable benchmark to this problem space and
- tagged data to determine applicability of benchmark concepts.

Milestone Re1–1: Overview of Ethical Dilemmata *PD 2021-11-05*

Definition of Done:

- Interview transcription on which ethical dilemmata apply to the development and training of the prototype and
- interview transcription on which ethical dilemmata apply to the autonomous exploitation of the prototype and derivative technology on real-world data.

Milestone De–1: Strawman Implementation *PD 2021-11-19*

Definition of Done:

- Prototype version containing at least stub implementations of all required functionality;
- evaluation of potential shortcomings of current prototype direction and
- informed decision on whether to continue this line of reasoning.

Milestone Ri3–2: Finalised Benchmark Strategy *PD 2021-11-19*

Definition of Done:

- Comparison of benchmarks based on tagged data from Ri3–1 and
- recommendation on benchmarking strategy to use on prototype and future development based on research.

Milestone Re2–1: Draft Ethical Recommendations *PD 2021-11-19*

Definition of Done:

- Overview of ethical frameworks as mentioned in Re2;
- draft application of frameworks on dilemmata identified in Re1–1.

Milestone De3–1: Evaluation of Benchmark *PD 2021-12-03*

Definition of Done:

- Implementation of benchmark on prototype output;
- documentation of results on prototype;
- documentation on design choices made in benchmark definition and
- documentation on how to apply benchmark on potential future projects.

Milestone Re2–2: Ethical Recommendations *PD 2021-12-03*

Definition of Done:

- Interview transcription on soundness of draft application as delivered in Re2–1.
- structured report on ethical recommendations based on these results.

Milestone De–2: Final Prototype *PD 2021-12-17*

Definition of Done:

- Demonstrable prototype including all required functionality;
- benchmarking of prototype according to results of De3–1;
- documentation on design choices made during prototype development and
- recommendations on how to improve / continue research direction or alternatively
- overview of lessons learned and recommendations on how to proceed.

Milestone The–1: Thesis Draft *PD 2021-12-17*

Definition of Done:

- Draft version of the final thesis for supervisor review.

Milestone The–2: Thesis *PD 2022-01-14*

Definition of Done:

- Final version of the thesis for presentation to the Utrecht University.

Milestone Pres–2: Final Research Group Presentation *PD 2022-01-21*

Definition of Done:

- Presentation on research execution and results to the AI research group of the host organisation.

Milestone Pres–3: Thesis Defence *PD 2022-01-28*

Definition of Done:

- Final presentation of research project and results to university supervision.

Bibliography

[AhS] Fareed Al-hindawi and D Saffah. Pragmatics and Discourse Analysis. 8:93–107.

[All] David Allen. *Getting Things Done. The Art of Stress-Free Productivity*. Penguin. ISBN-10: 0142000280 ISBN-13: 978-0142000281. URL: http://www.davidco.com/.

[Aus] John Langshaw Austin. *How to Do Things with Words*. Clarendon Press.

[AX] Felipe Almeida and Geraldo Xexéo. Word Embeddings: A Survey. Comment: 10 pages, 2 tables, 1 image. URL: http://arxiv.org/abs/1901.09069, arXiv:1901.09069.

[BB] Patrick Blackburn and Johan Bos. Computational Semantics. 18:27–45. arXiv:23918435.

[BBvB+] Kent Beck, Mike Beedle, Arie van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, Jon Kern, Brian Marick, Robert C. Martin, Steve Mellor, Ken Schwaber, Jeff Sutherland, and Dave Thomas. Manifesto for agile software development. URL: http://www.agilemanifesto.org/.

[BF] Luciano Barbosa and Junlan Feng. Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics.

[Bra] Tai-Danae Bradley. At the Interface of Algebra and Statistics. Comment: 135 pages, PhD thesis. URL: http://arxiv.org/abs/2004.05631, arXiv:2004.05631.

[BS] Adam Bermingham and Alan F. Smeaton. Classifying sentiment in microblogs: Is brevity an advantage? In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1833–1836. Association for Computing Machinery. doi:10.1145/1871437.1871741.

[BZ] Marco Baroni and Roberto Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1183–1193. Association for Computational Linguistics.

[Cha] Subashis Chakraborty. Curry-Howard-Lambek Correspondence. page 15.

