An agent-based opinion dynamics model with a language model-based belief system

Bridging language and opinion modeling

Django E.R. Beek

Master of Science Thesis

TU Delft Delft University of Technology

Delft Center for Systems and Control

An agent-based opinion dynamics model with a language model-based belief system

Bridging language and opinion modeling

Master of Science Thesis

For the degree of Master of Science in Systems and Control at Delft University of Technology

Django E.R. Beek

November 25, 2021

Faculty of Mechanical, Maritime and Materials Engineering (3mE) · Delft University of Technology

Delft University of Technology
Department of
Delft Center for Systems and Control (dcsc)

The undersigned hereby certify that they have read and recommend to the Faculty of Mechanical, Maritime and Materials Engineering (3mE) for acceptance a thesis entitled

An agent-based opinion dynamics model with a language model-based belief system

by

Django E.R. Beek

in partial fulfillment of the requirements for the degree of

Master of Science Systems and Control

Dated: November 25, 2021

Supervisor(s):

Dr. Ing. Sergio Grammatico

Dr. Jie Yang

Reader(s):

Dr. Ing. Sergio Grammatico

Dr. Jie Yang

Dr. Matin Jafarian

Abstract

As society's mutual problems grow, there is an increasing demand for understanding the intra- and inter-cultural differences. Rising polarization within national borders and stranded dialogues between nations on mutual problems, daily reach the headlines. It is argued that current models of opinion only scratch the surface of actual human opinion formation with the traditional value based exact approach. Developments in the domain of natural language processing highlight the overlap between real world cultural biases and biased language models. Although within their own domain these biases are seen as problematic, it is argued that it is exactly this associative prejudice that can be used to model human like opinion formation. This idea is emphasized by an conceptual approach on opinion, belief and knowledge, in which only the probability of an association distinguishes between 'objective' and 'subjective'. Through this concept, and by using identified significant cognitive tendencies, a framework is developed for inferring an opinion from text. Using this framework, an opinion model based on a language model is proposed. The contribution of this thesis is two-fold. Firstly, the proposed model provides evidence that agent perception through language is a significant model element, in which the bias in a language model can possibly be exploited to approach cultural perception. Secondly, as the bridge between the two domains has not yet been build, the findings of the results as well as the research process give direction for much needed future research.

Table of Contents

List of Figures

List of Tables

Preface

Before I started with Systems and Control engineering, I was already intrigued by the mind and its workings. Perhaps it is just my 'hammer', but with each course I felt the resemblance grow between the domains of engineering, psychology and philosophy. The idea that how we think and speak now, has major influence on how we will think and speak tomorrow I find fascinating and troubling at the same time. This auto-regressive characteristic of society got me thinking. What if we could predict and control society's ideology, belief and opinion, knowing that people tend to act according to their belief? How can diverging ideologies find common ground on problems that are mutual to all of us?

The deeper I dug, the more complex and unimaginable the interplay between language, thought and our current mathematical and engineering techniques seemed to become. If my original aim was for the moon, the resulting findings might have brought us only an inch closer. However, I firmly believe that language and opinion modeling will get more intertwined over the coming years, alongside other similar domains that the process itself will reveal. I hope that the findings in this research will not only spark the interest in related research domains, but also serve as a starting point for research that will change how we look at the world and each other. Perhaps not anymore 'I think, therefor I am', but more 'We are, what would we like to belief'?

I would like to thank my supervisors for letting me pursue these ideas, and their patience during this process.

"One may even say, strictly speaking, that almost all our knowledge is only probable; and in the small number of things that we are able to know with certainty, the principle means of arriving at the truth—induction and analogy—are based on probabilities"

— *Pierre Simon Laplace*

"If they have begun to do this as one people speaking the same language, then nothing they devise will be beyond them."

— *A bible*

Chapter 1

Introduction

In the past decades, the number of opinionated interactions on the internet has grown to staggering heights. The companies facilitating these interactions are more and more pressured to be transparent about the means with which they facilitate, and not without reason; polarisation between opposing views have never been so present and so destructive as it currently is in the age of algorithmic personalisation and false information.

Although it has long been known that it is in the nature of people to search for and prefer to interact with like-minded people; this nature present in our cognitive processes has, in all those thousands of years of society, not been tested to the extent with which it is being tested now. It is thus not unreasonable to question whether our society is at all capable of discovering and dealing with the consequences of unleashing a world wide opinionated web, especially in its current intransparent controlled form. It has brought to the public, what has long been acknowledged in psychological research, the staggering capacity of people acquiring foundationless knowledge and to quite blindly believe the unbelievable.

Furthermore, with the growing rate of computational resources, in especially the past decade, researchers have come increasingly close to creating systems that can process and hold textual information that can barely be distinguished from humans. They are called language models. Trained on vast amounts of texts, which are only found on the internet, they are shown to hold biases that go beyond flawed models and the selectivity and availability of information; they seem to hold cultural biases which can to a large extent be accounted for in written language. They are trained on what is the legacy of thousands of years of society, among others ingrained in the symbolic representation that language is. Although there are many domains in which we have established a wide consensus on the knowledge we call 'objective', to quote the famous mathematician Pierre Simon Laplace: "One may even say, strictly speaking, that almost all our knowledge is only probable; and in the small number of things that we are able to know with certainty, the principle means of arriving at the truth—induction and analogy—are based on probabilities".

The clues given on the apparent relationship between language, and opinion, belief and knowledge, suggest that a language model might be closely related to opinion modeling. However,

as the domains have long been developing independently, much research is needed to close the gap. For example, introducing an element as ambiguous as language in a traditionally exact system theoretic domain poses an obstacle. Also, the mentioned gap is filled with many unknowns that are so existential of nature, that they heavily rely on imagination. Bringing these domains together, might lead to answers on questions that are historically dubbed 'philosophical', but are more then ever in need of a realistic approach.

This thesis contributes to closing that gap by proposing a model of opinion based on a language model. Given the apparent cultural bias present in language models, the hypothesis is that the element of cultural perception can be modeled. In other words, that the initial prejudice carried by a group of similar speaking agents influences the opinion formation and propagation on a certain topic. Through this approach, the known and unknown biases in language models are embraced instead of labeled as problematic.

In this report, first the necessary background information is given. Through the concept of association, an abstract definition of opinion, belief and knowledge is discussed. Which is henceforth put into the context of human cognition. After, in Chapter 3, the domain of opinion modeling is introduced. It is shown how research actively borrows theories from cognitive sciences to approach human like opinion formation. In Chapter 4, language modeling is introduced. Finally, in Chapter 5, the background knowledge is brought together to formally introduce the research proposal and methodology.

During the research, first in Chapter 6, the concept of opinion and belief is put into the context of language models. Through these insights, a framework is designed for inferring an opinion from text on the basis of a known significant cognitive tendency. The framework is used to propose the core of the opinion model; the belief system. After in Chapter 7, significant identified cognitive and social tendencies are used to propose the dynamics surrounding this belief system. Finally in Chapter 8, the proposed model elements are brought together in an attempt to find evidence for accepting the hypothesis on the significance of agent perception and an associative (language based) belief system. These conclusions are summarized in Chapter 9. Lastly, the obstacles identified during the research are elaborated in a discussion to provide a real contribution in bridging opinion and language modeling.

Chapter 2

Association and Cognition

In the following section, first the principle of association is discussed. It represents a conceptual definition for knowledge, belief and opinion, and what we refer to as 'objective' and 'subjective'. The goal of this first part, is to shed light on the relativity of knowledge and the lack of a difference between knowledge, opinion and belief. Also, to show how society has formed a 'truth' through mutual associations and language, despite the apparent absence of an absolute truth. The line of thought described here, will lay the foundation for the subsequent sections to follow. It must be noted that this part on association is not meant to be read as 'scientific proven fact', as it remains a concept. Nevertheless, as will be discussed later on, it can be seen as a tool to mathematically approach knowledge, belief and opinion.

Afterwards, (neuro-)psychological literature is reviewed to see how the concept fits into what is known about the human brain. Specifically, the principle modes of information processing are discussed and a framework is presented that accounts for the most dominant characteristics of cognition.

The goal of this chapter is to provide background and context to the seemingly trivial but actually highly complex subject of opinion, belief and knowledge.

2-1 Association, knowledge and opinion

Long before (about 2000 years) the research domain of psychology found scientific ground, it was Aristotle who introduced (or elaborated on the works of Plato) the idea of associationism. Aristotle proposed four laws of association in context of human memory and opinion:

1. The law of contiguity. Things or events that occur close to each other in space or time tend to get linked together in the mind. E.g. with a cup one can think of coffee or tea.

2. The law of frequency. The more often two things or events are linked, the more powerful will be that association.

3. The law of similarity. If two things are similar, the thought of one will tend to trigger the thought of the other.

4. The law of contrast. Perceiving or remembering something can trigger a thought of the complete opposite.

Many philosophers and pioneering psychologists built on this idea of association, however although early theories of associationism refer to the workings of specifically the human mind, the principle of association represents a general framework for all that is given meaning. It describes how an organism with a(ny) sensing capability, is able to draw a conclusion from very limited (sensed) data in an ever complex environment. There is no right or wrong, there is just the conclusion, i.e. that what 'bias' (the current state of associations) has led to. Leaving the relative complexity of a task aside, a 'simple' organism sensing the same species to reproduce with, or better formulated, sensing something that leads to the conclusion of an organism of the same species being in the vicinity, is not very different from a human concluding that it is seeing a chair. In both cases something is sensed and a conclusion is formed by the existing associations with that sensed information, whether those associations are instinctive and/or 'hardwired', or (seemingly logically) reasoned to.

The (classifying) conclusion of an association cannot be a definitive or absolute truth, as its foundation is a correlation of associations. This principle concept is abstractly illustrated below in Figure 2-1. Suppose in the context of transportation, the different modes are set out on being a positively or negatively associated with. That what makes cycling to be on the positive side, could be attributed to health and climate benefits. However, the notion of 'positiveness' is in itself defined by health and climate benefits; i.e. 'positive' and 'negative' are in themselves classifications of correlated associations. It is a classic chicken and egg problem in which associations are created, strengthened and weakened with the addition of new information. How new information is (supposedly) perceived and processed by humans for associations to rise and fall is discussed later. Here it is important to notice, through this concept of correlated associations, the lack of a difference between (the forms of information) "knowing", "believing" and "finding" something.

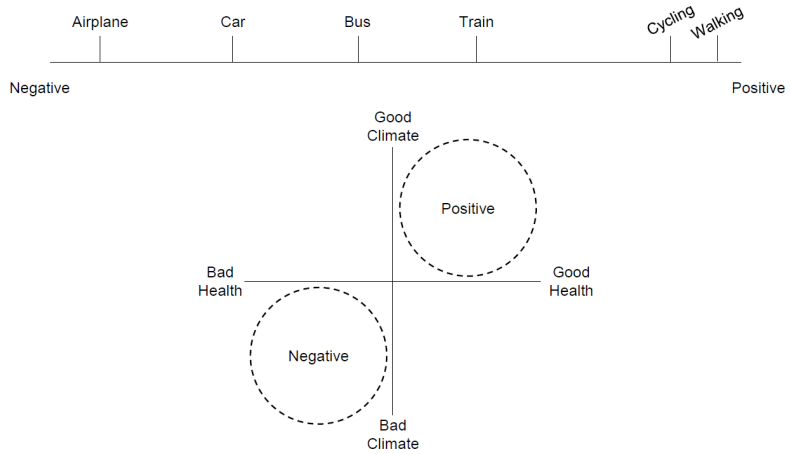

Figure 2-1: Abstract depiction of meaning through correlated associations.

Leaving aside any 'hardwired' evolutionary (and genetic) traits or biological mechanisms aiding (and preventing) an efficient learning and decision process, knowledge and information, can said to be built on the principle pillar of association. Inherent to the driver behind associations, namely (any form of) experience, is the unique 'reality' of every individual, on the premise of a unique sequence of experiences. The principle also gives way to not being able to know what one does not know, or in an opposite formulation, one only sees what one can see, which is discussed in more detail later through the context of cognition. Again leaving certain genetic traits aside, although it is established that every individual can have a completely different view, opinion or belief, societies have formed in which a certain ground truth can be captured. A truth that within a is seen as objective, even though that such objectiveness on the bases of the above, does not exist. Before elaborating on this characteristic particularly in context of current societies, let's discuss how the foundation of common truth came to stand; language.

2-2 Society and language

Quoting the Oxford dictionary, language is 'a system of communication used by a particular community'. It is the very moment when two people agree upon the same sound (or any means of information like a gesture) for the same perceived observation and thus the same association, that language is born. It is the consensus on which (combination of) word(s) should be associated with what circumstance, that is existential to the formation of language. Many researchers have tried giving substantial theoretical body to the evolution of language, among others Steven Pinker [4] and Terrance Deacon [5]. However, the extreme complexity arises from the co-evolution of the human brain, society and language.

The power of society on the formation of language and (individual) associations and vice versa, and the intertwinedness of knowledge, belief and language, is a concept hard to grasp and arguably impossible to experience. George Orwell described a particular view on this concept in his book '1984', in which he proposed a thought experiment concerning a totalitarian society, up to the point that even individual thought is subject to the control of the state. He argued that by putting censorship on language, 'free choice' and individual thought could be controlled. By censoring language, the 'range of consciousness' is limited to what the state desires, making even a thought-crime impossible. The idea of language influencing the way one thinks about reality is known as the Sapir–Whorf hypothesis, or linguistic relativity.

This idea of society being the foundation for the (non-)existence of a concept, is also explored by Deacon. One can learn a dog to fetch the paper, namely to condition it to perform a sequence of actions after a certain trigger, powered by association. For the dog, the cognitive models surrounding the concept of fetching is embedded in the objects and circumstances of its environment. For the concepts to become abstract, it has to rest on other concepts, a hierarchy of concepts. After years of training, primates could possibly learn (only) two levels of (conceptual) hierarchy [5]. This difficulty, and the feasibility with respect to humans, could lie in that the concepts are grounded, through language, as a mutual understanding between people [6]. The cognitive hierarchy on which concepts rest is supported by (the mutual reality of) language hierarchy. It could thus be said that the mutual reality is our principal tool to assign probabilities to associations and such create a rigid ground for abstract meaning.

Having said, arguably just as much as a cause as an effect, a common truth or reality (up to a certain degree) between communicators is needed for there to be any 'usable' or effective information transfer. Firstly, because of cognitive processes enabling or disabling effective communication, which is discussed in later. Secondly, because of a straightforward deduction of the simple associative principles; if definitions on which a transfer of information rests are not mutual (up to a certain point), information will not be acquired as intended or even not at all. These hold for complex abstract definitions (e.g. discussing the origin of gravity on earth with someone who believes in a 'flat earth') and for two different languages and their (co-evolved) corresponding culture with norms and values. David Axelrod [7] describes this fundamental principle of (human) communication as follows: "the likelihood that a given cultural feature will spread from one individual (or group) to another depends on how many other features they may have in common".

2-3 Bias

Having (lightly) treated the origin and the inherent significance of a common truth, it is evidently of importance to define certain terms used throughout this research, such that they are interpreted (or associated with) as intended.

The term 'bias' has been around for a long time, however its 'popularity' has seen a significant increase the past decade, with an ever growing number of (heavily) opinionated interactions on the world wide web, and the contributing factors of large tech companies facilitating these being more and more brought to light. Most people can straightforwardly distinguish the 'subjective' part in a sentence, especially if it is explicitly expressed as opinion, e.g. "I think red hats are ugly". If this sentence was presented as a claim, i.e. "Red hats are ugly", most people would still categorize this to be 'subjective'. Then why would most people argue that the statement "Computers are machines", is 'objective'? Or the nitpicking ones arguing that it depends on the definition of a computer and a machine?

Having previously elaborated on the relativity of knowledge and how it has 'come to be', namely as the consensus on associations within a group of people (up to a society), the difference between the above 'subjective' and 'objective' reasoned statement is that only the latter is built on a wide(r) consensus in society. In other words, both statements are 'biased', in that bias is just the current state of associations of an individual forming the respective conclusion, also with respect to ('truthful') knowledge. Thus, bias itself is neutral.

It is nevertheless useful to make a distinction between 'objective' and 'subjective' bias. Epistemological bias (named after the philosophical study concerned with the 'truth' carried by a society), accounts for the belief, opinion and knowledge that a group of people holds as 'objective' [8]. It captures all possible scopes of 'objectivity', e.g. locally with friends and family, culture wide in a country, or worldwide in a scientific community. This could indeed be extended up to the point of an individual, but as we are concerned with language, which by definition is not individually defined, there is no need.

Moving on in the context of language, on top of the epistemological, stands framing bias, that what is commonly called 'subjective'. Framing bias can be very subtle, only a small difference in formulation and word choice can lead to very different (possible) interpretations, for example a news headline with "a person has died" vs. "a person was killed". Notice that

how one perceives the different connotations of the two framed headlines is a function of ones current state of associations. It is important to note that without an epistemological consensus, there is no framework to discuss framing subjectivity; a person living a lifetime in a village with only red front doors, will (possibly) have an epistemologically grounded attitude towards blue doors and state that as a form of truth. A person having seen all kinds of doors would (possibly) categorize the statement as framing or subjective.

2-4 Cognition

Above, the concept of association in a general and fundamental sense was discussed. The following part sheds light on how this concept fits into the current view of the human brain, specifically cognition, with relevant literature from among others (neuro-)psychology. First, the 2 major modes of information processing are distinguished. After, a framework for describing the most dominant mode is given, in which the concept of association appears to play a significant role.

2-4-1 Associative vs. propositional processes

Cognition is the collective term describing the mental processes guiding in- and outgoing information. Daniel Kahneman, in a widely popular book (*Thinking Fast and Slow*) elaborated on two systems of information processing: one being fast, unconscious and effortless, the other controlled, conscious, intentional and rational. He argued that the former (system 1) is responsible for about 98 percent of our thinking, the latter (system 2) comprising only 2 percent.

That first (most used) system is driven by associative processes. From a conceptual point of view, the principle of associations can account for many of these processes, which will be discussed shortly. A key characteristic of these processes, and also inherent to the principle, is that they are independent of truth values [9]. Association is per definition only based on similarity. For example, certain (associative) stereotyping thoughts might come up while you know (or can reason) them to be inadequate or false. This latter propositional thinking of system 2, in contrast, can be characterised by the application of syllogistic rules to generate or interpret declarative knowledge. Syllogism is what we refer to as logical reasoning. For example:

Dogs are nice + This animal is a dog $\rightarrow$ This animal is nice

Although we generally refer to a proposition to be true or false, propositional thinking is not bound to this (simplistic) dichotomous conceptualization (i.e. having only two possible outcomes). It can also be interpreted in terms of a probabilistic approach, with a likelihood of truth, based on our current state of associations and what one perceives to be epistemologically objective. A longstanding theory also carried by Pierre Laplace as the introductory quote showed.

2-4-2 Associative processes: cognitive bias

Kahneman [10] introduced the idea of cognitive bias, a term used to address the observed phenomena of the cognitive part of our brain. Often it is incorrectly confused with a biased opinion. Cognitive bias is about the convergence aiding (or hindering) mechanisms at work that help to translate sensed information from a (perhaps infinitely) complex environment to a conclusion. It is about the biased dynamics guiding information in and out, not the biased state of the brain.

The lack of (exact) understanding of cognitive biases, leaves us to talking about tendencies; observations that may not be true for all (humans), but nevertheless are far too common to be the result of chance. This is illustrated by one of the end products of the research domain, namely a list of human (cognitive) tendencies, currently listing more than 180 and growing. These are about the tendency to reach a certain (cognitive) biased conclusion through systematic simplifications and deviations from rationality [11]. For example the superstition bias. It describes the tendency of people to keep performing certain actions e.g. before a sports- or gambling game, however unrelated to the outcome, because of a few accidental co-occurrences of the action performed just before a favourable outcome of the game.

These shortcuts towards what might be a suboptimal conclusion or decision are commonly referred to as heuristics, a term used to describe a fast but (possibly) suboptimal optimization. As to the origin of these cognitive processes, just as the evolution of language and thought in the previous section, multiple viewpoints exist in literature. However, given the systematic and consistent persistence of evidence of heuristics guiding the human decision process, it is out of scope to discuss the source. From an engineering perspective, the dynamics of the cognitive processes can be seen as time invariant, 'linearized around the current place in evolution', thus only parameterized with the influences of opinion (and society).

Framework for cognitive bias

Korteling [11] proposed a unifying framework for describing the most prominent clusters of cognitive biases, held together by the fundamental and basic 'binding' principle of association. The framework, consisting of four principles, "forms the basis for our tendencies to associate (unrelated) information, to give priority to information that is compatible and consistent with our present knowledge, opinion and expectations, to retain given information that sometimes better could be ignored, and to focus on dominant information while neglecting relevant information that is not directly available or recognized".

First and foremost, they describe the association principle, having ground in both fundamental and (neuro)biological context. For the latter, using correlation and coincidence, the brain associatively ingrains information through basic operations of neural functioning. Notable theories on such operations are for example Hebb's rule [12], famously summarized as "cells that fire together wire together", or the well known Pavlovian conditioning [13].

Secondly, they propose the principle of compatibility, stating that associative conclusions of the mind are highly determined by their compatibility with the current state of associations; such that one sees information according to its consistency with what one already knows. It makes that within a certain context and abundance of information, that what is 'selected

and processed' is compatible with the brain's momentary state, resulting in the tendency to predominantly see what one expects to see.

They continue with the retainment principle, which essentially describes that one cannot 'unsee'. False, irrelevant or counter-productive information that is associatively captured in the brain's neural circuitry cannot simply be ignored when making a decision. The brain cannot completely disregard associations, in a sense that the state of associations is more like a bucket of mixed (inseparable) fluid than a structure of reconfigurable Lego blocks. This is coined the wet mind [14]. This way of 'storing' information for example makes it impossible to leave out (new) pieces of information to recreate a reasoning of before the new information was included.

Lastly, they introduce the focus principle, stating "that the brain focuses associatively on dominant information, i.e. dominant 'known knowns' that easily pop up in the forming of judgments, ideas and decisions". Again building on the principle of association, only the information that is associated with, is (un)consciously 'seen' and thus 'exists'. Other information that might be lying in front of you is either ignored or not recognized, i.e. the 'known unknowns' and 'unknown unknowns'.

Except for perhaps association, none of the above described principles stand on their own and few cognitive biases can exclusively be attributed to just one. The significant red line through these induced findings concerning human perception, is that it is not so much 'what we see that we get', it is more 'what we get that we see'.

Cognitive dissonance

There is one cognitive tendency to specially highlight, namely cognitive dissonance [15]. It describes the negative feeling one has when holding contradictory beliefs, resulting in the tendency to minimize diverging ideologies or beliefs. This process is called cognitive dissonance reduction or minimization. Festinger's (1957) cognitive dissonance theory suggests that we have an inner drive to hold all our attitudes and behavior in harmony and avoid disharmony (or dissonance). This concept has come be known as cognitive consistency. Although his original theory is aimed at the relation between behaviour (an action) and knowledge (cognition), over the years the theory was expanded. It is even argued that cognitive consistency is fundamental to social cognition as well as general human information processing [16]. The significance of these insights will become more apparent in later sections.

2-5 Conclusion

In this section the fundamental concept of association was introduced as a general conceptual framework for all that is given meaning. Through the concept of (correlated) associations, there appears to be a lack of a difference between (the forms of information) "knowing", "believing" and "finding" something. It is the consensus in society on which (combination of) word(s) should be associated with what circumstance, that places those forms of information under the classes objectivity and subjectivity. Also, there seems to be a special, not (yet) entirely describable relationship between language, and our knowledge and beliefs.

Furthermore, it was established that interaction with our environment creates a state of associations, which are used to give meaning to our environment and circumstances. Although any exact distinction between nature vs. nurture with respect to human behaviour can still not be made, it can be stated that we are in many different ways subject to the concept of association. These mechanisms (cognitive biases) help our brain in learning and translating our complex environment to a(ny) perception, but definitely not all support a logical thought process. One of these significant mechanisms is called cognitive dissonance minimization, ensuring in cognitive (, knowledge and belief) consistency.

Chapter 3

Opinion Dynamics

In the previous section a widely carried but still particular view on human knowledge, belief and opinion was presented. Through this concept of association, abstractly, the interaction with our environment makes the associations and the probabilities we assign to them. Also, cognition was introduced and it was shown how this concept for knowledge, belief and opinion fits into our current view of the brain. The following two sections are to show the engineering (mathematical) approaches on respectively opinion formation and propagation, and language. After, the concept of association is put to use to form a bridge between the two research domains. From here on, the terms opinion, knowledge and belief are used interchangeably.

The domain traditionally concerned with opinionated interaction between people is that of opinion dynamics. It is about the dynamics underlying the (change of) opinion and influence of groups of people and/or individuals. For modeling opinion, mainly two approaches can be identified: agent-based, and physics- or kinetic-based modeling. The former will be further elaborated. The latter is about using statistical mechanics and kinetic theory from gases and fluids to model the 'diffusion' of opinion. As the research is focused on agent-based modeling, elaboration is out of scope.

In general, research shows two incentives for creating a model of opinion. One is to be able to predict the opinion (behaviour, decision etc) of people, for which many examples of applications can be thought of. The other is the ability to describe observations in the real world, for example to study problematic trends such as increasing polarization of opinion. The domain, such as many other sciences, tries to provide a mathematical foundation for the observations made in society, in an attempt to model the dynamics on how collective (and individual) opinions emerge, propagate and change. The research domain is subject to inherent difficulties of (truly) validating proposed models. As such, often proposed models are implicitly or explicitly assumed to be correct under certain conditions, in order to actually formulate a conclusion.

3-1 Agent-based modeling

The agent-based model can be stripped down to show what in essence the related research domain is trying to capture. A single agent and its respective environment is shown in Figure 3-1. The agent, an opinionated person, can essentially be described by two elements: its 'genetic traits' and its (current) associations. The agent's cognition, which is parameterised by these elements, represents the dynamics of the formation of opinion (how does the agent learn which beliefs to what extent), and the propagation (what beliefs are shared and with whom) through the agent's surroundings. In other words, an agent can be seen as a system with inputs and outputs from and to its environment, which is governed by the dynamics of cognition. The interaction of these local dynamics account for the beliefs of individual agents and act as a building block for (describing) collective opinion, i.e. the macro dynamics.

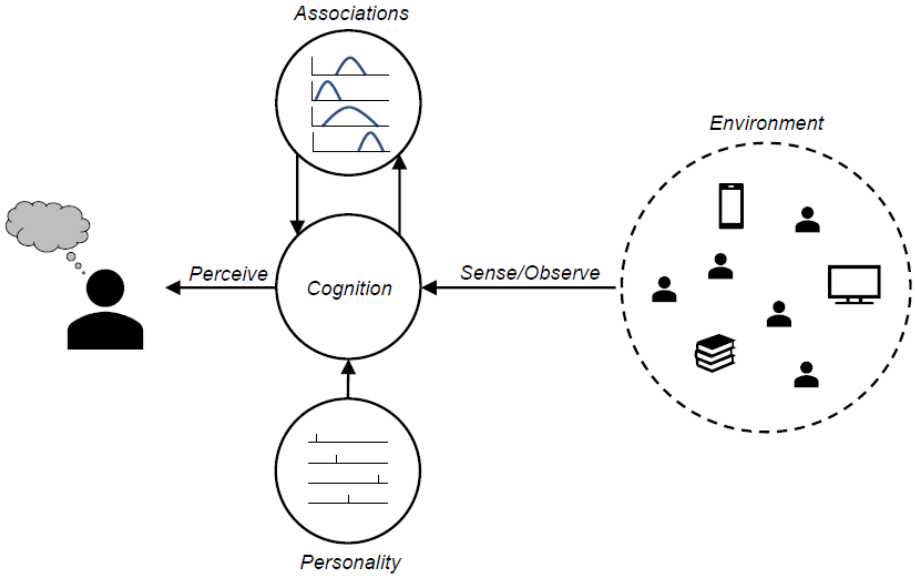

Figure 3-1: Abstract overview of elements considered in modeling opinion.

Important to note, although much research within opinion dynamics has been attributed explicitly to social impact, i.e. people interacting, the influencing environment can essentially hold anything to which an association can be attributed, thus from (the opinion of) opinionated agents to any observable representation of information.

