FROM USER-GENERATED TEXT TO INSIGHT
CONTEXT-AWARE MEASUREMENT OF SOCIAL IMPACTS AND INTERACTIONS USING NATURAL LANGUAGE PROCESSING

BY

REZVANEH REZAPOUR

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Information Sciences
in the Graduate College of the
University of Illinois Urbana-Champaign, 2021

Urbana, Illinois

Doctoral Committee:

Associate Professor Jana Diesner, Chair
Professor Roxana Girju
Professor Karrie Karahalios
Professor Ted Underwood

ABSTRACT

Recent improvements in information and communication technologies have contributed to an increasingly globalized and connected world. The digital data that are created as the result of people's online activities and interactions consist of different types of personal and social information that can be used to extract and understand people's implicit or explicit beliefs, ideas, and biases. This thesis leverages methods and theories from natural language processing and social sciences to study and analyze the manifestations of various attributes and signals, namely social impacts, personal values, and moral traits, in user-generated texts. This work provides a comprehensive understanding of people's viewpoints, social values, and interactions and makes the following contributions.

First, we present a study that combines review mining and impact assessment to provide an extensive discussion on different types of impact that information products, namely documentary films, can have on people. We first establish a novel impact taxonomy and demonstrate that, with a rigorous analysis of user-generated texts and a theoretically grounded codebook, classification schema, and prediction model, we can detect multiple types of (self-reported) impact in texts and show that people's language can help in gaining insights about their opinions, socio-cultural information, and emotional states. Furthermore, the results of our analyses show that documentary films can shift peoples' perceptions and cognitions regarding different societal issues, e.g., climate change, and using a combination of informative features (linguistic, syntactic, and psychological), we can predict impact in sentences with high accuracy.

Second, we investigate the relationship between principles of human morality and the expression of stances in user-generated text data, namely tweets. More specifically, we first introduce and expand the Moral Foundations Dictionary and operationalize moral values to

enhance the measurement of social effects. In addition, we provide detailed explanation on how morality and stance are associated in user-generated texts. Through extensive analysis, we show that discussions related to various social issues have distinctive moral and lexical profiles, and leveraging moral values as an additional feature can lead to measurable improvements in prediction accuracy of stance analysis.

Third, we utilize the representation of emotional and moral states in texts to study people's interactions in two different social networks. Moreover, we first expand the analysis of structural balance to include direction and multi-level balance assessment (triads, subgroups, and the whole network). Our results show that analyzing different levels of networks and using various linguistic cues can grant a more inclusive view of people and the stability of their interactions; we found that, unlike sentiments, moral statuses in discussions stay balanced throughout the networks even in the presence of tension.

Overall, this thesis aims to contribute to the emerging field of "social" NLP and broadens the scope of research in it by (1) utilizing a combination of novel taxonomies, datasets, and tools to examine user-generated texts and (2) providing more comprehensive insights about human language, cultures, and experiences.

To my beloved family.

ACKNOWLEDGMENTS

This Ph.D. was a life-changing experience that has been nothing short of amazing. Throughout this journey, I have received great support and guidance from many remarkable individuals whom I wish to acknowledge.

First and foremost, I would like to express my sincere gratitude to my advisor and committee chair, Dr. Jana Diesner, for her continuous support throughout my Ph.D. study. During my tenure, she contributed to a rewarding graduate school experience by motivating me to aim for high-quality research, giving me intellectual freedom in my work to find my own path, and allowing me to grow as a researcher. I am indebted to her for her constant patience, assistance, encouragement, and guidance, as well as the opportunities she provided me throughout my doctoral studies. I could not have imagined having a better advisor and mentor. I would also like to express my deepest appreciation of my other committee members, Prof. Roxana Girju, Prof. Karrie Karahalios, and Prof. Ted Underwood, for their insightful feedback, comments, and encouragement, as well as their helpful career advice, support, and suggestions. I am honored to have them on my committee.

I am grateful to all of those with whom I have had the pleasure to work during this project and other related ones. I would like to thank my past and current labmates, Jinseok, Julian, Shubhanshu, Craig, Ly, Janina, Lan, Kanyao, Jay, Apratim, Pingjing, and Ming, for the stimulating discussions, for the sleepless nights we were working together to meet the deadlines, and for all the fun we have had in the last couple of years. Some parts of this thesis are based on collaborative work with Saumil Shah and Dr. Samin Aref. I am sincerely grateful for their tremendous contributions. I am also thankful to have had the opportunity to work with Dr. Rosie Jones, Dr. Sravana Reddy, and Dr. Ann Clifton at Spotify Research as my mentors. I have learned a lot from all these incredible people.

Additionally, I would like to extend my sincere gratitude to the faculty and staff at the School of Information Sciences at UIUC, including Dr. Catherine Blake, Dr. Masooda Bashir, members of the Research Services, and the Information Technology staff. I want to give special thanks to Meg Edwards and Penny Ames for being patient with my questions and always being there to help me with all academic matters.

Throughout my Ph.D. I have received supports from different institutions. I would like to thank the Ford Foundation (Grant no. 0155- 0370), the National Center for Supercomputing Applications (NCSA) at UIUC, the Army Research Laboratory (Grant no. W911NF-17-2-0196), the German Federal Ministry of Education and Research (Grant no. 01IO1634), the John D. and Catherine T. MacArthur Foundation, and the iSchool and Graduate College Fellowships for their generous support.

I would also like to thank my friends who supported me through thick and thin. First, I would like to express my gratitude to Ruohua, Dianah, Noah, and Michael, my Ph.D. cohort, with whom I started this journey and shared moments of deep anxiety and excitement. My sincere thanks go to Dr. Laila Hussein Moustafa, Dr. Azam Feiz, Dr. Parisa Kordjamshidi, and Dr. Jennifer Zavaleta for their help, encouragement, and advice. I would like to thank Ying, Priscilla, and Linh for their friendship and moral and emotional support. Also, I wish to thank Jessica for her trust and help in starting the UIUC ACM-W and co-organizing so many great events.

I am forever grateful to my beloved family, my mom and dad, my dearest sister Shabnam, my brothers Rashid and Ramin, and my in-law family, for supporting me spiritually throughout this journey and my life in general.

Finally, I want to thank the most important person in my life, my best friend and husband, Ali, to whom I dedicate this dissertation. It is difficult to express my appreciation because your support and love are so boundless. You have been a true and great supporter of me all these years. Thank you for always inspiring me to do better. These past several years have not been an easy ride, both academically and personally. Thank you for taking your valuable time to listen to me whenever I needed it and giving me support and comfort. I am super excited that we are both at the end of our Ph.D. journeys, and I cannot wait to start a new chapter with you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

A language is not just words.
It's a culture, a tradition, a unification of a community,
a whole history that creates what a community is.
It's all embodied in a language.
-Noam Chomsky

1.1 Motivation

Today's world is more globalized and connected than it was in previous decades. People have access to a huge amount of online and offline information in the form of texts, videos, and images. In addition, through microblogging and social networking sites like Twitter and Facebook, people with different types of socio-cultural backgrounds, such as first languages, cultures, and personal values, can connect, tell their stories, and discuss topics related to their personal lives and social issues around the world.

The digital data that are created as the result of people's online interactions consist of different types of personal and social information that can be used to capture attributes and signals, such as personal opinion, social status, personality, sentiment, and moral traits [PP11, GK09, BWP+15, GHK+13, SD14, SGQ+19]. Moreover, language is one of the most powerful means by which people demonstrate their implicit and explicit beliefs, ideas, social biases, and stereotypes [Fis93, Tri89, HF19, HS16]. Cultural and socio-economic knowledge about (online) users is essential to comprehend the content that they share, as this information can lead to a deeper understanding of people, their feelings, and discourse [SEK+13, SWM19]. In addition, these attributes can provide insights into users' attitudes

and moral judgments towards social issues and events in societies, which may help prevent phenomena such as race or gender biases, filter bubbles, and discrimination, while improving the efficiency of communication [SEK+13, SGQ+19]. Furthermore, understanding cultural and moral differences and commonalities within and across communities can help mitigate conflicts and biases that emerge on online platforms [HG07, GHK+13], better explain social and individual differences, and build more purposeful social platforms and products that empower people to express themselves.

In recent years, the increasing amount of publicly available online data and the evolution of computational models have provided researchers unique opportunities to study various types of digital trace data and user-generated contents to better understand people, their opinions, and their interactions [BWP+15, PSE+15, LPA+09]. Furthermore, social media data such as tweets and Facebook posts and likes have been used to gain insights into people's personality [Sch10], gender [SPE+14], and political orientation [RWAD17]. In addition, content on Amazon, Yelp, and Reddit has been utilized to understand customers' opinions and sentiments toward various products [LZ16, JWOH20]. For instance, Park and colleagues (2015) utilized social media data as well as self-reported personality traits to predict personality of users and found that language can provide signals to extract different personality traits [PSE+15]. Similarly, using Facebook posts and interactions, Boyd and colleagues (2015) analyzed users' core values to understand how they affect people's online behavior and thoughts [BWP+15]. Schwartz and colleagues (2013) used social media data to find salient words and topics that are associated with people's gender, age, location, or psychological characteristics [SEK+13]. Mislove and colleagues (2011) used social media data, specifically from Twitter, to extract the geography, gender, and race/ethnicity of the users in order to analyze how social media represent society [MLA+11]. In addition, researchers leveraged these user-generated data to study problems related to social good, such as controversy [ARAD17], fake news [SSW+17], and hate speech [SW17], among other issues.

Recently, the advent of large language models such as BERT [DCLT18] and GPT3 [BMR+20] resulted in tremendous improvements in natural language processing (NLP) applications such as text classification, question answering, and text generation. However,

despite the power and success of NLP methods, their limited assumptions regarding human-generated language caused them to ignore some embedded information in texts, which represent *humans* and *their social contexts* [HYS21]. Although different fields such as (socio)linguistics, psychology, and philosophy [BES14, Tri89, Eck12] discuss the importance of social constructs and factors, the majority of systems in NLP still do not go beyond just words [BK20]. As Clark and Schober noted [CS92], "*It is a common misperception that language use has primarily to do with words and what they mean. It doesn't. It has primarily to do with people and what they mean. It is essentially about the speakers' intentions.*"

Furthermore, leveraging social views and concepts from fields such as sociolinguistics, (computational) social science, and humanities has resulted in developing more "*socially-aware*" natural language processing models [LPA+09, HS16]. In recent years researchers have been more attentive toward social contexts and user-informed attributes when studying data and analyzing problems, such as cultural and personal values in user-generated texts [BWP+15, Sch04, WWH05, SGQ+19], gender biases and sexism in language models [MR17, BCZ+16, HF19, BMP21], polarized views and stances in social media [MKS+16, MSK17, BAB+11], emotion and sentiment detection [WWC05, HBR19, ZF20], and analyses of moral values in texts [DJH+16, GHK+13, SD14]. For example, Danescu-Niculescu-Mizil (2013) and colleagues developed a classifier to predict politeness and used this model to study the association between social factors like power, social status, community membership, and politeness [DNMSJ+13]. Del Tredici and colleagues (2019) analyzed how leveraging the social network (graph) of users can enhance the prediction of sentiment, stance, and hate speech in texts [DTMWF19]. Similarly, Aldayel and Magdy (2019) leveraged users' social networks to enhance the prediction of stance in tweets [AM19]. Shu and colleagues (2019) leveraged users' social contexts such as their social engagement and the relation between news, publishers, and users to analyze fake news and showed the usefulness of these features in enhancing the prediction of false information and news [SWL19]. Sap and colleagues (2019) analyzed hate speech in annotated tweets for toxicity and found that lack of awareness regarding social norms and taboos can lead to propagating bias in computational models [SCG+19]. Shen and colleagues (2019) used a lexicon-based approach to measure personal values in texts from different countries around the world in order to quantify cultural differences

in the expression of personal values on blogs [SWM19]. Dehghani and colleagues (2014) leveraged topic modeling to analyze online debates and demonstrated that the formation of moral values and beliefs is different between liberals and conservatives [DSSG14]. Finally, Pei and Jurgens (2020) studied intimacy in texts and showed that the "intimacy level of language reflects cultural norms of masculinity and femininity," and that social distance impacts intimacy; i.e., close friends and strangers may receive more intimate questions than acquaintances [PJ20].

