BRIAN VAN DER BIJL

TWEET, REACTION, ACTION

Abstract

ToDo: abstract

Board of Examination

This thesis was submitted on April 22, 2024. ToDo: Input correct date when this is known

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: Lasha ABZIANIDZE PhD

Latest update

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

Contents

1 Introduction 1

2 Background 5

3 Research Methodology 11

4 Model Design and Architecture 17

5 Data Acquisition and Analysis 29

6 Application Concerns 39

7 Conclusion 41

8 Discussion 43

Bibliography 45

1 Introduction

On the 6th of January 2021, an angry mob attacked the United States Capitol Building in an attempt to overturn the defeat of then-current president Donald Trump by disrupting and threatening the joint session of Congress convened to formally count the Electoral College votes. This event has been described as a momentous event in US history, marking the first occupation of a government building by insurrectionists in recent history[1].

[1] Exact dates are difficult to establish without establishing more exact definitions. The most recent similar event would be either the 1954 shooting by Puerto Rican Nationalists[**?**], or the coup d'etat of 1898[**?**].

Insurrections and civil unrest have always been challenging societal problems to accurately predict, with relevant communication taking place on both public and private channels, thereby providing all parties involved with a fragmented picture of current threats and the overall situation. Private channel communication is hard to track and outside the scope of this research. Public communication will generally be less sensitive and informative, yet its inclusion is still necessary to achieve the critical mass required to have any societal impact. The problem in monitoring this public communication resides not in the information being hidden away or inaccessible, but instead is hidden in plain sight, in the sheer volume of non-relevant communication sharing the same channels.

ToDo: rewrite this section to continue from above 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 created a world where 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].

ToDo: later 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.

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[2] 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.

[2] 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

The ultimate desired outcome of the larger research 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 length 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.

As will become clear in the following sections, the goals of this project are too broad in scope to address within the limited amount of time and other resources available for this thesis project. As such, the scope of the portion described in this document has been adjusted at several points during the execution of the project. The following chapter will address the theoretical background, which will lead to a more substantiated presentation of the final scoping of the project in Chapter 3.

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.

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. It will subsequently explore some related work on analysis of Twitter data, which mainly focuses on sentiment analysis, and briefly discuss why this approach falls short for the matter of interest of this research.

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 4.1.3. 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.

Additional Definitions

The following terms are used in the following chapter discussing the research methodology. These terms generally refer to entire areas of study, the relevant parts of which will be further explored when they arise. This section provides a rough definition on some unfamiliar terms so that these may be used in formulating research questions and methodological discussion.

Semantics Within the context of linguistics and philosophy, semantics refers to the study of meaning and truth assigned to words and sentences.

Discourse Pragmatics The two related areas of study *Discourse Analysis* and *Pragmatics* are often used interchangeably to refer to the study of meaning of natural language beyond literal semantic meaning by regarding context and the way multiple sentences may be combined to evoke meaning beside their individual contents.
ToDo: Add some references for basic definitions - SEP

Significance and Related Work

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-Ibánez, 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[2]:

[2]

	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.

Research Question

In order to structure and scope this project, a main research question has been formulated to translate the relevant (as determined by the scope of this sub-project) parts of the eventual desired outcome — a prototype AI able to determine sentiment — into a research oriented project. This project is aimed at providing not only a prototype model for capturing intent, but also demonstrating its adequacy and the constraints placed on its responsible real-world application by the nature of the data used to feed the model. Given this, the main research question is posed as follows:

How can a sufficiently complete model be designed to capture intent from series of short informal text messages with minimal redundancy, in such a way as to be applicable to responsibly predict incitement based on their conversational structure?

Subquestions

This research follows three separate but codependent phases, each principally attached to one of the cycles as described by Hevner: (i) Collecting domain knowledge and formulating a model artefact [rigour/design cycle], (ii) Verification of the artefact [rigour cycle], and (iii) Application context and concerns [relevance cycle] The rest of this section will discuss each cycle in more detail and formulate the relevant subquestions.

Model Design and Architecture

RI1: "WHAT INDICATORS FROM DISCOURSE PRAGMATICS AND SENTIMENT ANALYSIS ARE APPLICABLE IN THE DETECTION OF INCITEMENT AND CATEGORISATION OF INTENT?"

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.

DE1: "HOW CAN THESE INDICATORS BE TRANSLATED INTO A MODEL ABLE TO CAPTURE INTENT? "

This question is central to the design cycle of this research, as its results describe the main artefact produced in this research project.

The answers to these questions, as well as a more detailed description of the factors forcing the inclusion of this extra question and the considerations taken into account in the process of answering it will be described in Chapter 4.

Data Acquisition and Analysis

The next set of questions are intended to verify the results of the main design cycle using rigour-based techniques of statistical analysis. These are meant to ensure the coverage and lack of redundancy required in the main research question.

RI2: "WHAT IS THE COVERAGE OF THE ACQUIRED DATASET WITH REGARD TO THE ENTIRE POSSIBLE SPACE OF INTENTIONS AS DETERMINED BY THE CHOSEN INDICATORS?"