3-2 Graph theory and convergence

Earlier, the terms local and macro dynamics were introduced. In control engineering, a typical problem of research is for interdependent (local) systems to individually or jointly get (converge) to a desired state. A key characteristic of such problems, is that the local systems have limited (to no) information about the state of all other systems. This should sound familiar with respect to humans in society; we can be described as entities (inter)acting according to our local dynamics, based on our own state and the limited information we have

of our (social) surrounding. Such problems are typically approached with graph or network theory. It must be noted that the theory consists of many classes and sub-classes of graphs, holding specific properties for specific use cases. Here, graph theory is only considered within the scope of opinion dynamics.

A graph consists of entities (the nodes) that are connected through so-called edges. These edges can be weighted and directed to represent a specific direction and weight of influence. For example, when modeling social media interactions, your followers are influenced by you and you are influenced by those you follow. A graph is denoted as

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}). \tag{3-1}$$

The agents (nodes) are part of the finite set $\mathcal{V} = \{v_i : i \in \mathcal{I} = \{1, ..., n\}\}$, with n number of agents. The set of edges is denoted as $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, in which the element $e_{ij} = (v_i, v_j)$ represent the directed outgoing connection between agent v_i to v_j. The weighted adjacency matrix $\mathcal{A} \in \mathbb{R}^{n \times n}$ is a matrix with non-negative elements a_{ij} whenever a directed connection exists, i.e. $a_{ij} > 0 \Leftrightarrow e_{ji} \in \mathcal{E}$.

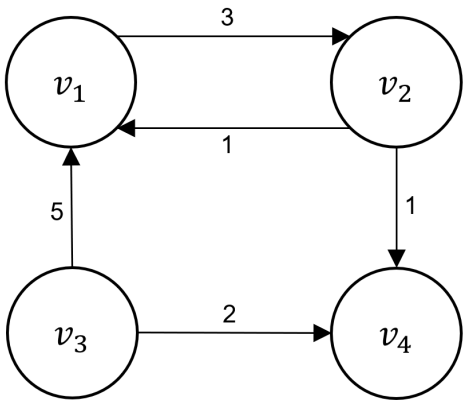

Figure 3-2: Opinion formation and propagation driven by dynamics of cognition.

As an example, suppose we want to represent Figure 3-2 as graph $\mathcal{G}$. Their are four agents ($n = 4$), represented in the set of nodes: $\mathcal{V} = \{v_1, v_2, v_3, v_4\}$. The set of five edges is denoted as $\mathcal{E} = \{e_{12}, e_{21}, e_{24}, e_{31}, e_{34}\}$. These can in turn be represented in the weighted adjacency matrix as follows:

$$\mathcal{A} = \begin{bmatrix} 0 & 3 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 5 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{3-2}$$

Notice how agent v_4 has only incoming connections, this is referred to as a sink. Having only outgoing connections, such as v_3, is called a source. These are straightforward examples of characteristics of a graph, which brings us to an important topic (in control engineering in general and) in modeling social networks, namely convergence.

Convergence, with respect to modeling social networks, is when the individual agents reach a stable state of belief. In other words, the agents have formed a belief, such that their beliefs do

no longer change due to interaction, without explicit new input or changes to the system. It does not necessarily mean that the agents converged to the same belief; convergence towards opposing views is also convergence. See Figures 3-3 and 3-4, to get an idea of what stability means in context of opinion dynamics. In this research three topics were simultaneously considered for propagation, the lines represent the individual agents and they are initialised with diverging opinions [1]. Figure 3-3 shows how these opinions at about $t \approx 20$ converge to a stable state. Figure 3-4 shows how an unstable system might behave; the opinions oscillate more heavily at every time step. Not only can their be nothing concluded from an unstable system of agents, a side from the instability, it also drives against our expectation of opinion behaviour in a society. The (un)stable behaviour of a network of interacting agents is defined by the local dynamics and the network topology (which can also be dynamic).

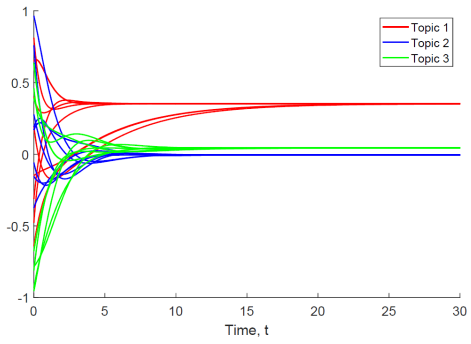

Figure 3-3: Stable system [1].

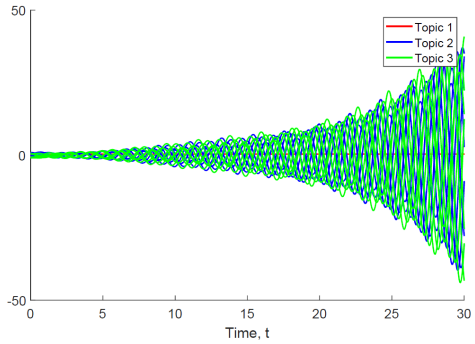

Figure 3-4: Unstable system [1].

3-3 Multi paradigm

The main trend in related research is concerned with proposing models for empirically observed 'human tendencies' (mostly through neuro- and social-psychological experiments) to incorporate into the abstract depiction given in Figure 3-1, in the context of agent interactions. The complexity of the task makes that many (mathematical and social) assumptions underlie the proposed models. Many, especially early, approached the subject with specifications of a social equilibrium rather than the social processes them selves. In other words, models were specifically designed to show expected general behaviour of a crowd, while neglecting possible flawed or over simplistic assumptions. Among others, [17] argued that relying on those equilibrium approaches creates multiple paradigms with which to address the dynamics of opinion formation and propagation, each with their own class of models. Focusing explicitly on the social process and 'fundamental' local dynamics instead, creates a single framework in which the dynamics of different models can be compared.

To underline this, suppose the following example. We are interested in modelling the fall of a rock when it is thrown of a building. We observe that it takes a certain time before it hits the ground, the 'behaviour'. Suppose we try to design two systems mimicking what was observed. One using only gravity, the other using only friction. As the systems are specifically designed and parameterised to show the expected behaviour, they will reproduce the observations. However, if the rock was thrown of a building on, let's say Mars, where both the different gravitational pull and the atmosphere result in a different observation

of the falling rock, both designed systems will not show the observed behaviour. It is the combination of such different dynamical approaches to explain observations that make the classes of models difficult to compare. The other (single framework) approach, is to focus on the (fundamental) local dynamics without designing it specifically to converge. So for the falling rock, focus on incorporating both gravity and friction. For opinion modeling, focus on local dynamics of cognition as depicted in Figure 3-5 below.

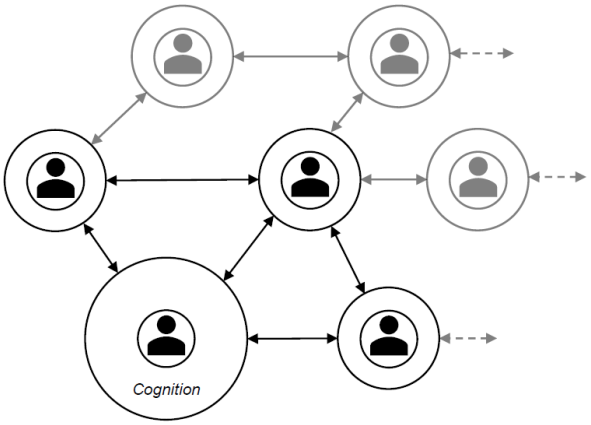

Figure 3-5: Opinion formation and propagation driven by dynamics of cognition.

Though, this single framework approach is only feasible if the dynamics can be approximated, which for the case of actual opinion formation, for example based on the previous section, is not (yet) feasible. What follows is an example of traditional agent based opinion modeling. Subsequently, how the domain borrows theories from cognitive sciences to incorporate into opinion models.

3-4 Traditional opinion modeling

The term belief system [18] describes a definition of the state of opinion and the processes (dynamics) that change this state. It refers to a general mathematical formulation of Figure 3-1. The combination of the circles *Associations* and *Cognition* can respectively be seen as the state of opinion and the dynamics. The genetic traits can be seen as parameters to tune certain characteristic dynamics of agents, such as stubbornness.

The depiction in Figure 3-1 however, is not how current belief systems are proposed. Traditionally, models of opinion are value based and as such, the interaction between agents as well. The value representing the state of opinion of an agent can be binary, one is for or against, real-valued between [-1,1], or a probability distribution. One of the earliest proposed opinion models is the classic DeGroot's model [19]. The opinions of the agents evolve simply by

$$x_i(k+1) = \sum_{j=1}^{n} a_{ij} x_j(k) \tag{3-3}$$

where the a_{ij} is an element from the adjacency matrix as in the example in 3-2. However, in the DeGroot's model, the adjacency matrix is normalized such that it is row stochastic. This

means that the new opinion $x_i(k+1)$ of agent i is the weighted average of the current opinions of all adjacent agents j that are directed at i. An important subsequent work proposed is the Friedkin-Johnson model [17]. The model expands the DeGroot's model by incorporating a personality trait which allows to model an agents susceptibility to influence, for example stubbornness. It can be formulated as follows:

$$x_i(k+1) = \lambda_i \sum_{j=1}^{n} a_{ij} x_j(k) + (1-\lambda_i) x_i(0). \tag{3-4}$$

The parameter $\lambda_i \in [0,1]$ models this specific trait. If λ_i is small, little emphasis is put on the neighbouring opinions. Though, the term $(1-\lambda_i)$ will be large, thus much emphasis is put on the initial opinion $x_i(0)$. Vice versa for a large λ_i. It intuitively models an individuals attachment to their prior belief. Both these (relative simple) models, with additional modest parameters, can be extended to account for more complex phenomena [20].

The mentioned models have a predefined network of connected agents, which is a specific approach. Differently are so-called bounded confidence models. Suppose $x_i \in [-1,1]$ represents the state of opinion of agent i. Bounded confidence means that two agents can interact if $|x_i - x_j| < d$, i.e. when the difference in opinion is smaller than parameter d. An example is the Deffuant-Weisbuch (DW) [21] model. The agents j that interact with agent i are defined by the set

$$I(i,x) = \{1 \leq j \leq n \ \wedge |x_i - x_j| \leq \epsilon_i\}, \tag{3-5}$$

which denotes all agents that have a similar opinion up to the (per agent) defined bound ϵ_i. The model can be formulated as follows:

$$x_i(k+1) = |I(i,x(k))|^{-1} \sum_{j \in I(i,x(k))} x_j(k) \tag{3-6}$$

Where $|I(i,x(k))|$ describes the number of adjacent agents. In other words, the next state of opinion of an agent is the average of all neighbouring agents, which are defined by $I(\cdot)$.

3-4-1 Modeling cognitive tendencies

As mentioned, over the years researchers proposed different models incorporating different cognitive tendencies. With the goal of studying polarization, a version of DeGroot's model was proposed [22] which included biased assimilation, as they argue that homophily[1] alone is not sufficient to polarize (modeled) society. Biased assimilation describes the tendency of individuals to perceive information in line with pre-existing bias, such that it increases or supports the current prejudice. They introduced a bias parameter which essentially represents to what extent the neighbouring opinions are assimilated, with the parameter equal to zero corresponding to the original DeGroot's model.

Closely related to the cognitive characteristic of biased assimilation is confirmation bias; the tendency to search for, interpret, focus on and remember information in a way that confirms one's preconceptions. For example [23] proposed an opinion revision rule which incorporates a projection operation, resulting in a neglection of the aspects of others opinions that deviate (strongly) from a certain reference. They also proposed a framework to include

[1] The tendency to interact with like-minded people.

cognitive dissonance, which as mentioned represents the negative feeling (experience) one has when holding contradictory beliefs, resulting in the tendency to minimize diverging ideologies. Mathematically, they presented it by an agent's opinion probability density having two or more (widely) separated peaks. The idea of using this tendency to minimize the number of peaks (dissonance reduction) as an objective for opinion propagation was explored by [24]. It was further elaborated by also incorporating dynamics derived from structural balance theory [25] [26]. It extends the theory of cognitive dissonance to explain the relations between individuals, namely that agents also try to minimize contradicting interpersonal connections and strive to reach an agreement with their neighbours.

3-5 Content-based

The earlier mentioned models are value-based. There is no relation between the actual topic and the value-based opinions. The belief systems of the models do not go further than representing an attitude towards a topic, which can be anything. The dynamics do not depend on the subject at hand. Though, in the earlier section, it was discussed how (for an apparent significant part) the state of associations, or a function of that state, makes for a opinion and belief. In other words, the content of the discussion does influence the opinion dynamics. This makes the earlier mentioned models only a very abstract approximation of actual opinion formation and propagation through society.

Recently, content-based approaches have slowly been gaining ground, as the research in expressing that content (such as language or pictures) in a mathematical formulation is moving forward fast paced. The means for representing language, language modeling, will be discussed in the next section. It presents a new dimension for opinion modeling. For example, that there exists a correlation between an agents style of (textual) expression and the influence it has on other agents [27].

Another recent work [28], and to the best of the authors knowledge still one of the very few, went one step further and investigated if language (models)[2] could naturally account for cognitive tendencies. They validated for example the conjunction fallacy [29], describing the tendency to attribute a higher probability to a more detailed or specific version of a statement (coherent with one's beliefs), than a single (although statistically more likely) statement. So for events A and B, one is more likely to choose the smaller or equal joint probability.

$$\Pr(A \wedge B) \leq \Pr(A), \;\; \Pr(A \wedge B) \leq \Pr(B).$$

They evaluated the agent with for example the following premise: 'Bill is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In school, he was strong in mathematics but weak in social studies and humanities.', and it has to choose between the options:

1. Bill plays jazz for a hobby.
2. Bill is an accountant who plays jazz for a hobby.
3. Bill is an accountant.

[2]Specifically, word embeddings. They are introduced in the next section.

The agents significantly selected the second option, which aligned with human experiments on the same topic. Not that in anyway the technical implementation resembles the human brain, the output shows similar characteristics.

3-6 Conclusion

The goal of this section was to give an idea of the engineering approach to opinion formation and propagation. The domain shows a clear means to an end, namely incorporating model elements that resemble cognitive (and social) tendencies towards approximating (predominantly) the propagation of attitudes through a network of agents. These local dynamics, described by a belief system, along with the network topology define the system's behaviour. This (model conditioned) behaviour can subsequently be used to create a mathematical foundation for observations made in society, with respect to for example polarization of opinion.

Most traditional value-based techniques, make assumptions on convergence of opinions. It is argued that this approach creates multiple classes of models, that cannot be directly compared. Also, building on the previous section, it does not hold any relation to actual opinion formation.

With the rise of mathematical representation of real (opinion) content, the domain is letting go of assumptions on convergence and moving towards a content-based approach. The combination of language and opinion models is touched upon and will be further elaborated in the next sections.

Chapter 4

Natural Language Processing

Mathematically handling a symbolic representation, such as language, has been a research topic since the introduction first computer. However, it is not until the past decade that the research truly matured with the phenomenal growth in computational resources. Before examining these breakthroughs and their implications, first the mathematical foundation for these systems is discussed, which actually appears to be a model for association. After, the approach for current language models is discussed. The goal of this chapter is to provide the necessary background knowledge on language models, such that it sheds light on the common ground with the earlier sections.

4-1 Distributed representation and distributional similarity

4-1-1 Distributed representation

Since mathematicians and computer scientists in the mid 20th century have invented means to mimic functions of the (biological) brain, we have been leaning more and more on its apparent associative and probabilistic capabilities. Recall the earlier section on the concept of association. If one has the task to classify an object to be a chair, the person will do so if the object and the context (of the person and the environment) is most associated with a chair. Suppose an object has three features: a seating surface, a backrest, and is positioned beside a table. If all three features are present, one could infer it to be a chair. Although the actual combination of patterns we use to form a conclusion on the classification remains largely a mystery, the above reasoning (in text) towards the chair seems logical. However, the three features are classifications on themselves, e.g. a surface can be considered a seating surface because it is part of a chair, and a backrest can be identified because of the seating surface. Moreover, these features will not be very useful when trying to point out an apple in a fruit basket. So how do we represent the chair and the apple, such that the (necessary) characteristics of both are captured, and a computer can mathematically reason towards the correct classification?

This representation is called an embedding. It is a vector in a vector space onto which the features of the object are projected. This distributed representation [30] has at its core that all objects (in question) are projected onto the same space, such that they can be related to one another.

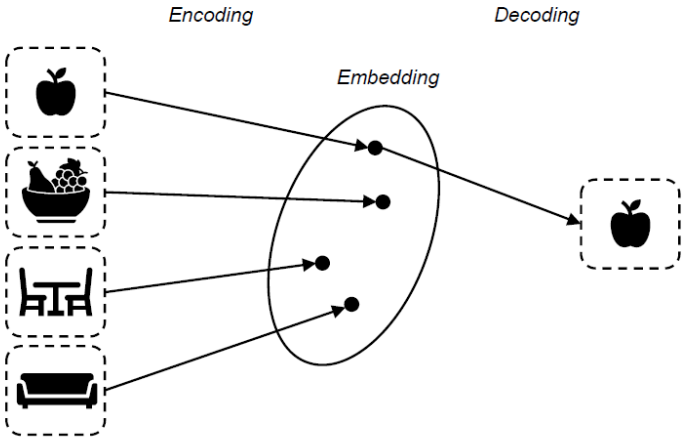

Figure 4-1: Explanatory depiction of encoder-decoder concept.

The mapping between the object (the observational space) and the embedding space and vice versa is referred to respectively as the encoding and decoding, this is illustrated in Figure 4-1. An early and straightforward method building on the idea of Hinton is One-Hot encoding. Say, you want to be able to classify 4 objects. You create an embedding space of 4 dimensions by representing every object with a vector of length 4, each with a single one and zeros at the other entries, as shown in Figure 4-2.

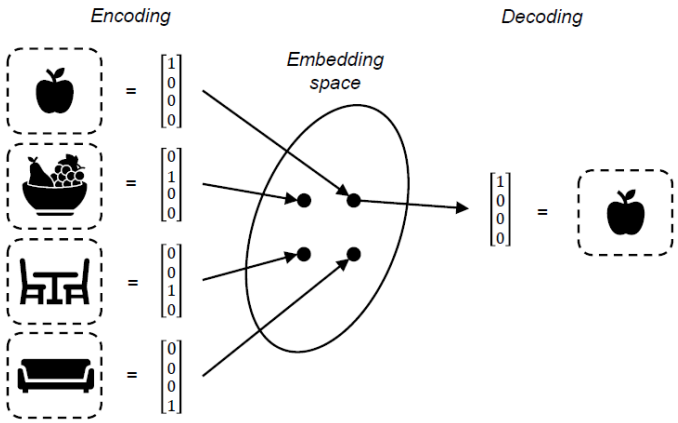

Figure 4-2: One-Hot Encoding example.

Although this practice theoretically creates the embedding that is needed, namely all objects represented within the same quantifiable space, some major implications can be identified. These become more apparent when translating the analogy of the chair and apple to the domain of language. The Oxford English Dictionary contains approximately half a million

unique words, not counting the inflicted forms (i.e. what 'drives' and 'drove' is to 'drive'). They have a meaning on their own, and within a specific context.

The first problem with the approach of One-Hot encoding is what is often referred to as the 'curse of dimensionality'. If we want to encode a vocabulary (V) of, say, a 100.000 words, and use the one-hot encoding to represent a sentence like "I ate an apple while sitting on a chair", we will get a matrix with 100.000x9 elements of which only 9 are non-zero. This is commonly referred to as a sparse matrix (or representation) and it suffices to say this is computationally not very efficient. Secondly, each word is embedded in 'isolation'. In other words, the terms "apple" and "fruit bowl" will be as equally spaced as "apple" and "chair", which implies they are equally associated with each other. In combination with the large dimension (of V), it becomes very difficult to extract ('correct' or useful) meaning. The question thus becomes, how do we reduce the dimensionality of the embedding, resolving the sparseness, and at the same time create a framework for extracting meaning?

4-1-2 Distributional similarity

The idea of extracting meaning from a vector space was originally introduced by Salton, 1975 as Vector Space Models (VSM). He proposed that the vector space spanned by the (embedded) words, also houses the vectors representing groups of words or documents. In other words, context can be seen as the chicken to the egg that is word meaning (and thus vice versa). It is built on the Distributional Hypothesis stating: "linguistic items with similar distributions tend to have similar meanings [or representations]" [31], implicitly referring to the notion of correlated associations in Figure 2-1.

It is important to note the difference between this distributional (semantic) similarity and the distributed representation introduced earlier. The latter is computed from the former; the underlying statistics in distributional similarity are translated into the probability of a representation. For capturing the distributional similarities of words, two trends can be identified, namely (more traditionally) context-counting and (more recent) context-predicting methods. Although both approaches essentially rely on the same premise, namely the probability of words occurring within a context, they can be distinguished by the method of retrieving those probabilities. Context counting methods use (co-)occurrence frequencies to establish matrices (of probabilities) from which, through numerous factorization methods, a 'meaningful' vector space model is derived. The context-predicting methods rely on neural networks to establish the probabilities of words occurring in context. The former method will not be further elaborated. The latter method of context-predicting will be extensively discussed and is referred to as language modeling.

4-2 Language Modeling

In this section, first the general idea behind language modeling is presented. Subsequently, how the design of the vocabulary and the choice and availability off training data play there part. Also, how that training data can lead to both problematic and essential biased language models. Lastly, a state-of-the-art model is discussed.

When the performance of language model is mentioned, is refers to the performance on a NLP task. These tasks have been created to test models on their textual competence, but also to give a practical use. Examples of such tasks are co-reference resolution (what is a piece of text referring to), natural language inference (logical reasoning) and sentiment classification (is the text positive or negative).

4-2-1 Context-predicting

As mentioned, one of the key questions NLP research has been concerned with is how to create a general meaning of a word in its respective context while ensuring a feasible dimensionality. The potential of distributed representation for words, or word embeddings, in language models was underlined by introducing them as off-the-shelf embeddings [32]. The ideas were not new, but earlier such (neural) approaches were computationally still infeasible. This approach can neatly be summarized this approach as follows [33]: (1) associate with each word in the vocabulary a distributed word feature vector (a real valued vector), (2) express the joint probability function of word sequences in terms of the feature vectors of these words in the sequence, and (3) learn simultaneously the word feature vectors and the parameters of that probability function. This essentially describes a solution to the chicken and egg problem mentioned earlier, namely to induce a distributed representation from distributional similarity.

An in-depth review of the training process of neural networks is out of scope for this research. However, it is important to get a grasp of what it (currently) means for a computer to learn a language. As mentioned, the goal of learning language, is to capture the probability of a word occurring with some other word, i.e. the distributional similarity. Suppose the context is "I am throwing a ...", then we want the probability of e.g. "ball" to be higher than "piano", i.e. $P(\text{ball}|\text{I am throwing}) > P(\text{piano}|\text{I am throwing})$.

The neural network is trained by presenting samples of context and than asking it to choose from its vocabulary, which word is most associated with this context. During training the parameters of the neural network are optimized, for which a widely used method is stochastic gradient descent. Informally, the model has the objective to minimize the amount of 'surprise' over all training samples, i.e. the number of wrongly predicted words. It is important to note that the model essentially learns a probability distribution of (groups) of words. Not just 'any' distribution, but the 'true' distribution of language as presented in the training samples. The minimizing objective function is referred to as perplexity (or cross-entropy). This function is a derivation of the concept of (Shannon) entropy of famous mathematician Claude Shannon, it is however out of scope here.

The probability of a word appearing in context is outputted by the model through a softmax function. The softmax function outputs a vector (of vocabulary length V) that represents the probability distribution (in that specific context), with all elements between [0,1] and summed up to 1. Or formally [33]:

$$\hat{P}\left(w_t \mid w_{t-1}, \cdots w_{t-n+1}\right) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \tag{4-1}$$

where w_t is the word in a sequence and the exponential ensures that all probabilities (y_{w_t}) are expressed as positive values. Dividing by the sum of all probabilities produces the characteristic output as described. It essentially expresses what word is most associated with the

given context. Eq. 4-1 is used for explanatory purposes; it only uses the past words to predict the next word, which is essentially a 'depreciated' method. Subsequent proposals introduced more elaborate prediction strategies, but further elaboration is out of scope here.

The general idea of these processes is abstractly shown through an example in Figure 4-3. Suppose the training sample is "I am throwing a [ball]", and the word between brackets must be predicted. Step (1) is to convert the context vector with strings (the actual words) into a vector with their respective index in the vocabulary, such that they can be looked up. In step (2), this vector is fed into the encoder function, outputting the respective embedding vectors per word of a (chosen) dimension. Then, in step (3), these embeddings are fed into the decoder function, which outputs the probabilities of all words in the vocabulary (in this context) via the softmax function. Lastly, the argmax function selects the word with the highest probability. During training, under the objective as earlier described, the model parameters are updated such that the likelihood of 'ball' in this context grows, and implicitly that of 'piano' decreases.

$$\text{(1)} \qquad [\text{I, am, throwing, a}]^{\text{T}} \xrightarrow{\text{convert to indices}} [2,\ 34,\ 3120,\ 64]^{\text{T}} \ \hat{=}\ \text{input}$$

$$\text{(2)} \qquad \begin{bmatrix} 2 \\ 34 \\ 3120 \\ 64 \end{bmatrix} \rightarrow [\textbf{ENCODER}] \rightarrow \overbrace{\begin{bmatrix} \begin{bmatrix} 1.32 & 0.08 & \cdots \end{bmatrix} \\ \begin{bmatrix} -0.34 & 0.68 & \cdots \end{bmatrix} \\ \begin{bmatrix} 4.87 & 2.45 & \cdots \end{bmatrix} \\ \begin{bmatrix} 1.54 & -1.43 & \cdots \end{bmatrix} \end{bmatrix}}^{\text{Embedding dimension}\rightarrow}$$

$$\text{(3)} \qquad \rightarrow [\textbf{DECODER}] \rightarrow \left.\begin{bmatrix} \text{P(ball): } 0.4 \\ \text{P(party): } 0.35 \\ \text{P(piano): } 0.0 \ldots 1 \\ \vdots \end{bmatrix}\right\} \text{Vocabulary length}$$

$$\text{(4)} \qquad \text{argmax(output)} = \text{ball}$$

Figure 4-3: Generalized example of context prediction. See text for explenation.

Early proposed models [34] [35], though ground braking at the time, were aimed at static word embeddings. Although the representation of the words were learned by the many contexts they can appear in, the resulting word embeddings are context independent. This limitation is shown in Figure 4-4. In 2018 ELMo was introduced [2], based on contextual embeddings. The idea of a contextual word embedding is that it is dynamic and dependent on the context surrounding the word, just as the word 'play' in Figure 4-4.

4-2-2 Vocabulary and tokens

The design of a vocabulary is crucial for having a reasonable computation time and capturing general meaning of language. Early language models used the full stem of words in their

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

Figure 4-4: Example of context (biLM) vs. word (GLoVe) embedding [2].

vocabulary, which has a strong limitation: the model can only represent those words that have been included. Later, the vocabulary evolved to subwords, parts of words. It gives the ability to capture the morphology of words and also deduce meaning of out-of-vocabulary words. Multiple methods have been proposed to improve the design of a subword vocabulary and they are used in the state-of-the-art systems. Notable methods are WordPiece [36] and Bytepair Encoding (BPE) [37]. The first iteratively adds characters and subwords to the vocabulary during training, while optimizing the likelihood (decreasing perplexity) on the training data. The algorithm is ended when either the likelihood stops increasing, or a maximum defined number of word units is reached. BPE is similar, but while WordPiece relies on maximizing the likelihood on the training data, BPE simply builds upon the most frequent symbol pairs. Both methods make out-of-vocabulary (unknown) words implicitly part of the vocabulary.

4-2-3 Training data

At the core of modeling language lies training data. Over the years, many different sets of training data have been proposed to learn the embeddings through the systems described above. They can roughly be divided into meticulously designed data sets for task specific training, and large heaps of unlabeled text for pre-training the systems on general semantics and grammar. Some notable examples are elaborated below.