The preceding concepts, studies, and observations have motivated this thesis to further investigate user-generated texts to gain a more comprehensive understanding of people and their online interactions and communications. Furthermore, exposure to various types of information and interacting with different communities and groups, online or offline, may change a person's values along with their knowledge and attitude towards various social phenomena. Since human behavior is social and adaptive, it is important and insightful to investigate how these interactions and exposures affect people's viewpoints, judgements, values, and behaviors. As the previous studies show, human-generated (real-world) language plays a crucial role in studying people and societies, as it can reveal different aspects in language and unfold embedded cultures, attitudes, and socio-economic differences or similarities [HS16, BES14]. This thesis contributes to the emerging fields of "social" NLP and computational social science by leveraging methods and views from natural language processing and social sciences to study and analyze the manifestations of signals and linguistic cues: namely, opinion, impact, personal values, and moral traits in user-generated texts.

While prior research has studied a diverse set of social phenomena in texts, the majority of work in this area has focused on benchmark taxonomies and datasets. The overarching goal of this work is to broaden the scope of research in these fields and demonstrate that utilizing a combination of novel taxonomies, datasets, and tools can enhance our understanding about human language and provide greater insights into their cultures and experiences. Following previous work that explored opinion, sentiment, and stance in user-generated texts, we extend these tasks in this thesis and propose new methodologies and approaches to (1) analyze how information products such as documentary movies (and books) affect a person's cognition, and how people express impacts of these products in texts such as (film)

reviews, (2) investigate the usefulness of personal values such as morality in the study of polarized viewpoints on Twitter, and (3) study how leveraging social and personal values such as emotional and moral states in texts can provide insights into people's interactions in social networks and their stability. Section (§1.3) provides more details about the proposed research questions.

1.2 Terminology Definitions

To facilitate the discussion, we first provide a brief explanation about key terms used in this thesis–namely *social impact*, *personal and moral values*, and *social interactions*–to clarify on the definitions, scopes, and meanings of these concepts in this work.

Social Impact: To investigate and analyze the social effects of events, projects, policies, and infrastructures, researchers and policymakers have been leveraging an analytical method known as a Social Impact Assessment (SIA) to understand the implications of their plans. Moreover, SIA is an instrument used to assess the sustainability of policies, projects, and plans in order to understand the depth and magnitude of their potential impacts and to increase fairness in communities. Due to its importance, SIA has been adopted in various disciplines, including psychology [Lat81], economics [S$^+$10], environmental studies [Van99], political studies [GS13], artificial intelligence [TCH$^+$20], natural language processing [HS16, JCT$^+$21, PRERSVC20], and media studies [BL08, Whi04], to name a few. The definition of social impact can vary based on the field, target, purpose, and application domain. However, the goal with SIA is to measure, understand, and anticipate the consequences of information or events on the cultures, behaviors, values, and beliefs of individuals, groups, communities, or societies [Lat81, Fin14]. Based on the target audience and affected groups, we can categorize the aspects of impact into three levels: macro (society), meso (group or community), and micro (individual). For instance, projects like autonomous vehicles can influence people, communities, and their environments [RBF$^+$20], while sources like media products and social media can affect people's perceptions, emotions, cognition, and attitudes [SVY14].

In this work, we define social impact as any changes (or reaffirmations of changes) in people's cognition, knowledge, behavior, or emotion that result from their exposure to information products such as documentary films and books. Our analysis is only focused on micro-level impact. To that end, we extract and analyze people's self-reported experiences, opinions, shifts, impacts, and changes after they watch a film or read a book. More details on the definition of impact are provided in Chapter 2.

Personal and Moral Values: People's values and beliefs affect their attitude, decision-making process, and what they perceive as good or bad, and moral or immoral. Personal values are defined as characteristics or behaviors that motivate people and guide them to take actions. Moreover, the core principles, customs, and cultures of our communities are among important factors that cultivate our values, knowledge, and behaviors toward various social concepts and phenomena. Therefore, as human beings, we each have different sets of moral characters and values that are formed by (or learned from) social norms, cultural values, educational systems, socio-cultural environment, and personal experiences. These differences are foundations to our moral values, ethical codes, and personality, in general, and define us, our ideologies, and judgments.

Following this definition, in this work, we leverage the Moral Foundations Theory [GHK+13] that categorizes human behavior into five basic principles 3.1 to analyze people's stances on several controversial topics. Our assumption is that the differences in people's values and idiosyncrasies manifest themselves in people's social discourse and everyday use of language, and we can operationalize (and extract) these values to better analyze, explain, and measure people's positions or perception about societal issues and (controversial) concepts. Chapter 3 provides more details about Moral Foundations Theory, our assumptions, and methodologies.

Social Interactions: Social interactions are known as any type of communication, relationship, or exchange between individuals through which they connect and share their experiences or influence (or are influenced by) other people [Mar07]. Social interactions are the basis of social structure and can be helpful in distinguishing the characteristics of dif-

ferent communities and networks [DNMLPK12]. For instance, using people's interactions, we can quantify the stability of a network and analyze its tendency toward maintaining balance or breaking off. Based on their social statuses and roles, people form different types of positive (affiliative) or negative (agonistic) relationships in their network.

In today's online world, with virtual connections, social interactions are mostly mediated by technology, e.g., texting, emailing, or skyping. In this work, we define social interactions as email communications between people in two different organizational settings. We further study and compare patterns of stability that govern the relationships between individuals in three different network levels: micro, meso, and macro. Chapter 4 provides details on how we analyze interactions and our methodology for studying structural balance.

1.3 Research Questions and Contributions

This thesis presents a systematic approach to studying user-generated texts by leveraging views from social science theories and methods from natural language processing, machine learning, and social network analysis. Moreover, by using computational approaches, this work introduces new concepts, methods, and approaches to describe and understand people, moral values, polarized viewpoints and stances, and social interactions in real-world settings. The following research questions (RQ) are going to be explored throughout this thesis:

- **RQ1: How Can We Leverage User-Generated Reviews to Analyze the Impact of Information Products on People's Behavior and Cognition?** *Detecting the Impact of Issue-Focused Documentary Films on People Using Reviews*

  Review mining is a well-known field of research that has been used to extract and study people's opinions and sentiments, mostly expressed in user-generated reviews [LZ12, ZJZ06]. This analysis has a broad range of applications, from understanding how people like and dislike products such as movies to predicting users' shopping habits. This research question extends the prior work in the area of review mining by looking into not just people's sentiment or opinion regarding a product but also how the products influenced people and if they had any impact on people's cognition,

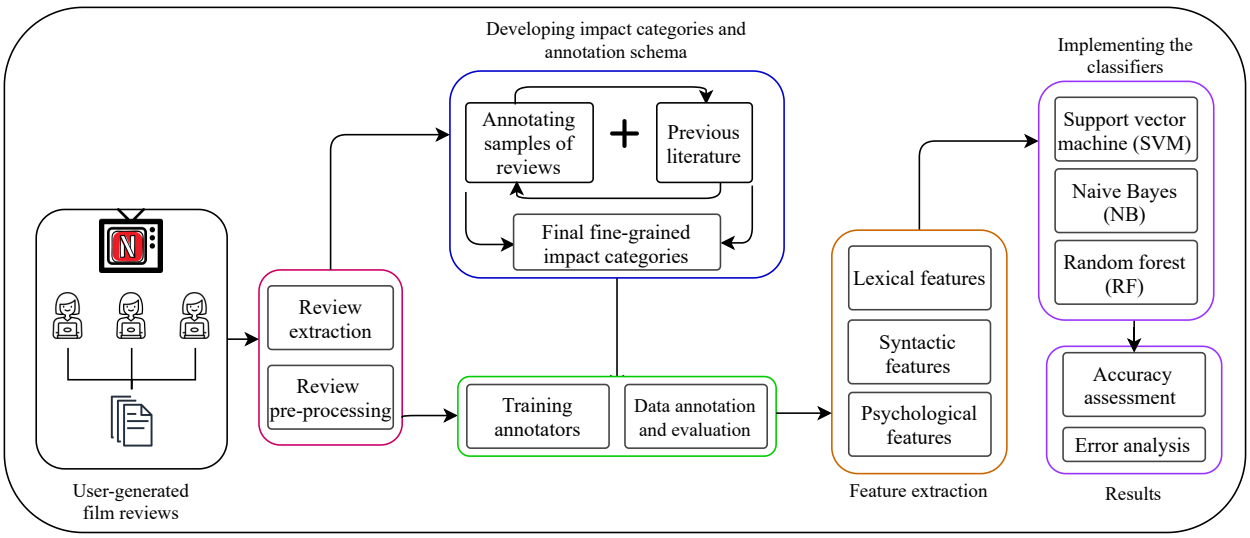


Figure 1.1: The overall workflow of extracting and analyzing the impact of information products in user-generated reviews

emotion, and behavior.

To study this problem, we focus on documentary films since they aim not only to tell a compelling story but also to engage the public and raise their awareness about specific social justice issues [DRJ16]. Traditionally, the impact of documentary films has been measured by conducting surveys and closed-group interviews. However, these methods are mostly limited to small groups of people, which may yield biased results. With the availability of texts from online sources, such as online reviews, and advancements of computational models, we have an opportunity to access and learn about a more diverse group of people.

In this work, we leverage user-generated film reviews and develop a computational model to predict if and how people discuss any types of impact when they review films. To categorize impact and find the influence of films on individuals, after closely reading reviews and reviewing the literature, we first develop a novel impact categorization schema consisting of nine different impact types. After labeling the data using these impact categories, we leverage three sets of features (linguistic, syntactic, and psychological) and a set of classic machine learning algorithms to predict the impact of information products on people's cognition and behavior. Figure 1.1 illustrates the

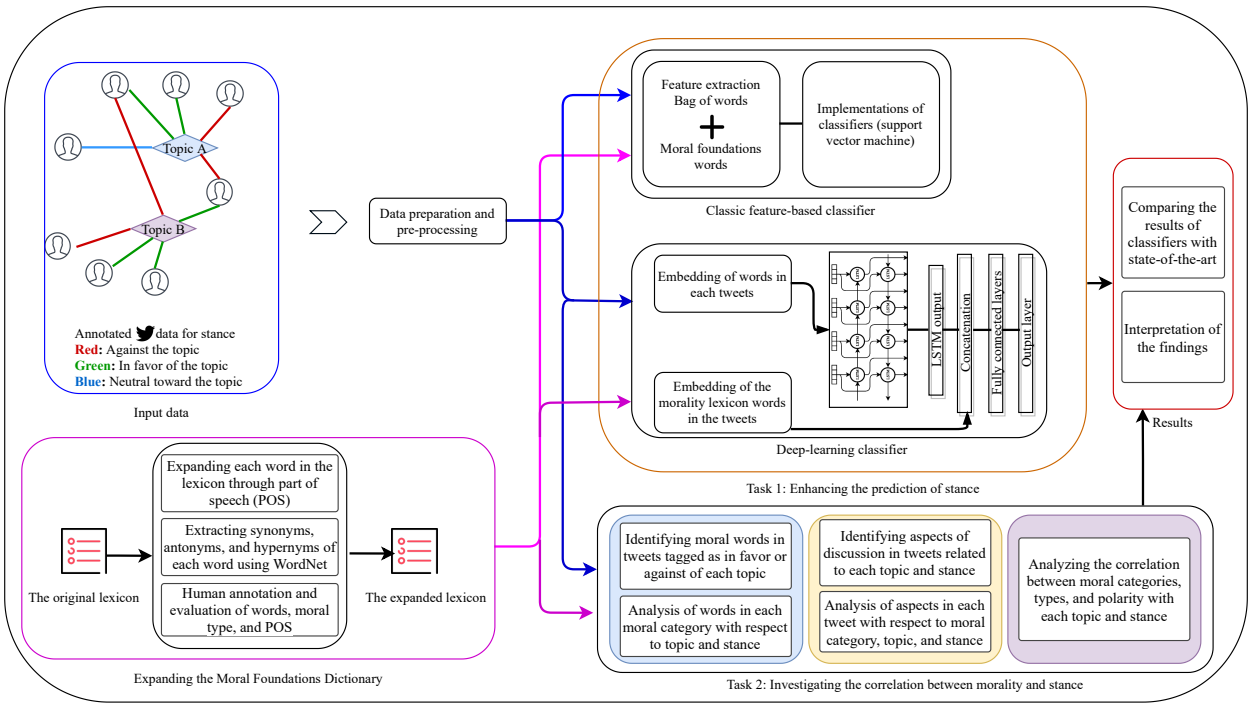


Figure 1.2: The overall workflow of incorporating morality in predicting and analyzing stance in tweets

overall workflow of our proposed model for impact analysis. Through this work, we aim to broaden the scope of research in the area of opinion and review mining and introduce a framework and methodology for studying the impact of (information) products on people's socio-cultural settings.