Here, the aim is to identify clusters within the space (again, using algorithms such as t-SNE) to determine the completeness of the data represented. The goal is to identify empty regions, attempt to explain why these gaps show up, and whether these represent gaps in the degree that the dataset represents relevant Twitter discourse. ToDo: This question should concern more with the model than with the specific dataset used to analyse the model. It is a prerequisite to answer the following question, although the way it is answered depends more on the specific data than the next question.

RI3: "WHAT IS THE LEVEL OF RELEVANCE OF THE INDICATORS IDENTIFIED IN (RI1) IN CAPTURING INTENT INFORMATION OF TWEETS AND PROVIDING EXPLAINABILITY?"

This question addresses the choices made in the design of the intent-space and its chosen basis vectors (axes). This question aims to determine how well these bases where chosen, and how the computational usability of the dataset could (if and when the usage warrants it) be increased without sacrificing the content present. To answer this question, explanatory factor analysis and the t-SNE cluster-preserving dimensionality-reduction algorithm are utilised. The goal is to determine to what extend dimensionality reduction

can be applied without damaging the data contained in the set, and preferably whilst maintaining explainable labels for the remaining axes.

These questions together aim to verify the choices made in the first phase of this project by subjecting the assumptions to real-world data. The answers to these questions, as well as a more detailed description of the factors forcing the inclusion of this extra question and the considerations taken into account in the process of answering it will be described in Chapter 5.

Application Concerns

The last two questions deal with the implications of the results of this research, which precautions should be considered in its application, and how to interpret results when using the model in prediction tasks.

RE1: "WHAT ETHICAL PITFALLS CAN BE ANTICIPATED IN THE APPLICATION OF THE DESIGN ARTEFACT?"

This question aims to place the model designed in this research project in the broader context of its application. As with all AI and data based projects, ethical considerations form a major concern informing how the results should be interpreted and applied to real-world situations.

RE2: "TO WHAT DEGREE CAN BIAS BE IDENTIFIED IN THE DATASET, AND HOW HAS THIS BEEN INTRODUCED BY THE TAGGING PROCESS USED IN ITS CONSTRUCTION?"

This question is twofold, as the second part only makes sense to ask if substantial bias can be shown to be present. The main goal of this question is to determine how biased (and biased how) the tags used to determine the embeddings of tweets within the space are. The approach here is to analyse the (co)variances on each axis within groups and compare this to the (co)variances between groups using methods such as Student's t-test. This question aims to identify and isolate these biases insofar as these can be correlated to known differences in the backgrounds of the taggers, in order to determine whether (and in what way) factors such as educational background (social vs applied exact science) colour the resulting tags and whether the average that determines the final embeddings should be weighed to better reflect a more general or specialised population.

These questions together aim to address the responsibility constraint present in the main research question. The answers to these questions, as well as a more detailed description of the factors forcing the inclusion of this extra question and the considerations taken into account in the process of answering it will be described in Chapter 6.

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 **??**.

4 Model Design and Architecture

This chapter details the results of the first hybrid rigour / design cycle. It elaborates the research done informing the choices made in the design of the artefact for this project. It will answer the first two research questions, Ri1 and De1.

Theoretical Review

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*. The former of these is explored in the following section, as the latter ultimately did not appear in the solutions explored in this research project.

Distributional Semantics

Translating the concept of semantics to computers and AI is challenging: The language used in general purpose computers is one of numbers[1], 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.

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 4.1 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

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

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.

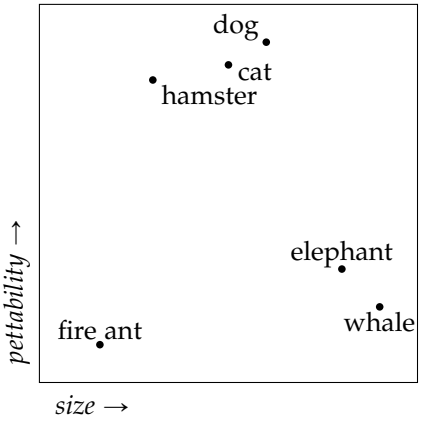

Figure 4.1: An example 2D word embedding space

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 4.1, 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 embeddings 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"[2], 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.

The field of distributional semantics and the concept of word embeddings deriving from it form the basis of many forms of tex-

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

tual 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].

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[3]). This flow could play a part in determining the intended effect of a series of statements.

[3] 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[4] or including a dessert, one breaks this maxim.

[4] 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[5] 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

[5] 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.

is a publicly visible tweet or thread[6] 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?

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

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 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[7]. 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[8] multiplied by the probability that any randomly chosen verb would turn out to be "consuming"[9]. 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.

Tagging Tweets and Conversation Trees

We tag each tweet with an initial intent/sentiment vector based solely on the tweet itself. Afterwards, we can estimate tags for the reply relations going back up. We view a conversation as a tree-shaped graph. Vertices corresepond to tweets, edges to reply relations.

Combining intent values on tweets is done in three steps. First, the tweets are analysed in isolation, so without consideration of the tweets it replies to, or the tweets that follow. This gives an initial

[7] 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.