A widely used training set for pre-training (among others used by state of the art) is for example the BookCorpus [38], consisting of thousands of books with genres ranging from romance to science fiction. A key feature of this corpus is the sequence length; giving training data with long stretches of continuous text, allowing the model to condition on long-range information [39]. Also (English) Wikipedia is a popular choice for pre-training. Attributing to its popularity is the large community guarding the strict guidelines concerning its content. An example is the NPOV (Natural Point Of View) principle, stating that all content should be written such that it is not multi-interpretable, making it one of the go to places for creating (largely) non-subjective training sets [8]. Quite the opposite is the Common Crawl dataset, which as the name suggests is an assembly of (among others) texts of the 'entire' internet consisting of billions of pages. Having such quantities of training data is much needed for when capturing the proper probabilities for general semantic knowledge concerning systems with state-of-the-art size. However, concerning actual human text, quantity without quality could

contribute to failing generative grammar, inaccurate or incorrect semantics [40]. Moreover, if it is exposed to harmfully biased (e.g. stereotypical) text, it will simply 'pick up on it'. This is an important topic on multiple levels, which are discussed below.

4-2-4 Biased systems

Recall the neutrality of bias itself as discussed in the first section. The harmful objective and subjective opinions present in real life are implicitly and explicitly represented in text. Although the choice on what is called harmful is up to individual interpretation (and epistemologically defined), an overarching definition for current research would be 'that what is not neutral' with respect to inclusivity and stereotyping in society.

All current language models are essentially build on the co-occurrence of words. So if 'he' and 'programmer' are more often found together than 'she' and 'programmer', the language model will be biased to predict a programmer to be a male. A classic example of stereotypical bias. With respect to static word embeddings, [41] showed that word embeddings trained on Google News exhibit female/male gender stereotypes. Pointing out that even in the 'neutrality' of news problematic biases exist, and also the amount to which these are amplified in the embeddings. They introduced a 'debiasing' algorithm, which first identifies the (semantic) subspaces representing gender in the encoding space. It then neutralizes the (handpicked) gender (to-be-neutral) words, such that they are zero in that gender subspace. Then these words are equalized between the different genders, such that for example 'homemaker' and 'programmer' are equally associated with 'he' and 'she'.

The differential bias [42] introduced, is a novel approach of handling problematic bias, although based on (the currently outpaced) static count based (GloVe [35]) embedding. They proposed a count-based approach to determine the 'bias gradient', with which the bias in embeddings would change after removing a small part of the training corpus. Massively speeding up the naive alternative of iterating the GloVe algorithm over a changing corpus and inspecting the difference. The idea is interesting as it touches upon the 'opinion delta' in opinion models, with the corpus being the environment and the GloVe algorithm representing a belief structure, i.e. analogous to how an agent's opinion (bias) would change after having read a certain document.

Both these methods, and all other current methods, are inherently bound to removing only known biases. Referring to the section on cognition, we can only 'see' the biases or prejudice that we 'get' as biased.

4-2-5 SOTA

For prediction and (deep) neural based methods, it must be noted that the learning task of the algorithm is of great influence on its performance on various downstream tasks. It is like learning for example a human to play a sport. Focusing on fundamental (foundational) knowledge of a ball and its spin, makes the step towards all sorts of ball related sports easier, while focusing on a kick will not help you much with a tennis game. In machine learning , using 'knowledge' of one data set to speed learning on another is called transfer learning [43]. To this end, if the goal of the neural network is to play both tennis and football, the architecture

of the (deep) network must also be able to capture both 'tasks', of what is essentially ('just') a function mapping the input to the output. The idea of language models having a single architecture for many to all possible (textual) tasks, has been getting a more ground with the growth of computational resources and with the introduction of the Transformer [44].

The Transformer is a specific type of neural network architecture of which all state-of-the-art language models are descendants. One of these proposed language models is called BERT, Bidirectional Encoder Representations from Transformers [45]. As the model architecture of the Transformer itself and BERT do not add to a greater understanding of the essence of language modeling, they are not elaborated here. BERT is trained to have a large general understanding of language, but is not directly able to perform a textual task. Just as with the example of ball sports, the idea is that with little training of the baseline model, it is able to perform all sorts of tasks.

4-3 Conclusion

The relativity of meaning in language represents a chicken and egg problem. It is tackled by inducing a distributed representation (or embedding) from the distributional similarity, based on the Distributional Hypothesis stating that linguistic items with similar distributions tend to have similar meanings. The general process of language modeling is capturing the the 'true' distribution of language as presented in the training data.

The resulting language model is essentially based on the concept of correlated association (Figure 2-1): which word is associated with what context, i.e. the co-occurrence. Language models are in a sense the embodiment of association. Furthermore, it appears that (known) stereotypical biases we have accumulated through experience (among others with our cognitive biases) are not only represented in (human) text, they are also captured by language models.

Chapter 5

Conclusion and research proposal

5-1 Conclusion and proposal

The goal of this background knowledge was to give an overview of the elements that have to be considered when modeling the formation and propagation of belief and opinion through society.

In the first section, the concept of association was discussed. It was shown that it can be used to form a definition of knowledge, belief and opinion. Also, through this concept, there seems to be a lack of a difference between these forms of thought or information. The distinction between what one finds 'objective' and 'subjective' can be attributed to the high probability one assigns to the associations, or similarly, the association have a high probability to be truth full or objective because of the associative strength. These associations are strengthened and weakened through the (re-assuring) interaction with our environment. That what is perceived from an observation appears for a significant part to be a function of those associations. Also, the majority of cognitive biases can be said to be built on the concept of association. Lastly, there seems to be a special relationship between language, and knowledge, belief and opinion.

In the second section, the domain of opinion dynamics was introduced. The research area tries to find mathematical foundations for observed (opinion) behaviour in society. It was discussed how this engineering domain borrows theories from neuro and social sciences to incorporate into the micro dynamics representing an individual. It was shown that the traditional approach of modeling opinion is predominantly based on a value-based attitude, instead of the associative structure an individual is expected to have as seen in Chapter 2. Also, most research is focused on achieving certain (macro dynamic) convergence properties, instead of the bottom up approach, namely approximating the human local dynamics without any assumption on convergence. Though, the trend of content-based opinion modeling is getting more ground with the advances in natural language processing

Subsequently, the domain of NLP was introduced. Here it was shown how the research domains tackles the chicken and egg problem of relative knowledge. Through the Distributional

Hypothesis, stating that linguistic items with similar distributions tend to have similar meanings, a distributed representation of words and context is created. From this embedding space, the (relative) meaning of language can be extracted. Furthermore, it was shown that the stereotypical biases seen in society, are also present in these embeddings inferred from real (human) text.

Concluding, (empirical) evidence shows that language models and opinion dynamics might be highly related research domains. However, as of yet, the combination of language models and opinion dynamics has not been explicitly researched. It is this gap to which the following research topic should contribute to:

> An agent-based opinion dynamics model with a language model-based belief system

Specifically, there are two hypotheses of interest. Firstly, including an associative belief system should introduce a known important aspect of observing (or sensing in general) information, namely perception, based on the current state of associations of an individual.

Secondly, on the premise of the above being true, with every epistemological scope, different associations (or biases) govern the propagation and formation of an opinion. An arbitrary example of these cultural biases, is that convincing a group of people that puppies are 'filthy' will be a harder (if not impossible) job than trying to convince them that rats are filthy. This implies that the formation and propagation of a belief is dependent on the actual subject of discussion and the initial prejudice of the individuals.

Parallel to the research of these hypothesis, an overarching goal is to emphasize the need for bridging the gap between language and opinion modeling. Also, to layout the obstacles ahead for a mathematical approach to researching something as ambiguous as language and opinion.

Having said, the goals for this research are as follows:

1. Proposing and validating an agent-based opinion dynamics model, that
   (a) uses a language model to infer the opinion of an agent,
   (b) enables the (agent's) individual perception of information.
2. Using the model's novelty, inspect the significance of an associative belief system in terms of
   (a) agent perception,
   (b) the dependency on the actual topic of discussion.
3. Layout both the relevance and the obstacles for bridging language and opinion modeling.

Towards these goals, the main research questions to be answered are:

1. How can the opinion of an agent be modeled using a language model? Such that,
   (a) agents perceive information individually based on their current state.

2. How can this belief system validated?
3. What processes of cognition govern the formation of an opinion over time and how should they be modeled?
4. What social and cognitive tendencies govern the interaction between agents an how should they be modeled? Such that,
   (a) agents can individually perceive information that is shared to them.
5. How can these dynamic processes be validated?
6. Can the model reproduce a known (opinion) outcome?
7. Can the significance of perception be shown?
8. Can the topic dependence be shown?

5-2 Research Methodology

The domain of opinion dynamics and social engineering in general is subject to an inherent problem: models are difficult to impossible to truly validate. We cannot look into a mind, and if someone is asked about their opinion, we cannot tell it is 'truly' what one finds or beliefs.

The goal of the research is to show a fundamental element is missing in current models of opinion, namely the associative nature of knowledge and opinion. As such, the approach is bottom-up. In other words, the focus lies on producing micro dynamics that are 'fundamental' to human opinion formation. It is the hope that the macro dynamic characteristics can be interpreted to show why it does or does not follow the results of earlier opinion dynamics research.

The research questions are spread out over three phases. Which are shortly described below.

Phase 1: Design of a language model-based belief system

In this first phase the first two research questions will be answered, namely

1. How can the opinion of an agent be modeled using a language model? Such that,
   (a) agents perceive information individually based on their current state.
2. How can this belief system validated?

The possible means that can be used to close this gap have only recently reached a mature phase, mostly due to increase in computational resources, namely language models. However, it presents no straightforward implementation, as they are not built to represent an opinion or individual.

In this phase the core of an agent will be proposed, namely the belief system. It represents a mathematical definition for inferring the opinion of an agent by means of a language model and text. By using the literature, a generic framework will be introduced to intuitively approach the concept of opinion, belief and knowledge, on which the belief system is based.

The proposed belief system will be tested in experiments to validate the parts that to some extent can be validated. Mostly, these are experiments where we have a clear expectation of the results.

Phase 2: Cognitive and social process design

In the second phase the dynamics that describe the change of opinion are designed, namely:

3. What processes of cognition govern the formation of an opinion over time and how should they be modeled?
4. What social and cognitive tendencies govern the interaction between agents an how should they be modeled? Such that,
   (a) agents can individually perceive information that is shared to them.

Here, again conclusions from the literature review are used, namely the (apparent) most dominant cognitive and social tendencies. Although the belief system is generic for all sorts of opinion model paradigms, these dynamics will be specified towards fixed topology networks. Validating these dynamics individually is not straightforward, as their true effect (during the process) is unknown. As such, it will be tested on a real world network, with a known (opinion) outcome, presented in the next phase.

Phase 3: Model validation and interpretation

In this last phase, the entire model is brought together. Numerical simulations are run on a widely known network of agents, made up of true connections identified by the researcher at the time. The results are used on the one hand to validate the proposed model, and on the other hand to show the significance of the novel elements. As will be shown, inherent to the unknowns with respect to opinion formation, the interpretation of the results is key for concluding anything about the model.

5. How can these dynamic processes be validated?
6. Can the model reproduce a known (opinion) outcome?
7. Can the significance of perception be shown?
8. Can the topic dependence be shown?

Chapter 6

Belief system design

In this section firstly, the (dis)similarity between language models and cognition is discussed. After, the concept of opinion, belief and knowledge is intuitively approached by means of text. From this intuition, a mathematical framework is introduced on which the proposal for the belief system is based. The proposed belief systems describes a possible approach for inferring an opinion from text, and what that opinion represents. After having shown a small numerical toy example, the belief system is tested on cases where there exists a clear expectation on the outcome. As no true validation can be inferred from the limited number of experiments, they are also about inspecting how to interpret the output of the belief system.

6-1 Opinion and language models

In the literature review, a possibly fundamental driver for general human information processing was identified, namely cognitive dissonance minimization. Also, with respect to that information, the principle of association suggests that there is no difference between a belief, opinion or knowledge. The review also touches upon the training process of language models, namely through minimizing the cross-entropy. This can be described as minimizing the surprise of a word appearing in a certain context. The apparent similarity between the concept of (Shannon) entropy and cognitive dissonance minimization has attracted much research over the years [46] [24]. However, no research has, to the best of the authors knowledge, yet included language models. As the incentive for this research is based on the apparent relation between language and opinion, first it is investigated to what extent (current) language models can be used to mimic human information processing. See Figure 6-1. It very abstractly indicates the analogy between the information processing of an opinionated agent vs. the training of a language model.

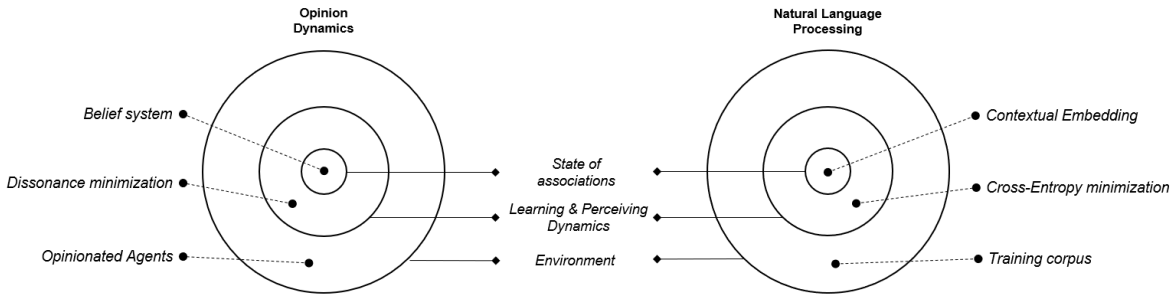

Figure 6-1: Abstract analogy between opinion and language modeling.

Obviously, human information processing and opinion formation cannot be brought down to just textual information. However, with the goal of using a language model, only text is considered. Suppose that the opinion of a person in the broadest sense can be characterised as follows:

(environment + personal) context + (denoted / real) subject → connotation + opinion.

The denoted topic is what could be said to be the sensed 'real' part, for example the printed letters on a paper. Although anyone could perceive it differently, the actual words in the sentence are the same for everyone. It is the personal context that makes how it is perceived, namely what connotations are added to the denoted words. That context can be anything: current state of mind, emotion, the person discussing the topic with, state of body, etc. Not only do these personal connotations make up for the opinion, belief and perceived truthfulness of a statement, for the majority of the time, a person only 'sees' the subject as it is experienced. This concept is shown in in Figure 6-2 below.

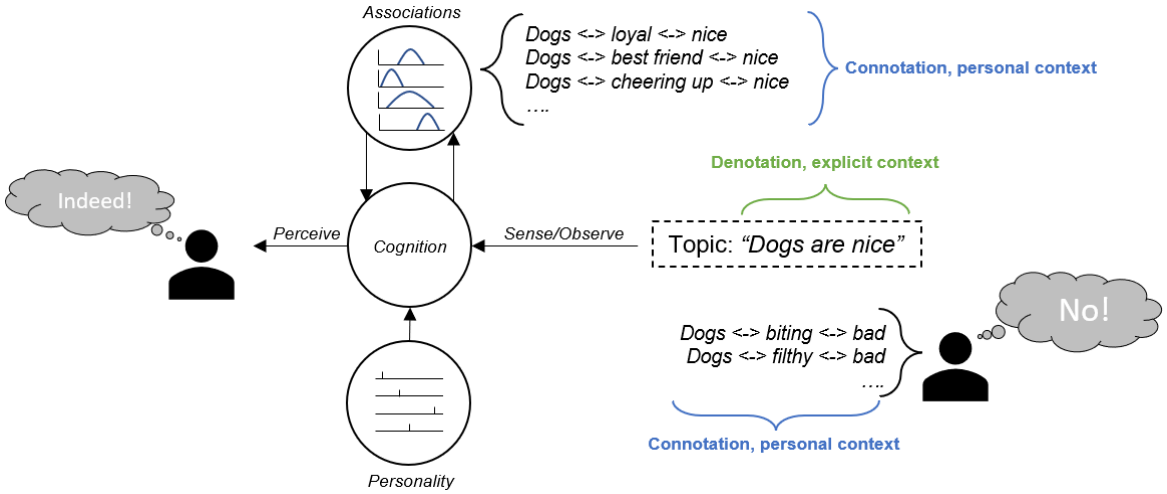

Figure 6-2: Association based opinion.

Suppose a language model is used to mimic the associations as shown in Figure 1. Recall the relative meaning of language. Language models essentially describe a projection operation; from text onto an isometric encoding space. The isometric property resulting from the way it is trained, makes that a single vector in the encoding space has no meaning on its own.

Nothing can be inferred from its absolute place, it can only be compared to another vector in that space. As such, the association between 'dogs' and 'nice' on itself says nothing. Only when that association is compared with that between 'dogs' and 'bad', something can be inferred. See Figure 6-3.

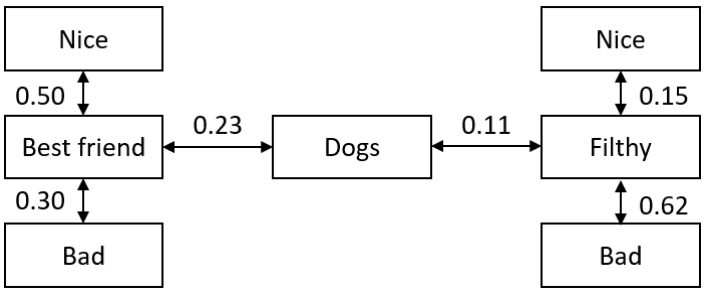

Figure 6-3: Relation between words captured by the language model.

These relations are determined through the cosine similarity. It is a method to express the similarity of two vectors in a Euclidean space by measuring the angle between the two vectors. Suppose two words (or possibly any length of context) are projected onto the encoding space as vectors $\vec{d}$ and $\vec{q}$, then their cosine similarity is defined as follows.

$$\begin{aligned} \text{CoSim}(\vec{d}, \vec{q}) &= \frac{\vec{d} \cdot \vec{q}}{\|\vec{d}\| \|\vec{q}\|} \\ &= \frac{\vec{d} \cdot \vec{q}}{\sqrt{\sum_i d_i^2} \sqrt{\sum_i q_i^2}} \end{aligned} \tag{6-1}$$

where $\|\vec{v}\| = \sqrt{\sum_i v_i^2}$ is the Euclidean length of $\vec{v}$.

Traditionally, as discussed in the literature, the 'associations' in Figure 6-3 are referred to as a bias in the language model. Having also discussed the essence of 'bias', on a first sight, the relation could also be said to be the opinion of the language model. However, in the example in Figure 6-2, the topic only denoted the word 'nice'. As just stated, the semantic meaning of 'nice' in a language model is only defined by its relative relation to other other words or context. The choice to compare it with 'bad' comes from the author. That choice is also based on association. If the context for that choice is not explicitly denoted, it is an implicit connotation. For example, one could also respond with 'Cats are nicer'.

So, it seems that for inferring an (association based) opinion through a language model, an contextual antonym is needed, i.e. the (biased) semantic opposite of a meaning given the context. It is this implicit context, or connotation, from which the associative attitude can be derived. However, as the model is build from co-occurring words and context, and not on 'semantic logic', the encoding space is not structured to extract semantic opposites. See the arbitrary example in Figure 6-4 below. It just finds resemblance in the word 'agree'.

Despite the apparent similarities between language model training and human information processing, current language models do not lend themselves directly for extracting an opinion as depicted. Furthermore, the concept of association mathematically lies in the domain $[0, 1]$;

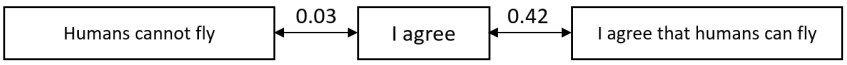

Figure 6-4: Arbitrary example of relations through co-occurrence of context.

i.e. it is a measure of similarity. In Figure 6-2, that means that the right individual will not disagree, but just agree very little, while the very idea of discussing a topic proposition, is that one can agree or disagree, find or not find, belief and disbelief. In other words, an expressed opinion should also be able the contradict.

To account for contradiction, a well known language model task is introduced, namely natural language inference. It is an extension of the language model that uses the relative relations captured in the model to infer (the probability of) a relation between two sentences. See Figure 6-5 below for an example.

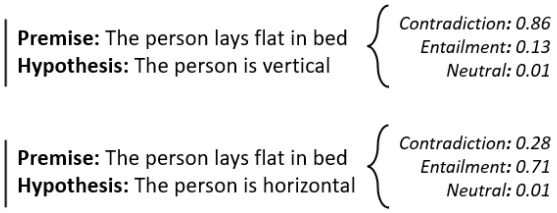

Figure 6-5: Arbitrary example of Natural Language Inference.

It is important to emphasize that this 'logic' is build on the associations (or bias) in the model; the concepts 'being horizontal' and 'laying flat' define each other. Just that one could perceive 'not flat' as 'vertical'. The idea is that all these possibly 'objective' associations, along with all the 'subjective' associations (as the example of the dog) could for a large part, account for the (English speaking) culture. That being said, current language models have no direct relation to human opinion formation. However, the indirect use of a language model could still provide the element of perception. In the next subsection, first, the above presented findings are used to form an intuitive framework for inferring an opinion from text. After, this framework is used, along with the mentioned language model tools, similarity and inference, to propose the belief system design.

6-2 Framework for opinion and proposal

The framework is built on the probabilistic paradigm of the mind, also seen in the quote of Laplace noted on the first page, namely that the mind and the its (un)conscious conclusions can be represented in terms of probability. The framework describes a possible approach on mainly two points of interest in inferring the believe of an agent from text: *What should the probability express? And, how is that probability formed?*

6-2-1 Intuitive framework

Intuition behind belief expression

Suppose the probability of an agent believing in a topic can be formulated as follows

$$\text{P(topic|agent experience)}$$

In other words, conditioned by earlier experiences, what is the probability that the agent beliefs or finds the topic formulation to be 'true'. So, for example:

$$\text{P(Dogs are nice|I only know dogs that are nice)} \rightarrow \text{High probability}$$
$$\text{P(Dogs are nice|I only know dogs that are not nice)} \rightarrow \text{Low probability}$$

Now suppose an agent has multiple experiences that hold some relation to the topic. More over, the experiences also hold a relation to each other. For example:

$$\text{P}\left(\text{Earth is round}\middle|\begin{bmatrix}\text{I have seen pictures and movies}\\ \text{Fundamental physics}\\ \text{...}\\ \text{I have never actually seen Earth from space}\end{bmatrix}\right) \rightarrow \text{Some probability}$$

Here the assumption is that experiences that are more believed in, have a larger influence on the opinion then experiences that are lesser believed in. Furthermore, the belief in an experience is a function of the other experiences. To explain this intuition, see Figure 6-6. Suppose a person that has accumulated many interdependent associations with respect to the Earth being round, e.g. photo's, physics theories, etc (accumulated in the blue mass), and very little of the Earth not being round (the orange mass). All the experiences in blue ball rest on each others shoulders, and enhance each other as it is congruent information. This person cannot belief in a flat Earth from one moment to another, because 'the stakes are too high'; all previous dependent 'knowledge' or belief would become 'invalid'. They would swap from being supporting to denying experiences, which would go against every 'fiber in its body'. This common expression can said to be a direct result of the concept of cognitive dissonance.

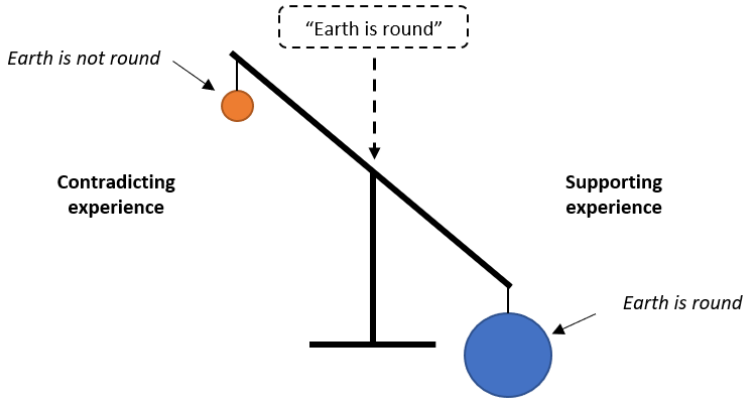

Figure 6-6: Analogy of masses hanging at a balance that represents the perceived opinion.

Now suppose in Figure 6-6 above, the topic changes to 'The Earth is flat'. Logically, the balance would look exactly the same way only mirrored, as all supporting experiences become denying and vice versa. However, perceiving 'a flat Earth' as the exact (and only) semantic or logical[1] opposite of 'a round Earth' is subject to the context of the discussion. In other words, 'the Earth is not round' could only be said to be equal to 'the Earth is flat' in this context. On the scope between round and flat, any astrophysicist would 'belief' in round. On an lesser or undefined scope, one could not belief in both, as the Earth is not a perfect sphere (apparently it is something called an oblate spheroid). This idea is shown in Figure 6-7, in the analogy of the balance.

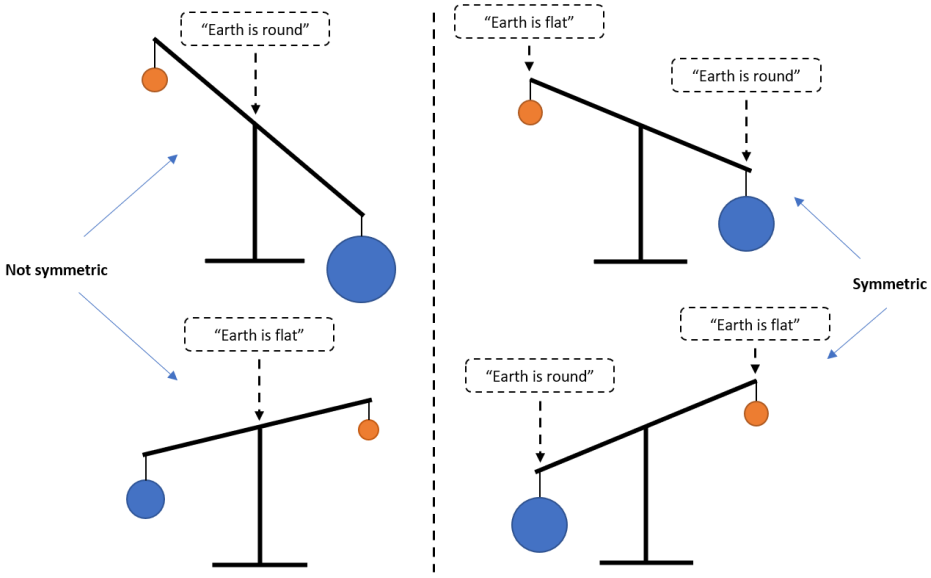

Figure 6-7: Depiction of non symmetry for an 'open' topic, and symmetry for an explicit topic.

As the agents will be expressing their belief towards an explicitly defined proposition, their

[1]Referring to literature, there might not be much difference between a logic or semantics, as logic can be seen a result of semantics.

is the need to overcome this described ambiguousness to ensure consistency in the agents belief. Therefor the notion of a conjugate topic is introduced, denoted as t^* (versus topic t). The conjugate topic should be chosen to define the scope on which the opinion is needed. As outlined in the example above; there could be a large difference between: 'Earth is flat' and 'Earth is not flat'; 'Earth is flat' and 'Earth is round'. How exactly the combination of the topic and its conjugate is used to infer a belief is shown in the proposal the next section. This kind of detail in defining an opinion is rather novel. The traditional opinion dynamics research is at most concerned with the general nature of the topic.

Concluding, the two beliefs that can be expressed by an agent with respect to a topic proposition are

$$\mathrm{P}(\text{topic}|\text{agent experience}), \tag{6-2}$$

without the conjugate information and coined the *implicit opinion*, and

$$\mathrm{P}(\text{topic}|\text{conjugate topic}, \text{agent experience}), \tag{6-3}$$

with the conjugate information, coined the *explicit opinion*. Both these opinion expressions will be worked out, to examine the difference. Having discussed the intuition behind what it means for an agent to belief in a topic, below the framework is presented for how this belief is inferred.

Intuition behind belief formation

Firstly, with respect to the implicit opinion. Keeping with the analogy of the balance, see Figure 6-8. An experience is represented by a mass. This mass is a measure of belief in the experience, which is defined by all other experiences. Depending on the relation (and the strength of that relation) with the topic, it will be hanged somewhere at the balance. So M_2 has a larger direct relation than the other experiences, and it is contradicting the topic along with M_3. While M_1 supports it. The idea is that every experience represents a force, and the combination of those forces result in an angle of the balance. The framework in words is as follows:

> The mass of the individual experiences is a function of all experiences.
>
> The belief 'force' is defined by the supporting experiences with respect to the topic.
>
> The disbelief 'force' is defined by the contradicting experiences with respect to the topic.
>
> The probability of an agent (dis)believing in the topic, is a function of all 'forces' acting on the balance.