- **RQ2: How Can We Utilize Personal Values to Study Social Effects?** *Investigating the Effect of Moral Values on Stance Prediction and Analysis in Tweets*

  When people are expressing their opinion regarding social issues, their values, biases, and cultures are involved. Therefore, these values can help us investigate people's language. We aim to explore if the consideration of personal and moral values can help in predicting and studying social effects, e.g., stance analysis. Stance is the way that speakers position themselves in relation to social issues like abortion [MKS+16]. When a speaker describes their position toward this issue, they are expressing their attitude and personal values. Therefore, using features that represent values can benefit stance prediction in theory.

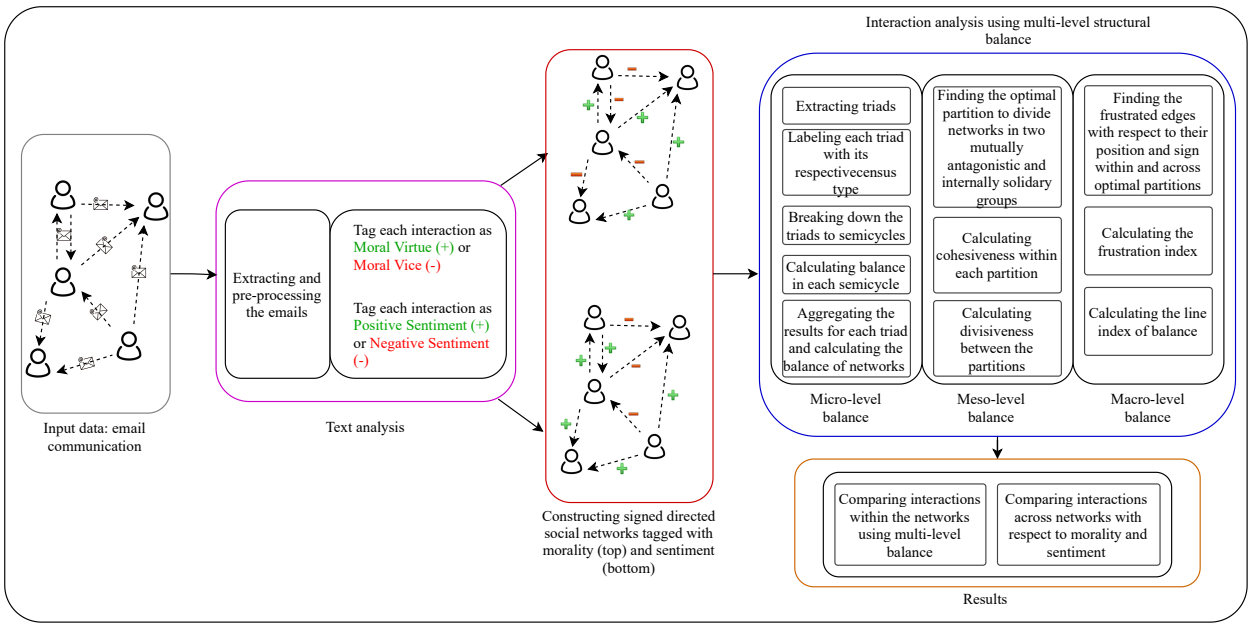


Figure 1.3: The overall workflow of leveraging linguistic properties and multi-level structural balance to examine social interactions

In this work, we use moral values to investigate this problem in detail. We leverage the Moral Foundations Theory [GHK+13] to first investigate the usefulness of moral values in enhancing the prediction of social effects. After that, we analyze stance with respect to morality and explain how morality and stance are associated in user-generated texts. Figure 1.2 shows the overall workflow of this study.

In this work we introduce and operationalize morality as a feature for natural language processing tasks, and we show that leveraging socio-cultural settings can result in a better understanding of the human language.

- **RQ3: How Can Social and Personal Values as Textual Properties Facilitate Studying People's Social Interactions?** *Exploring Social Networks by Extracting Sentiment and Moral Values from Email Communications*

  Real-world communication, verbal or nonverbal, written or visual, involves various types of explicit and implicit relationships, such as like versus dislike and trust versus distrust. To study real-world communication, we need to have access to people's networks, interactions, and perceptions. To create such real-world communication

networks, we can collect data by observing people and their interactions, surveying them, or analyzing digital trace data such as texts.

In this study, we leveraged email communications between people in organizational settings to extract sentiment and moral values, build communication networks, and analyze people's interactions with respect to these two features (Figure 1.3). Furthermore, to create the networks, we analyzed communication (i.e., emails) between every two individuals and used morality and sentiment to label these textual interactions. One of the basic dynamics and theories that explain interactions is structural balance [Hei44], which models how stable the relationships between individuals are in a network. Thus, after creating the networks and labeling the edge signs, we leveraged a multi-level structural balance analysis to analyze our networks.

This study advances the research in the area of network analysis by (1) extending the theory of structural balance to operationalize signed digraphs where both transitivity and sign consistency are required and considered for calculating balance in triads and beyond, and (2) leveraging natural language processing to infer social aspects from texts and using them to analyze and compare interaction and its balance in social networks.

To answer the proposed questions, this thesis utilizes existing publications in which I contributed as the first author. The following list provides the references to these publication and my contributions. It is noteworthy that the use of these manuscripts is in line with the publisher's thesis reuse guidelines [1].

- Rezapour, R., & Diesner, J. (2017). Classification and detection of micro-level impact of issue-focused films based on reviews. Proceedings of *20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)* (pp. 1419-1431). Portland, OR. ACM. [RD17]

  Contributions: I co-developed the research questions and research design, developed the impact categories, data annotation, and verification, designed and implemented

[1] ACM: `https://authors.acm.org/author-resources/author-rights`
ACL: `https://www.aclweb.org/anthology/faq/`
Scientific Reports: `https://www.nature.com/srep/journal-policies/editorial-policies`

the NLP models to extract the feature, implemented the machine learning models, evaluated the results, and prepared the manuscript.

- Rezapour, R., Shah, S., & Diesner, J. (2019). Enhancing the measurement of social effects by capturing morality. Proceedings *of the 10th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA). Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (pp. 35-45). Minneapolis, MN. [RSD19]

  Contributions: I extended the Moral Foundations Dictionary, co-developed the research questions, developed the algorithms for feature extraction and machine learning, evaluated the results, and prepared the manuscript.

- Rezapour, R., Dinh, L., & Diesner, J. (2021). Incorporating the measurement of Moral Foundations Theory into analyzing stances on controversial topics. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media (HT '21)* (pp. 177-188). Virtual Event, Ireland. [RDD21]

  Contributions: I developed the research questions, implemented and evaluated the moral analysis, executed and evaluated the topic extraction, evaluated the statistical analysis (Ly Dinh calculated the correlation), and prepared the manuscript.

- Aref*, S., Dinh*, L., Rezapour*, R., & Diesner, J. (2020). Multilevel structural evaluation of signed directed social networks based on balance theory. *Scientific Reports*, 10(1), 1-12. (*authors have equal contributions) [ADRD20]

  Contributions: This work was a group effort and Samin Aref, Ly Dinh, and I have equally contributed to this work. More specifically, I developed and implemented the algorithm of the micro-level analysis. Samin Aref developed and implemented the meso- and macro-level algorithms. All the authors equally contributed in evaluating the results and preparing the manuscript.

- Dinh*, L., Rezapour*, R., Jiang, L., & Diesner, J. (In preparation). Structural balance in signed digraphs: considering transitivity to measure balance in graphs constructed

by using different link signing methods.(*authors have equal contributions) [DRJD20]

Contributions: Ly Dinh and I have equally contributed to this work. I mainly assisted in extending the balance theory to include direction and transitivity. I designed and implemented the algorithms for analyzing the emails, including sentiment and moral analysis. I also designed and implemented the computational framework for analyzing networks and calculating balance in Enron and Avocado networks (Lan Jiang helped in extending the implementation and running the models). Ly Dinh and I equally contributed to evaluating the results and preparing the manuscript.

1.4 Ethical Consideration and Impacts

This section discusses ethical frameworks and broader impacts of the NLP models and their outcomes.

1.4.1 Biases in Data and Methods

As shown in previous studies, leveraging users' information has resulted in various types of biases when used in NLP systems and in a broader sense in artificial intelligence (AI) models [Nob18, HS16, Geb20, PRT08]. What we mean by bias is "a skew that produces a type of harm" for a specific community or group [Cra17]. Moreover, bias can result in two types of harms: *harms of allocation* and *harms of representation*. In the former type of harm, certain groups or communities are harmed since they are not represented fairly in the systems, e.g., due to their limited access to resource. For the latter one, systems "reinforce the subordination of some groups along the lines of identity"[2] [Cra17].

Due to these harms, while data-driven methods and tools have transformed people's lives, there are concerns about ethical and societal issues and impacts that can arise as the result of accessing and analyzing users' data [Wal14, HS16, Bru02, BC12]. When studying user-generated texts, information such as users' demographics (e.g., age, gender, and ethnicity) as well as their interactions and connections with other communities may be embedded in

[2] https://machinesgonewrong.com/bias_i/

the (responsibly) collected data. To make use of this source of information and to minimize harms, researchers leverage various methods, e.g., removing named entities and neutralizing texts to eliminate traceable information [BCD17, CBDC19]. However, it is challenging to properly anonymize data and remove personal identifiers and signals when studying data and sharing it for the purpose of reproducibility. Furthermore, studies show that marginalized communities such as people of color and women suffered the most as the results of automated decision-making tools [Nob18], e.g., search engines [Nob18, Swe13], recommendation systems and models [SSS+16], machine translations [HBF20], ad targeting algorithms [Swe13, ELS17], face recognitions and surveillance [GWGK19, CPCO20], and large language models and embeddings [BMP21, BCZ+16].