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

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

label to each vertex within a conversational tree. After this, we will traverse each path in the tree from top to bottom, which we call the forward search step. Here, the relation between intent of a parent tweet and its child tweet are analysed. This step yields labels for the edges of the tree. Finally, we combine the results bottom to top — the backward consolidation step. Here, the intent given to child-nodes and the δ-intent given to vertices combine to form an updated valuation for each non-leaf node in a conversation tree. The reasoning behind this step is that if a tweet, which in itself cannot be classified as inciteful, nonetheless solicits replies that are largely recognised as inciteful, the parent tweet should be flagged as interesting as new inciteful replies are more likely to appear in the future.

NLP / SA on Dutch Tweets

State of the art seems to be mostly centred around the BERT model. For the initial sentiment training, we require a model that

- Understands Dutch
- Outputs sentiment/intent vectors

which does not appear to exist. Transfer learning allows us to use pre-trained models and fine-tune for specific languages and use cases. I'm still figuring that out.

It appears Bert accepts labels as potential output, cannot find anything on using vector embeddings. Labels can be translated to 1-hot encoded vectors in a sentiment/intent space which would allow further processing. — Further reading: this is due to softmax, let's see if we can disable that!
ToDo: Need to find tagged data

Note on reuse of existing models in other languages[dVN].

Intent Embeddings

In order to tag data for training BERT, we need a standard for how an intent vector space looks. The suggested format is to use $\mathbb{R}^{12}$, specifically $[0,1]^{12}$ for manual tagging of data. The resulting training data vectors will be normalised per participant and then averaged between participants before being used for training BERT. The structure of the intent space for this tagging is hand-crafted, as the initial data is entered by hand instead of learned automatically from context. Table 4.1 details the meaning assigned to each basis vector. This might be supplemented by other tweet statistics such as number of likes/retweets and/or author statistics such as number of followers. Main problem with this is that these values are subject to change and add additional requirements on keeping the local data storage up to date. Initial trial will leave these out but this might be added in the future. It might also make more sense

to include these statistics in a later stage, when viewing tweets in context instead of in isolation.

e_0	**Assertiveness**	As described by Searle[Sea]
e_1	**Directiveness**	As described by Searle[Sea]
e_2	**Commissiveness**	As described by Searle[Sea]
e_3	**Expressiveness**	As described by Searle[Sea]
e_4	**Declarativeness**	As described by Searle[Sea]
e_5	**Naive Sentiment**	Positive wording as in regular SA
e_6	**Sadness**	Tweet expresses sadness / is a call for emotional support
e_7	**Hostility towards Target**	(person, group, situation)
e_8	**Hostility towards Reader**	(general public reading the tweet)
e_9	**Call for Action**	Expresses specific incitement towards (violent) action
e_{10}	**Use of coded language**	Use of language indecipherable to out-group
e_{11}	**Use of euphemism**	Known dog whistles or metaphor
e_{12}	**Use of sarcasm**	
e_{13}	**Question**	
e_{14}	**Contradiction**	Tweet appears to directly contradict another tweet or news item

Table 4.1: Proposed basis vectors for intent embedding space

Tweet Dataset Acquisition

Single search query in period of one hour. Coronapas? Then follow each thread up in case of replies. Finally, download entire tree for each tweet.

5 Data Acquisition and Analysis

ToDo: Meer introducerend (previously, on ..) - we hebben nu een model van 15 dimensies, belangrijke termen /getallen herhalen
This chapter details the approach and results of the data acquisition step of this project. The chapter is split in five sections, the first of which describes the acquisition process and the choices made therin. The next three sections endeavour to answer the subquestions Ri2 and Ri3. Finally, the last section will conclude this chapter.

Data Acquisition approach

ToDo: doel (data nodig), probleem (geen data), oplossing(verzamelen)
The initial scoping of this project assumed the availability of tagged data from a previous research project. This data soon turned out to be untagged and missing the structure required for the intended strutcural analysis. As of such, the data available was determined to be insufficient, which added the requirement of collecting a set of Twitter-conversations, including text and structure, and subsequently tagging each tweet using the dimensions described in chapter 4.
ToDo: Hier ongeveer: plaatje visueel proces Twitter -> Data trees - resultaat beschrijven in plaats van focus op proces. Data nodig, was er niet, hoe is dit verzameld?
ToDo: doel (threads), probleem (twitter), oplossing(bottom up algo) To build the initial dataset, ToDo: 856 - getal eerder noemen, later terugkomen tweets were read using the Twitter API and stored locally. Threads are accessed using the API and include a field referencing $0 - 1$ parents. Threads are then retrieved from the bottom up ToDo: Why - top down sensible but impossible by repeatedly following the parent reference of a tweet. Building a conversation tree down from an initial tweet is turned out to be more convoluted. The only viable approach to this is as follows:

(i) ToDo: pseudocode

(ii) Search all tweets addressed to the author of an original tweet.

(iii) For each tweet found:

  (a) Check whether the parent reference matches the id of the original tweet.

(b) If so, add the tweet to the tree structure.

(c) If not, the tweet remains in consideration as it might be a response to a response.

(iv) After all tweets have been tried, repeat to place grand-children, etc, until nothing changes.

(v) Now, repeat the process for each newly found node, until no new tweets are found.