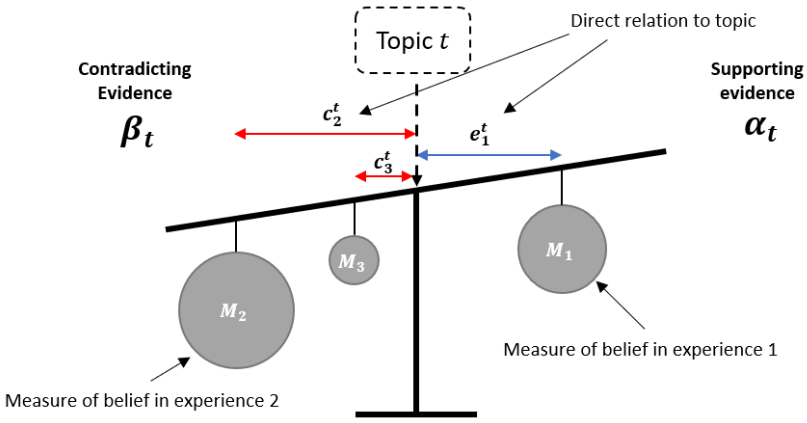

Figure 6-8: Schematic depiction of the intuition of the balance.

The probability of belief in the topic ('the angle of the balance') is represented by the expectation of a Beta distribution, parameterized by a summation of the (dis)belief forces. The implicit opinion can be represented as follows

$$\mathrm{P}(\text{topic}|\text{experience}) := \mathrm{E}[\mathrm{Beta}(\alpha_t, \beta_t)] \cdot 2 - 1 \in [-1, 1] \tag{6-4}$$

with

$$\alpha_t(\cdot) = \sum_{i=1}^{n} e_i^t m_i \tag{6-5}$$

$$\beta_t(\cdot) = \sum_{i=1}^{n} c_i^t m_i \tag{6-6}$$

The Beta distribution is a natural way of counting evidence for (α) and against (β), and outputs the probability between $[0, 1]$. For emphasizing the difference between belief and disbelief, the output is scaled to reside in the domain $[-1, 1]$. An in depth elaboration the Beta distribution is out of scope.

e_i^t and c_i^t represent respectively the (direct) entailing and contradicting relation of experience i (of all n experiences) to the topic t. m_i represents the mass of the respective experience. It basically is a summation of all the influences of the individual experiences, that have an 'arm' and a 'mass'.

The explicit belief of Eq. 6-3, follows the same idea of the balance. However, instead of using the inferred contradiction with respect to the topic, the entailing relation with respect to the conjugate topic is taken. This is illustrated in Figure 6-9 below.

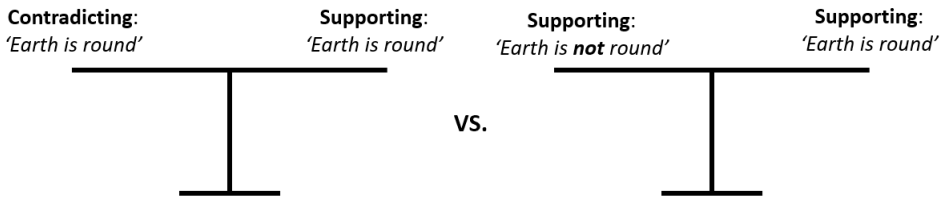

Figure 6-9: Difference between the implicit opinion (left) and the explicit opinion (right).

The resulting framework of the balance is nearly identical, with the exception of the added script t^* denoting the explicit use of the conjugate topic information.

$$\mathrm{P}(\text{topic}|\text{conjugate topic}, \text{experience}) := \mathrm{E}[\mathrm{Beta}(\alpha_{t,t^*}, \beta_{t,t^*})] \cdot 2 - 1 \in [-1, 1] \tag{6-7}$$

with

$$\alpha_{t,t^*}(\cdot) = \sum_{i=1}^{n} e_i^{t,t^*} m_i \tag{6-8}$$

$$\beta_{t,t^*}(\cdot) = \sum_{i=1}^{n} c_i^{t,t^*} m_i \tag{6-9}$$

Having elaborated the framework and the intuition of the balance behind it, the proposed belief system in the next section is a proposal for specifically: a function defining the direct relations e_i^t and c_i^t (from Eq. 6-19 and 6-20), and e_i^{t,t^*} and c_i^{t,t^*} (from Eq. 6-8 and 6-9), and a function for defining the measure of belief in an experience m_i.

6-2-2 Proposed belief system

The belief system of an agent is characterised by three elements: (1) its current experiences $\mathcal{S}$ filled with n unique textual experiences, i.e. $n \in \mathbb{N}, \{s_1, ..., s_n\}\mathcal{S}$, with $n > 1$. (2) A bias matrix $B \in \mathbb{R}^{n \times n}$, and (3) the parameters of the language model Θ, which is the equal for all agents. Below, first the implicit opinion expression will be proposed, thereafter the explicit opinion.

The implicit opinion is denoted as x^t. Note that t is interchangeable for any topic proposition, including the conjugate t^*. Within the framework, the probabilistic output of the first approach is inferred through

$$x^t = \mathrm{E}[\mathrm{Beta}(\alpha_t, \beta_t)] \cdot 2 - 1 \in [-1, 1] \tag{6-10}$$

The belief in the conjugate topic is simply by swapping t for t^*. The parameters of the Beta distribution are defined as follows

$$\alpha_t(\mathcal{S}, B, \Theta) = \sum_{\forall i \in \mathcal{S}} e_i^t m_i \tag{6-11}$$

$$\beta_t(\mathcal{S}, B, \Theta) = \sum_{\forall i \in \mathcal{S}} c_i^t m_i \tag{6-12}$$

with

$$m_i = \mathrm{E}\Big[\mathrm{Beta}\Big(E_{i,:} \cdot \Big(B_{i,:} \sum_{j=1}^{n} \Phi_{i,j}\Big), C_{i,:} \cdot \Big(B_{i,:} \sum_{j=1}^{n} \Phi_{i,j}\Big)\Big)\Big] \in [0, 1] \tag{6-13}$$

The measure of belief in an experience (the mass) m_i is equal for both types of belief expression. It is only the relation with respect to the topic and the conjugate where they differ. Those direct relations will be proposed shortly, first, the mass of an experience is elaborated. It is defined as a probability distribution as well. Let's breakdown the α and β part of this

Beta distribution in Eq. 6-13. The center dots represent dot products and can be rewritten to a summation for explanation. For the α part:

$$E_{i,:} \cdot \Big(B_{i,:} \sum_{j=1}^{n} \Phi_{i,j}\Big) \hat{=} \sum_{\forall j \in \mathcal{S}} \underbrace{E_{i,j}}_{\mathbf{A}} \Big(\underbrace{B_{i,j} \sum_{k=1}^{n} \Phi_{j,k}}_{\mathbf{B}} \Big)$$

Here the same intuition is used, with an 'arm' (**A**) and a 'mass' (**B**) as force acting on a balance. The direct relation shown with the brace **A**, is calculated with the inference NLP task. Suppose sentence s_i (the premise), sentence s_j (the hypothesis), and Θ (the trained LM parameters). The function representing the inference task can be formulated as:

$$f(s_i, s_j, \Theta) = \begin{bmatrix} P(Contradiction|s_i, s_j) \\ P(Entailment|s_i, s_j) \\ P(Neutral|s_i, s_j) \end{bmatrix} = \begin{bmatrix} C_{ij} \\ E_{ij} \\ N_{ij} \end{bmatrix} \tag{6-14}$$

In which

$$E_{ij}, N_{ij}, C_{ij} \in [0, 1] \quad \text{and} \quad E_{ij} + N_{ij} + C_{ij} = 1 \tag{6-15}$$

The terms under brace **B** represent a measure of the mass. $\Phi_{i,j}$ represents the cosine similarity between sentences i and j, as earlier presented. For completeness:

$$\Phi_{i,j} = \cos(\theta) = \frac{s_i^{ENC} \cdot s_j^{ENC}}{\left\|s_i^{ENC}\right\| \left\|s_j^{ENC}\right\|} \tag{6-16}$$

where the superscript *ENC* denotes that these concern the encodings of the sentences. The summation of all elements in row i should represent a measure of how much similar experiences the agent has. $B_{i,j}$ represents an element from the bias matrix B, which will later be used to increase or decrease the association between experience i and j, and in turn lead to a changed opinion. The contradicting β part, is setup up the same way, only it uses the contradicting value $C_{i,j}$ of the inference. Note that here two experiences can have a supporting and contradicting relation at the same time.

For the implicit opinion, the direct relation between the experiences and the topic are defined as follows

$$e_i^t = \max\Big(0, E_{i,t} - C_{i,t}\Big) \tag{6-17a}$$

$$c_i^t = \max\Big(0, C_{i,t} - E_{i,t}\Big) \tag{6-17b}$$

where $E_{i,t}$ and $C_{i,t}$ represent the relation between the (premise) experience and the (hypothesis) topic. The difference between the supporting and contradicting relations are clipped at zero, such that every experience is exclusively supporting or contradicting. This is to ensure the agents can reach the full domain between [-1,1].

The explicit opinion is denoted as x^{t,t^*}, and analogue to the first approach, namely with

$$x^{t,t^*} = \mathrm{E}[\mathrm{Beta}(\alpha_{t,t^*}, \beta_{t,t^*})] \cdot 2 - 1 \in [-1, 1] \tag{6-18}$$

with

$$\alpha_{t,t^*}(\mathcal{S}, B, \Theta) = \sum_{\forall i \in \mathcal{S}} e_i^{t,t^*} m_i \tag{6-19}$$

$$\beta_{t,t^*}(\mathcal{S}, B, \Theta) = \sum_{\forall i \in \mathcal{S}} c_i^{t,t^*} m_i \tag{6-20}$$

Where m_i is defined as in Eq.6-13. Note that, t and t^* are interchangeable as well. For this explicit opinion, as mentioned, instead of the inferred contradiction with respect to the topic t, the entailing relation with respect to the conjugate topic t^* is taken. So, instead of $C_{t,i} = \mathrm{P}(\text{contradiction } |s_i, t)$, this approach uses $C_{t,i} = \mathrm{P}(\text{entailment } |s_i, t^*)$.

This results in the following definitions of the direct relations:

$$e_i^{t,t^*} = \max\left(0, \left(\Phi_{i,t} E_{i,t}\right) - \left(\Phi_{i,t^*} E_{i,t^*}\right)\right) \tag{6-21a}$$

$$c_i^{t,t^*} = \max\left(0, \left(\Phi_{i,t^*} E_{i,t^*}\right) - \left(\Phi_{i,t} E_{i,t}\right)\right) \tag{6-21b}$$

The conjugate information is also used as measure of relatedness with $\Phi_{i,t}$ and Φ_{i,t^*}. These represent the cosine similarity between the experience i and the topic. Note that, just as the symmetry shown on the right side in Figure 6-7:

$$e_i^{t,t^*} \hat{=} c_i^{t^*,t} \qquad \text{and} \qquad e_i^{t^*,t} \hat{=} c_i^{t,t^*}$$

Note that although the bias could change the weighing of individual experiences, and with more similar and supporting experiences they gain mass, the direct relation between the topic and the experiences does not change. Besides that this is a simplification, and that for example updating the parameters for each agent individually is a research topic further down the line; in terms of cultural perception it is not a senseless model assumption. Suppose someone does not want to get vaccinated against the corona virus, because the person believes it is a way for the government to chip all of its civilians. Here the negative relation between getting vaccinated and receiving an unwanted micro-chip is not illogical. If it was indeed the case, a lot fewer people would probably be willing to get the vaccine. So, it is no so much the argument (the direct relation) that is perceived differently; it is the belief in the existence of the argument that makes that the majority of the people think otherwise. This is precisely what the proposed belief system describes.

Language model used

The proposed belief system uses two types of pre-trained language models. For the inference task, a cross encoder architecture is used which directly takes in the sentences as strings. For encoding and similarity, a sentence transformer is used that is pre-trained on semantic similarity. Both models are decedents from the BERT model introduced in the literature review. The inference model is taken from *nli-distilroberta-base*[2]. The semantic similarity model is taken from *stsb-roberta-large*[3]

[2]Pre-trained, taken from the HuggingFace Sentence Transformer repository.
[3]Pre-trained, taken from the HuggingFace Sentence Transformer repository.

6-3 Numerical toy example

As this example considers a single agent, their is no script denoting the agent.

Agent initialization with experiences and neutral bias:

$$\mathcal{S} \leftarrow \begin{Bmatrix} \text{The car has a great engine} \\ \text{The car seats are not comfortable} \end{Bmatrix} \qquad B = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

Through equations 6-14 and 6-16, the following relations and similarities are obtained.

$$E = \begin{bmatrix} 9.8e-1 & 6.3e-3 \\ 1.1e-3 & 9.8e-1 \end{bmatrix} \qquad C = \begin{bmatrix} 4.4e-4 & 8.9e-1 \\ 9.3e-1 & 5.0e-4 \end{bmatrix} \qquad \Phi = \begin{bmatrix} 1 & 0.52 \\ 0.52 & 1 \end{bmatrix}$$

The resulting mass of the experiences is

$$\begin{aligned} m_1 &= \mathrm{E}\Big[\mathrm{Beta}\Big(E_{1,:} \cdot \Big(B_{1,:} \sum_{j=1}^{n} \Phi_{1,j}\Big), C_{1,:} \cdot \Big(B_{1,:} \sum_{j=1}^{n} \Phi_{1,j}\Big)\Big)\Big] \\ &= \mathrm{E}\left[\mathrm{Beta}\left(\begin{bmatrix} 9.8e-1 \\ 6.3e-3 \end{bmatrix} \cdot \begin{bmatrix} (0.5)(1.52) \\ (0.5)(1.52) \end{bmatrix}, \begin{bmatrix} 4.4e-4 \\ 8.9e-1 \end{bmatrix} \cdot \begin{bmatrix} (0.5)(1.52) \\ (0.5)(1.52) \end{bmatrix} \right)\right] \\ &= 0.523 \end{aligned}$$

And in similar fashion:

$$m_2 = \mathrm{E}\left[\mathrm{Beta}\left(\begin{bmatrix} 1.1e-3 \\ 9.8e-1 \end{bmatrix} \cdot \begin{bmatrix} (0.5)(1.52) \\ (0.5)(1.52) \end{bmatrix}, \begin{bmatrix} 9.3e-1 \\ 5.0e-4 \end{bmatrix} \cdot \begin{bmatrix} (0.5)(1.52) \\ (0.5)(1.52) \end{bmatrix} \right)\right] = 0.512$$

Define the topic and its conjugate, and calculate the inferred relation and the cosine similarities with respect to t and t^*:

$$\begin{matrix} t = \text{The car is good} \\ t^* = \text{The car is not good} \end{matrix} \qquad f(s_1, t, \Theta) = \begin{bmatrix} C_{1,t} \\ E_{1,t} \\ N_{1,t} \end{bmatrix} = \begin{bmatrix} 7.2e-4 \\ 0.92 \\ 7.5e-2 \end{bmatrix}$$

And in similar fashion:

$$f(s_2, t, \Theta) = \begin{bmatrix} 0.98 \\ 4.5e-3 \\ 1.7e-2 \end{bmatrix} \qquad f(s_1, t^*, \Theta) = \begin{bmatrix} 0.99 \\ 8.3e-4 \\ 2.1e-3 \end{bmatrix} \qquad f(s_2, t^*, \Theta) = \begin{bmatrix} 5.4e-4 \\ 0.97 \\ 3.3e-2 \end{bmatrix}$$

The cosine similarity is determined through Eq. 6-16, after they are encoded by the language model.

$$\Phi_{1,t} = 0.92 \qquad \Phi_{1,t^*} = 0.67 \qquad \Phi_{2,t} = 0.53 \qquad \Phi_{2,t^*} = 0.73$$

From here, the direct relations with respect to t and t^* can be calculated. These are determined through Equation 6-17:

$$\begin{aligned} e_1^t &= \max\Big(0, E_{1,t} - C_{1,t}\Big) \quad = \max\Big(0, 0.92 - 7.2e-4\Big) \\ &= 0.92 \end{aligned}$$

In similar fashion:

$$e^{t} = \begin{bmatrix} 0.92 \\ 0 \end{bmatrix} \qquad c^{t} = \begin{bmatrix} 0 \\ 0.97 \end{bmatrix} \qquad e^{t^*} = \begin{bmatrix} 0 \\ 0.92 \end{bmatrix} \qquad c^{t^*} = \begin{bmatrix} 0.96 \\ 0 \end{bmatrix}$$

Also the explicit relations can be determined through Eq. 6-21:

$$\begin{aligned} e_1^{t,t^*} &= \max\left(0, \left(\Phi_{i,t} E_{i,t}\right) - \left(\Phi_{i,t^*} E_{i,t^*}\right)\right) \\ &= \max\left(0, (0.92)(0.92) - (0.67)(8.3e-4)\right) \\ &= 0.85 \end{aligned}$$

$$e^{t,t^*} = \begin{bmatrix} 0.85 \\ 0 \end{bmatrix} \qquad c^{t,t^*} = \begin{bmatrix} 0 \\ 0.70 \end{bmatrix}$$

Then, the probability of belief in t can be determined through Eq. 6-10 and 6-18:

$$\begin{aligned} x^{t} &= \mathrm{E}[\mathrm{Beta}(\alpha_t, \beta_t)] \cdot 2 - 1 \\ &= \mathrm{E}[\mathrm{Beta}(0.92)(0.52) + (0)(0.52), (0)(0.52) + (0.97)(0.52))] * 2 - 1 \\ &= -0.016 \end{aligned}$$

In similar fashion:

$$x^{t^*} = -0.025 \qquad x^{t,t^*} = 0.10 \qquad x^{t^*/t} = -0.10$$

As expected, both implicit and explicit opinion are close to neutral. Now suppose the agent receives the following information or experience: 'The power of a car is most important'. In similar fashion as calculated above, the belief in the same topic changes as follows.

$$\mathcal{S} \leftarrow \left\{ \begin{array}{c} \text{The car has a great engine} \\ \text{The seats are not comfortable} \\ \text{The power of a car is most important} \end{array} \right\} \qquad \rightarrow \qquad \begin{bmatrix} x^{t} \\ x^{t^*} \\ x^{t,t^*} \\ x^{t^*,t} \end{bmatrix} = \begin{bmatrix} 0.34 \\ -0.57 \\ 0.41 \\ -0.41 \end{bmatrix}$$

And one more new experience: 'The car is fantastic in the corners'.

$$\mathcal{S} \leftarrow \left\{ \begin{array}{c} \text{The car has a great engine} \\ \text{The seats are not comfortable} \\ \text{The power of a car is most important} \\ \text{The car is fantastic in the corners} \end{array} \right\} \qquad \rightarrow \qquad \begin{bmatrix} x^{t} \\ x^{t^*} \\ x^{t,t^*} \\ x^{t^*,t} \end{bmatrix} = \begin{bmatrix} 0.71 \\ -0.79 \\ 0.75 \\ -0.75 \end{bmatrix}$$

Now suppose with that same experience, the topic is changed to explicitly question the power of the car in a popular manner.

$$\begin{array}{c} t = \text{Loving the power} \\ t^* = \text{Hating the power} \end{array} \qquad \rightarrow \qquad \begin{bmatrix} x^{t} \\ x^{t^*} \\ x^{t,t^*} \\ x^{t^*,t} \end{bmatrix} = \begin{bmatrix} \text{-0.02} \\ 0.17 \\ 0.73 \\ \text{-0.73} \end{bmatrix}$$

This is a nice example of the needed opinion 'scope' as discussed in the intuition. The implicit opinion is not in line with the expectation (indicated by green and red). The explicit opinion expression does show what is expected.

6-4 Static testing and validation

In this section, the goal is to empirically inspect and validate the proposed belief systems on numerical experiments with an expected outcome. In this context, static testing means that no dynamic processes yet play a part, and that the time invariant state (of belief) is tested. The agents will be confronted with text, which as discussed is defined in 1024 dimensions where a single spacing in a sentence could lead to a different meaning. So besides validating what can be validated, a parallel goal is to get to know the behaviour of the belief system. Here for, the implicit and explicit belief with respect to both a topic and a conjugate topic is included.

The proposed belief system is generic, in the sense that it could be fitted into different paradigms of opinion modeling. As such, the experiments outlined in the following section are generic as well. In the next section, when the dynamic processes are proposed, a more concrete paradigm is chosen in order to validate the entirety of the system.

6-4-1 Experiment data

To make any conclusion with respect to the opinion behaviour of the agent, the text presented to it, should at least be able to 'correctly' be interpreted by human inspection. Considered were among others Twitter and Reddit data. However, on inspection, the majority of the textual data was very ambiguous and littered with hashtags and mentions that is needed context to induce what is meant. Social media texts, especially scraped, appear to have very few information in them. Most of the clear and sound messages expressing an opinion on the platforms, are long and elaborated texts, which are not suited to the current setup of the agent. Although, later on social media data might be of interest, for these initial experiments, a data set is used from [47]. It is a collection of reviews about products and features of those products. Mainly two sets are used: (1) reviews with respect to car seats, (2) reviews with respect to hotel service. The majority of the reviews is about the specific feature, but not exclusively. Adding a touch of randomness. Examples of reviews are:

> There is a great deal of road noise in the cabin and the seats are very low quality
>
> We appreciated the elegant service and manner of the staff

To create an expectation of belief when presenting data to the agent, the reviews are split into positive and negative with sentiment classification[4]. The sentiment split is binary based, i.e. the inference task could find contradicting and entailing parts in both sentiment lists. The resulting data sets are: 44 positive and 37 negative car seat reviews, and 114 positive and 84 negative hotel service reviews.

Note the difference between general sentiment classification and a supporting or contradicting argument. For the remainder of the thesis, when referring to positive or negative reviews, the on sentiment classified positivity or negativity review is implied. Within the paradigm of opinion and the proposed belief system, the relations are referred to as supporting (entailing) or contradicting.

[4]Pre-trained taken from the Hugging Face repository (*sentiment-roberta-large-english*).

6-4-2 Expected belief

In this first experiment, the agents are provided with only positive or negative reviews such that there is a clear expectation on the resulting belief. To (partially) neglect the influence of the content in the experiences, the belief output is averaged over 10 random draws from the respective data set. Averaging over more iterations does not show significant differences. For each random draw of experience, the belief system is evaluated on both the topic t and the conjugate t^*. The results are shown in the Table 6-2. The column names refer to the car/hotel reviews as C/H, the P/N between brackets denotes whether the agent is positively or negatively initiated. The topics and conjugate topics are chosen to be:

C_1	The seats in the car are good
C_1^*	The seats in the car are not good
H_1	The service in the hotel is good
H_1^*	The service in the hotel is not good

Table 6-1: Used topics. Asterisk refers to the conjugate topic.

	Q.	$C_1[P]$	$C_1^*[P]$	$C_1[N]$	$C_1^*[N]$	$H_1[P]$	$H_1^*[P]$	$H_1[N]$	$H_1^*[N]$
x^t	1	0.998	-1.000	-0.613	0.985	0.748	-0.804	-0.239	-0.191
	2	0.958	-0.759	-0.918	0.750	0.992	-0.786	-0.232	-0.021
	5	0.799	-0.823	-0.858	0.694	0.990	-0.940	-0.465	0.492
	10	0.919	-0.939	-0.922	0.772	0.995	-0.997	-0.691	0.329
	20	0.916	-0.966	-0.915	0.783	0.995	-0.989	-0.648	0.446
	30	0.950	-0.976	-0.936	0.821	0.988	-0.984	-0.769	0.636
x^{t,t^*}	1	0.994	-0.994	-0.804	0.804	0.806	-0.806	-0.200	0.200
	2	0.919	-0.919	-0.769	0.769	0.796	-0.796	-0.763	0.763
	5	0.812	-0.812	-0.846	0.846	0.913	-0.913	-0.696	0.696
	10	0.916	-0.916	-0.896	0.896	0.987	-0.987	-0.671	0.671
	20	0.920	-0.920	-0.890	0.890	0.973	-0.973	-0.669	0.669
	30	0.944	-0.944	-0.916	0.916	0.964	-0.964	-0.799	0.799

Table 6-2: Attitude results for the implicit and explicit opinion. On topics given in 6-1.

The results are shown in Table 6-2. Firstly, clearly both agents reflect the belief that is expected from their experience. This gives ample confidence to say that the few results that deviate from what is expected (x^t with little experience subject to $H_1^*[N]$), are a result of the (sentiment) heuristic with which the data sets were set up. Furthermore, as expected, the results from the explicit opinion x^{t,t^*} are symmetric between the topic and conjugate topic.

6-4-3 Belief consistency

This second test concerning belief consistency, is about testing whether the agent also beliefs in a paraphrased formulation of the topic, having approximately the same meaning. It provides a first insight into the influence of the topic formulation in combination with the

given experience of the agent. For small amounts of experience, the results are clearly less consistent so only 10, 20 and 30 experiences are given. The data is again randomly picked and averaged over 10 iterations. The considered topics are shown below.

H_1	The service in the hotel is good
H_1^*	The service in the hotel is not good
H_2	The hotel staff is nice
H_2^*	The hotel staff is unfriendly
H_3	The service is lovely
H_3^*	The service awful

Table 6-3: Used topics. Asterisk refers to the conjugate topic.

	Q	$H_1[P]$	$H_1^*[P]$	$H_2[P]$	$H_2^*[P]$	$H_3[P]$	$H_3^*[P]$
x^t	10	0.991	-0.969	0.901	-0.999	0.952	-1.000
	20	0.994	-0.982	0.871	-0.998	0.974	-0.999
	30	0.979	-0.968	0.899	-0.999	0.943	-0.997
x^{t,t^*}	10	0.922	-0.922	0.997	-0.997	0.997	-0.997
	20	0.956	-0.956	0.998	-0.998	0.993	-0.993
	30	0.929	-0.929	0.998	-0.998	0.985	-0.985

Table 6-4: Results positively initiated on hotel seats reviews, on topics 6-3.

	Q	$H_1[N]$	$H_1^*[N]$	$H_2[N]$	$H_2^*[N]$	$H_3[N]$	$H_3^*[N]$
x^t	10	-0.721	0.509	-0.814	-0.968	-0.681	-0.454
	20	-0.664	0.569	-0.702	-0.977	-0.775	-0.363
	30	-0.725	0.424	-0.662	-0.988	-0.866	-0.384
x^{t,t^*}	10	-0.770	0.770	0.131	-0.131	-0.452	0.452
	20	-0.789	0.789	0.439	-0.439	-0.482	0.482
	30	-0.710	0.710	0.537	-0.537	-0.613	0.613

Table 6-5: Results negatively initiated on hotel seats reviews, on topics 6-3.

The results shown in Tables 6-4 and 6-5. Although obvious, it shows that the topic formulation matters for the resulting belief. As the agents are designed to specifically (dis)belief in what the topic stands for, this is a fine result. Also, it gives a first insight into the ambiguousness of the opinion 'scope' earlier discussed. On closer inspection of the data, there are few to none negative reviews that explicitly talk about the friendliness of the staff. Here the difference between the two opinion expressions becomes more obvious. The output of x^t could be interpreted as not finding the staff very nice, but also not really unfriendly. The explicit opinion x^{t,t^*} however does not have that freedom, it is forced to pick a side and favors that the staff is nice, because apparently the 'evidence' suggests that. However, there is a clear difference between the negatively and positively initiated agents. The belief in a negative topic in combination with negative experience is less 'confident'. The results of the car reviews are shown in the Appendix Table A-2 and A-3, and although the difference is much less, the trend

is the same.

These are first signs that both opinion expressions are less inclined to associate negative experiences with a negative topic proposition. However, more experiments are needed to inspect whether this comes from the language model, the designed belief system or biased data sets. Later experiments will shed more light on this.

6-4-4 Conjunction fallacy

Here, a known cognitive tendency is tested, namely the conjunction fallacy. As earlier mentioned, it describes the cognitive tendency of expressing a higher belief in the conjunction of multiple 'events', then the individual events alone. The 'fallacy' comes from the statistical fact that the joint probability of multiple events is by definition smaller or equal than the probability of the events themselves. In other words, for two events A and B:

$$\Pr(A \wedge B) \leq \Pr(A), \;\; \Pr(A \wedge B) \leq \Pr(B).$$

In terms of text, the expectation is that when the agent has more information to associate with, the belief will be higher. To check this expectation, the following topics are considered. The experiences are again randomly drawn and averaged over 10 iterations.