Some methods and datasets used in NLP applications and in this work may raise concerns about bias, especially in the development of more human-centric and socially-aware models, which are informed by users' social contexts and attributes [Fle20, HYS21, Geb20]. We acknowledge this as a challenge for data-driven models that consume users' data to comprehend language. Overall, this challenge is twofold. On the one hand, utilizing users' information may result in more personalized and fair systems that better understand everyone's language [Fle20, HYS21]. Moreover, the majority of NLP tasks, such as part of speech tagging, sentiment analysis, and dependency parsing, are *homogenized* and developed for languages like English that are used by the majority of people. As the result of this skewed representation, the majority of models (may) lack performance when used for analyzing marginalized languages, including variations and dialects, and may result in misclassifying or ignoring these communities [Fle20, HS16]. Having access to demographics and locations can assist in debiasing the data and increasing the representation of different communities in the training sample [HS16, BES14, SPE+14]. On the other hand, there are ethical concerns and questions about accessing this information, users' privacy, and how algorithmic or data biases affect people's lives.

To reduce any type of bias, following previous literature and research, we have the following suggestions. First, it is important to develop taxonomies, ethics guidelines, and infrastructures for the research communities to ensure that users' sensitive information will not be used in any unfair systems [RSW+20]. In addition, strengthening the guidelines and

regulations for ethically collecting, storing, and sharing users' data will help in minimizing stereotypical profiling and over- (or under-) representing communities, languages, and races [MWZ+19]. Furthermore, distinguishing algorithms and models that are or should be kept data agnostic would be helpful in understanding the subjective or objective nature of the models and their applications in NLP and beyond [Fle20, HS16].

Apart from the list above, emphasizing interdisciplinary collaboration and providing proper education about possible impacts and harms of skewed data representations and biased computational models can enhance and enable a more rigorous and targeted testing of the systems. In addition, diversification and the inclusion of marginalized communities will assist in better scrutinizing the models for any embedded societal issues and stereotypical profiling when analyzing user-generated data and making decisions (see [Geb20] for more detailed discussion). Finally, different fields such as information retrieval and library and information sciences have demonstrated that considering users' needs can increase the efficiency of the models as well as users' satisfaction [MR11, KBQ+19, STZ05, MOPS18]. Leveraging the experiences of and drawing insights from these fields may help in finding solutions for this challenge and developing more ethical models when leveraging users' information.

This thesis follows the practices discussed above; e.g., removing personal identifiers from the analysis and rephrasing the content if it is quoted in the texts. We also reviewed our studies with the Institutional Review Board (IRB) and followed their guidelines, when needed, to reduce any data misuse.

1.4.2 Reliability of NLP Models

AApart from issues with data and a lack of heterogeneous representation of communities, it is also important to be aware of the potential overgeneralization and unreliability of the results when analyzing the models. Regarding the former issue, it is crucial to first analyze the error rates and cost of false positives (and false negatives) before interpreting the results [HS16]. Moreover, in specific scenarios, relying on models that produce false positives may not be sufficiently sensitive, e.g., junk email filtration. However, there are cases in which

using such models results in harm and leads to overgeneralization or confirmation biases. Therefore, it is crucial to carefully explore and quantify the direct individual and societal impacts of NLP models before generalizing the results [SCNP21].

Furthermore, with the pervasive use of NLP models in our everyday lives, it is crucial to develop effective frameworks to test the robustness and reliability of these models and their outcomes. Through reliability testings, we can measure the "degree to which a system, product or component performs specified functions under specified conditions for a specified period of time" [80117] to exploit the negative impact of the systems. Previous research showed that using adversarial attacks [TJB+21] and implementing different types of perturbation methods and datasets [MS21] can help in analyzing the robustness of the NLP models. In this thesis, we diversified our studies by leveraging different generic and benchmark datasets to represent various topics and domains. While we performed multiple error analyses and tested the robustness of our models, we believe that our results and findings should not be further generalized since more in-depth analysis is needed to measure the reliability of our models using different real-world datasets.

1.5 Thesis Organization

The following chapters present the main motivations, backgrounds, methodologies, and findings of the explorations of the research questions enumerated above. The chapters are organized as follows:

In *Chapter 2*, we aim to answer RQ1 by analyzing user-generated reviews to extract and describe the impact of information products, namely documentary films, on people's cognition and behavior. More specifically, We will describe the development of our novel impact categories, data annotation, and the implementation of machine learning models to predict impact in reviews. *Chapter 3* focuses on leveraging and extending the Moral Foundations Dictionary for studying polarized viewpoints and stances in tweets (RQ2). *Chapter 4* investigates people's social interactions in networks and shows how we can use social and moral values extracted from user-generated data to analyze people's interactions (RQ3). Finally, *Chapter 5* presents overall conclusions, limitations, and future directions.

CHAPTER 2

LEVERAGING USER-GENERATED REVIEWS TO ANALYZE THE IMPACT OF INFORMATION PRODUCTS ON PEOPLE'S BEHAVIOR AND COGNITION

Contents of this chapter is based on the following paper (contributions are listed in §1.3): Rezapour, R., & Diesner, J. (2017). Classification and detection of micro-level impact of issue-focused films based on reviews. Proceedings of *20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, (pp. 1419-1431), Portland, OR. ACM.

2.1 Introduction

A recent improvement and advancement in the field of impact assessment is the consideration of the engagement of users with information products, such as scholarly publications [THLS13, PTGN10], and media content [NA09, Whi04]. We contribute to this line of work by developing a theoretically grounded classification schema for assessing the impact of media products on individuals. We bring this task to the domain of issue-focused documentaries, where funders and filmmakers are interested in knowing how their products engage communities, impact society, and raise awareness for the issues addressed in their films [KJ11, CA11]. Social impact assessment (SIA) has been practiced for more than five decades in different sectors [Van99, Bec01]. There is also a long lineage of work on measuring impact and public opinion in academia [Nap14, CD14]. In psychology, social impact is defined as the effect of an individual or group on other people [Lat81]. Also, impact assessment has a long tradition in the fields of environmental and political science [Van99, Bec01]. While the definition and naming of the concept of SIA may vary across fields and application domains, the goal with SIA is typically to measure, understand, and anticipate the consequences of information or events on individuals, groups, or society [Lat81].

There are various sources of stimuli which can trigger a change or confirmation of a person's behavior, mindset, or emotions. Examples of these sources from the area of information products entail books, TV series, and films, including documentaries. Documentary films aim not only to tell a compelling story [Ros12], but also to engage the public as well as to raise awareness about social justice issues, among other goals [NA09, CA11]. According to George Stoney, a film pioneer and professor at New York University, "fifty percent of the documentary filmmaker's job is making the movie, and fifty percent is figuring out what its impact can be and how it can move audiences to action" [KJ11]. Researchers at the University of Ohio conducted a case study where they compared the knowledge gained between two groups of students who watched a motion picture film versus a documentary film about the same issue [NA09]. They found that increased awareness and knowledge were higher among the participants who watched the documentary film.

The practical relevance of understanding the effects of information as represented in media products on people has motivated researchers from different fields to identify and measure the types and magnitude of these effects by using qualitative and quantitative methods [BHNS16, NA09, Nap14, CD14]. Access to user-generated as well as professionally-generated reflections on media products in the form of reviews has provided new opportunities for strategically generating and disseminating information, reaching people even in remote areas, and mapping the public opinion on various topics. With impact assessment becoming an increasingly important step for monitoring the post-production evolution of films [Nap14, CA11, BL08], scholars and practitioners have been developing strategies for increasing the engagement of individuals (micro-level), groups and organization (meso-level), and society (macro-level) with themes and stakeholders, and measuring the effectiveness of this process. Several (normative) frameworks for this process have been proposed, but practical implementations are lagging behind [Nap14, CD14, BL08]. In this chapter, we address this gap by developing a computational solution for discovering impact from user-generated reviews.

Previous studies analyzed the impact of documentary films on society by leveraging techniques from network analysis and natural language processing [DR15, JJ11]. We extend this work by turning our attention to the micro-level impact of documentaries. Moreover, people

express their opinions on review sites, and these reviews are valuable sources of information for both commercial and research applications. Reviews may demonstrate different types of impact on a person. In fact, the act of writing a review is already an indicator of impact. Since social movements often start with engagement on the personal level, understanding the type and magnitude of this type of impact can also contribute to better model macro-level effects. In this chapter, we leverage this idea and present a novel classification schema and method for measuring the impact of documentaries on people by using review data.

Traditionally, micro-level impact has been measured by conducting surveys and closed-group interviews [Whi04, JJ11, Lei04]. These methods are limited to small groups of people as study populations. In addition, based on the nature of surveys and interviews, the questions are sometimes not broad enough or limited to closed question responses, which can lead to biased results, and may lack the explanatory details necessary to capture different levels of impact. If collected and used ethically, large-scale corpora of written accounts of user perceptions from online sources and databases allow us to overcome these limitations. As a necessary precondition for our study, we first obtained permission for collecting a corpus of user-generated documentary reviews (no personally identifiable information was collected). Using online sources also gives us more opportunities to gather data from users from different locations, ethnicities, and educational backgrounds; potentially resulting in a diverse set of opinions considered for analysis.

With the work presented in this chapter, we have made the following contributions. First, we defined a categorization schema for micro-level impact based on a systematic review of different applicable bodies of literature (psychology, media studies), and close readings of samples from our data. Second, we developed a codebook for annotating reviews for these categories, and trained two individuals to apply the codebook to the data. Third, we analyzed the annotated data, selected features for training a classifier that predicts the defined impact categories (guided by prior work in review mining and our data analysis), and conducted experiments to evaluate the predictability of our impact types. We performed a detailed analysis on both results from human annotators and automatic prediction. We found that the sentence structure and tone in reviews are suitable features.

The knowledge gained with this work may be informative for future studies of impact in

different fields as it allows researchers to focus on tagging and measuring micro-level impact efficiently, even for large corpora, and with relatively high accuracy. Our work might also inspire new research in review mining, which has been traditionally focused on sentiment analysis and opinion extraction, predicting ratings and helpfulness, and text summarization. Finally, the gained insights may be useful to filmmakers, funders, and outreach teams for understanding individual impact on a more fine-grained level.

2.2 Literature Review

2.2.1 Review Mining

The large body of work in this area can be classified into three categories; 1) rating and helpfulness prediction [GI10], 2) summarization [GI10, HL04, ZJZ06], and 3) opinion and polarity extraction [Tur02, WLZ10, ML13, CMD06, DA07, DAPGD11]. Our application (impact detection) is marginally related to the third category, i.e., opinion and sentiment analysis. In that field, researchers have tried to identify the users' opinion about (specific features of) products, and categorized the users' sentiment about an object as being for example positive, negative, or neutral [LZ12]. Different methods have been used in this area, such as supervised and unsupervised learning techniques [GI10, WLZ10], sometimes combined with ontology-based approaches [ZL09].

Prior research on rating and helpfulness prediction has identified subjectivity or objectivity of the reviews as a useful feature for these tasks [GI10]. Other typically helpful characteristics for prediction include text meta-data features, e.g., the average length of sentences, lexical features, e.g., top $tf-idf$ n-grams, and syntactic features, e.g., counts of part of speech and parse tree constituents [KPCP06, ZV06]. In addition, Ng and colleagues found that using top unigrams is a prominent feature for separating reviews from other types of texts [ZV06].

The work in this chapter leverages prior insights on features and training algorithms from review mining, but differs from previous studies in that we aim to detect and classify the impact of films on peoples' cognition, emotions, and behavior.

2.2.2 Impact Assessment and Media Effects

Impact assessment (IA) of media focuses on the influence of information on people. Prior assessment of documentaries used a variety of methods, e.g., conducting surveys, analyzing screening metrics, and applying text-mining methods to user-generated and professionally-generated reflections on films [Whi04, JJ11, Lei04].

For example, Leiserowitz studied the impact of a Hollywood film about climate change ("The Day After Tomorrow") by quantitatively analyzing news articles before and after the release of the film, surveys, and interviews. His results showed both an impact on individuals' risk perception and an increase in the number of news articles by a factor of ten [Lei04]. Whiteman analyzed the relationship between films and social movements [Whi04]: He proposed a coalition model to assess the political impact of activist films and their role in social movements and public discourse by studying three successful films using interviews, participant-observation, and content analysis. His findings suggest that the new model broadens the range of impact after release.