ToDo: Could be -> discussion / future? THIS requires potentially filtering an repeatedly re-scraping a lot of tweets, especially for older tweets[1]. Additionally, it became clear during testing that Twitter's API, specifically the search part, did not produce reliable results: Searching via the API and via the web-interface did not produce the same results, with the API apparently omitting large amounts of information. As the process of searching for replies depends on recursively searching replies and each failure compounds, the process of automatic tweet mining was abandoned in favour of an approach including some manual labour.

[1] This could be mitigated by limiting reply-searching to replies made no more than n days after the original. THIS seems reasonable, as a conversation progressing at a slower pace is less likely to be a heated discussion.

In this final approach, paid volunteers were asked to traverse Twitter using the web interface and note down URLs for tweets regarding specific subjects. They were instructed to focus on longer threads and/or broader conversations, noting the URLs for the last tweet in a thread (after which the thead upwards could be accessed using the API). THIS approach potentially introduced some a priori bias ToDo: this will be examined in sec X in the dataset, as singular tweets were ignored and the selection was limited to tweets recommended by the Twitter algorithm that caught the attention of the volunteers, all within the context of specific topics. One obvious result of this is that the resulting dataset mostly contains tweets within a select topic of conversation with a likely overrepresentation of divisive ToDo: betere term of introduceren - wat is divisive tweets. THIS is of course by design and appropriate for the context of the intended usage, but should nonetheless be noted when considering its applicability to more diverse domains.

The topics considered for the first batch were the Dutch approach to the COVID-19 crisis, specifically the discussion surronding the "Coronapas" ToDo: citation - nrc/nu.nl/whatevers as a means of reopening Dutch society at the tail-end of the crisis. Later, a smaller second set has been constructed using the same procedures, this time focusing on the Western response to the Russian invasion of Ukraine ToDo: citation krant . By focussing on the Dutch language domain, most tweets in this set specifically consider the Dutch reactions.

A second artefact of this data-gathering approach is the limited quantity of data. Combining the results of this step with the few larger clusters of tweets mined during the initial attempt resulted in a dataset of 500 tweets, comprising 19 clusters of average size 26.316. The second set adds 356 tweets.

Tagging

The subsequent step was tagging the data using intent-embeddings as described above. In order to accomplish this, 6 paid participants have been asked to tag each tweet on the 15 established parameters. THIS was done using a web-interface which showed the participant a single tweet at a time, and recorded the ratings in a database. The tweets were shown in a semi-random ordering, with the same tweet never appearing twice ToDo: doel (uniform ipv normaal verdeeld), probleem, oplossing for the same participants and tweets with less tags shown with increased probability[2]. In addition to the paid participants, individuals have been asked to participate by tagging a small number of tweets on a voluntary basis, again being served semi-random tweets favouring those that had been rated the least.

For the following sections, some nomenclature is introduced. Each tweet is tagged by each participant at most once. THIS mapping of $(\text{tweet}, \text{participant}) \to [0,1]^{15}$ will be referred to as a *tag*. The aggregation of all tags pertaining to the same tweet, usually via plain average unless otherwise specified, is referred to as that tweet's *embedding*. Finally, similarly aggregating the tweets per participant yields that participant's *profile*. ToDo: definities - format ToDo: 15 magic number?

[2] Specifically, the probability of a tweet i appearing is $P(i) = 1 - \frac{c(i)}{499t}$ with $t = \sum_{i=0}^{500} c(i)$ and $c : \mathbb{N} \to \mathbb{N}$ mapping a tweet to the number of tags for that tweet.

Dataset Coverage

ToDo: coverage is het begin, waarom?! data quality, coverage is een aspect

Given the existence of a tagged dataset, the next logical step ToDo: WHY? is to provide a basic understanding of how the data is structured, and whether it can be considered sufficiently representative of the norms and sensibilities we wish to extract from the data and use to train an artificial intelligence based solution.

gaten waarschijnlijk want: Kleine dataset en gezocht obv query, meer niet

ToDo: Discussie The number of tweets available at this point is, as described, lower than what was initially planned and desired. THIS will limit the effectiveness when using this data to train machine-learning algorithms, especially large and complicated models such as BERT[TOD]. Despite this limitations, we believe that the data gathered so far is enough to perform some level of exploratory analysis.

ToDo: gaten -> missen -> betrouwbaarheid model A major concern is the degree of coverage the dataset provides: given that 856 tweets are embedded in a 15 dimensional space, it is unlikely (why) that each point within the unit 15-cube of possible embeddings is sufficiently (wat is sufficient?) close to a known tweet, which implies the existence of holes within the dataset. These holes may be either representative of unlikely combinations of indicators, in which case they represent holes in the manifold of possible tweets,