H_1	The service in the hotel is good
H_1^*	The service in the hotel is not good
H_2	The rooms are clean and the service is good
H_2^*	The rooms are not clean and the service is not good
H_3	The hotel staff provides good service and is friendly
H_3^*	The hotel staff provides bad service and is unfriendly

Table 6-6: Used topics. Asterisk refers to the conjugate topic.

	Q	$H_1[N]$	$H_1^*[N]$	$H_2[N]$	$H_2^*[N]$	$H_3[N]$	$H_3^*[N]$
x^t	10	-0.658	0.422	-0.896	0.277	-0.800	-0.881
	20	-0.759	0.548	-0.924	0.415	-0.884	-0.936
	30	-0.718	0.577	-0.867	0.413	-0.878	-0.905
x^{t,t^*}	10	-0.645	0.645	-0.721	0.721	-0.124	0.124
	20	-0.782	0.782	-0.807	0.807	0.072	-0.072
	30	-0.723	0.723	-0.718	0.718	0.016	-0.016

Table 6-7: Results of the hotel reviews, negatively initialized. Using topics 6-6.

To show the behaviour of this cognitive tendency, the agents should show an increase in (dis)belief for topics 2 and 3, compared to 1. As can be seen in Table 6-7 and 6-8, some results show this behaviour and some do not. On closer inspection, and with multiple runs, it appeared that the agents only show the wanted behaviour if they have an experience to associate the 'event' with. So, the agents will only show more (dis)belief in topic 2 when they have experiences about the cleanness of the room. Thus, the agents do show a form of the conjunction fallacy, however it is not implicitly through the language model, but explicitly through the experience.

	Q	$H_1[P]$	$H_1^*[P]$	$H_2[P]$	$H_2^*[P]$	$H_3[P]$	$H_3^*[P]$
x^t	10	0.976	-0.955	0.401	-0.894	0.914	-1.0
	20	0.992	-0.983	0.598	-0.964	0.928	-1.0
	30	0.995	-0.990	0.658	-0.984	0.944	-1.0
x^{t,t^*}	10	0.863	-0.863	0.519	-0.519	0.996	-0.996
	20	0.946	-0.946	0.773	-0.773	0.997	-0.997
	30	0.982	-0.982	0.895	-0.895	0.997	-0.997

Table 6-8: Results of the hotel reviews, positively initialized. Using topics 6-6.

6-4-5 Neutrality

When evaluating an agent on its belief, besides for and against, it must also be able to express neutrality or indifference. The question on whether 'true' neutrality exists can be ignored, as this experiment is to get to know how the output of the agent must be interpreted to infer a neutral standpoint. Although, both the implicit and explicit belief output is defined over the domain $[-1, 1]$, it remains to be seen if neutrality is expressed by 0. Specifically, we are interested in two possibilities: (1) When there is no relation at all between topic and experience, (2) when there is equal contradicting and supporting experience.

The first situation is tested as follows. The agent is given randomly picked car reviews, and is evaluated on the arbitrary topics: 'The dog is cute' (t_1), 'The dog is not cute' (t_1^*), and 'The floor is lava' (t_2), 'The floor is wood' (t_2^*). The results shown below.

	t_1	t_1^*	t_2	t_2^*
x^t	-0.999	-0.999	-0.999	-0.999
x^{t,t^*}	-0.993	0.993	-0.999	0.999

Table 6-9: Results on unrelated topics.

A clear trend can be concluded from the results. When presented with two completely unrelated sentences, the language model is inclined to infer it as more contradicting than entailing. This explains why for every topic, the implicit opinion has a very consistent disbelief. The explicit opinion seems unjustly highly confident in its belief. It appears that only the implicit belief is able to express indifference with respect to an unrelated topic, and both the topic and conjugate are needed. The negative bias towards unrelated topics is mostly due to 'wrong' inference outputs. Although the used language models are close to state of the art, they are not perfect at all. Figure 6-10 below shows an example, made with the used language model.

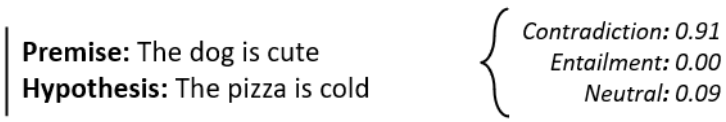

Figure 6-10: Example of 'wrong' (unwanted) natural language inference.

The second situation is where there is equal supporting and contracting experience. Earlier it could already be seen that the positive and negative data sets might not strictly be supporting or contracting. The same is observed when the agents are given an equal amount of positive and negative reviews; the agents consistently believe in the positive topic formulation. In order to come close to a neutral standpoint, double the negative reviews had to be presented to the agents. The considered topics are shown below. The idea is to check how the agents respond to the earlier used general topic formulation between 'good' and 'not good', a less convincing formulation as 'OK' and 'less then OK', and a neutral formulation. Though, the concept of a proposition does not lend itself easily to expressing neutrality, unless it proposes neutrality. The topics considered are shown below.

C_1	The seats in the car are good
C_1^*	The seats in the car are not good
C_2	The seats in the car are OK
C_2^*	The seats in the car are less then OK
C_3	I am neutral about the seats in the car
C_3^*	I am not neutral about the seats in the car

Table 6-10: Used topics. Asterisk refers to the conjugate topic.

	Q.	C_1	C_1^*	C_2	C_2^*	C_3	C_3^*
x^t	10	-0.404	-0.376	-0.601	-0.810	-0.971	0.981
	20	-0.488	-0.380	-0.563	-0.835	-0.950	0.990
	30	-0.424	-0.500	-0.475	-0.788	-0.955	0.989
x^{t,t^*}	10	0.021	-0.021	0.356	-0.356	-1.0	1.0
	20	-0.065	0.065	0.415	-0.415	-1.0	1.0
	30	0.014	-0.014	0.431	-0.431	-1.0	1.0

Table 6-11: Results car reviews w.r.t. topics in 6-10, initialized with both positive and negative experiences.

The results show that for neutrality in this second situation, both opinion expressions show an indifferent standpoint. Also, the difference between the considered topics is also as expected. For the first, neutrality should be between 'good' and 'not good', as it does. For the second, both x^t and x^{t,t^*} show a higher belief in a more neutrally formulated topic. For the third, actually none of the experiences are neutrally formulated, so although this output might not be how a person would respond, it is a logic result.

The results in Table A-6 show that the belief system does not resemble the concept of the wet mind[5], as the experiences individually add up to the equation, they do not 'blend'.

6-4-6 Influence of experience

Here a closer look is taken into how exactly the experiences influence the opinion. Through the earlier proposed framework, an experience gains mass with the addition of similar and

[5] As mentioned in the section 2-4-2.

supporting experiences. Also, an experience loses mass with the addition of contradicting experiences. The direct relation of the individual experiences with respect to the topic at hand does not change.

To show the influence of the individual experiences, the agent is initialized with 5 negative car reviews and supplemented with 15 positive. With every additional experience, multiple elements are evaluated. Firstly, the belief with respect to the topic 'The car seats are good'. Secondly, the alpha and beta (from Eq. 6-19, 6-20, 6-8 and 6-9) that parameterize the Beta function are set out. Lastly, the weight of the individual experiences, divided by the maximum at each time step. Recall the difference between the implicit and explicit expressions, namely only the direct relation is calculated differently. So the mass of the experiences is equal for both expressions, but can change with every additional experience as it a function of the experiences. The direct relation is the same over all time.

In Figure 6-11 the belief and the alpha and beta values for both belief expressions is shown. Regions of interest here are the initially decreasing difference between alpha and beta, the point of intersection where the belief is neutral, and the increasing difference between both lines. This already clearly shows how the addition of positive reviews, besides the obvious increase of the alpha value, also decreases the belief in the contradicting beta value. That the intersection happens nicely around the point where there are as many positive as negative experiences, is mostly a coincidence. As was discussed in earlier experiments.

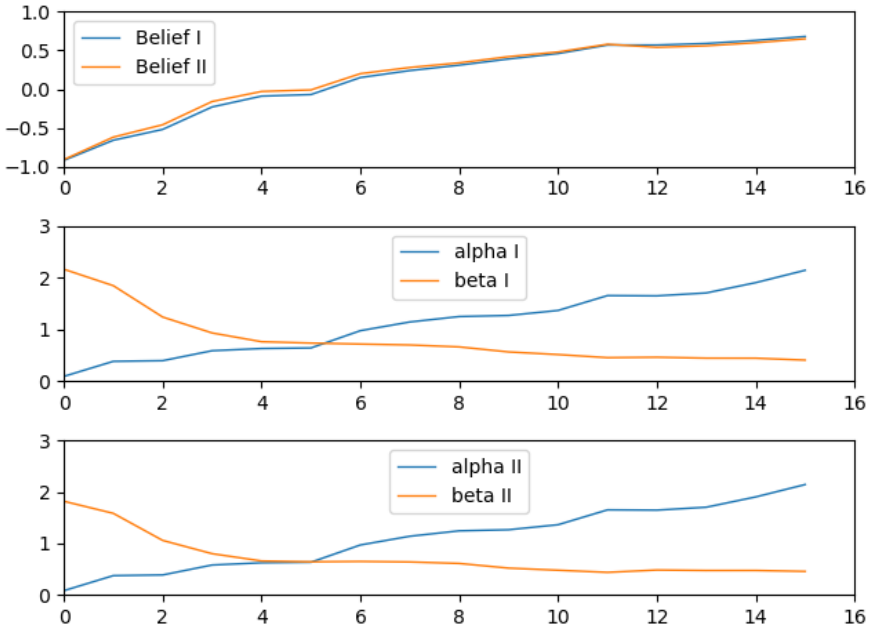

Figure 6-11: Values for α_t and β_t of the implicit opinion (I), α_{t,t^*} and β_{t,t^*} of the explicit opinion (II).

Figure 6-12 shows on the left the mass of the experiences normalized at each addition. The initial 5 negative experiences are the 5 blocks on the left. It clearly shows how with the addition positive experiences, the negative become less influential, as the belief in those decreases. The right of the figure shows the direct relation for the implicit opinion (0) and the explicit (1). There is only a marginal difference between them as expected from Figure

6-11, and also as these topics have shown to be consistent. In Appendix Table A-1 and A-2, two similar figures are shown. The agents are given the same experience, but evaluated on a more ambiguous topic. Again it can be seen that the explicit expression is able to isolate the expected belief, while the implicit opinion does not manage at all.

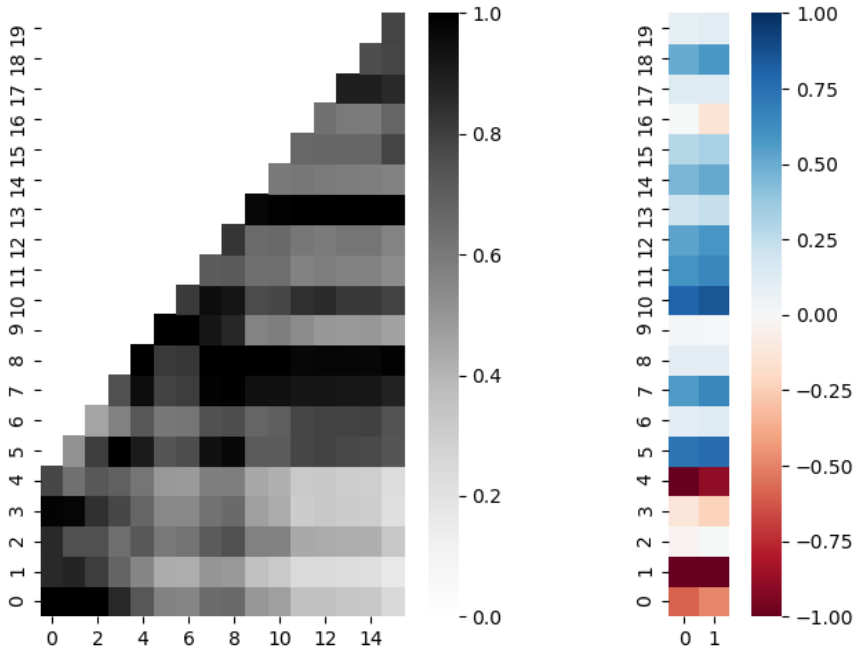

Figure 6-12: Left: mass of each experience. Right: the direct relation of the experience; (0) represents x^t, (1) x^{t,t^*}. See explanation in text.

6-4-7 Bias

Both opinion expressions quite consistently favor the positive topic over the negative. It has already been concluded that the negative reviews are less negative, than the positive reviews are positive, or at least that is how they are interpreted by the language models. As earlier mentioned, it is precisely that bias in the models that could result in a (average English speaking) human interpretation. So although it will probably not be a perfect representation, it is chosen to be embraced.

As for the data set, it will be biased by definition; 'unbiasedness' can only be concluded by a biased measurement. Furthermore, an agent will be initialized with experiences based on the belief output, which can be steered towards the desired belief by picking different experiences at random or by heuristic strategy.

Identifying biases in a language model is inherently limited to the 'known knowns', and as such will always be incomplete. Also, the very idea is to use the unknown associations, thus exploring the biases in the language models is not part of this research. However, it would be nice to have a hypothesis with respect to the significance of topic dependence. Below the distance (cosine similarities) between the general topics and the respective product is shown. For the car, the model is slightly negatively biased. For the hotel, the model shows

significant signs of positive bias. Keeping in mind that these are only minor indications, the earlier experiments are in line with these similarities. These biases, or associations, are also implicitly used when inferring a contradicting or supporting relationship, and thus also seep through to the inference task [6].

$$\text{The car is good} \overset{0.5823}{\longleftrightarrow} \text{The car} \overset{0.5855}{\longleftrightarrow} \text{The car is bad}$$

$$\text{Good} \overset{0.1540}{\longleftrightarrow} \text{The car} \overset{0.1928}{\longleftrightarrow} \text{Bad}$$

$$\text{The hotel is good} \overset{0.6548}{\longleftrightarrow} \text{The hotel} \overset{0.5065}{\longleftrightarrow} \text{The hotel is bad}$$

$$\text{Good} \overset{0.4143}{\longleftrightarrow} \text{The hotel} \overset{0.2551}{\longleftrightarrow} \text{Bad}$$

In order to gain a better understanding of the bias, both in the data and model, the following test is performed. The agent is initialized with only positive or negative reviews, and then supplemented with respectively negative or positive reviews. The results are shown in Figures 6-13 and 6-13. The test is performed 10 times, each with random selection of both initial and supplemented experience. After each addition of a new experience, the agent is evaluated on both the topic and its conjugate. These are as before: 'The hotel service is good' and 'The hotel service is not good'.

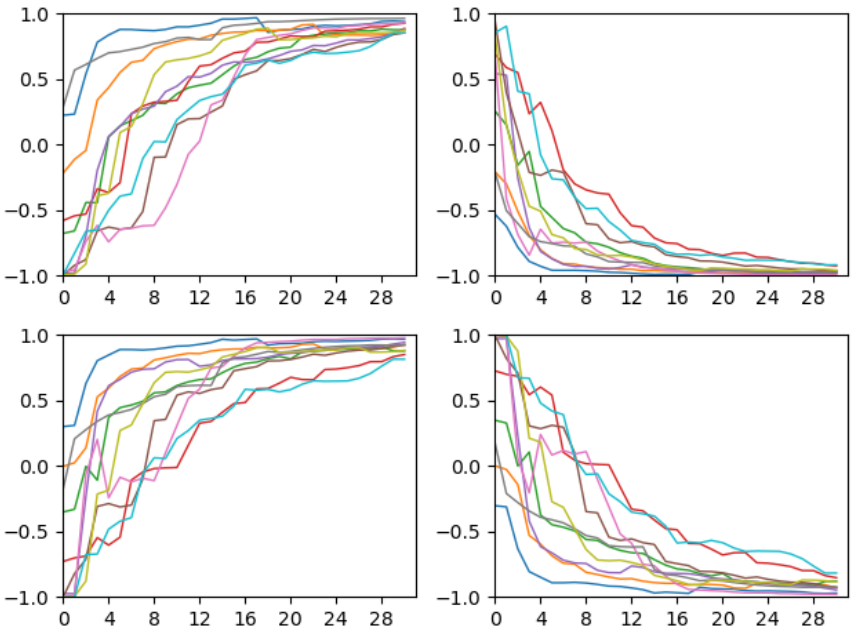

Figure 6-13: Implicit opinion (top), Explicit opinion (down). Evaluated on t (left) and t^* (right). Negatively initiated.

In multiple ways, these graphs show what is expected. Firstly, the general trend for each is logically correct. There initial beliefs are conform the experience they are initialized with, and the change in belief is conform the added experience. Secondly, the difference between the

[6] At least that is assumed, as they are different language models

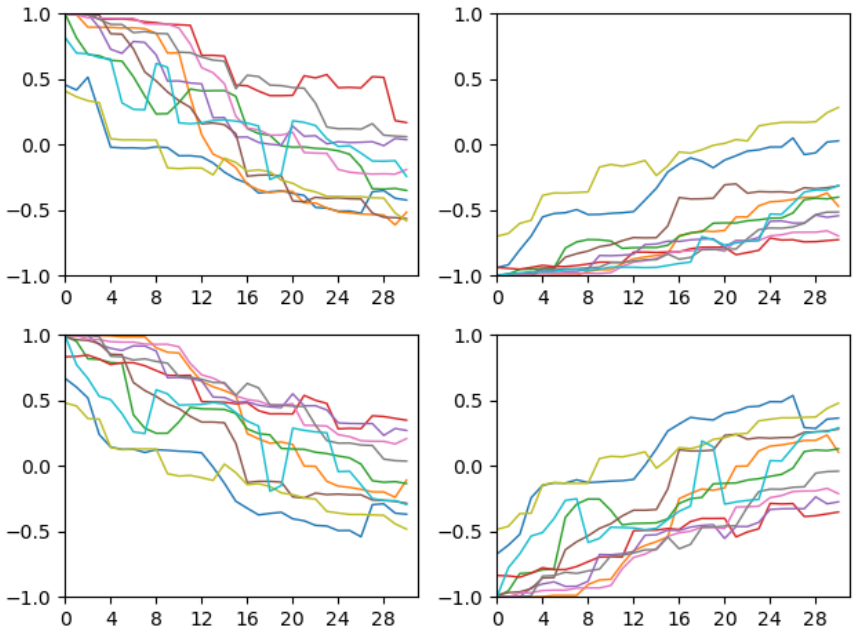

Figure 6-14: Implicit opinion (top), Explicit opinion (down). Evaluated on t (left) and t^* (right). Positively initiated.

multiple runs is entirely because of the content that is presented. This makes that the change of opinion is in not monotonic. Furthermore, the bias identified before, especially for the topic of the hotel, very clearly comes forward. The negatively initiated agent quickly becomes positive, while the positively initiated agent in some runs does not even reach neutrality. Also in line with this expectation are the results of the car topic shown in Appendix Table A-3 and A-4. They show far more symmetry with respect to positively or negatively initiated experience.

It should be noted that the bias shown in the figures cannot be exclusively attributed to a bias in the language model. It could just as well be biased towards a certain style of expression, word choice, sentence length or some completely unknown factor. However, the assumption is that the reviews to some degree match real world statements about the topic. More on the data set bias in the sections on results and discussion.

Note that the belief expressions do not yet need to be able to reach the full domain between $[-1, 1]$. These results are entirely content depended. In the next section, when dynamics are introduced the full domain will become reachable.

6-5 Conclusion

It was discussed that an opinion cannot directly be inferred from current language models. However, the tools of natural language inference and similarity can be used to approach the idea of an opinion. Through the concept of cognitive dissonance, an intuitive framework for defining the concept of an opinion is setup. From here, the belief system, the core of the

agent, is proposed. It describes how the opinion of an agent can be inferred from textual experiences using a language model.

Two opinion expressions are considered. The implicit opinion expresses the belief only with respect to a single topic proposition. The explicit opinion also uses a introduced conjugate topic. The conjugate topic describes the explicit scope for which the opinion is needed.

It is seen that both opinion expressions output results that are in line with the expectation. Both expressions by definition, and seen in the results, have different purposes. However, the explicit opinion appears better equipped to counter both ambiguous topic formulations and experiences. As such, the dynamical processes in the following section are build around the explicit opinion.

Chapter 7

Cognitive and social process design

In the previous section, a proposal was given for a definition of a belief system on the basis of textual experiences. The output of the belief system, the belief in a topic, represents the state of the agent. In this section, a definition of the processes that change that state of belief are proposed, the micro dynamics. As the parameters of the language model are taken as constant, only the bias matrix B and the experiences in $\mathcal{S}$ make up the formation and change of belief. The processes influencing those elements of an agent are divided into an intra- and inter-agent process. The former describes the cognitive tendencies that could be said to be independent of the agent's environment. In Figure 7-1 it is represented by the self loop. The inter-agent process describes when and how agents interact, represented by the arrow between the agents.

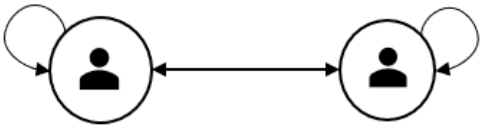

Figure 7-1: Connected agents, with self loop denoting intra-agent process.

Firstly, the intra-agent process will be proposed and described. It is build on a single known cognitive tendency, namely cognitive dissonance reduction. Secondly, the agent's social behaviour will be proposed. Following from the conclusion in the previous section, the processes will be aimed at the explicit opinion expression.

7-1 Intra-agent: Cognitive dissonance reduction

The micro dynamics are build on a single conceptual cognitive process: cognitive dissonance reduction. The idea of dissonance minimization as an agent (micro-dynamic) objective is among others researched in [24]. As previously mentioned, cognitive dissonance is defined as the unpleasant feeling one experiences when being confronted with contradicting beliefs. As

a consequence, the human brain tries reduce or minimize contradicting beliefs. It might be best described as the natural process of building a network of congruent knowledge, opinion and belief, such that one experiences as little surprise as possible with respect to ones own thought and ones environment, as mentioned in the previous Chapter.

To explain this concept, and how it can be seen as an (implicit) objective for the agent, see the analogy of the balance Figure 7-2. The topic of discussion is 'I should get vaccinated'. The weight of the experiences and the resulting angle of the balance is the state of the agent at $t = 0$ (neglecting how it has arrived here). The process of reducing cognitive dissonance is the step from $t = 0$ to $t = 1$. Just as the concept of correlated associations, the process of reducing cognitive dissonance is a solution to a chicken and egg problem. It is because the agent beliefs it should not to get vaccinated, that the experiences underlying that opinion are in time more and more believed in. While the agents belief is a function of those experiences. At the same time, the experiences contradicting the current belief loose weight.

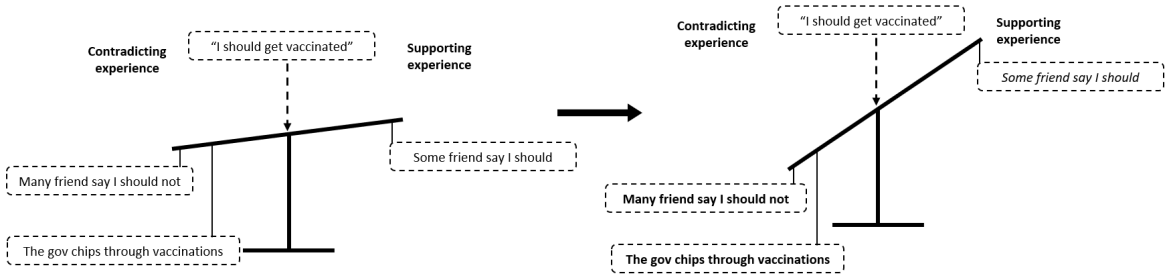

Figure 7-2: Depiction of dissonance reduction. The bold experiences are more believed in after a step of dissonance reduction. The italic experience decrease in belief mass.

A straightforward way to design this process would be to just penalize or emphasize the direct relation of the experiences with respect to the topic. However, through the concept of association, cognitive dissonance reduction can theoretically be seen as a global optimization. In other words, the relation between all the individual experiences change as well, such that a change of the relation between experiences also changes the belief in any topic proposition.

7-1-1 Proposed definition

The proposed definition for the process of cognitive dissonance reduction is as follows. Suppose an agent at a certain time has a certain bag of experiences $\mathcal{S}$, a bias matrix B and a language model parameterised by Θ. Also, suppose this results in a positive belief in topic t. For every experience $s_i \in \mathcal{S}$, the direct relation to the topic is inferred through the earlier proposed definition. If it is a supporting relation (in line with the agent's belief), the association between all other experiences ($s_j \in \mathcal{S}$) is updated such that those which contradict s_i have less influence on s_i, and those that support s_i have more influence on s_i. If the direct relation of s_i is contradicting with respect to topic t, the update of the associations is the other way around. The association or bias between experiences is captured in bias matrix B.

The change of an increasing association between two experiences s_i and s_j is defined as

$$B_{i,j} = B_{i,j} + r(1 - B_{i,j}). \tag{7-1}$$

The change of a decreasing association is defined as

$$B_{i,j} = B_{i,j} - r(B_{i,j}). \tag{7-2}$$

With $r \in (0, 1)$ the association rate, which parameterises the strength or speed with which the associations grow or diminish. An overview of the proposed algorithm is shown in Algorithm 1. Note that if the agent would explicitly disbelief topic t, and therefor explicitly belief in the conjugate topic t^*, the relations with respect to the topic swap. In other words, process will reduce belief in t and increase belief in t^*. This is represented by the first if-statement in the algorithm. $E_{i,j}$ and $C_{i,j}$ are the relations as seen in Eq. 6-14. e_i^{t,t^*} and c_i^{t,t^*} as seen in 6-8 and 6-9

Algorithm 1 Cognitive dissonance reduction.
input agent: $\{\mathcal{S}, B, \Theta\}$, t, t^*, $r \in (0, 1)$ (Association rate)

if $x^{t,t^*} - x^{t^*,t} < 0$ **then**
  $t \leftarrow t^*$
  $t^* \leftarrow t$
end if
for all $s_i \in \mathcal{S}$ **do**
  if $e_i^{t,t^*} > 0$ **then**
    for all $s_j \in \mathcal{S}$ **do**
      if $E_{i,j} - C_{i,j} < 0$ **then**
        $B_{i,j} \leftarrow B_{i,j} - r(B_{i,j})$
      else
        $B_{i,j} \leftarrow B_{i,j} + r(1 - B_{i,j})$
      end if
    end for
  else if $c_i^{t,t^*} => 0$ **then**
    for all $s_j \in \mathcal{S}$ **do**
      if $E_{i,j} - C_{i,j} <= 0$ **then**
        $B_{i,j} \leftarrow B_{i,j} + r(1 - B_{i,j})$
      else
        $B_{i,j} \leftarrow B_{i,j} - r(B_{i,j})$
      end if
    end for
  end if
end for

7-1-2 Numerical toy example

Suppose the agent has the following initial experience and bias.

$$\mathcal{S} \leftarrow \begin{Bmatrix} \text{The car has a great engine} \\ \text{The seats are not comfortable} \\ \text{The power of a car is most important} \\ \text{The car is fantastic in the corners} \end{Bmatrix} \qquad B = \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 \end{bmatrix}$$

Which, as the earlier toy example showed, has the following belief:

$$\begin{matrix} t = \text{The car is good} \\ t^* = \text{The car is not good} \end{matrix} \quad \rightarrow \quad \begin{bmatrix} x^t \\ x^{t^*} \\ x^{t,t^*} \\ x^{t^*,t} \end{bmatrix} = \begin{bmatrix} 0.54 \\ -0.99 \\ 0.54 \\ -0.54 \end{bmatrix}$$

Now the agent does one step in cognitive dissonance reduction, with respect to the denoted topic. The association rate is $r = 0.05$. The agent beliefs in the topic (and not in the conjugate), so $x^{t,t^*} - x^{t^*,t} > 0$ and t and t^* do not have to be swapped in the first if-statement. Then for s_1:

$$e^{t,t^*} = 0.85 > 0$$

So s_1 supports the topic, and should increase in (belief) mass. To achieve this, the association between all experiences that contradict s_1 in decreased, while the supporting ones are increased. So, for each sentence $j \in \mathcal{S}$, check $E_{i,j} - C_{i,j}$:

$$E_{1,:} - C_{1,:} = \begin{bmatrix} 0.98 \\ -0.89 \\ 7.4e-2 \\ 0.17 \end{bmatrix} = \begin{matrix} > 0 \\ < 0 \\ > 0 \\ > 0 \end{matrix}$$

Then for each contracting, in this case only s_2, the association is decreased towards:

$$B_{1,2} = B_{1,2} - r(B_{1,2}) = 0.5 - (0.05)(0.5) = 0.475$$

The supporting ones gain association, with s_3 as example:

$$B_{1,3} = B_{1,3} - r(B_{1,3}) = 0.5 + (0.05)(1 - 0.5) = 0.525$$

This process repeats for all i and all j, which results in the following new bias matrix, and the changed belief in the topic:

$$B = \begin{bmatrix} 0.525 & 0.475 & 0.525 & 0.525 \\ 0.525 & 0.475 & 0.525 & 0.525 \\ 0.525 & 0.475 & 0.525 & 0.525 \\ 0.525 & 0.475 & 0.525 & 0.525 \end{bmatrix} \qquad \begin{bmatrix} x^t \\ x^{t^*} \\ x^{t,t^*} \\ x^{t^*,t} \end{bmatrix} = \begin{bmatrix} 0.58 \\ -0.99 \\ 0.58 \\ -0.58 \end{bmatrix}$$

Note that the direct relation towards the topic have not changed, only the belief in the experiences them self. The belief masses of the experiences have changed as

$$m^{before} = \begin{bmatrix} 0.64 \\ 0.25 \\ 0.75 \\ 0.63 \end{bmatrix} \qquad \rightarrow \qquad m^{after} = \begin{bmatrix} 0.66 \\ 0.23 \\ 0.77 \\ 0.65 \end{bmatrix}$$

With no other influences, over time, these will become:

$$\lim_{t \to \infty} m = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

7-1-3 Testing and validating

To validate the proposal on the intra-agent process of cognitive dissonance reduction, a straightforward assumption is tested. The assumption is that if an agent in its current state is put in a room such that it can only interact with itself, the agent will believe more and more in its current ideologies. More over, the experiences underlining that belief should a have more and more influence on the topic at hand and related subjects.