[CSC] Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical Foundations for a Compositional Distributional Model of Meaning. Comment: to appear. URL: http://arxiv.org/abs/1003.4394, arXiv:1003.4394.

[DDF+] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. page 17. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[Duf] I.S. Duff. A survey of sparse matrix research. 65(4):500–535. doi:10.1109/PROC.1977.10514.

[DWP] David R. Dowty, Robert E. Wall, and Stanley Peters. A Higher-Order Type-Theoretic Language. In David R. Dowty, Robert E. Wall, and Stanley Peters, editors, *Introduction to Montague Semantics*, Studies in Linguistics and Philosophy, pages 83–111. Springer Netherlands. doi:10.1007/978-94-009-9065-4_4.

[Fel] Ronen Feldman. Techniques and applications for sentiment analysis. 56(4):82–89. doi:10.1145/2436256.2436274.

[Fet] Anita Fetzer. 2. Conceptualising discourse. In *Pragmatics of Discourse*, pages 35–62. De Gruyter Mouton. doi:10.1515/9783110214406-003.

[GBH] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. 150.

[GIMW] Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 581–586. Association for Computational Linguistics.

[Gri] H. P. Grice. Logic and Conversation. pages 41–58. doi:10.1163/9789004368811.

[GS] Robert E. Goodin and Michael Saward. Dog Whistles and Democratic Mandates. 76(4):471–476. doi:10.1111/j.1467-923X.2005.00708.x.

[Hev] Alan R Hevner. A Three Cycle View of Design Science Research. 19:7.

[HMPR] Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. Design Science in Information Systems Research. pages 75–106.

[JMa] Melissa N. P. Johnson and Ethan McLean. Discourse Analysis. In Audrey Kobayashi, editor, *International Encyclopedia of Human Geography (Second Edition)*, pages 377–383. Elsevier. doi:10.1016/B978-0-08-102295-5.10814-5.

[JMb] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice-Hall, Inc., 2 edition.

[JP] Paul Johannesson and Erik Perjons. *An Introduction to Design Science*. Springer International Publishing. doi:10.1007/978-3-319-10632-8.

[KH] Istvan Kecskes and Laurence Horn. Pragmatics, discourse and cognition. In Stephen R. Anderson, Jacques Moeschler, and Fabienne Reboul, editors, *The Language-Cognition Interface*, pages 353–375. Librairie Droz.

[KSP] Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. A Unified Sentence Space for Categorical Distributional-Compositional Semantics: Theory and Experiments. page 10.

[Kup] Julian Kupiec. Robust part-of-speech tagging using a hidden Markov model. 6(3):225–242. doi:10.1016/0885-2308(92)90019-Z.

[Lama] J. Lambek. Type Grammar Revisited. In Alain Lecomte, François Lamarche, and Guy Perrier, editors, *Logical Aspects of Computational Linguistics*, Lecture Notes in Computer Science, pages 1–27. Springer. doi:10.1007/3-540-48975-4_1.

[Lamb] Joachim Lambek. The Mathematics of Sentence Structure. 65(3):154–170. doi:10.1080/00029890.1958.11989160.

[Lan] J M Landsberg. Tensors: Geometry and Applications. page 83.

[Lea] Anti Defamation League. Hate on Display™ Hate Symbols Database. URL: https://www.adl.org/hate-symbols.

[Lev] Stephen C. Levinson. *Pragmatics*. Cambridge University Press.

[Lyn] Marc Lynch. After the Arab Spring: How the Media Trashed the Transitions. 26(4):90–99. doi:10.1353/jod.2015.0070.

[Mar12] Angel R. Martinez. Part-of-speech tagging. 4(1):107–113, January/February 2012. doi:10.1002/wics.195.

[MB] Subhabrata Mukherjee and Pushpak Bhattacharyya. Sentiment Analysis in Twitter with Lightweight Discourse Analysis.

[MW] Michael Moortgat and Gijs Wijnholds. Lexical and Derivational Meaning in Vector-Based Models of Relativisation. Comment: 10 page version to appear in Proceedings Amsterdam Colloquium, updated with appendix. URL: http://arxiv.org/abs/1711.11513, arXiv:1711.11513.

[NGM+] Faria Nazir, Mustansar Ali Ghazanfar, Muazzam Maqsood, Farhan Aadil, Seungmin Rho, and Irfan Mehmood. Social media signal detection using tweets volume, hashtag, and sentiment analysis. 78(3):3553–3586. doi:10.1007/s11042-018-6437-z.