or they may be introduced by the selection process and thus indicate potential shortcomings of the dataset as a projection of the possible-tweet manifold. In the latter case, these ommissions can be further subdivided into those that are desirable given the intended use of the dataset (for example, pictures of pets, which are unlikely to contribute to the emergence of incitement) and those that require attention.
ToDo: confident -> we weten niets, mogelijk niet uniform verdeeld, willen we weten ToDo: inital test: mean/var per dimensie, makkelijk aantoonbaar niet uniform? In order to assess the way holes within the dataset are justifyable, the first step is to discover where such holes exist. We can confidently hypothesise the existence of holes as their absence would mean that the data is distributed evenly over the 15 dimensional space, which is statistically exceedingly unlikely. In order to find the distribution of these holes, we can first look for the inverse and identify which clusters exist. Unfortunately, due to the high dimensionality of the space most naive clustering algorithms like k-means and Gaussian mixture models fall short as their implied usage of distance metrics yields large Euclidean distances even for points relatively close to eachother[TOD]. ToDo: plaatje ToDo: On the other hand, dim red heeft dit prob.ee m niet ToDo: an example is ... , we used this ToDo: wat is tsne, wat hebben we er aan? To address this, a form of dimensionality reduction is required which will preserve cluster distinctions. To this end, the t-Distributed Stochastic Neighbour Embedding[vdMH] (t-SNE) has been used on the embeddings of the initial selection of 866 tweets of the dataset (Figure 5.1). The results clearly suggest the existence of clusters, an observation that could also inform future work on learning conversational flow, but unfortunately the loss of information makes it infeasible to reconstruct holes from these results. ToDo: Hieruit weten we dat er clusters zijn, dingen dicher bij elkaar dus ook verder uit elkaar (niet uniform) dus holes.

Monte Carlo for finding negative space

A possible approach to answer this (which?) question is to apply a Monte Carlo based approach and randomly sample the space of possibile tweet embeddings and calculate the distance to the nearest point. Lemley, Jagodzinski and Andonie suggest such an approach[LJA]. In their research, they specifically target detecting axis-aligned hyper-rectangles, which is prudent from a representation standpoint but limiting within the context of intent-spaces where partial correlation between axis might make it beneficial to support different shapes and orientations of holes.

Ideally, we would like to determine the existence of *negative clusters*, i.e. point-clouds of points outside of the dataset having the property of being close together and not being close to points within the dataset. These negative clusters could then be analysed

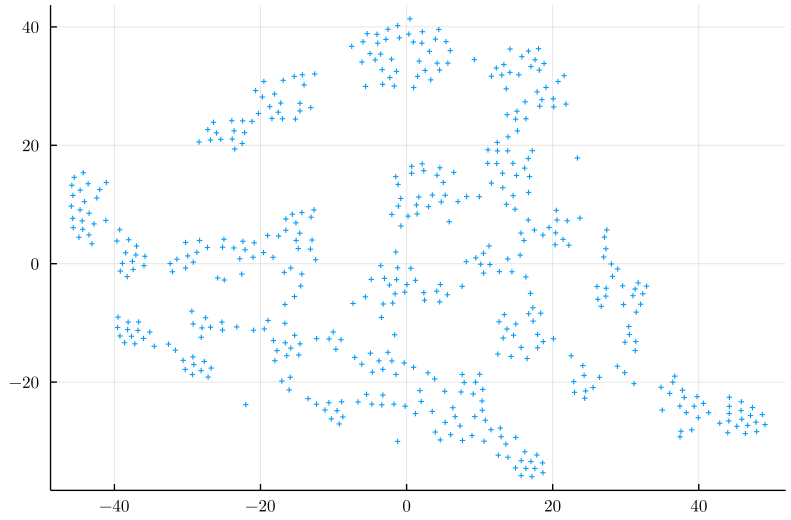

Figure 5.1: 2 dimensional cluster preserving image of the set of embeddings fitted using t-SNE

to get a metric of size via the hyper-volume of the complex hull[3]. Relatively small clusters likely result from the small number of datapoints relative to the set of potential embeddings, larger clusters warrant closer inspection on what hypothetical tweet would project onto the centroid of the cluster as these represent combinations of values not accounted for within the dataset.

The main problem with this approach is that in order too work with negative clusters, we need to generate a finite random approximation of the entire possible space. Given the 15 dimensions, even a sparsely saturated set of embeddings quickly becomes impractically large. Using Monte Carlo, we can generate points as far as it remains feasible, but the resulting set is unlikely to adequately represent the entire space. For example, if one were to discretise the space by dividing each axes into 5 segments, this would result in $5^{15} = 30517578125$ subvolumes. In the hypothetical scenario that the randomly generated points would be distributed in a perfectly uniform manner, generating one 64-bit floating point vector for each subvolume would require approximately 3.33 TiB of memory/storage, without considering overhead or any metrics beyond the location of each point. As this exceeds what is feasible within the scope of this project, multiple runs would need to be performed and aggregated, with consideration on how to guarantee results which are significant.

ToDo: concl: more practical approach required

[3] Any further reference to the size of negative clusters should be taken to mean this volume, not the number of points in the cluster.

Negative space

The main insight on the shortcomings of the application of clustering algortithms on this problem is that we are interested not in groupings of datapoints, but in large areas of their absence. In essence, we want to apply clustering to all the points not in the dataset. THIS is clearly infeasible, as this space is in theory un-

countably infinite[4]. Even discretising the space using a fixed number of segments for each axis and generating one 64-bit floating point vector for each subvolume would require large ammounts of resources. A low-end reasonable goal of 5 segments per axis would entail $5^{15} = 30517578125$ subvolumes, requiring approximately 3.33 TiB of memory/storage for 15 64-bit floating point numbers — without considering overhead or any metrics beyond the location of each point. THIS still greatly excedeeds what is feasible within the scope of this project.