Firstly, a single topic is considered. An agent is initialized with an equal amount (10) of randomly drawn positive and negative reviews. As seen before, because the belief is inferred from the actual content of the experiences, chances are very small that the initial belief is exactly zero. The agent takes 50 steps of cognitive dissonance reduction, through the algorithm outlined above, with $r = 0.1$. The results are shown below in Figure 7-3.

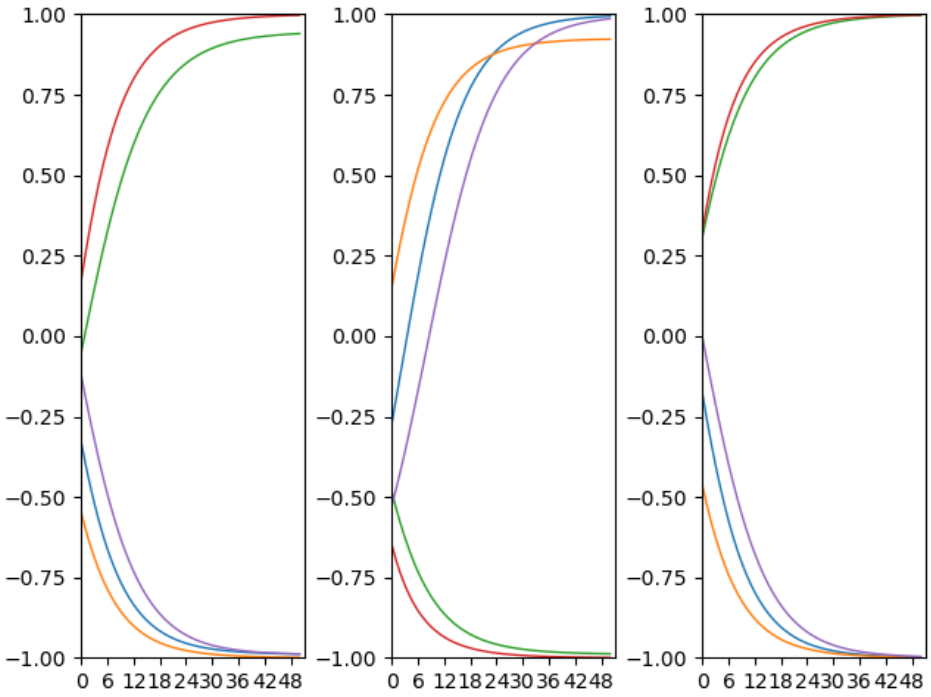

Figure 7-3: Left: implicit opinion wrt t. Middle: implicit opinion wrt t^*. Right: explicit opinion wrt t.

As expected, none of the agents start from the same point and the beliefs evolve as expected. As mentioned, the dissonance reduction is performed with respect to the explicit opinion. This explains the difference between the explicit and implicit opinions. The change of influence of the experiences of a single agent is shown in Figure 7-4 below. The influence is normalized over the columns, which represent the 20 experiences. Note that the influence can diminish to zero, but can only grow to the maximum of its direct relation to the topic. This is why, after normalizing, the influence of experience 6 does not seem to change. Red indicates contradiction, blue supporting influence. The grey color indicates influence close to zero.

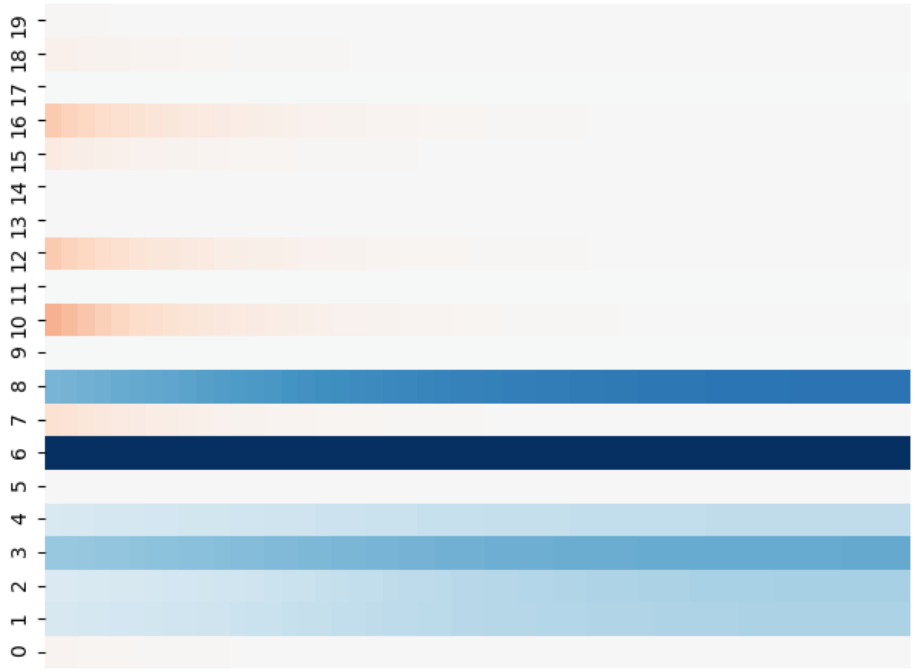

Figure 7-4: Influence of the experience over time (x-ax). Normalized by the maximum.

7-2 Inter-agent: Interaction

The design of the agent interaction can be divided into two basic questions. When do agents interact? And when they do, how do they interact? The first question is very much related to the network design choice which will be presented first. After, the actual definition of interaction is proposed. The proposal will be tested and validated in the next section, where all elements come together.

7-2-1 Network paradigm choice

There are multiple many paradigms of network design. Depending on the research topic (what kind of observed social behaviour is of interest) a certain paradigm can be chosen to fit the requirements. These agent-based paradigms can roughly be divided into two groups of networks: fixed or not fixed topology. Both are concerned with the question of when agents interact. When the connections between agents are not fixed, a rule or definition is setup that defines when agents interact. As mentioned in the literature review, a classic example is so-called bounded confidence. Agents will interact if they up to a certain point, the bound, think alike. Obviously, this goes hand in hand with a defined measure of 'thinking alike'. With a fixed topology, the agents and their connections to other agents are explicitly defined in a graph, as earlier discussed.

The proposed belief system and intra-agent process are generic in the context of these opinion model paradigms. The to be proposed definition of agent interaction will be aimed at an

undirected[1] fixed topology of connected agents. However, with minor alteration, the same elements could be used for a bounded confidence model.

7-2-2 Interaction design proposal

Within a fixed topology, agents can only interact with those they are connected with. The agents will interact with every network update. Therefor, the first question posed earlier (when do agents interact) is answered. Next will be a proposal for the actual interaction design.

The concept of cognitive dissonance and the process of minimizing the dissonance has been thoroughly discussed earlier. However, the concept can be extrapolated to social interaction as well, namely structural balance theory [26]. It describes that the same kind of internal process, namely the process of creating a congruent set of knowledge and belief (read without conflict), also appears externally, between social interactions. In other words, that one implicitly (or explicitly) has the objective to surround oneself with people that have knowledge and beliefs congruent to ones own knowledge and beliefs, such to avoid tensions and conflicting relations. This concept falls very nicely in the concept of cultural dissemination of Axelrod and the relation between language and thought, both discussed in the literature review. Belief, opinion and knowledge has influence on the language one speaks (in the broader sense), and also how that language is interpreted on the receiving end. People that better understand each other, are more likely to interact and transfer knowledge. For example, communicating an idea or thought to a best friend might be as easy as a facial expression, while for persons outside your close circle you perhaps need to elaborate in words.

This concept in its true nature is, just as earlier mentioned concept of the mind, a chicken and egg problem. People get like minded and attached when they have much interaction, and people are more likely to interact if they are like minded. In the case of the fixed topology network, the interactions are decided already. Thus, the idea is that connected agent's grow to be attached to each other and become more like minded.

Having said, this concept is translated into the (directed[2]) influence one agent has over the other. It describes to what extent the opinion, and the experience underlining that opinion, of one agent is shared to the other agent. Also, to what extent that opinion is 'taken in' by the other agent. The definition for this coined influence factor will be introduced first. After, the actual interaction design is proposed.

Influence factor

For this proposal, we consider the influence of agent a_i to agent a_j. The design of the influence factor, is based on three factors: (1) a measure of directed like-mindedness as perceived by a_j, (2) a measure of (self)confidence of agent a_j in its own opinion, and (3) a measure of authority of a_i. The three elements and the resulting influence factor are described below.

Like-mindedness

[1] Interaction between connected agents goes both ways.

[2] Directed means two connected agents can have different influence on each other.

The measure of like-mindedness is defined as follows:

$$\mathrm{lm}_{i\leftrightarrow j} = \max\left(\frac{\sum_{\forall s\in(\mathcal{S}_i\cap\mathcal{S}_j)}|m_s^i - m_s^j|}{\max(\mathrm{n}(\mathcal{S}_i\cap\mathcal{S}_j),1)}, 1e-4\right) \in [1e-4, 1] \tag{7-3}$$

The numerator in the fraction describes the absolute error between all common experiences between agent A and B. The denominator divides this by the number of common experiences, such that it lies between 0 and 1. When there are no common experiences (yet), the max operator in the denominator ensures it is not divided by 0. The max operator over the fraction and $1e-4$ means that the agents cannot have zero like-mindedness. This is, as the double arrow subscript denotes, an undirected measure of like-mindedness. The directed (or perceived) like-mindedness that is used in the influence factor is based on this undirected measure. See Figure 7-5. Let $\mathcal{E}^i$ be the set of all connected agents to agent a_i, of which a_j

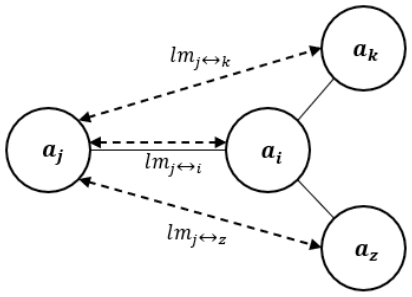

Figure 7-5: Depiction of undirected like-mindedness used for directed like-mindedness.

is part of, and $\mathcal{E}^i\backslash\{a_j\}$ be that same set excluding a_j. Then the directed like-mindedness is formulated as

$$\mathrm{dlm}_{i\to j} = \frac{\left(\sum_{\forall a\in\mathcal{E}^i\backslash\{a_j\}} g\cdot\mathrm{lm}_{a\leftrightarrow a_i}\right) + \mathrm{lm}_{i\leftrightarrow j}}{1 + g\cdot\mathrm{n}(\forall a\in\mathcal{E}^i\backslash\{a_j\})} \in [1e-4, 1] \tag{7-4}$$

Essentially, it is a weighted average of the like-mindedness between agent a_i and a_j, and agent a_j and all other connections of agent a_i. With $g\in[0,1]$. The idea is that it encourages the agents to belong to a group, rather than only their first degree connections.

Confidence

The (self) confidence of agent a_j is defined as follows.

$$\mathrm{confidence}_j = \frac{1}{2} + \frac{1}{2}\frac{\sum_{\forall k\in\mathcal{S}_j^+} m_k}{\mathrm{n}(\forall k\in\mathcal{S}_j^+)} - \frac{1}{2}\frac{\sum_{\forall k\in\mathcal{S}_j^-} m_k}{\max(\mathrm{n}(\forall k\in\mathcal{S}_j^-),1)} \in [0,1] \tag{7-5}$$

with $\mathcal{S}^+$ the set of all experiences that support the current opinion, and $\mathcal{S}^-$ the set of all experiences that contradict the current opinion. And as earlier presented in Eq. 6-13, $m_k\in[0,1]$. The term $\frac{\sum_{\forall k\in\mathcal{S}^+} m_k}{\mathrm{n}(\forall k\in\mathcal{S}^+)}$ describes a summation of all belief masses that support the current opinion, divided by the amount of those experiences. If the agent is completely confident in

its opinion, all these masses should be 1. The last part describes a summation of all belief masses that contradict the current opinion. These should be 0 for complete confidence. Note that when the agents are initialized neutrally, and the fractions are near to equal, the agent is still is confident by the that first half, which also ensures the domain $[0, 1]$. The max operator ensures that the denominator is not zero, when there are no contradicting experiences.

Authority

Finally, there is the authority of an agent. This is defined by number of agents they are connected with, the so-called degree, normalized by the maximum degree of all agents. Let maxdegree($\mathcal{E}$) be the maximum degree of all agents in the network, then

$$\text{auth}_i = \frac{\text{n}(\forall a \in \mathcal{E}^i)}{\text{maxdegree}(\mathcal{E})} \in [0, 1] \tag{7-6}$$

These elements come together in the following definition if the influence factor, defining the influence of agent a_i on agent a_j.

$$\text{IF}_{i \to j} = (1 - \text{confidence}_j) \cdot (\text{dlm}_{i \to j}) \cdot (\text{auth}_i) \in [0, 1] \tag{7-7}$$

Having defined the influence factor, next the interaction is proposed.

Interaction

As mentioned, the interaction between agents is undirected. This means that the information flow between connected agents a_i and a_j is back and forth. The interaction between connected agents is chosen to exist of two elements: (1) the possibility of sharing an experience, and (2) equalizing the belief mass for experiences agents have in common. In the final model implementation, first all agents will interact, and after all agents will update. Therefor, the interaction (the shared signals) outlined in Algorithm 2, and the agent update (the update on those signals) outlined in Algorithm 3, are separated. Though, for clarity in this proposal, the interaction and update are jointly explained.

Experience sharing is outlined in Algorithm 4. The probability that this algorithm is executed in Algorithm 2, makes that the agent with the highest influence will always share an experience. This is less certain the for the least influential of the two. Algorithm 4 for sharing an experience can be summarized as follows. A sample is drawn from the Beta distribution of agent a_i. If the sign of that sample is on the same side as the the expectation of the distribution (the current opinion), the agent will share an experience supporting its opinion. If the sign of the sample is different, the agent shares an experience that contradicts its own opinion. Then, agent a_i shares the experience from the respective set that has the largest influence on its opinion, that is not included in the set $\mathcal{S}^j$ of agent a_j.

Agent a_j updates its experience through Algorithm 5. First the new experience is included in $\mathcal{S}^j$. Then it initializes the newly added bias rows and columns in B^j, with the elements of another experience that has the largest entailing probability. The new element in B^j on the

Algorithm 2 Agent Interaction:
input a_i: $\{\mathcal{S}^i, B^i, \Theta\}$, a_j: $\{\mathcal{S}^j, B^j, \Theta\}$, t, t^*.

$\text{IF}_{i\to j} \leftarrow$ Equation 7-7
$\text{IF}_{j\to i} \leftarrow$ Equation 7-7
interactions[i] (saved interactions for update of agent a_i)
interactions[j] (saved interactions for update of agent a_j)
Algorithm 4 $\leftarrow (a_i, a_j, t, t^*)$ with probability $\frac{\text{IF}_{i\to j}}{\max(\text{IF}_{i\to j}, \text{IF}_{j\to i})}$
interactions[j] $\leftarrow$ **Algorithm** 4
Algorithm 6 $\leftarrow (a_i, a_j, t, t^*)$
interactions[j] $\leftarrow$ **Algorithm** 6
Algorithm 4 $\leftarrow (a_j, a_i, t, t^*)$ with probability $\frac{\text{IF}_{j\to i}}{\max(\text{IF}_{i\to j}, \text{IF}_{j\to i})}$
interactions[i] $\leftarrow$ **Algorithm** 4
Algorithm 6 $\leftarrow (a_j, a_i, t, t^*)$
interactions[i] $\leftarrow$ **Algorithm** 6

Algorithm 3 Agent Update:
a_i: $\{\mathcal{S}^i, B^i, \Theta\}$, interactions[$i$]

Algorithm 5 $\leftarrow (a_j, s_z, \text{IF}_{j\to i}) \leftarrow$ interactions[i]
Algorithm 7 $\leftarrow (a_j, s_{z^+}, s_{z^-}, \text{IF}_{j\to i}) \leftarrow$ interactions[i]

diagonal (basically the 'self bias' of the experience) is filled with the influence factor. This update resembles a specific cognitive tendency, namely biased assimilation [3] [22].

The second element of interaction is outlined in Algorithm 6. Agent a_j updates the bias matrix for two experiences that both agents have in common. The first common experience is one that supports the opinion of agent a_i, for which the belief mass has the largest difference compared to agent a_j. This bias towards this experience is updated such that agent a_j will be believe more in the experience. This update is outlined in Algorithm 7. The process is analogue to the concept of cognitive dissonance reduction outlined in Algorithm 1. Here however, it concerns the update of the bias matrix for a single experience. Also, instead of the association rate, it is updated with the influence factor multiplied by a scaler[4]. The second common experience is one that contradicts the opinion of agent a_i. It follows the same steps, only the bias of agent a_j is updated towards lesser belief in the experience.

[3] Biased assimilation means that (newly) observed information is interpreted and assimilated (included) through the pre-existing bias.

[4] It was found through experiments that scaling the influence factor increases the consistency of the model. More on this in the next section.

Algorithm 4 Experience sharing $a_i \rightarrow a_j$
input a_i: $\{\mathcal{S}^i, B^i, \Theta\}$, a_j: $\{\mathcal{S}^j, B^j, \Theta\}$, t, t^*
output a_j, s_z

$x_i^{t,t^*} \leftarrow \mathrm{E}[\mathrm{Beta}(\alpha_{t,t^*}(\mathcal{S}^i, B^i, \Theta), \beta_{t,t^*}(\mathcal{S}^i, B^i, \Theta))] \cdot 2 - 1$
$X_i^{t,t^*} \leftarrow X_i^{t,t^*} \sim \mathrm{Beta}(\alpha_{t,t^*}(\mathcal{S}^i, B^i, \Theta), \beta_{t,t^*}(\mathcal{S}^i, B^i, \Theta)) \cdot 2 - 1$
$\mathrm{IF}_{i \rightarrow j}$
if $X_i^{t,t^*} >= 0$ **then**
  if $x_i^{t,t^*} >= 0$ **then**
    $p \leftarrow 1$
  else
    $p \leftarrow 0$
  end if
else
  if $x_i^{t,t^*} >= 0$ **then**
    $p \leftarrow 0$
  else
    $p \leftarrow 1$
  end if
end if
if $p = 1$ **then**
  $\mathcal{S}^{+,i} \leftarrow$ experiences that support current opinion of a_i
  $\mathcal{S}^{+,i-j} \leftarrow \forall s \in \mathcal{S}^{+,i} \not\subset \mathcal{S}^j$
  if $\mathrm{n}(\mathcal{S}^{+,i-j}) > 0$ **then**
    $s_z \leftarrow \max(|m_s(e_s^{t,t^*} - c_s^{t,t^*})| \; \forall s \in \mathcal{S}^{+,i-j})$
    output $\leftarrow (a_j, s_z)$
  end if
else
  $\mathcal{S}^{-,i} \leftarrow$ experiences that support current opinion of a_i
  $\mathcal{S}^{-,i-j} \leftarrow \forall s \in \mathcal{S}^{-,i} \not\subset \mathcal{S}^j$
  if $\mathrm{n}(\mathcal{S}^{-,i-j} > 0$ **then**
    $s_z \leftarrow \max(|m_s(e_s^{t,t^*} - c_s^{t,t^*})| \; \forall s \in \mathcal{S}^{-,i-j})$
    output $\leftarrow (a_j, s_z)$
  end if
end if

Algorithm 5 Experience update:
input a_j: $\{\mathcal{S}^j, B^j, \Theta\}$, s_z, $\mathrm{IF}_{i \rightarrow j}$

$\mathcal{S}^j \leftarrow s_z$
$s_k \leftarrow \mathrm{argmax}(E_{z,:})$
$B^j_{z,:} \leftarrow B^j_{k,:}$
$B^j_{:,z} \leftarrow B^j_{:,k}$
$B^j_{z,z} \leftarrow \mathrm{IF}_{i \rightarrow j}$

Algorithm 6 Experience belief mass sharing:
input a_i: $\{\mathcal{S}^i, B^i, \Theta\}$, a_j: $\{\mathcal{S}^j, B^j, \Theta\}$, t, t^*, $\text{IF}_{i\to j}$
output (a_j, s_{z^+}, s_{z^-})

$E_{z,k}$, $C_{z,k}$ (relations between between experiences s_z, s_k)
$\mathcal{S}^{+,i} \leftarrow$ experiences that support current opinion of a_i
$\mathcal{S}^{-,i} \leftarrow$ experiences that contradict current opinion of a_i
$\mathcal{S}^{+,common(i,j)} \leftarrow (\mathcal{S}^{+,i} \cap \mathcal{S}^j)$
$\mathcal{S}^{-,common(i,j)} \leftarrow (\mathcal{S}^{-,i} \cap \mathcal{S}^j)$
$\Delta^+_{i,j} \leftarrow \{m^i_s - m^j_s\}\ \forall s \in \mathcal{S}^{+,common(i,j)}$ /w $m^i_s > m^j_s$
$\Delta^-_{i,j} \leftarrow \{m^i_s - m^j_s\}\ \forall s \in \mathcal{S}^{-,common(i,j)}$ /w $m^i_s < m^j_s$
if $\text{n}(\Delta^+_{i,j}) > 0$ **then**
  $s_{z^+} \leftarrow \text{index}(\text{argmax}(\Delta^+_{i,j}) \in \mathcal{S}_j)$
end if
if $\text{n}(\Delta^-_{i,j}) > 0$ **then**
  $s_{z^-} \leftarrow \text{index}(\text{argmax}(\Delta^-_{i,j}) \in \mathcal{S}_j)$
end if
output $\leftarrow (a_j, s_{z^+}, s_{z^-})$

Algorithm 7 Experience belief mass update:
input $(a_j$: $\{\mathcal{S}^j, B^j, \Theta\}$, s_{z^+}, s_{z^-}, $\text{IF}_{i\to j})$

$r_{IF} \in [0, 1]$: scaling factor
$E_{z,k}$, $C_{z,k}$ (relations between between experiences s_z, s_k)
for all $s_k \in \mathcal{S}_j$ **do**
  if $E_{z^+,k} - C_{z^+,k} < 0$ **then**
    $B_{z^+,k} \leftarrow B_{z^+,k} - r_{IF}\text{IF}_{i\to j}(B_{z^+,k})$
  else
    $B_{z^+,k} \leftarrow B_{z^+,k} + r_{IF}\text{IF}_{i\to j}(1 - B_{z^+,k})$
  end if
end for
for all $s_k \in \mathcal{S}_j$ **do**
  if $E_{z^-,k} - C_{z^-,k} >= 0$ **then**
    $B_{z^-,k} \leftarrow B_{z^-,k} - r_{IF}\text{IF}_{i\to j}(B_{z^-,k})$
  else
    $B_{z^-,k} \leftarrow B_{z^-,k} + r_{IF}\text{IF}_{i\to j}(1 - B_{z^-,k})$
  end if
end for

7-2-3 Numerical toy example

As mentioned, in the final implementation of the model, first all interactions will take place, and after all agents will update. Here the interaction and update of both elements are jointly shown. Suppose the following network of agents:

This will be an example of the interaction between the denoted agents a_1 and a_2. Let us first

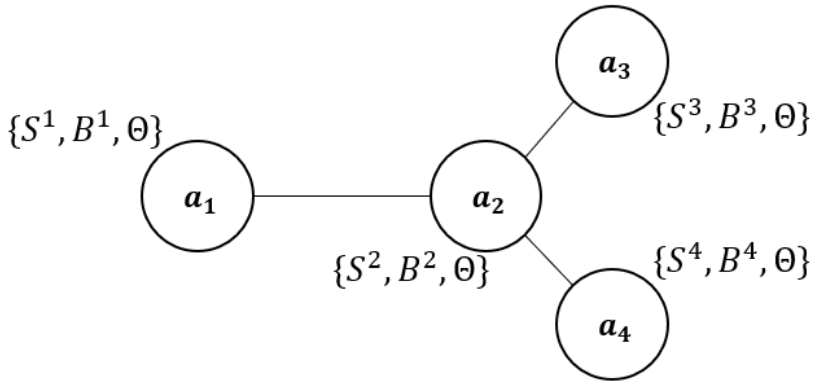

Figure 7-6: Network considered in example.

define the these agents. For agent a_1

$$\mathcal{S}^1 \leftarrow \begin{Bmatrix} \text{The car has a great engine} \\ \text{The seats are not comfortable} \\ \text{The power of a car is most important} \\ \text{The car is fantastic in the corners} \end{Bmatrix} \qquad B^1 = \begin{bmatrix} 0.66 & 0.33 & 0.66 & 0.66 \\ 0.66 & 0.33 & 0.66 & 0.66 \\ 0.66 & 0.33 & 0.66 & 0.66 \\ 0.66 & 0.33 & 0.66 & 0.66 \end{bmatrix}$$

And for agent a_2

$$\mathcal{S}^2 \leftarrow \begin{Bmatrix} \text{The car has a great engine} \\ \text{The seats are not comfortable} \\ \text{The color of the car is dreadful} \\ \text{The steering of the car is bad} \end{Bmatrix} \qquad B^2 = \begin{bmatrix} 0.36 & 0.64 & 0.64 & 0.64 \\ 0.36 & 0.64 & 0.64 & 0.64 \\ 0.36 & 0.64 & 0.64 & 0.64 \\ 0.36 & 0.64 & 0.64 & 0.64 \end{bmatrix}$$

Agent a_3 and a_4 have arbitrary experiences that are not outlined in the example. The explicit opinions of agent a_1 and a_2 are:

$$x_1^{t,t^*} = 0.95; \qquad x_2^{t,t^*} = -0.83$$

First, the influence factors are calculated according to Equations 7-5, 7-4, 7-6 and 7-7.

$$\begin{aligned} \text{IF}_{1\to2} &= (\text{confidence}_2) \cdot (\text{dlm}_{1\to2}) \cdot (\text{auth}_1) \\ &= (0.23) \cdot (0.39) \cdot (0.33) \\ &= 0.031 \\ \text{IF}_{2\to1} &= (\text{confidence}_1) \cdot (\text{dlm}_{2\to1}) \cdot (\text{auth}_2) \\ &= (0.13) \cdot (0.61) \cdot (1) \\ &= 0.078 \end{aligned}$$

Then Algorithm 2 is followed, with $a_i = a_1$ and $a_j = a_2$. The probability that a_1 executes Algorithm 4 is 0.39. Let's presume that the agent gets the chance to share its opinion. Following Algorithm 4, the sampled opinion is

$$X_1^{t,t^*} = 0.99$$

Resulting in $p = 1$. So it will share an experience in favor of its current opinion. This set of experiences is, ordered by their influence on the opinion of a_1:

$$\mathcal{S}_{+,1} = \left\{ \begin{matrix} \text{The car has a great engine} \\ \text{The car is fantastic in the corners} \\ \text{The power of a car is most important} \end{matrix} \right\}$$

The first most influential experience that a_2 does not yet have is the 'The car is fantastic in the corners', which will be included in $\mathcal{S}^2$. The new row and column of the bias matrix B^2 will be initialized with the elements of the most entailing other experience in $\mathcal{S}^2$. This is 'The car has a great engine', so the new bias matrix will be

$$B^2 = \begin{bmatrix} 0.36 & 0.64 & 0.64 & 0.64 & 0.36 \\ 0.36 & 0.64 & 0.64 & 0.64 & 0.36 \\ 0.36 & 0.64 & 0.64 & 0.64 & 0.36 \\ 0.36 & 0.64 & 0.64 & 0.64 & 0.36 \\ 0.36 & 0.64 & 0.64 & 0.64 & 0.03 \end{bmatrix}$$

The new diagonal element takes on the value of the influence factor. For the bias update, Algorithm 5 is followed. We take $r_{IF} = 1$. The agents have only one supporting and one contradicting experience in common, the just received experience is ignored. The sets containing the difference in belief mass between the two agents are set up as follows.

$$\begin{aligned} \Delta^+_{1,2} &= \{m^1_s - m^2_s\} \forall s \in \mathcal{S}^{+,common(1,2)} / \text{w } m^1_s > m^2_s \\ &= \{0.66\} = \mathcal{S}^{+,common(1,2)} \\ \Delta^-_{1,2} &= \{m^1_s - m^2_s\} \forall s \in \mathcal{S}^{-,common(1,2)} / \text{w } m^1_s < m^1_s \\ &= \{0.53\} = \mathcal{S}^{-,common(1,2)} \end{aligned}$$

Now, analogue to cognitive dissonance reduction, agent a_2 will update the bias for both these experiences. For the experience in $\Delta^+_{1,2}$ bias increases, for the one in $\Delta^-_{1,2}$ it decreases. The new bias matrix becomes:

$$B^2 = \begin{bmatrix} 0.37 & 0.62 & 0.62 & 0.62 & 0.37 \\ 0.37 & 0.62 & 0.62 & 0.62 & 0.37 \\ 0.35 & 0.64 & 0.64 & 0.64 & 0.35 \\ 0.35 & 0.64 & 0.64 & 0.64 & 0.35 \\ 0.35 & 0.64 & 0.64 & 0.64 & 0.03 \end{bmatrix}$$

The new explicit opinion of agent a_2 is

$$x_2^{t,t^*} = -0.61$$

7-3 Conclusion

Firstly, a single cognitive process was identified that governs the micro dynamics of an agent. The process is named cognitive dissonance reduction (or minimization) and is a direct result of the concept of cognitive dissonance, on which the proposed belief system is based. The

process describes how an agent internally updates the belief in its experiences. It can be extrapolated to the concept of structural balance, describing the influence of other agents.