Researchers from the John S. and James L. Knight Foundation published a report in which they made the case for using content analysis and sentiment analysis to analyze reviews written by attendees of screenings [KJ11]. They also developed and used a new metric called "key indicator points" (KIP), which considers and employs factors such as audience, content, sustainability, and social media by monitoring websites to measure the impact of media [JJ11]. In another study, a new set of metrics to measure reach, impact, the influence of media and engagement of the audience both online and offline was developed [KK$^+$11]. Researchers also used online surveys to measure the amount of knowledge that each audience could absorb [KK$^+$11, SZ15].

The Norman Lear Center in collaboration with the University of South California and the Knight Foundation are among the active research centers for finding new methods and metrics to evaluate the impact of different kinds of media. For instance, they have conducted an impact assessment study of a well-known, Oscar-nominated documentary film, "Food Inc.", where they used a combination of quantitative and qualitative methods. Based on their report, they compared two groups of individuals as viewers and non-viewers, and conducted

a survey with some open-ended questions [BHNS16]. Their findings showed that the group of viewers gained knowledge and intended to change their behavior as the result of the film's message. Beside quantitative analysis, they used the answers to open-ended questions to conduct a qualitative analysis by using open coding for each answer, and reported the ratio of the perception of the viewers around the main concept of the film. The result of this study indicates the capability of films in changing people and improving societal knowledge.

As mentioned in several of these reports, website traffic data is insufficient to show or measure users' attitudes. Therefore, it is necessary to use a combination of qualitative and quantitative techniques as well as other data-mining methods to better identify and analyze different types of impact of information products. Overall, IA of documentaries is a quickly evolving field. So far, basic text analysis techniques have been explored, but we argue that advanced data analytics can help in gaining a more comprehensive understanding of the influence of an information product on micro, meso, and macro levels [DR15].

2.3 Data

Our choice of data type, i.e., reviews, is driven by our goal, i.e., measuring the impact of documentaries on individuals. Reviewers can be divided into two groups depending on whether their contributions are intrinsically motivated, which is associated with voluntarily provided or user-generated content, versus extrinsically motivated, which typically applies when people write reviews as part of their job (professionally generated content), e.g., expert film reviews. This chapter is focused on the former type. As a data source, we chose to use Amazon, because their product reviews seem to attract a large population of content providers. The following sections provide details on data collection and data annotation process.

2.3.1 Data Collection

Based on our prior collaboration with a foundation, we chose eight documentary films related to different social justice issues: "Fed Up," "This Changes Everything," "Pray the Devil

Back to Hell," "Through a Lens Darkly," "Pandora's Promise," "Solar Mamas," "The House I Live in," and "Pay to Play." After obtaining permission from Amazon for our work, we collected 2,290 reviews. The films that relate to health and healthcare (Fed Up) and environmental issues (This Changes Everything) received the highest number of reviews (1,263, 664), which may suggest that individuals connect more with problems related to their everyday life compared to other social problems, e.g., criminal justice [DR15]. We randomly selected 1,000 reviews for labeling to keep manual annotation manageable. Very short and very long texts were excluded. The remainder, about 870 reviews, were annotated based on our codebook, which we introduce next.

2.3.2 Defining Impact and Data Annotation

What types of impact can an information product have on individuals? We use a data-driven and a theoretically-grounded approach to develop a practical solution to this question. We randomly selected a small sample of our review corpus for close reading. With the help of a linguistics student, we qualitatively and collaboratively explored types of influence reflected in reviews.

To verify and expand the set of the identified categories, we reviewed prior work from media studies and psychology [NA09, Whi04, KJ11, Lat81, BL08, Lei04, GS03, Vil01, GP13, ZBT+14, ST06]. Media can have substantial short-term and long-term influence [Vil01]. In a study conducted on children and adolescents, it was concluded that different kinds of media, such as movies, games, advertisements, and music, have significant influence on the behavior and attitude of viewers in different age groups [Vil01]. Media products, such as films and social media, can influence the way of thinking, social relationships, brain activity, and human identity [ZBT+14]. Besides raising awareness, documentary films can have an impact on individuals, society, and policies [BL08]. The impact of documentary films can be direct, indirect, or cumulative. Direct impact includes changes in individuals, and cumulative impact consists of changes in groups, systems, and conditions [GP13]. The level of impact on individuals varies, but based on different studies, media can change the behavior, cognition, belief, attitude, and emotion of a person [NA09, BL08, Vil01, GS03].

	Impact Types	Definition and Examples
Rank 1	Change in Behavior	A person indicates that they have changed their lifestyle or actions after viewing a documentary; person is influenced by the movie, e.g.: "Changed my lifestyle"; "I am doing more reading nowadays"; "buying healthier alternatives"
	Change in Cognition	A person changes their beliefs or way of thinking; a person clearly indicates that they have learned something new from the documentary and/or perceive something differently as a result, e.g.: "makes a person look at a problem from a new perspective"; "I knew so little!"
	Intention to Change	A person shows interest in changing their lifestyle in the near future; person is convinced by the movie enough to want to change something, e.g.: "I plan to use..."; "within a few years , I hope to do..."
	Change in Emotion	A person indicates that they experienced an affective change because of the documentary; person reacts emotionally to the general theme of the film or topics discussed in the film, e.g.: "The issue of...made me feel..."
Rank 2	Reaffirm Behavioral State	A person indicates that their behavior after viewing a documentary remains the same; person may have been influenced by a movie or a pre-existing experience, e.g.: "That is too bad that we will never be able to do anything about it..."
	Reaffirm Cognitive State	A person indicates that their cognition/knowledge after viewing a documentary remains the same; person may have been influenced by a movie or a pre-existing experience, e.g.: "I have had my experiences, and I opted to sober up of my own volition..."
	Reaffirm Emotional State	A person indicates that their emotion(s) after viewing a documentary remain the same; person may have been influenced by a movie or pre-existing experience, e.g.: "I am sick and tired of seeing my money go to waste"; "I felt like it would be such a downer. There is no doubt that lots of this is depressing".
Rank 3	Personal Opinion	A person expresses the general idea or opinion about a film without confirming any changes to them, person mentions other movies/ books that they find relevant, or suggests a documentary to others. The opinion can be positive or negative, e.g.: "This is an important issue and an important book"; "a must read"; "it does a good job of..."
Rank 4	Impersonal Report	Person summarizes the documentary and does not share any personal thoughts or opinions; information that the reviewer provides is from the film or addresses artistic or technical features of the film, e.g.: "the author...suggests that only national..."; "tells story of how..."; "the authors wrote in the introduction..."; "the film is executive produced by... "

Table 2.1: Excerpt from impact codebook

We conducted a three-step procedure for developing an impact codebook. First, we defined six impact types: "change in cognition," "change in attitude," "change in emotion," "change in behavior," "personal opinion," and "impersonal report." We wrote a codebook with precise definitions and examples, and trained two human annotators to label 50 reviews. Once completed, we closely studied the annotations and discussed the weaknesses and shortcomings of the codebook with the annotators. Based on their feedback, we found sentences related to "change in attitude" closely related to cognition and behavioral change. We also found that, in some cases, individuals talk about their future plans to change their behavior. To address these findings, we excluded "change in attitude" from the codebook, added a new class called "intention to change" to reflect the future plans, refined the codebook accordingly, and labeled a new set of reviews. We iterated through these steps (4 times) until we were sufficiently certain that the labels were comprehensive enough to cover different types of impact. Based on this process, we found that, in some cases, people also indicate previous influences from other sources, and reaffirm prior changes or current states. We accounted for these situations in the codebook.

The final category schema has nine types of impact: change in cognition, change in behavior, intention to change, change in emotion, reaffirm cognitive state, reaffirm behavioral state, reaffirm emotional state, personal opinion, and impersonal report (summary). We further grouped these nine types into four ranks that indicate the decreasing significance of impact. The codebook contains specific definitions and example sentences (short overview in Table 2.1). Examples are quoted from selected Amazon reviews.

2.3.3 Data Labeling

A review can entail none, one, or multiple types of impact. For example, a reviewer might start with a short summary of a film, then talk about their personal opinion, and later on mention the influence of the film on their personal life. To capture all these types of impact, we decided to label the reviews on the sentence level. To label the sentences, we first explained the task to two annotators individually, and asked them to annotate 10 reviews based on the codebook. After getting their results, we went through each sentence, discussed

Importance	Impact Types	#Sentences
	Change in Behavior	46
	Change in Cognitive	470
	Intention to Change	77
	Change in Emotion	170
Rank 2	Reaffirm Behavioral State	22
	Reaffirm Cognitive State	48
	Reaffirm Emotional State	0
Rank 3	Personal Opinion	2,060
Rank 4	Impersonal Report	831
–	NA	248

Table 2.2: Number of sentences of each type of impact

the chosen categories, resolved emerging issues, and gave each annotator more data to label.

Following the example of prior work, we had 10% of the data labeled by both coders [Neu16]. In addition to cross-annotation, we also designed three check points during the process to get feedback from the annotators, resolve any issues, and check if they still have a good understanding of the task and codebook.

Since the annotators came from different educational and cultural backgrounds, they had different interpretations of some labels. For example, one misunderstanding between the annotators was about "change in emotion." While one annotator marked sentences with emotional words such as "love" and "like" as "change in emotion," the other one labeled them as "general opinion." They also found the distinction between the classes of "general opinion" and "impersonal report" somewhat confusing without having a basic knowledge about the films. We resolved these issues through a detailed discussion.

To calculate the agreement between the coders, we used weighted Cohen's Kappa, since it is mainly designed to be used for categorical data. In the primary stage, the average inter-coder reliability was around 45%. The lowest agreement was related to reaffirmations, and the highest was related to "change in behavior". We understand that annotating sentences with 9 levels of impact can be a cognitively demanding task for the coders, especially in the beginning. As shown in our codebook (Table 2.1), this task requires a high level of pragmatic knowledge of tags and sentences. After discussing the misunderstandings and resolving the confusions, the average agreement increased to 97%, with the lowest being

related to "personal opinion" and "change in emotion". To achieve 100% agreement, either the codebook developer picked the final tag for the 3% disagreements, or they were excluded from the dataset. It is necessary to mention that after resolving the final issues and misunderstandings, the annotators were asked to review their tags and revise the annotated sentences accordingly. At the end, 300 reviews were checked by the codebook developer to assess the correctness of the assigned labels. Overall, the process of labeling sentences and making revisions took around 90 days.

We found that some sentences in the reviews do not belong to any of our defined categories. e.g., statements about experiences with delivery time or the quality of a DVD box. We labeled these sentences as "Not Applicable" (NA). They were later excluded from the data set because they have no impact weight. In some studies, NA sentences can be used as negative examples for learning, especially when building binary classifiers. We did not choose this option for our work.

Overall, we labeled 3,972 sentences. Table 2.2 shows the number of instances for each type of impact. Only 6% of the sentences do not feature any of our defined types of impact (NA). The majority of sentences, around 51%, are related to general opinions. We could not find any instances of "reaffirming emotional state" in the studied dataset.

2.4 Method

2.4.1 Feature Selection

As mentioned in the background section, we build our models based upon previous work. Therefore, we decided to use a combination of features suggested in the literature, namely lexical features, linguistic features, and psychological features. Lexical features can help us find words that are both highly salient and highly informative in texts. This process also entails the removal of a) dominant (with respect to the cumulative power law distribution of word frequencies in texts) yet not content bearing words, and b) highly rare words in a collection. Linguistic features entail the consideration of relation between words and their role in a sentence, subjectively connoted adjectives and other modifiers, punctuations as

determiners of sentence type (such as declarative or exclamatory), and the ratio of different parts of speech in a sentence. In NLP, these characteristics are known to be standard features for learning. In addition to these two features, we found some specific words to be uniquely indicative of (certain types of) impact in our data, e.g., authentic words. We refer to these features as psychological features and leverage prior work to capture them. In the following section, we provide more details regarding calculating each of these features.