One option to approximate the desired negative space up to a given level of coverage would be to use a Monte Carlo approach to generate random samples, but this would still require a significant amount of data per point as well as a large number of points before even a perfectly uniform distribution would be sufficiently saturated.

ToDo: plaatjes / pseudocode! Ultimately, we opted for an alternative using binary partition. The entire space is subdivided into 2^{15} subvolumes, with a central point added for each subvolume and evaluated. We calulate the minimal distance to the set of existing points, and keep n points with the largest distance. THIS process can be performed in a reasonable amount of time, but in itself lacks granularity. By repeating this proces recursively, considering each subvolume in turn and dividing this into 2^{15} parts, we can iteratively refine the cast of the set of embeddings. The intermediate results are, in contrast to the randomly generated points, cheap to store, as we can define a bijection of each partition to 32-bit integers. After each recursive generation, the best found points are kept.

TODO - how many are kept / what does that matter.

Determining a threshold for closeness to existing points

- Preliminary runs with fractions, time and storage - Stats for trials, average

Validity of chosen indicators

In the design phase of this research project, 15 indicators were proposed to structure the tagging of tweets and to use as a basis for predicting conversational flow. One prime concern is to verify whether these 15 indicators behave independently, or whether the correlation between certain subsets of these indicators allow for some dimensionality reduction. Using Principal Component Analysis (PCA), it can be shown that the set of 15 dimensions can be reduced to 10 whilst retaining 95% of the total variance accounted for ToDo: Update , or 6 whilst retaining 80%. These new sets of dimensions are determined using an optimisation algorithm mainly intended for compression, and yield a more compact space at the cost of explainability. The 15 axes initially chosen have been shown

[4] In practice, the space is finite, as each of the 15 dimensions is discretised as a 64-bit floating point value. Roughly 37.5% ToDo: check of floating point representable numbers fall within the considered interval $[0, 1]$ yielding approximately $(2^{64} * 0.375)^{15} \approx 4 \cdot 10^{282}$ points for consideration.

by the tagging process to be sufficiently intuitive for human participants to estimate, which allows more easy verification of predictions made based on this dataset. Reorienting these dimensions and discarding some effectively generate a new basis for the embedding space, without obvious labels for each of the axes. Instead, *Explanatory Factor Analysis* is applied to make more informed decisions about the combining of axes to reduce dimensionality. This approach allows for a more bespoke combining of indicators where latent factors are first identified and then associated to existing indicators, after which the process is repeated until all (or a specific amount of) variance is accounted for in the chosen factors.

We first plot the correlation and covariance matrices of the entire set of (at this point 1553) tags to see whether one or more axes are obviously redundant and can be manually removed or combined. Based on the current state of the dataset, the highest correlation is between action-call and declarative, which makes sense. The correlation between these two is (at the time of writing) 0.515, which suggests some dependence between at least these axes.

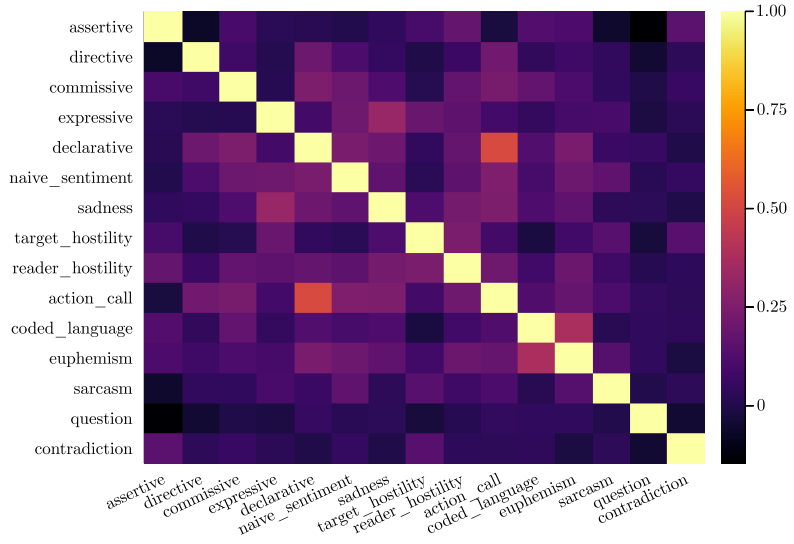

Figure 5.2: Correlation matrix for all tags

We can repeat the same procedure after aggregating the tweets into embeddings, as this is the form in which the final data will exist. The results of this are shown in Figure 5.3. After aggregation, correlations are visibly much higher suggesting some degree of variance in the different tags for each tweet, suggesting some tagger bias may indeed be represented in the data. Figure 5.4 shows the covariance matrix for the set of embeddings, which will be used in Exploratory Factor Analysis.