From these concepts, both intra- and inter-agent dynamics are proposed. These processes will be validated in the following chapter, where all model elements come together in a simulation. Parallel to the validation, it is inspected how the novel elements of the model, agent perception and topic dependence, play a part opinion formation and propagation.

Chapter 8

Results, validation and understanding

In this section, a real world network of agents is introduced, which outlines the social experiment to be used for validating and understanding the proposed model. The elements of proposed model are validated by reproducing the true outcome of the real world network.

The results are used as an example to show the added value of perception and language against what traditional value-based methods. Furthermore, it is inspected where and why the results deviate from the real outcome to gain model understanding. Besides validation, the results will be used to show the contribution of the proposed model as well as its limitations.

The proposed model to the best of the authors knowledge, is a first step in bringing language models and opinion dynamics together. As such, the last part is about laying out the obstacles yet to overcome for closing that gap.

8-1 Experiment design

In this section the experiment is designed. First, the network choice is presented. After, the proposed model elements in earlier sections are brought together to form the experiment setup.

8-1-1 Real world network

As outlined in the literature review, there are multiple paradigms of network design. Depending on the research topic (what kind of observed social behaviour is of interest) a certain paradigm can be chosen to fit the requirements. This research has a bottom up approach, with the goal of showing the significance of a possibly fundamental element of association. As with many to all research in opinion modeling, true validation remains impossible. However, over the years a hand full of networks have been proposed that abstractly represent real people with real connections, of which the opinion outcome is known. A widely used network with a real known opinion outcome is known as Zachary's karate club [48].

From 1970 to 1972, Zachery studied the social interaction between 34 members of a karate club, where he determined the social connections between the members. Whom interacts with whom. During the study a conflict arose between the president of the club and the karate instructor. The conflict resulted in a split of the group. Half of the members joined the new karate club of the instructor, the other half stayed at the original club with the president. Zachary correctly predicted the choice to join one of the two groups for all members except number 9. The method Zachary used is based on classical network theory and is out of scope to discuss here. Though, his measured outcome is years later very much welcomed by the community of social engineering. The connections and the outcome are shown in Figure 8-1 below. Node 1 represents the instructor, node 34 the club president.

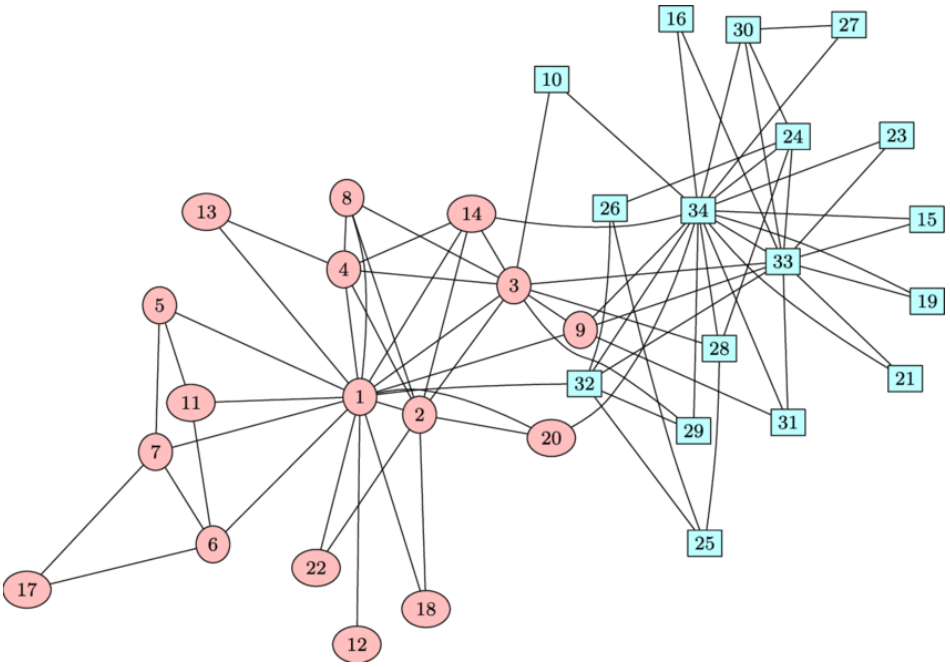

Figure 8-1: Depiction of Zachary karate club network [3]. The color and shapes represent the final groups formed.

The network has a fixed, unweighted and undirected topology.

8-1-2 Proposed model setup

Before introducing the simulation setup, one last parameter must be introduced: stubbornness. Or in other words, the amount to which an agent is susceptible to another agents opinion. This is an essential agent trait that must be included to a polarizing model such as the Zachary network. The karate instructor and the club president not only have opposing views, they are also stubbornly attracted to their opinion. The trait is modeled through the definition of self confidence. Seen in Eq.7-7, a high self confidence makes the agent less susceptible to another opinion. The trait stubbornness makes use of that as follows:

For stubbornness parameter $stubbornness \in [-1, 1]$ and $stubbornness < 0$:

$$\text{confidence}_i = \text{confidence}_i - (stubbornness)(\text{confidence}_i) \in [0, 1] \tag{8-1}$$

For stubbornness parameter $stubbornness > 0$:

$$\text{confidence}_i = \text{confidence}_i - (stubbornness)(1 - \text{confidence}_i) \in [0, 1] \tag{8-2}$$

In Algorithm 8 the entire model simulation is outlined. For readability, the earlier proposed algorithms are referenced to. Within these proposed definitions, a number of parameters is introduced. The choice on their values in the experiment is elaborated in the next section. The algorithm shows one new parameter: r_d. It is used to increase the association strength r for the process of cognitive dissonance reduction in Algorithm 1. This ensures that over time, the agents will always converge to -1 or 1. However, convergence for the simulation is defined as follows: when after a minimal of 5 steps, for two consecutive time steps, all agents have $|x_i^{t,t^*}| \geq 0.5$. The simulations showed that the opinions do not change in opposite direction after this point. This is also computationally beneficial when running many simulations. More on computation and convergence in the next section.

The data used is equal to the ones used in the previous sections. As mentioned, the split between negative and positive reviews is made with sentiment classification, which holds no real relation to how the belief system interprets them on the basis of a topic. The agents are initialized as follows. The instructor, agent a_1, has 20 experiences. 17 drawn from the positive classified reviews and 3 from the negative. These are iteratively randomly drawn, until the explicit opinion of the agent is between $0.85 \leq x_1^{t,t^*} \leq 0.95$. Similarly for the club president, agent a_{34}, but towards an opposite attitude, so $-0.95 \leq x_{34}^{t,t^*} \leq -0.85$. The idea is that they also have some experiences against their own opinion.

The other agents, a_2 to a_{33}, follow the same routine for 5 positive and 5 negative classified reviews. The reviews are drawn until their initial explicit belief is close to neutral, namely between $[-0.1, 0.1]$.

The 10 steps of cognitive dissonance reduction (Algo 1) for agent 1 and 34 before the simulation start, is to ensure these agents are heavily biased towards their initial opinion. Then, for each simulation step, first all connected agents interact, then all agents are updated. The update for each agent consists of 1) updating on all the interactions through Algorithm 3, 2) a step of cognitive dissonance minimization through Algorithm 1 and 3) an increase of the association rate.

Algorithm 8 Model simulation
intput t, t^*, s^+ (positive product reviews), s^- (negative product reviews)

$G = (\mathcal{V}, \mathcal{E}) \leftarrow$ Zachary graph
$\{a_1, ..., a_{34}\} \in V$
a_1: ($stubborn = 0.8$, $r = 0.05$, $\mathcal{S} := \{17 \times s^-, 3 \times s^+\}$, $B_{i,j} = 0.5\ \forall i, j$)
a_{34}: ($stubborn = 0.8$, $r = 0.05$, $\mathcal{S} := \{3 \times s^-, 17 \times s^+\}$, $B_{i,j} = 0.5\ \forall i, j$)
$a_2 : a_{33}$: ($stubborn = -0.2$, $r = 0.05$, $\mathcal{S} := \{5 \times s^-, 5 \times s^+\}$, $B_{i,j} = 0.5\ \forall i, j$)
$g \leftarrow 0.5$
$r_{IF} \leftarrow \frac{1}{15}$
$r_d \leftarrow 0.02$
for 10 steps **do**
  Algorithm 1 $\leftarrow$ (a_1, t, t^*, (confidence$_1$) $\cdot$r$_1$)
  Algorithm 1 $\leftarrow$ (a_{34}, t, t^*, (confidence$_{34}$) $\cdot$r$_{34}$)
end for
while not converged **do**
  for each edge e in $\mathcal{E}$ **do**
    a_i, $a_j \leftarrow e$
    interactions[i], interactions[j] $\leftarrow$ Algorithm 2 $\leftarrow$ (a_i, a_j, t, t^*)
  end for
  for each agent $a_i \in \mathcal{V}$ **do**
    interactions[i] $\leftarrow a_i$
    for each interaction[i] in interactions[i] **do**
      Algorithm 3 $\leftarrow$ (a_i, interaction[i])
    end for
    Algorithm 1 $\leftarrow$ (a_i, t, t^*, (confidence$_i$) $\cdot$r$_i$)
    $a_i[r] \leftarrow a_i[r] + r_d(1 - a_i[r])$
  end for
end while

8-2 Results and model understanding

First a single simulation run in detail, to show and to some degree validate the workings of the proposed model elements. After, more simulations will be run to show how the opinion formation in the network is influenced by the experiences and the topic. Lastly, the characteristics and limitations of the model are discussed.

8-2-1 Model Validation

First, a single simulation is run in detail. The agents are initialized as outlined earlier. The data set used is that of the car seat reviews, with the topic 'The car seats are good' and the conjugate 'The car seats are not good'. The initial network is shown below in Figures 8-2 and 8-3. Note that the count starts at 0, representing agent 1.

The simulation is run over 16 time steps. A three fold of data is collected to inspect the evolution of the network and the workings of the proposed model elements. In Figure 8-4 the

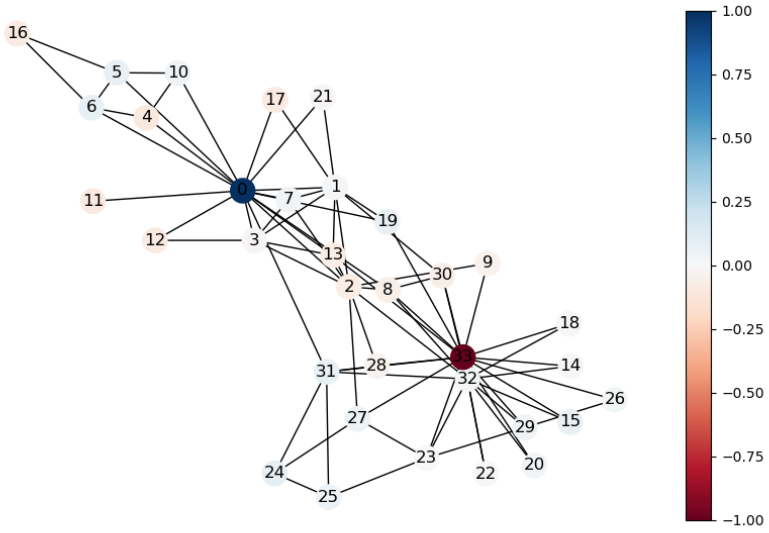

Figure 8-2: Initial network. The colors indicate their explicit opinion. Node 0 is agent 1, node 33 is agent 34.

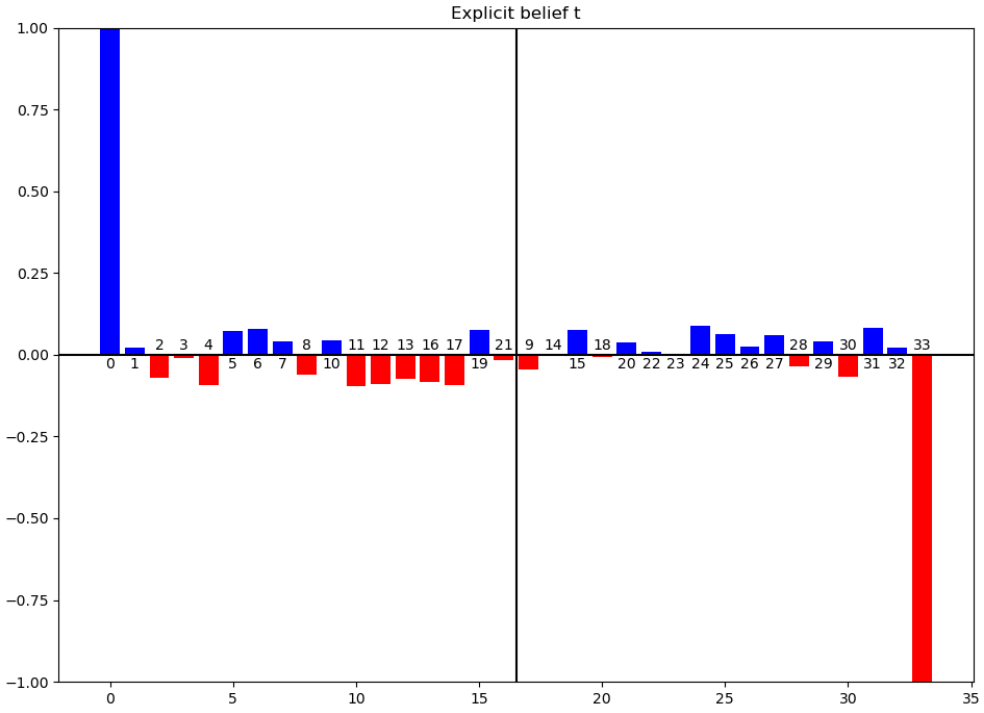

Figure 8-3: Initial (random) beliefs as bar chart. Node 0 is agent 1, node 33 is agent 34.

implicit and explicit beliefs of the agents over time is shown. The agents are not individually labeled. The orange lines stand for agent 1 and its (true) followers, the green lines for agent 34 and its (true) followers. The bold lines represent the group leaders.

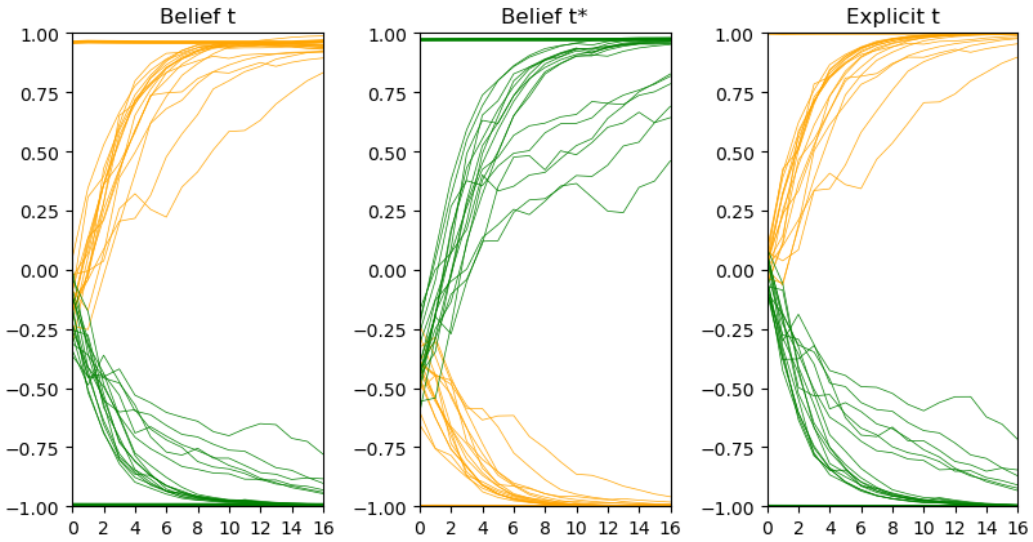

Figure 8-4: Left/middle: Implicit belief in t and t^*. Right: Explicit belief in t. Orange: group of agent 1. Green: group of agent 34.

Furthermore, the undirected like-mindedness is measured at each time step according to Eq. 7-3. In Figure 8-5, this is shown for agent 1 (0) and agent 34 (33). The color coding is equal to previously mentioned. The bold lines for the leaders are barely visible as they are very much not like-minded, and stay that way. It clearly shows how within the formed groups, agents come to think alike.

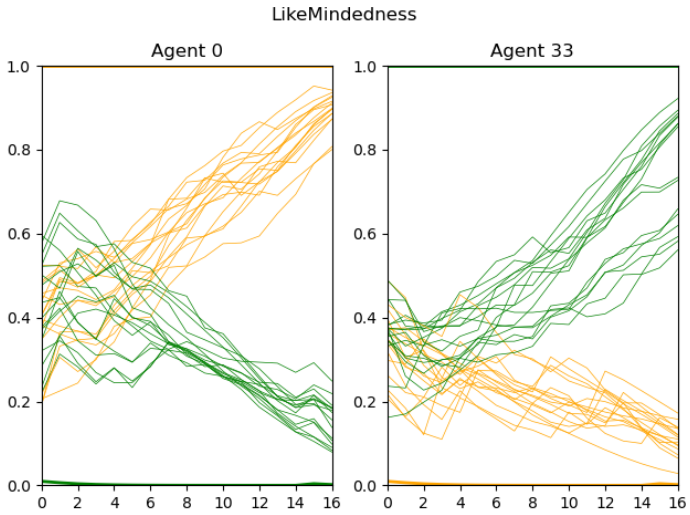

Figure 8-5: Undirected Like-mindedness (Eq. 7-3) for all agents w.r.t. agent 1 (0) and agent 34 (33).

Lastly, in Figure 8-6, the confidence of the agents over time is shown. As expected, when neutrally initialized, the agents start close to a half. The group leaders start highly confident, and with the exception of a minor dip, remain that way.

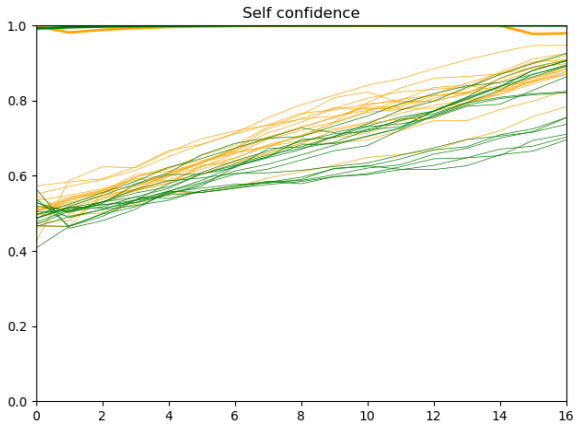

Figure 8-6: Confidence of the agents over time.

Below in Figure 8-7 the final states of the agents are shown after 16 simulation steps. The true groups are separated by the vertical bold line. The bar chart shows nicely how the final states are aligned with the true outcome of the discussion.

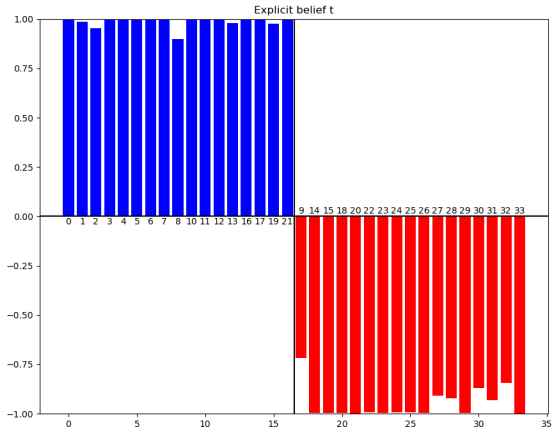

Figure 8-7: Agent states after 16 time steps.

The figures show that the model is able to reproduce the known outcome of the Zachary network. However, this is not sufficient for validating the model, as will be shown in the next section. The measures of confidence and like-mindedness, do show that the micro dynamics behave as they are designed to do.

8-2-2 Contribution of perception through language

The real decisions of the individuals in the Karate club might not be the only possible outcome. Although any value-based method could implement a form of probability, they are bound to

the dimensions of their class of models. The probability of an agent having a certain initial experience, or having a neighbour with a certain experience, along with a possible cultural bias is not part of that class.

In Table 8-1 and 8-2 the probabilities are shown for the agents to belong to the group as in the real outcome, after 100 simulations. Group 1 has the instructor as leader (agent 1), group 2 has the president as leader (agent 34). Each simulation, the agents are initialized with random experience through the same process as described earlier. All other parameters are unchanged. In the results in Table 8-1, agent 1 is positively initialized and agent 34 negatively. For the results in Table 8-2 the other way around.

Group 1		Group 2	
Agent	Pr.	Agent	Pr
1	1.0	10	1.0
2	1.0	15	1.0
3	0.8	16	1.0
4	0.9	19	1.0
5	1.0	21	1.0
6	1.0	23	1.0
7	1.0	24	1.0
8	0.9	25	0.7
9	0.2	26	0.8
11	1.0	27	1.0
12	1.0	28	1.0
13	1.0	29	1.0
14	0.8	30	1.0
17	1.0	31	1.0
18	1.0	32	0.8
20	0.8	33	1.0
22	1.0	34	1.0

Table 8-1: Probabilities of agents belonging to group. Agent 1 is positively initialized, agent 34 negatively.

Group 1		Group 2	
Agent	Pr.	Agent	Pr
1	1.0	10	1.0
2	1.0	15	1.0
3	1.0	16	1.0
4	1.0	19	1.0
5	1.0	21	1.0
6	1.0	23	1.0
7	1.0	24	1.0
8	1.0	25	0.8
9	0.7	26	0.8
11	1.0	27	1.0
12	1.0	28	1.0
13	1.0	29	0.8
14	0.9	30	1.0
17	1.0	31	0.9
18	1.0	32	0.9
20	0.9	33	1.0
22	1.0	34	1.0

Table 8-2: Probabilities of agents belonging to group. Agent 1 is negatively initialized, agent 34 positively.

First, looking only at the results in Table 8-1, they show that the initial (explicit) content matters for the resulting opinions, as the initial beliefs remain in the same range. It is not only the content of the experiences for an agent itself, also the surrounding agents and to some extent all other agents. Unfortunately, the ambiguousness of textual experiences makes it difficult to find exact causal relations; more on this in the section on understanding and limitations. Still only looking at Table 8-1, a number of agents show a deviation from the real outcome. For example agent 9 for the majority of the simulations ends up following agent 34. Seeing its position and neighbours in Figure 8-1, this is not surprising. As for agent 25, 26, and 32, they appear to form a subgroup on their own. On closer inspection, when early in the simulation agent 32 receives a heavy argument from agent 1, supporting the topic, it is past on to agents 25 and 26. Through the element of biased assimilation and like mindedness, they are less influenced by the contradicting experiences shared by their negatively opinionated

neighbours, and in time lean more on each other.

These insights underline the time variant characteristic of the model; previous interactions heavily way on the future evolution of the network. The initialization itself can be seen as their current state of association in a point in time. It is an expected and desired characteristic through among others biased assimilation. Though, the weight of this time variance on the opinion formation can be disputed, and with that, the amount to which an agent attaches to its prejudice over time.

Now the two tables 8-1 and 8-2 are compared. Recall that the language model is slightly negatively biased towards the topic of cars. This is clearly visible in the difference between the two tables. In Table 8-2, Agent 9 now follows agent 1 for the majority of the simulations. Also, here agents 3 and 4 follow agent 1 irrespective of the distribution of initial experiences. Under the assumption that the language model indeed carries general cultural bias, this is the first evidence that the actual topic proposition of a discussion, and the stance towards it, affects the formation of opinion in a network of agents.

8-2-3 Topic dependence

To underline that significance, the same simulations are also run for the topic of the hotel. Recall that the language model was significantly positively biased towards hotels. Figure 8-8 shows the results of Table 8-1 and 8-2 in a bar chart. Figure 8-9 shows the results of the simulations with respect to the hotel topic. The bars represent the probability that an agent will follow the opinion of the instructor or the president, i.e. agent 1 and 34. These are denoted as respectively Group 1 and 2. The agents are arranged according to their true outcome. So the left annotated agents correspond to the agents in Group 1 in Table 8-1.

In Figure 8-8, it clearly shows that the final opinion of agent 9 is highly dependent on the stance towards the car topic. One could argue that the agent balances between the two groups due to the topology, and that the biased stance makes it favor one over the other.

The difference between the discussion about a car and about the hotel in Figure 8-9 clearly shows the topic dependence. Keeping in mind that the topology is the same, agent 9 shows the same characteristic, but now towards the opposite stance. Moreover, the group leader with the positive stance very consistently attracts more followers. Which is expected not so much because of the 'absolute' bias identified, but more because of the difference compared to the bias towards the car.

Figure 8-8: Results car topic and data.