Lexical Features

We considered salient unigrams, bigrams, and trigrams. After preprocessing the data, removing stop words, and the words with less than five occurrences, we selected the top 450 unigrams, top 300 bigrams, and top 100 trigrams based on their $tf - idf$ values (Eq.(2.1)).

$$tf - idf(t, d) = tf_{(t,d)} \times \log \frac{n}{1 + \mathrm{df}(t)}. \tag{2.1}$$

where $tf_{(t,d)}$ is the frequency of term t in document d, n is the total number of documents, and $df(t)$ is the number of documents in the document set that contain term t.

Linguistic Features

We considered a) grammatical features, i.e., presence of different parts of speech, b) sentence-level information, such as number of different punctuations, and length of sentences, c) sentiment of the sentence (computed as the ratio of positive and negative words to find the polarity of a string), d) ratio of dictionary words, i.e., words that can be found in a dictionary, and function words, i.e., words with less of a lexical meaning, but importance for sentence formation, and e) time orientation of sentences, conceptualized as past, present, and future, calculated by using different verb tenses and related adverbs. We used a combination of the Apache OpenNLP library [Apa14] and the "Linguistic Inquiry and Word Count" tool (LIWC 2015) [PBJB15] to extract the linguistic features. LIWC is a validated and broadly used tool, which classifies words into categories based on proprietary, embedded dictionaries. To be consistent with the outcome of LIWC, we normalized our ratios by sentence length.

We ended up with 45 attributes for linguistic features.

Psychological Features

Given the nature of this study, which is focused on personal effects of information products, we used LIWC's set of psychological features, which are compound metrics (descriptions adapted from LIWC): a) "Cognition Processes", which are words related to causation, discrepancy, tentative, differentiation, and certainty, b) "Informal Language Markers", such as assents, fillers, and swears words, c) "Core Drives and Needs", such as words that are related to personal drives like affiliation, power, achievement, reward, and risk, d) "Biological Processes", which are words related to health, body, and ingestion, e) "Perceptual Processes", such as words that refer to multiple sensory and perceptual dimensions associated with the five senses, f) "Social Words", which are words related to family and friends, g) "Clout", i.e., words related to the social status, confidence, or leadership of individuals presented in the text, h) "Tone", i.e., words related to the emotional tone of the writer, which are a combination of both positive and negative sentiment terms, i) "Authentic", which are words related to the real personality of the writer, and j) "Analytical Thinking", which comes from the words reflecting the experiences and logic of the writer. Overall, we considered 45 attributes for the psychological features set provided in LIWC.

2.4.2 Dealing with Imbalanced Class Distributions

As shown in Table 2.2, the high-ranking impact classes have fewer instances than ranks 3 and 4. This imbalance can bias the classifier such that ranks 3 and 4 get predicted with higher accuracy. To mitigate this problem, different approaches have been proposed. In addition to cost-sensitive learning, methods such as over sampling, under sampling, and combinations of the two have been used [CJK04, CBHK02]. Based on prior work, oversampling and using a combination of different techniques can result in a better outcome compared to cost-sensitive learning [CJK04].

To balance our dataset, we used a combination of two methods: oversampling for classes with small numbers of instances, and under sampling for large classes. For the first case,

Importance	Impact Types	After Balancing
Rank 1	Change in Behavior	276
	Change in Cognitive	940
	Intention to Change	462
	Change in Emotion	850
Rank 2	Reaffirm Behavioral State	110
	Reaffirm Cognitive State	288
	Reaffirm Emotional State	0
Rank 3	Personal Opinion	990
Rank 4	Impersonal Report	831

Table 2.3: Number of sentences of each type of impact after balancing

we used a method called Synthetic Minority Over-Sampling Technique (SMOTE). In this method, new instances are synthetically created using the k nearest neighbors. This method has a better performance compared to oversampling with replacement [CBHK02]. According to the number of instances of each class, a range between 100 to 500% was chosen using k=5 nearest neighbors to minimize the risk of over-fitting the classifiers. After oversampling and randomizing the data, we used random undersampling with the ratio of 9:1 to reduce the size of the large classes. These algorithms were implemented using WEKA's Java API. Table 2.3 shows the new number of instances after balancing the dataset. As shown in the table, the difference between the instances is minimized.

2.4.3 Classification

To classify the sentences, we used three different learning algorithms: Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB). We implemented the classifiers using WEKA's Java API [HFH+09] and conducted 10-fold cross validations.

To find the best combination of features, we 1) built a baseline model using the unigrams, 2) added bigrams, and 3) added trigram to complete the linguistic features. We then 4) added psychological features, and 5) linguistic features separately to the linguistic features. Finally, we 6) combined all three feature types. Before classifying the sentences, we chose and ranked

the best attributes using Information Gain (Eq.(2.2)) [RKC06].

$$InfoGain(Class, Attribute) = H(Class) - H(Class|Attribute) \tag{2.2}$$

For assessing prediction accuracy, we used the standard metrics of precision(Eq.(2.3)), recall(Eq.(2.4)), and F-score (with $beta = 1$)(Eq.(2.5)). The results for each feature and classifier are listed in Table 2.5.

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.4}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.5}$$

2.5 Result

2.5.1 Class Distribution

Based on the labeled data, we found that around 51.8% of the sentences contain general opinions, 20.9% provide summaries, and 6% do not contain any impact types (Table 2.2). Overall, approximately 20.9% of the sentences in our corpus feature emotional, cognitive, and behavioral impact (change or reaffirmation). This finding supports our effort to build a classifier that enables the detection of more fine-grained levels of micro-level impact of information products. As the number of instances for each class shows (Table 2.2), the ratio of "intention to change" is higher than "change in behavior," which suggests that people are more prone to plan to change their course of action or way of thinking compared to actually implementing these changes. Both "change in cognition" and "change in emotion" have the highest number of instances. Also, "reaffirming emotional state" has no instances, in contrast to the other two types of impact in rank 2, which may indicate that individuals

	Change Behavior	Change Cognition	Change Emotion	Intention Change	Reaffirm Behavior	Reaffirm Cognitive	Personal Opinion	Impersonal Report
Solar Mamas	**16**	0	**16**	0	0	0	48	20
Fed Up	2.4	**19.78**	4.81	1.58	**1.37**	1.92	49.31	18.82
The House I Live In	0.61	10	5.91	2.73	0	1.36	58.18	21.06
Pray the Devil Back to Hell	0	9	3.32	1.9	0	0	45.97	39.81
Pandora's Promise	0	8.64	0.82	1.23	0	**2.06**	63.79	23.46
This Changes Everything	0	7.01	4.24	2.37	0.2	0.59	**63.97**	21.42
Pay 2 Play	0	6.35	4.76	**4.76**	0	0	38.1	46.03
Through a Lens Darkly	0	1.89	3.77	3.77	0	0	41.51	**49.06**

Table 2.4: Different types of impact across each film (values are percent, the highest value of each column is highlighted)

may seldom feel (or be motivated to express) a confirmation of emotional states compared to their cognitive and behavioral states.

In addition to the comparative ratio of each impact type, we also analyzed the amount of different types of impact across each film to find out to what extent a film moved and motivated individuals (Table 2.4). Our results show that "Solar Mamas", a film about women, education, and mitigating poverty, changed the behavior rather than the cognition of reviewers. After reading the labeled sentences, we found that people stated that they donated money to charitable organizations, which indicates a positive influence of the film. In addition, we found that "Solar Mamas" and "The House I Live In", a film related to minimum mandatory sentencing, affected individuals' emotions more than other films. "Fed Up", a film related to health, sugar, and taxes, changed viewers' cognition and behavior. People also indicated more reaffirmation of behavioral and cognition states in the aforementioned film compared to the other considered movies. Overall, when compared to affecting change in cognition, fewer films could change the behavior of reviewers. This finding indicates that a) it is difficult to change peoples' behavior and more stimuli are needed for this purpose, and b) not every film is capable and/or aims to change the behavior on the micro-level. For instance, for "This Changes Everything," a film related to capitalism and environment, "change in cognition" is more desired than "change in behavior," since the former one can

Features		SVM			Random Forest			Naive Bayes		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Lexical	Unigram (Baseline)	53.3	46.4	47.3	63.8	61	61.3	50.9	49.2	49.3
	Unigram+Bigram	57.4	51.2	52.5	67.4	64.7	65	55.2	53.1	53.1
	Unigram+Bigram+ Trigram	57.3	51.5	52.7	67.7	65.2	65.3	56.1	54.4	54.3
Lexical + Psychological		71	70.6	70.6	80.2	79.2	79.5	55.2	52.8	52.5
Lexical + Linguistic		72.7	72.5	72.5	**81.4**	**80.8**	**81.1**	**64.4**	**64.1**	**63**
Lexical + Psychological + Linguistic		**73**	**73.1**	**73**	80.5	79.9	80.2	58.6	56.9	56.4

Table 2.5: Result of three classifiers using 10-fold cross validation (highest value per column in bold)

have a more long-lasting impact and lead the person to change of actions. In contrast, in "Fed Up," one would like to see both. These findings are shown in Table 2.4. In summary, we found that information products can change individuals' perception of social justice problems, raise awareness in society, and move people to act. These findings are aligned with the results obtained by others, such as the Norman Lear Center [BHNS16], where researchers interviewed people and used quantitative analysis to identify micro-level impact. This shows that our codebook and classification algorithms can capture some dimensions of the impact of information products.

2.5.2 Classification

As shown in Table 2.5, we first created a baseline model by using the top salient unigrams. This baseline is needed to enable the assessment of the influence of added features on the models. The best performance with the baseline was achieved with the RF classifier. Adding in bigrams and trigrams increased the performance of all three classifiers by around 5% (for all three accuracy metrics). Combining lexical (salient top unigrams, bigrams, and trigrams) and linguistic features further boosted the performance of all three classifiers. As shown in Table 2.5, accuracy increased by approximately 10-15%. Adding psychologically connoted terms to the set of lexical features also resulted in a considerable jump in the performance of SVM and RF. All metrics for these classifiers increased by nearly 15%. However, for

NB, adding the psychological features led to a drop in performance by roughly 2%. Finally, the combination of all three set of features improved the performance of SVM. However, the performance of NB and RF slightly decreased compared to the performance with the combination of lexical and linguistic features. Overall, RF outperformed SVM and NB when using lexical plus linguistic features, with an overall F1-score of 81%. However, SVM benefitted the most from combining all three feature sets, with a final F1-score of 73%.

In the following section, we analyze the performance of the classifiers in more details by 1) examining the top attributes for each feature type, and 2) conducting an error analysis.

2.5.3 Feature Analysis

To identify the most contributing attributes of each feature type, we calculated information gain to rank each attribute (Eq.(2.2)). The most informative attributes per feature type are listed in Table 2.6. As shown in Table 2.6, the best attributes of the lexical features come from the unigrams. Bigrams are rare in that set, and trigrams do not feature there. The combination of lexical and psychological features mostly benefitted from attributes of the latter one. Clout, tone, and analytical thinking are the top attributes, while the presence of lexical features is limited to one word, namely "people." However, this set is joined by "change" and "food" in the combination of lexical and linguistic features. With respect to psychological features, 1st person singular pronouns ("I"), sentiment words, pronouns, and time orientation of the sentences had a significant role in both, the lexical and linguistic set, and the lexical plus linguistic plus psychological set. Finally, the consideration of all features benefitted from the combination of top psychological and linguistic features, where attributes of the latter set are more highlighted than the former one. Based on these findings, we conclude that using linguistic and psychological features was beneficial for this task. As the analysis of the top informative attributes has shown, the structure of the sentences, grammatical indices, subjective words, and the tone of sentences are useful for predicting the impact.