ToDo: Work in progress

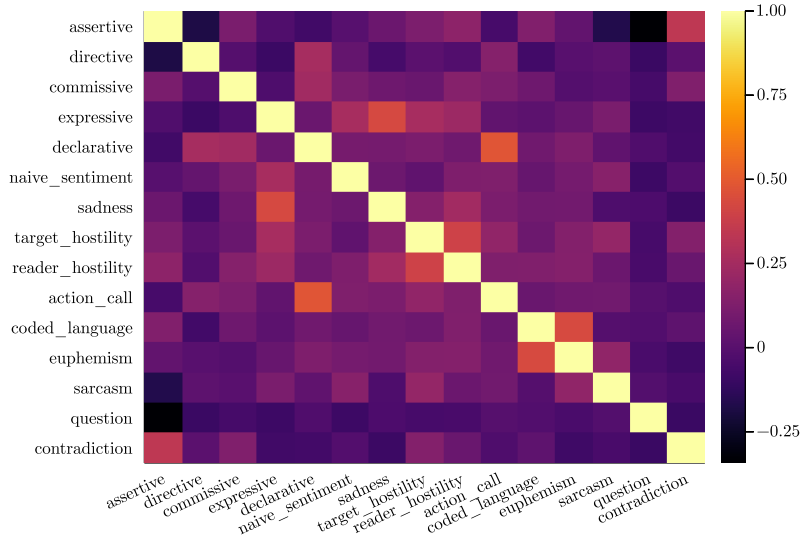

Figure 5.3: Correlation matrix for all embeddings

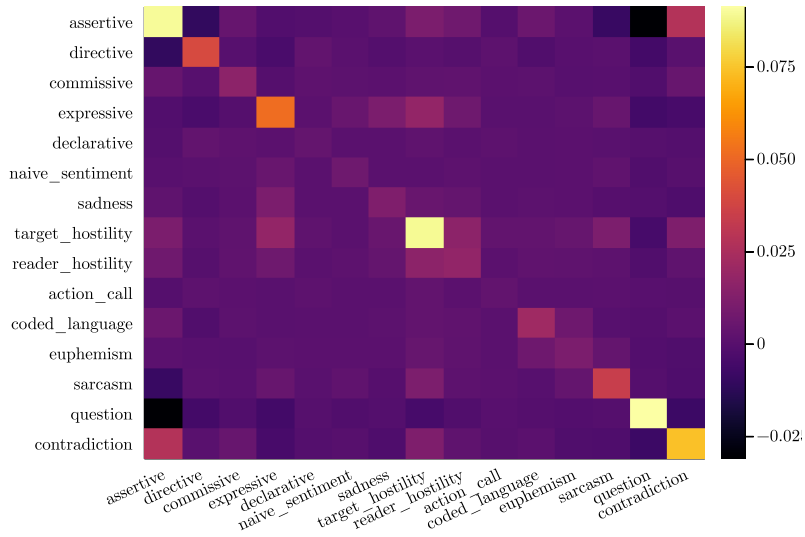

Figure 5.4: Covariance matrix for all embeddings

Biases

During the tagging process, each participant was associated with a participant id to ensure no repeats were being shown. The partial tags of the dataset, as provided by volunteers, are assigned randomly and anonymous unless a volunteer shares their generated token. For the paid participants, the participant id was logged providing some insights into the full dataset tags. This allows for some analysis to determine whether specific biases can be identified based on the background of the tagger. The 5 partcipants can be grouped into two categories based on their educational / professional sector, with 3 participants from the *journalism* and *creative media design* bachelors, representing a background in applied social science and humanities, and 2 participants from the *computer engineering* bachelor, representing a background in applied formal

and natural science. This allows us to provide a rough estimate on whether the final embeddings, which are essentially an average of the collected tags, can be considered representative of general societal norms and sensibilities, or whether large differences can be found between the two groups of participants, which would imply some bias which needs to be accounted for. ToDo: Can the third group (anonymous) be used for verification somehow?

Conclusion

This section provides a description of the techniques used to answer the questions raised in the previous section, and the answers projected based on the current state of the data ToDo: keep updated and rephrase . The code for generating these stats provided as a Pluto notebook (for now in a private repo, in the future (after name removal) publicly). At the time of writing, the total number of tags is 851. This includes one full tagging of the entire set of 500 tweets. 294 tweets have been tagged more than once, with 241 tagged twice, 50 tagged thrice, and 3 tagged more often than that. In total, 9 different participants appear to have contributed, based on 9 distinct participant ids.

6 Application Concerns

This chapter details the implications of the results of this research. It discusses which precautions should be considered in its application, and how to interpret results when using the model in prediction tasks. It will provide answers to subquestions Re1 and Re2.

7 Conclusion

8 Discussion

Future Work

Finetuning BERT

To predict intent vectors to individual tweets

- Why BERT?
- Pros - state of the art
- Cons - intensive
- Consideration - training intent vectors -> no out of the box solution

BERT accuracy after training

Learning Tweet Relations

- Goals: 1-to-many prediction
- Predicting what a next tweet may look like
- Alert if alarming
- Alert if replies strongly deviate from prediction
- Predicting expected structure of tree (thread/shallow) and note if deviant
- Choice of algorithm (TBD), considerations
- Mayhaps try different algos

Accuracy on next-reply prediction (results chapter?)