[Olu] Emmanuel Olang' Oluoch. Sentiment analysis model for detection of radicalization on twitter. URL: https://su-plus.strathmore.edu/handle/11071/12100.

[Onr] Onrustig begin avondklok, corona teststraat in brand gestoken op Urk, ME opgeroepen in Stein. URL: https://tinyurl.com/audps9kc.

[Par] Barbara H. Partee. Semantics. In *The MIT Encyclopedia of the Cognitive Sciences*. The MIT Press. URL: http://people.umass.edu/partee/docs/Partee%20MITECS%20Semantics.pdf.

[Rea] Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48. Association for Computational Linguistics.

[Res] Research Group Artificial Intelligence | Hogeschool Utrecht. URL: https://www.internationalhu.com/research/artificial-intelligence.

[RH] Ian Michael Ratner and Jack Harvey. Vertical Slicing: Smaller is Better. In *2011 Agile Conference*, pages 240–245. doi:10.1109/AGILE.2011.46.

[Sch] Hinrich Schutze. Automatic Word Sense Discrimination. 24(1):28.

[Sea] John R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

[SKB] Mehrnoosh Sadrzadeh, Dimitri Kartsaklis, and Esma Balkır. Sentence entailment in compositional distributional semantics. 82(4):189–218. doi:10.1007/s10472-017-9570-x.

[ST] Erhan Sezerer and Selma Tekir. A Survey On Neural Word Embeddings. Comment: 33 pages, 2 figures, 8 tables. URL: http://arxiv.org/abs/2110.01804, arXiv:2110.01804.

[Tim] Wibke K. Timmermann. *Incitement in International Law*. Routledge. doi:10.4324/9781315769516.

[TN] Hirotaka Takeuchi and Ikujiro Nonaka. The new new product development game. URL: http://apln-richmond.pbwiki.com/f/New%20New%20Prod%20Devel%20Game.pdf.

[Tse] Alexander Tsesis. Social Media Accountability for Terrorist Propaganda. 86:605. URL: https://heinonline.org/HOL/Page?handle=hein.journals/flr86&id=623&div=&collection=.

[Twi] Twitter: Monthly active users worldwide. URL: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

[VHLH] Cynthia Van Hee, Els Lefever, and Véronique Hoste. Exploring the fine-grained analysis and automatic detection of irony on Twitter. 52(3):707–731. doi:10.1007/s10579-018-9414-2.

[War] Dan R. Warren. *If It Takes All Summer: Martin Luther King, the KKK, and States' Rights in St. Augustine, 1964*. University of Alabama Press.

[Wei] Edda Weigand. *Language as Dialogue: From Rules to Principles of Probability*. John Benjamins Publishing.

[Wor] In a world becoming more polarized, Europe must stay united. URL: https://www.weforum.org/agenda/2019/01/in-a-world-becoming-more-polarized-europe-must-stay-united/.

[WS] W. John Wilbur and Karl Sirotkin. The automatic identification of stop words. 18(1):45–55. doi:10.1177/016555159201800106.

ToDo

- [3] Huib: terugwijzen op wat je eerder geschreven hebt in introductie: bedoel je hier met een letterlijke terugverwijzing of door voorbeelden te herhalen? Is het OK om vanuit deze sectie naar elders te verwijzen of moet dit deel opzichzelf staan (dacht ik namelijk)?
- [3] Huib: nu zonder exploration, maar gaat het zo niet meer lijken op een project opdracht dan een academische exercitie? Ik zoek een beetje naar de sweet spot tussen beide uitersten.
- [11] Huib: ik heb hier nog wat meer context toegevoegd mbt formal semantics; hiermee wordt de voornaamste tekortkoming van het systeem ook beter uitgelicht
- [17] Huib: is dit waar je op doelde?
- [18] Huib: sectie hierboven is uitgebreid, hypothese valt hier logisch maar goed om deze in BG te noemen?
- [20] Huib: deze sectie en die daaronder zijn (met grotendeels dezelfde inhoud) geherstructureerd - leest dit wat jou betreft beter zo? De citaties zijn naar voren gehaald, de voetnoot over sentiment versus intent daaronder. Daarnaast ter overzicht de voorbeelden ook in een tabel gezet