Lexical	Unigram (baseline)	change, food, sugar, people, movie, documentary, film, hope, eat, climate, years, war, healthy, book, life, watch, sad, real, nuclear, industry
	Unigram+Bigram	change, food, people, movie, hope, film, documentary, sugar, this movie, kids, eat, years, book, climate, war, life, i hope, sad, problem
	Unigram+Bigram+Trigram	change, food, movie, kids, people, this movie, hope, film, sugar, years, documentary, eat, problem, war, perspective, book, climate, i think, life, i hope
Lexical + Psychological		tone, clout, analytical thinking, biological process, discrepancy, ingestion, social words, authentic, relativity words, causation, tentative, differentiation, people, insight words, drives words, perceptual processes
Lexical + Linguistic		1st person singular , negative words, personal pronouns, overall sentimental words, focus on past, articles, all pronouns, length of sentence, verb, dictionary words, focus on future, positive words, change, people, function words, adjectives, food, adverbs
Lexical + Psychological + Linguistic		1st person singular, tone, clout, personal pronouns, negative words, analytical thinking, total sentimental words, length of sentence, focus on past, all pronoun, discrepancy, articles, verbs, biological process, social words, function words, anxiety words, focus on future

Table 2.6: Most informative attributes of each feature set (top 20 or less)

2.5.4 Error Analysis

In addition to analyzing the contribution within and among features classes, we also studied the confusion matrix of the classifiers to find patterns in misclassifications. We chose the confusion matrix of the SVM because of its comparatively higher accuracy scores when using all sets of features. Table 2.7 shows the classified instances per impact category. As this matrix shows, "impersonal report," "personal opinion", and "change in cognition" are the most misclassified categories. While the first two classes have the lowest accuracy rate and the highest number of wrongly predicted instances, i.e., they are the least orthogonal to other classes and/or the least predictable with the features we used. In fact, the highest error for these two classes comes from predicting the two other class. This finding is consistent with the feedback from our human annotators, who found it hard to distinguish "personal opinion" from "impersonal report" without prior knowledge about a given film. After studying sentences in these two classes, we found them to be very similar to each other in sentence structure and lexicon use. The overlap occurs in cases where people tend to agree with a

a	b	c	d	e	f	g	h	
58.7	26	0.7	8.5	0.6	4.2	1.2	0	**a =Impersonal Report**
19.3	**57.2**	1.7	11.3	1.1	5.3	3.6	0.5	**b=Personal Opinion**
0.7	5.4	**92**	1.8	0	0	0	0	**c=Change in Behavior**
7.7	17.1	1.9	**66.6**	1.5	2.4	2.3	0.4	**d=Change in Cognition**
1	8	0	2.8	**87.5**	0.7	0	0	**e=Reaffirm Cognitive state**
1.8	6	0	2.6	0.4	**89.1**	0.2	0	**f=Change in Emotion**
0.6	5.8	0	2.2	0	0.2	**91.1**	0	**g=Intention to Change**
0	2.7	0	2.7	0	0	0	**94.5**	**h=Reaffirm Behavioral State**

Table 2.7: Confusion matrix of SVM classifier (values are percent)

concept in the film or want to add their own ideas to the concept. Furthermore, "change in cognition" has been misclassified as "personal opinion" more often when compared to the other high-ranked impact categories.

To further analyze this problem, for each of these classes, we randomly selected 30 sentences from different reviews, took them out of their original contexts, removed the labels, and asked the human annotators to label them again. Table 2.8 shows the underlying ground-truth result of the misclassified sentences labeled by the human annotators. The first column shows the original labels, column two provides the new labels selected by the human annotators during error analysis, and column three indicates the ratio of the labels. As the results of this case study show, human coders make similar mistakes as the classifier. This finding, which is also consistent with the confusion matrix, shows that some sentences are, in nature, hard to categorize, and more pragmatic and contextual analysis might be needed to solve this problem. Based on our discussion with the human annotators, we found

Initial Tags	**Secondary Tags**	**Ratio**
Personal Opinion	Personal Opinion	68%
	Change in Cognition	16%
	Impersonal Report	16%
Change in Cognition	Change in Cognition	36%
	Personal Opinion	54%
	Impersonal Report	10%
Impersonal Report	Impersonal Report	53%
	Personal Opinion	46%

Table 2.8: Error analysis: example for misclassified instances and human annotation

that being able to see preceding sentences in a review and familiarity with the content of the film would lower these errors.

2.6 Discussion

In this study, we developed a theoretically grounded and data-driven classification schema, related codebook, corpus annotation, and prediction model for detecting multiple types of impact of documentaries (as a specific instance of information product) on individuals based on user-generated content (reviews).

Our analysis of a set of reviews showed that information products can change peoples' conception of an issue, and can be associated with changes in attitudes toward societal problems. This finding is a meaningful outcome for sponsoring organizations, such as foundations, and filmmakers, as it demonstrates the potential impact of documentary films, and highlights the importance of assessing impact beyond frequency metrics. The data annotation and analysis procedures also showed that user-authored reviews contain or represent different types of impact, which justifies the development of a classification schema of micro-level impact types as well as the suitability of using reviews as a data source for studying impact. To identify and define impact types, and generate a codebook, we used a combination of reviewing prior work from media studies and psychology on the effects of print and social media on individuals, and qualitative exploration through close reading techniques by an interdisciplinary team that included a linguist. Our resulting categorization schema is composed of four levels: (1) (intent to) change and (2) reaffirmation in cognition, behavior, and emotions, as well as (3) personal opinions, and (4) impersonal reports (Table 2.1). Around 20.9% of the sentences in our corpus indicate high impact (type 1 and 2), 6% do not contain any impact type considered herein, and around 72.7% show lower levels of user engagement (types 3 and 4) (Table 2.2). Sentences of types 3 and 4 are often the focus of review mining studies that aim to predict ratings and sentiment. Our work builds upon and expands this line of research by separating impact into practically relevant and theoretically supported types.

To build classifiers, we worked with three sets of features: lexical, linguistic, and psy-

chological ones. We trained three commonly used types of classifiers, i.e., Support Vector Machines, Random Forest, and Naive Bayes. We first built a baseline model using top unigrams, gradually added the other feature types, and measured the incremental contribution of each type. The classification results (Table 2.5) showed that the combination of all three sets of features was most beneficial for SVM, where it improved the performance from 51% (baseline) to 73% final model (F1 score). The Random Forest classifier outperformed the other two models, and achieved the best overall performance, but did so by using only a combination of lexical and linguistic features (from 63% for baseline to 81%). Naive Bayes also performed well with a combination of lexical and linguistic features only, however, its score for F1, recall, and precision was lower than those for the Random Forest. The comparison of the top attributes of each set revealed that using informative attributes from the linguistic and psychological feature sets were helpful in building impact prediction model (Table 2.6). We also conducted an error analysis of misclassified instances, finding that sentences related to "personal opinion" and "impersonal report" are very similar to each other in structure and lexical profiles, making it challenging for the classifiers to distinguish the two (Table 2.7). Furthermore, human coders experienced similar challenges with these two types of impact, especially when respective sentences were presented out of context, and these difficulties carry through to the labeling and learning steps (Table 2.8).

In contrast to similar research in the field of review mining, where it is a common goal to identify user opinions about products, we categorized and based on that predicted different types of impact that an information product can have on individuals with relatively high accuracy. The findings from this work can advance review mining research by introducing a classification schema for micro-level impact assessment. Our outcomes may also be informative for sponsors, makers, and producers of documentaries as we provide a detailed yet comprehensive understanding of citizen engagement with issue-focused films. This might offer support in developing strategies for improving user engagement, and raising awareness for social justice issues. As shown in Table 2.4, the proposed impact codebook is helpful for formalizing and exemplifying a documentary film's various types of influence. As mentioned, some films can influence people to change their behavior and take action, e.g., by donating money or supporting a movement, while other films aim to raise awareness and

change the cognition of (re)viewers. Our findings might also help social movements to better understand the kind of impact that outreach work on certain topics can have. The potential future contributions of our codebook and classifiers are not limited to finding the impact of information products. These tools can also broaden our understanding of an individual's interactions with online communities, and the impact of the information products on individuals' everyday lives. Respectively, researchers and practitioners from different application domains of impact assessment can leverage our codebook to find the influence of policies or projects on the micro-level in their contexts. Our codebook can be domain-adapted and expanded to be applicable in other sectors [RBF+20]. In addition, in the era of Big Data, gaining better knowledge of online reviews can be useful to both academia and the corporate sector.

2.7 Conclusion, Limitations, and Future Work

The outcomes of this work confirm that documentary films can have different types of impact on individuals, and that these types can be identified from reviews. The developed codebook can advance research in review mining such that these types of impact can also be considered or be used as features. Our work might also improve research and methodology on impact assessment in different fields, from environmental studies to economics, in two ways. First, by advancing our knowledge about micro-level impact. Second, by increasing our understanding of the different types and magnitude of influence that various products or themes can have on individuals.

To study impact on the micro-level, we used online reviews instead of surveys and interviews. One advantage with this data source is a possible reduction in bias in comparison to results based on analyzing (text) data obtained through questionnaires and surveys. However, we do not know whether a review was only based on the impression that a person got from watching a film (if they watched it at all), and/or also by other information sources. In the future, it would be insightful to compare the types of impact that can be identified from interviews and surveys (offline sources) to those found in reviews (online sources, subject of this chapter). In addition, we limited our study to exclusively finding the impact of

documentary films, which often aim to raise awareness and affect the behavior, knowledge, or opinion of viewers. However, in addition to succeeding at the box office, some motion pictures might have similar goals. In our future work, we plan to identify and compare the types of impact of documentaries versus motion picture films on the same topics on the micro-level.

Our proposed method is not complete and has some shortcomings. Both humans and the predictor had difficulties with distinguishing two of the classes when labeling was done on a sentence level out of the review context. This problem will be further explored in the future by leveraging contextual analysis and more advanced techniques, such as pragmatic and deep syntactic analysis. Another challenge that we faced with this project was understanding and implementing regulations, (local) norms, and (cultural) expectations for accessing, collecting, and using review data in a lawful and ethical manner. The fact that some of these data are publicly available does not necessarily mean that one has permission to collect and analyze them. We obtained permission from Amazon for this process, but cannot share our corpus due to regulatory reasons and terms of service. However, the categorization schema can be used and further tested by others.

Acknowledgment

This work was supported by the FORD Foundation, grant 0155- 0370, and by a faculty fellowship from the National Center for Supercomputing Applications (NCSA) at UIUC. We are grateful to Amazon for giving us permission to collect reviews from their website. We thank Professor Corina Roxana Girju from the Linguistics department at UIUC for her helpful insights and advice for developing the codebook. We also thank Ming Jiang, Sandra Franco, Harathi Korrapati, and Julian Chin from the iSchool at Illinois for their help with this work.

CHAPTER 3

UTILIZING MORAL VALUES TO ANALYZE STANCE IN USER-GENERATED TWEETS

Contents of this chapter is based on the following papers (contributions are listed in §1.3):
Rezapour, R., Shah, S., & Diesner, J. (2019). Enhancing the measurement of social effects by capturing morality. In *Proceedings of the 10th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA). Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (pp. 35-45). Minneapolis, MN.
Rezapour, R., Dinh, L., & Diesner, J. (2021). Incorporating the measurement of Moral Foundations Theory into analyzing stances on controversial topics. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media (HT '21)* (pp. 177-188). Virtual Event, Ireland.