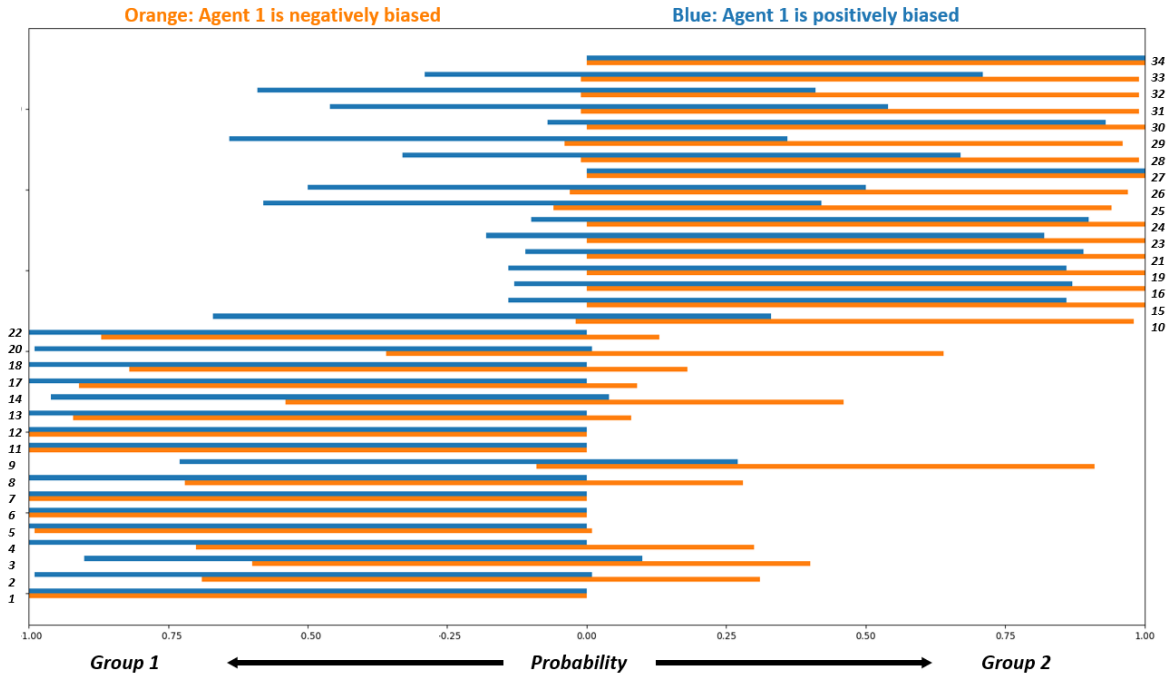

Figure 8-9: Results hotel topic and data.

8-2-4 Model understanding and limitations

The opinion formation in a network is clearly dependent on the topic and the formulation of the content (or experiences) about that topic. This dependence shows that an associative language-based structure for the belief system could be important for modeling 'true' opinion formation. However, it adds a complex, nonlinear and unpredictable new dimension. Results cannot be exclusively attributed to either the parameterized model elements, the network topology, nor the initial experience of the agents. Any attempt in finding a correlation between them did not yield in consistent results. Only the group factor set at $g = 0.5$ and the use of authority were identified to aid convergence towards the true outcome of the Zachary network.

Having said, convergence cannot be proven under traditional system theory. The only option would be a very large number of numerical experiments, which for two reasons is infeasible. Firstly, because the current model setup is computationally very expensive. The model complexity grows with approximately $\mathcal{O}(ns^2)$, with n number of agents and s the number of unique experiences present in the network. Although much can be gained from calculating the inference and saving the results between all experiences against each other prior to simulating, this is only feasible for relatively small data sets. Secondly, there exists no 'true' objective to optimize the proposed elements and parameters towards. In other words, although the true outcome of the Zachary network is known, there is nothing known about the underlying probability distribution, nor how this came to be.

Furthermore, the model is highly dependent on how the textual content is formulated. On the one hand this is a necessary evil, as just as in the real world a single word can entirely change the meaning of a sentence. On the other hand, this makes the model perhaps to much dependent on the textual data that is used. Although the very idea is that real data is used, the distribution of the experiences in the network, along (theoretically all) the dimensions of language, should match the distribution of those experiences in the real world. Similar to how a language model is biased towards the distribution of language in its training data. In this context, the proposed opinion model follows the same rule of large numbers: the more experiences used in the network, the more likely that the distribution is similar to that of real society. Also, the influence of individual experience formulation will have less weight. However, in light of the computational complexity, this is not feasible with the current approach.

Moreover, suppose it is assumed that the data indeed represents the 'true' distribution of (expressed) arguments concerning the subject. With respect to the hotel reviews, the language model bias makes that an agent needs on average almost twice as many contradicting (with respect to positively formulated t) reviews in order to get close to a neutral initial opinion. Consequently, on average the entire network also has about twice as much unique contradicting experiences. In real life, we would expect the bias of the language model to be proportional to the distribution of experiences. Although the results still show the expected behaviour, currently the distribution of experiences is inversely proportional to the bias in the language model.

Lastly, the associative power of a language model is not fully used. One of the key characteristics of an entropy-wise (globally) trained language model, is that the change on one association (co-occurrence) leads implicitly to an change on all associations. In other words,

that a discussion on one topic, also changes the attitude towards another topic. Although the concept of a belief mass as a function of all experiences approaches this idea, the characteristic is not explicitly used. Furthermore, the use of the still flawed inference makes that the belief system cannot handle diverging or unrelated experiences. The mentioned key characteristic can therefor not be researched with the current setup.

Chapter 9

Conclusion

The literature review showed empirical evidence that language models and opinion dynamics are possibly two highly related research domains. Therefor the research focused on highlighting this apparent relation through the proposal of an opinion model with a language model-based belief system. The hypothesis was that language models could serve as the associative structure to model the element of perception, specifically cultural perception. Also, on that premise, that the formation and propagation of a belief is dependent on the actual subject of discussion and the initial prejudice of the individuals. This research shows evidence that this is indeed the case. Also, it emphasizes the importance of closing the gap between language and opinion modeling.

The proposed opinion model is designed with a bottom-up approach. Therefor, first, significant cognitive processes were identified. The initial framework for the belief system is based on the concept of cognitive dissonance. The dynamic processes influencing the belief are based on cognitive dissonance reduction and biased assimilation. The behaviour of the belief system and the dynamic processes are tested and inspected through numerous experiments.

Using the proposed model, numerical experiments are conducted on a polarizing real world network. The model is able to reproduce the real outcome of the network. Furthermore, the results show that the prejudice carried by the language model comes forward in both the topic as well as the stance towards that topic. The leading agent with a stance in favor of the language model bias, is consistently able convince more agents than vice versa. Especially the already, through the topology resulting indecisive agents are sensitive to this prejudice. Though, this cannot be proven to be exclusively attributed to a topic, because of the complexity the novel element of language brings. Also, the textual data used might not be exemplary for how those 'experiences' are distributed in society.

The proposed model has many known and unknown limitations. However, it does not dispute the empirical evidence that suggests language models are closely related to opinion modeling.

Chapter 10

Discussion and future research

First, some discussion points with respect to the proposed model are given. After, remarks are given outside the scope of the proposed model, with hope that future research can efficiently build on this thesis.

In the current design, all agents speak the same language. The direct inferred relation between an experience and the topic of discussion is equal for all. The belief in that experience is individually defined, but over the domain [0,1], i.e. the polarity of an experience with respect to the topic does not change. This simplification was explained in the context of corona and chipped vaccinations. However, for researching significant cognitive tendencies such as the backfire effect, individual interpretation is needed. It would be very interesting to see how the associations of the agents evolve if they individually updated their language model on the received experiences. Also, if the intra-agent dissonance reduction is combined with a language model update, modeling the recall of a thought or experience (memory reconsilidation). An individually updating language model could replace the currently proposed bias matrix.

The belief state of an agent is determined by accumulating the belief masses and direct relations in a Beta distribution. It represents the probability distribution for the belief of the agent in the topic. Besides the expectation to define the state of belief, only sampling is used in interaction. Other characteristics of the distribution are not explicitly used. It was examined whether the variance could be used to scale the relation between experiences and the topic, such that uncertain relations have less weight. However, this did not lead to a clear improvements. Also, for a meaningful implementation of the distribution characteristics such as the variance, there is the need to define bounds on the alpha and beta values that parameterise the Beta function. The extent to which these values can grow is indirectly determined by the number of unique experiences in the network. This is on itself a point of discussion; determining a suitable number of experiences in the network is an ambiguous task with the interplay between the belief system, the surrounding dynamics and the formulation of the experiences. Also, without a clear objective to optimize towards.

One requirement that can be given, is for the distribution of experiences to be similar to that in society. However, the resulting distribution after neutral initialization of the agents

seems to be inversely proportional to the bias in the language model. If this is purely a topic bias, it could be a welcoming feature for bias mitigation. However, the bias could lie under every (un)imaginable rock in language (models). Furthermore, with the use of experiences, comparing different topics or different languages on the same topic is not straightforward. They cannot be properly compared and the list of unknowns will only grow with respect to data biases.

The proposed model operates in its own paradigm. As such, further research aimed specifically at the proposal might not be fruitful, as unknown and complex parameters are possibly unique to this approach. The proposed framework and used NLP tools can be seen as a way to circumvent the structure of current language models. For future research it is recommended to specifically question how a language model can be restructured to extract an attitude or opinion towards a subject. This with the idea of creating a single paradigm of language and opinion models. Such future research could start by focusing on (biased contextual) antonymy. Little research has been conducted in this area, while (relative) opinion and knowledge could be attributed to a subjective, contextual and mostly unconscious comparison.

Another recommended spear point for future research with overlap in both language and opinion modeling is the concept of analogy. Hidden inside that concept lies an important key to communication between agents (or actual individuals) whom 'do not speak the same language'. Suppose a vector in the contextual embedding space represents an abstract meaning, concept or even a 'feeling', for example 'the feeling of not belonging to the group'. In another (differently trained) language model, that meaning also exists, but is based on different associations. What context or words should be used for the former language model to express that meaning, such that the latter language model will know what is meant? For example with the analogy of the Ugly Duckling. As language and opinion modeling poses to be a great tool in bringing people together, the concept of analogy does not yet have the attention is seems to deserve.

Lastly, language, knowledge and opinion are formed by and in society. The idea that a single language model, which can inherently only have one opinion (if always the maximum probability is taken), can account for all possible contextual opinions, might be too simplistic. Or the other way around, a single model being a bridge to far. Language could said to be formed over two levels of distributed representations: the representations formed at individual level (analogue to contextual embeddings), and the distribution of those (distributed) representations in society. Currently, the idea is that massive heaps of text could implicitly account for both. However, it blends the two (conceptual) distributions to the extent that biases remain ambiguous and uncontrollable. As such, it is also recommended for future research to inspect how the distributed nature of agent based opinion models could be used to control the learning and training of language models.

Appendix A

Other results belief system design

Belief consistency: car results

C_1	The seats in the car are good
C_1^*	The seats in the car are not good
C_2	The seats in the car are fine
C_2^*	The seats in the car are awful
C_3	The car is comfortable
C_3^*	The car is uncomfortable

Table A-1: Topics used.

BS	Q	$C_1[P]$	$C_1^*[P]$	$C_2[P]$	$C_2^*[P]$	$C_3[P]$	$C_3^*[P]$
x^t	10	0.853	-0.972	0.888	-0.997	0.856	-0.952
	20	0.941	-0.986	0.938	-0.999	0.894	-0.988
	30	0.948	-0.987	0.959	-0.999	0.911	-0.990
x^{t,t^*}	10	0.927	-0.927	0.984	-0.984	0.919	-0.919
	20	0.979	-0.979	0.996	-0.996	0.975	-0.975
	30	0.978	-0.978	0.995	-0.995	0.979	-0.979

Table A-2: Results positively initiated on car seats reviews, on topics A-1

BS	Q	$C_1[N]$	$C_1^*[N]$	$C_2[N]$	$C_2^*[N]$	$C_3[N]$	$C_3^*[N]$
x^t	10	-0.910	0.612	-0.916	-0.276	-0.931	0.782
	20	-0.910	0.542	-0.915	-0.153	-0.930	0.662
	30	-0.931	0.596	-0.936	-0.220	-0.946	0.750
x^{t,t^*}	10	-0.834	0.834	-0.629	0.629	-0.793	0.793
	20	-0.779	0.779	-0.613	0.613	-0.729	0.729
	30	-0.814	0.814	-0.629	0.629	-0.805	0.805

Table A-3: Results negatively initiated on car seats reviews, on topics A-1

Conjunction fallacy

C_1	The seats in the car are good
C_1^*	The seats in the car are not good
C_2	The seats in the car are good when driving
C_2^*	The seats in the car are not good when driving
C_3	The seats in the car are good and comfortable
C_3^*	The seats in the car are not good and not comfortable

BS	Q	H_1	H_1^*	H_2	H_2^*	H_3	H_3^*
x^t	10	0.864	-0.934	0.813	-0.916	0.772	-0.968
	20	0.916	-0.961	0.871	-0.949	0.828	-0.981
	30	0.951	-0.977	0.916	-0.969	0.884	-0.990
x^{t,t^*}	10	0.862	-0.862	0.855	-0.855	0.932	-0.932
	20	0.926	-0.926	0.908	-0.908	0.963	-0.963
	30	0.946	-0.946	0.941	-0.941	0.976	-0.976

Table A-4: Results car reviews, pos init

BS	Q	H_1	H_1^*	H_2	H_2^*	H_3	H_3^*
x^t	10	-0.888	0.729	-0.892	0.768	-0.891	0.657
	20	-0.902	0.778	-0.916	0.812	-0.922	0.715
	30	-0.931	0.803	-0.944	0.835	-0.943	0.764
x^{t,t^*}	10	-0.807	0.807	-0.822	0.822	-0.745	0.745
	20	-0.885	0.885	-0.903	0.903	-0.835	0.835
	30	-0.904	0.904	-0.928	0.928	-0.875	0.875

Table A-5: Results car reviews, neg init

Ambiguous topic: better captured by *explicit opinion*

Car reviews. Topic: 'I like the car seats', conjugate: 'I do not like the car seats'.

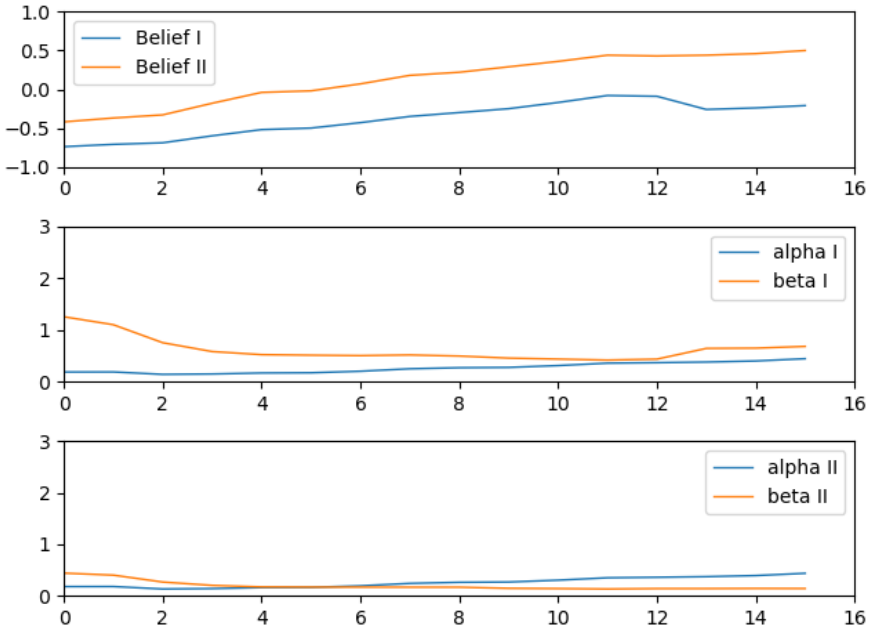

Figure A-1: Values for α_t and β_t of the implicit opinion (I), α_{t,t^*} and β_{t,t^*} of the explicit opinion (II).

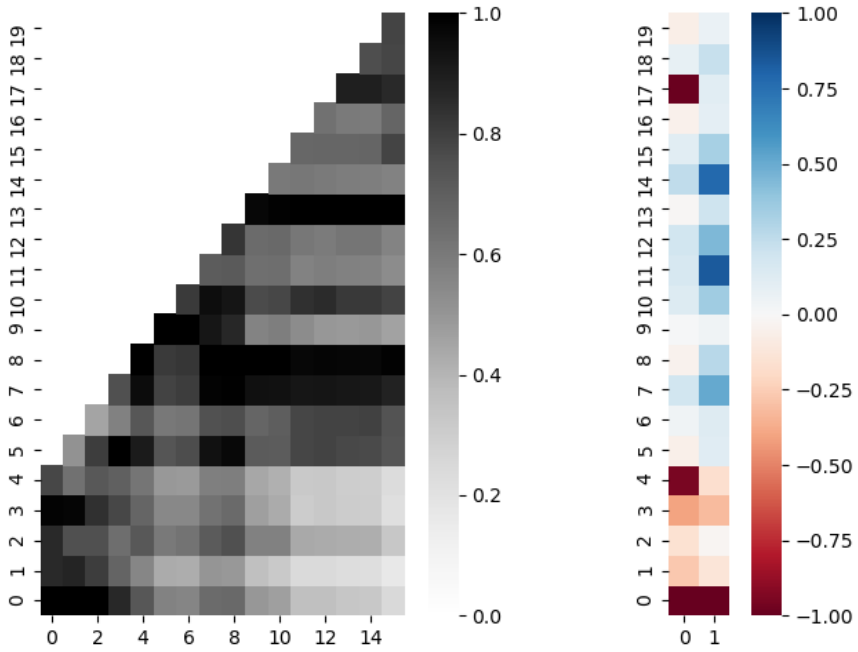

Figure A-2: Left: mass of each experience. Right: the direct relation of the experience; (0) represents x^t, (1) x^{t,t^*}.

Neutrality, second situation, Hotel results

H_1	The service in the hotel is good
H_1^*	The service in the hotel is not good
H_2	The service in the hotel is OK
H_2^*	The service in the hotel is less then OK
H_3	The service in the hotel is neutral
H_3^*	The service in the hotel is not neutral

BS	Q	H_1	H_1^*	H_2	H_2^*	H_3	H_3^*
x^t	10	-0.047	-0.459	-0.641	-0.787	-0.925	0.944
	20	0.093	-0.557	-0.563	-0.753	-0.969	0.953
	30	0.105	-0.554	-0.515	-0.673	-0.929	0.957
x^{t,t^*}	10	0.061	-0.061	0.289	-0.289	-0.996	0.996
	20	0.184	-0.184	0.359	-0.359	-1.000	1.000
	30	0.217	-0.217	0.448	-0.448	-0.995	0.995

Table A-6: Results hotel reviews

Adding experience: car reviews

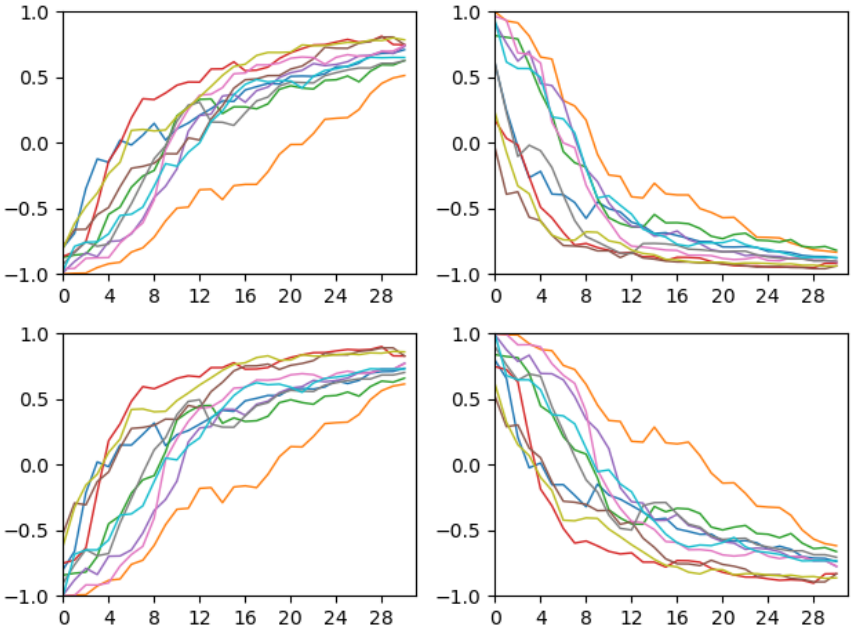

Figure A-3: Implicit opinion (top), Explicit opinion (down). Evaluated on t (left) and t^* (right). Negatively initiated.

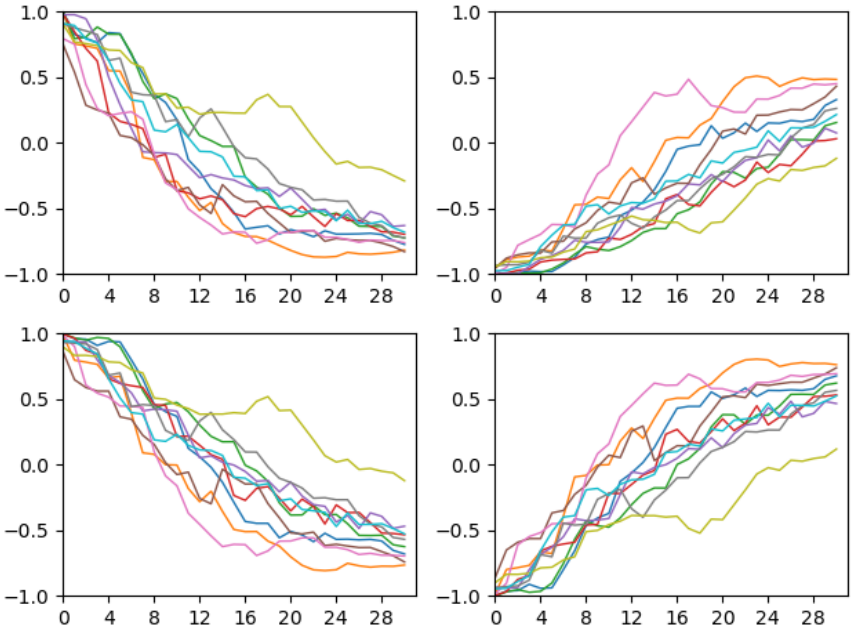

Figure A-4: Implicit opinion (top), Explicit opinion (down). Evaluated on t (left) and t^* (right). Positively initiated.

Bibliography

[1] M. Ye, M. H. Trinh, Y. H. Lim, B. D. Anderson, and H. S. Ahn, "Continuous-time opinion dynamics on multiple interdependent topics," *arXiv*, no. January, 2018.

[2] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 2227–2237, 2018.

[3] J. B. Leger, C. Vacher, and J. J. Daudin, "Detection of structurally homogeneous subsets in graphs," *Statistics and Computing*, 2014.

[4] S. Pinker and P. Bloom, "Natural language and natural selection," *Behavioral and Brain Sciences*, 1990.

[5] T. W. Deacon, *The Symbolic Species: The Co-Evolution of Language and the Brain.* 1997.

[6] L. I. Perlovsky, "Toward physics of the mind: Concepts, emotions, consciousness, and symbols," *Physics of Life Reviews*, vol. 3, no. 1, pp. 23–55, 2006.

[7] R. Axelrod, "The dissemination of culture: A model with local convergence and global polarization," *Journal of Conflict Resolution*, vol. 41, no. 2, pp. 203–226, 1997.

[8] M. Recasens, C. Danescu-Niculescu-mizil, and D. Jurafsky, "Linguistic models for analyzing and detecting biased language," *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, vol. 1, pp. 1650–1659, 2013.

[9] B. Gawronski and F. Strack, "On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes," *Journal of Experimental Social Psychology*, 2004.

[10] D. Kahneman and A. Tversky, "Judgment Unüer Uncertainty:Heuristics and Biases," *Science*, vol. 185, no. November, pp. 1124–1131, 1974.

[11] J. E. Korteling, A. M. Brouwer, and A. Toet, "A neural network framework for cognitive bias," *Frontiers in Psychology*, vol. 9, no. SEP, pp. 1–12, 2018.

[12] D. O. Hebb, "Organization of behavior. new york: Wiley, 1949, pp. 335," *Journal of Clinical Psychology*, 1950.

[13] I. P. Pavlov, "Classics in the History of Psychology – Pavlov (1927) Lecture 2," 1927.

[14] S. M. Kosslyn, "On cognitive neuroscience," 1994.

[15] A. Pepitone and L. Festinger, "A Theory of Cognitive Dissonance," *The American Journal of Psychology*, 1959.

[16] B. Gawronski and F. Strack, "Cognitive Consistency as a Basic Principle of Social Information Processing," *Cognitive Consistency: A Fundamental Principle in Social Cognition*, 2012.

[17] N. E. Friedkin and E. C. Johnsen, "Social influence and opinions," *The Journal of Mathematical Sociology*, 1990.

[18] P. E. Converse, "The nature of belief systems in mass publics (1964)," *Critical Review*, vol. 18, no. 1-3, pp. 1–74, 2006.

[19] M. H. Degroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.

[20] B. D. Anderson and M. Ye, "Recent Advances in the Modelling and Analysis of Opinion Dynamics on Influence Networks," *International Journal of Automation and Computing*, vol. 16, no. 2, pp. 129–149, 2019.

[21] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, "Mixing beliefs among interacting agents," *Advances in Complex Systems*, vol. 03, no. 01n04, pp. 87–98, 2000.

[22] P. Dandekar, A. Goel, and D. T. Lee, "Biased assimilation, homophily, and the dynamics of polarization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 15, pp. 5791–5796, 2013.

[23] A. E. Allahverdyan and A. Galstyan, "Opinion dynamics with confirmation bias," *PLoS ONE*, vol. 9, no. 7, 2014.

[24] P. Groeber, J. Lorenz, and F. Schweitzer, "Dissonance Minimization as a Microfoundation of Social Influence in Models of Opinion Formation," *Journal of Mathematical Sociology*, vol. 38, no. 3, pp. 147–174, 2014.

[25] S. Schweighofer, D. Garcia, and F. Schweitzer, "An agent-based model of multidimensional opinion dynamics and opinion alignment," *Chaos*, vol. 30, no. 9, pp. 1–34, 2020.

[26] D. Cartwright and F. Harary, "Structural balance: a generalization of Heider's theory," *Psychological Review*, vol. 63, no. 5, pp. 277–293, 1956.

[27] C. Chen, Z. Wang, and W. Li, "Tracking Dynamics of Opinion Behaviors with a Content-Based Sequential Opinion Influence Model," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 627–639, 2020.

[28] B. C. Runck, S. Manson, E. Shook, M. Gini, and N. Jordan, "Using word embeddings to generate data-driven human agent decision-making from natural language," *GeoInformatica*, vol. 23, no. 2, pp. 221–242, 2019.

[29] A. Tversky and D. Kahneman, "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment," *Psychological Review*, 1983.

[30] G. E. Hinton, J. L. Mcclelland, and D. E. Rumelhart, "Distributed representations (memory storage)," *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, pp. 77–109, 1986.

[31] Z. S. Harris, "Distributional Structure," *WORD*, 1954.

[32] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: A simple and general method for semi-supervised learning," *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, no. November, pp. 384–394, 2010.

[33] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A Neural Probabilistic Language Model," in *Journal of Machine Learning Research*, 2003.

[34] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations ofwords and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013.

[35] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014.

[36] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google ' s Neural Machine Translation System : Bridging the Gap between Human and Machine Translation," pp. 1–23, 2016.

[37] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2016.

[38] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 19–27, 2015.

[39] A. Radford and T. Salimans, "Improving Language Understanding by Generative Pre-Training (transformer in real world)," *OpenAI*, 2018.

[40] Radford Alec, Wu Jeffrey, Child Rewon, Luan David, Amodei Dario, and Sutskever Ilya, "Language Models are Unsupervised Multitask Learners," *OpenAI Blog*, 2019.

[41] T. Bolukbasi, K.-w. Chang, J. Zou, V. Saligrama, and A. Kalai, "Debiasing Word Embedding," *30th Conference on Neural Information Processing Systems*, no. NIPS 2016, pp. 1–9, 2016.

[42] M.-e. B. Colleen, A.-h. A. Anderson, and R. Zemel, "Understanding the Origins of Bias in Word Embeddings," 2019.

[43] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," 2020.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[45] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, 2019.

[46] A. E. Allahverdyan, A. Galstyan, A. E. Abbas, and Z. R. Struzik, "Adaptive decision making via entropy minimization," *International Journal of Approximate Reasoning*, 2018.

[47] K. Ganesan, C. X. Zhai, and J. Han, "Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions," in *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*, 2010.

[48] W. W. Zachary, "An Information Flow Model for Conflict and Fission in Small Groups," *Journal of Anthropological Research*, 1977.