Alternative Avenues

Sheaves

Inspired by the work of Hansen[HG] we explore the utilisation of a sheaf structure to our conversation tree. In it ToDo: cite more? , Hansen explores cellular sheaves where each vertex in a graph is assigned associated data, and each edge corresponds to a transformation between the data associated by the two endpoints of

the edge. This appears to be more than is required for trees. The difficulty addressed by cellular sheaves is the requirement for consistency when multiple paths exist between nodes — the diagram formed should commute, and places where this requirement is not met for the actual data are considered with special interest. As a tree does not allow for multiple paths, this extra power appears to be unnecessary. Still, the concept, and that of sheaves in general, helps to inform the way tweets and conversations are associated to numerical data within the structure imposed by Twitter.
ToDo: how do quotes and retweets figure into conversation trees

Recurrent Neural Networks

Recurrent Neural Networks and its variants Long Short-Term Memory and Gated Recurrent Units are of interest in analysing sequential data. In the case of twitter conversations, we can use this method to traverse threads of tweets in the forward search step. Although a conversation tree is non-linear, each path within the tree is linear and as such this step does not require anything more fancy.

Recursive Neural Tree Networks

Recursive tree networks can be utilised to do statiscal semantic analysis on sentences without identifying the parse tree beforehand ToDo: Maybe there's a library get this part covered / focus on conversations? . Maybe the same principle can be applied to twitter conversation trees, although the structure is apparant in this scenario beforehand and the data is not structured in binary trees but rose trees instead.

Bibliography

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

[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.

[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.

[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.

[dVN] Wietse de Vries and Malvina Nissim. As Good as New. How to Successfully Recycle English GPT-2 to Make Models for Other Languages. pages 836–846. arXiv:2012.05628, doi:10.18653/v1/2021.findings-acl.74.

[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.

[HG] Jakob Hansen and Robert Ghrist. Opinion Dynamics on Discourse Sheaves. 81(5):2033–2060. doi:10.1137/20M1341088.

[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.

[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.

[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.

[LJA] Joseph Lemley, Filip Jagodzinski, and Razvan Andonie. Big Holes in Big Data: A Monte Carlo Algorithm for Detecting Large Hyper-Rectangles in High Dimensional Data. In *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 563–571. doi:10.1109/COMPSAC.2016.73.

[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.

[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.

[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.

[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.

[TOD] TODO. URL: ToDo:CitationNeeded.

[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/.

[vdMH] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-SNE. 9:2579–2605.

[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

- [ii] abstract
- [ii] Input correct date when this is known
- [1] rewrite this section to continue from above
- [1] later
- [2] 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)?
- [8] Add some references for basic definitions - SEP
- [13] This question should concern more with the model than with the specific dataset used to analyse the model. It is a prerequisite to answer the following question, although the way it is answered depends more on the specific data than the next question.
- [23] Huib: is dit waar je op doelde?
- [25] Huib: sectie hierboven is uitgebreid, hypothese valt hier logisch maar goed om deze in BG te noemen?
- [27] Need to find tagged data
- [29] Meer introducerend (previously, on ..) - we hebben nu een model van 15 dimensies, belangrijke termen /getallen herhalen
- [29] doel (data nodig), probleem (geen data), oplossing(verzamelen)
- [29] Hier ongeveer: plaatje visueel proces Twitter -> Data trees - resultaat beschrijven in plaats van focus op proces. Data nodig, was er niet, hoe is dit verzameld?
- [29] doel (threads), probleem (twitter), oplossing(bottom up algo)
- [29] 856 - getal eerder noemen, later terugkomen
- [29] Why - top down sensible but impossible
- [29] pseudocode
- [30] Could be -> discussion / future?
- [30] this will be examined in sec X
- [30] betere term of introduceren - wat is divisive
- [30] citation - nrc/nu.nl/whatevers
- [30] citation krant
- [31] doel (uniform ipv normaal verdeeld), probleem, oplossing
- [31] definities - format
- [31] 15 magic number?
- [31] coverage is het begin, waarom?! data quality, coverage is een aspect
- [31] WHY?
- [31] Discussie
- [31] gaten -> missen -> betrouwbaarheid model
- [32] confident -> we weten niets, mogelijk niet uniform verdeeld, willen we weten
- [32] inital test: mean/var per dimensie, makkelijk aantoonbaar niet uniform?
- [32] plaatje

- [32] On the other hand, dim red heeft dit prob.ee m niet
- [32] an example is ... , we used this
- [32] wat is tsne, wat hebben we er aan?
- [32] Hieruit weten we dat er clusters zijn, dingen dicher bij elkaar dus ook verder uit elkaar (niet uniform) dus holes.
- [33] concl: more practical approach required
- [33] check
- [34] plaatjes / pseudocode!
- [34] Update
- [35] Work in progress
- [37] Can the third group (anonymous) be used for verification somehow?
- [37] keep updated and rephrase
- [43] cite more?
- [44] how do quotes and retweets figure into conversation trees
- [44] Maybe there's a library get this part covered / focus on conversations?