3.1 Introduction

User-generated text data are used in various fields to study, analyze, and extract people's cultures, behaviors, opinions, and emotions. The access and popularity of social media platforms such as Twitter attract individuals to participate in online discussions or share their points of view. To solve social issues, legislative changes or gradual reforms at the individual, organizational, and societal levels might be needed. Online conversations can serve as one among several sources for recognizing opposing points of view, also referred to as stances, and are a necessary ingredient for bridging gaps between groups, facilitating constructive conversations, and mitigating biases. However, different beliefs and perspectives on social, political, economic, and other potentially controversial issues can lead to debates or conflicts among groups, and can result in arguments, abusive discussions, and segregated

communities [CRF+11, PP10].

Given this type of behavior on online platforms, researchers have been investigating the relationship between basic principles of human values and the expression of opinions in user-generated text data by using (lexical) resources developed for this purpose and domain. This work is done as part of stance analysis [MKS+16], analysis of controversial topics [ARAD17], sentiment analysis [WWH05], and other standard NLP tasks.

Following this line of research, in this chapter, we operationalize and extract morality as a basic principle of human decision-making and an interaction guideline for people, e.g., when expressing themselves in relation to social or political topics. Our research is based on the assumption that people's backgrounds, cultures, and values affect their perceptions and expressions of knowledge and beliefs about everyday topics. These personal idiosyncrasies and differences manifest themselves in people's social discourses and everyday uses of language [Tri89], and can be helpful in analyzing or measuring people's positions or values regarding various social issues.

Measuring concepts such as morality is challenging, as it requires reliable operationalization and identification of regularities and accounting for context and meaning [Bat00]. To measure such concepts, we need to make sure that our results are-as much as possible-a reflection of the behavioral effect we want to study, not of the tools we use. The same is true for a wide range of social concepts that have been measured by applying lexicons to text data, such as opinion [WWC05], emotions [MMSP14], sentiment [PL+08, RWAD17], and culture [VHJC+13]. Moreover, natural language text data are inherently ambiguous, and signals relevant for detecting personal characteristics and social effects are sparsely distributed across text data. Therefore, we can make the basic assumption that the reliable measurement of human behavior based on text data requires robust, reliable, and transparent tools to measure any effects in a credible fashion [DR15].

This chapter contributes to this challenge by improving an off-the-shelf lexicon, known as the Moral Foundations Dictionary (MFD) [HJ04, GHK+13, GHN09], and mitigating biases in measurement by expanding and validating the lexicon (enhanced MFD) with multiple strategies and datasets. To achieve this goal, we performed a quality-controlled, semi-automated, and human-validated expansion of the original MFD (from 324 to 4,636 syntactically dis-

ambiguated entries) (discussed in §3.3). We then used the morality lexicons in two different tasks (as shown in Figure 1.2):

In the first task, we used the original and extended MFDs to show their effectiveness as features for supervised learning in order to predict two social effects: (1) stance, and (2) individual value or morality. To make a clear distinction between the two lexicons used in this chapter, from this point we refer to the original MFD as MFDO and to the enhanced lexicon as MFDE. Moreover, we aim to answer the following research questions (RQ) in task 1:

- **RQ1.1**: Can we enhance the prediction of social effects by leveraging user-informed language properties such as morality?

- **RQ1.2**: How is expanding the Moral Foundations Dictionary useful in the prediction task, and What insights can we gain from such analysis?

For predicting stance, we used the Semeval 2016 stance detection benchmark dataset [MKS+16]. In addition, we leveraged the Baltimore protest benchmark dataset [MHL+17] created for predicting people's morality in tweets. The stance detection task is relevant to our assumption, since individual differences in stance may relate to cultural differences. Therefore, we believe that the MFDE can be of assistance in improving the predictability of stance in user-generated texts. We found the Baltimore dataset relevant to our task since it comes from the same domain as our lexicons, is annotated on morality, and can show the usefulness of the MFDE lexicon. Our prediction models show that the MFDE outperformed the MFDO as a feature of prediction. Using morality as a feature increased the performance of both classical feature-based (93%) and deep-learning models (85.7%) in the majority of test cases. From that result, we conclude that morality can be a useful feature for detecting social effects in text data. In addition, we observe that lexicon expansion is worthwhile as it improves prediction accuracy in the majority of experiments on both morality and stance prediction.

After demonstrating the effectiveness of MFDO and MFDE for analyzing stance, in the second task, we further analyze the SemEval dataset to identify the type and magnitude of moral values that people draw from when expressing their opinions (stances) about social

issues. In brief, Moral Foundations Theory (MFT), which is discussed in more detail in the background section (§3.2), postulates that several innate and universal psychological features are the foundation of "intuitive ethics." Each culture then constructs narratives, norms, and institutions that are influenced by these foundations, thereby creating the unique value systems we see around the world that sometimes trigger conflicts among groups [GHK+13]. Because of this theory, we applied MFDE to the stance tweet corpus to measure and understand basic differences between opposing sides, regardless of political orientation [FCUPP16]. After categorizing each social issue with respect to morality types, we extract the most salient terms from both sides of the discussion (stance as favor and against). Overall, in the second task, we address the following research questions:

- **RQ2.1**: What basic morality types are contained in tweets about social issues?
- **RQ2.2**: What are the characteristics associated with each morality dimension, given there are opposing sides (known as stance) to a social issue?
- **RQ2.3**: What are the correlations between each morality dimension and stance?

Our results show that each social issue has different "moral and lexical profiles." While some social issues project more authority-related words (Donald Trump), others consist of words related to care and purity (abortion and feminism). Our correlation analysis of stance and morality revealed notable associations between stances on social issues and various types of morality, such as care, fairness, and loyalty, hence demonstrating that there are certain morality types that are more attributed to stance classification than others. Overall, our analysis highlights the usefulness of considering morality when studying stance. The differences observed in various viewpoints and stances highlight linguistic variation in discourse, which may assist in analyzing cultural values and biases in society.

This study makes several contributions. First, we introduce and operationalize morality as a feature for NLP tasks, and show that incorporating this information can lead to measurable improvements in prediction accuracy of social effects such as stance. Second, we apply the morality lexicon not only for morality prediction, but also for stance prediction, and this out-of-domain test enhances the robustness of our findings. Third, we improve the

Dimensions		Explanation
Virtue	**Vice**	
Care (CareVirtue)	Harm (CareVice)	Protecting versus hurting others
Fairness (FairnessVirtue)	Cheating (FairnessVice)	Cooperation/ trust/ just versus cheating in interaction with objects and people
Loyalty (IngroupVirtue)	Betrayal (IngroupVice)	Ingroup commitment (to coalitions, teams, brands) versus leaving group
Authority (AuthorityVirtue)	Subversion (AuthorityVice)	Playing within the rules of hierarchy versus challenging hierarchies
Purity (PurityVirtue)	Degradation (PurityVice)	Behavioral immune system versus spontaneous reaction

Table 3.1: Principles of Moral Foundations Theory

accuracy and transparency of measuring morality based on text data and provide a rigorous and reusable strategy for lexicon expansion and validation. In addition, our in-depth analysis of stance and morality enhances the status quo of knowledge about the application of the MFT to empirical data about controversial issues of general interest and independent of political orientation. We believe that the MFT is a suitable framework for examining morality and stance as it establishes the correlation among five fundamental moral foundations (care, fairness, ingroup, authority, purity) and moral behaviors (showing support/against a certain issue) [GHK+13]. Secondly, we advance psycho-linguistic understanding of how individuals' personal beliefs and values such as morality and stance can be manifested through language and discourse [Jas99]. Moreover, we expand the scope of stance analysis, which traditionally focuses on identifying binary polarization in discussions, by examining the narratives on either side of the topic in more depth and identifying patterns that describe moral foundations across topics. Finally, characterizing each moral foundation via aspects (key terms) from empirical data provides a window into individual values that are used when discussing controversial social issues.

3.2 Literature Review

3.2.1 Moral Foundations Dictionary

Moral Foundations Theory (MFT), considers four sources of individual moral judgment: 1) innate features, 2) human learning, based on the cultural context in which people are

embedded, 3) judgment based on situational intuition, and 4) pluralism of moral primitives [HJ04, GHK+13, GHN09, HJ+07].

Based on the MFT, the Moral Foundations Questionnaire (MFQ) was developed to facilitate measuring people's spontaneous morality [GHK+13]. Such standardized questionnaires are often used by researchers to conceptualize morality and elicit information about moral reasoning from individuals in a lab or remote settings. Socio-demographic characteristics (e.g., age, gender) and personal characteristics (e.g., educational level, political orientation, religiosity) were often used to aggregate and compare the results of these questionnaires. While questionnaires and lab experiments provided valuable information, they entail some shortcomings such as high costs, limited scalability, mock-up setups, and reliability issues of self-reported data [HWBS14]. Furthermore, alternative approaches like enhancement of a user study with neuro-physiological measures [DMK11], AI-based simulations [PS07], and extracting signals about morality from text data were used to address these shortcomings. In addition, text-mining techniques have been used to study user-generated, empirical data while eliminating issues with artificial lab settings and self-reported data.

The majority of prior studies that use NLP to study morality has focused on analyzing rhetorical aspects. Sagi and Dehghani [SD14] used the MFDO to measure the moral loading of news data by analyzing articles about socio-political conflicts (World Trade Center before and after 1993 and 9/11 attacks, Ground-Zero Mosque and abortion) from the New York Times. Moreover, The MFDO associates text terms with basic moral principles (see Table 3.1), and also includes a 6th "miscellaneous" category, which is a collection of morally relevant terms that could not be mapped to any of the considered categories. In another study, Kaur and Sasahara [KS16] leveraged a combination of the MFDO and latent semantic analysis to measure morality in tweets about different social issues, such as homosexuality and immigration. They found two dimensions, namely purity and care, to be dominant in conversations focused on immorality. Moral values have also been predicted using background knowledge and textual features. Lin and colleagues [LHPW+18] proposed a context-aware framework to aggregate external knowledge with text and improve morality prediction by 13.3% compared to the baseline. Garten and colleagues [GHJ+18] used a Distributed Dictionary Representations (DDR) approach to measure semantic similarity be-

tween dictionaries and text instead of using word counts. The DDR model was further used for predicting moral values of Twitter data related to Hurricane Sandy. Mooijman and colleagues [MHL+17] evaluated the relation between online moral rhetoric and violent protests by applying Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) to a Baltimore Protests dataset. Dehghani and colleagues [DJH+16] used the MFT to understand homophily, and found that people whose tweets are highly indicative of purity tend to be more like-minded. Finally, Fulgoni and colleagues [FCUPP16] leveraged the MFDO to analyze polarized debates in news sources. Their analysis showed different moral dimensions in liberals and conservatives conversations, where the former group favored care/harm and fairness, and the latter one focused on authority and loyalty.

Overall, a few studies have extended the MFDO using variations of word embedding models and calculating the cosine similarities between moral foundations context vectors and word vectors [KS16]. Our work builds upon prior studies of MFDO expansions, but differs from them in that we evaluate the semi-automated and human-validated expansion of the original lexicon as a feature for NLP prediction problems. Our ultimate goal is not to improve morality prediction or stance detection (though we do by a small margin), which are intensively studied problems in NLP. Instead, we aim to provide a rigorous strategy for lexicon expansion, and based on that a generally useful lexicon that can serve as a feature for a variety of information extraction and classification tasks. This can particularly be useful for people who want to use reliable resources. Moreover, our work advances the understanding of potential relationships between moral foundations expressed in public discourse and stances on social issues. We establish the connection between morality and stance from a psycholinguistics perspective [Jas99], in which an individual's moral values and beliefs are manifested through language and expressions in favor or against (also known as "stance") a certain topic [BRS02]. In the next section, we review prior literature on stance classification and establish the importance of understanding and measuring moral values in stance detection tasks.

3.2.2 Stance Analysis

Stance has been defined as the overall position of a person towards an idea, object, or proposition [SW09, BF88]. Extant literature has studied the relationship between morality and stance through the lens of "moral politics" [Lak95] and moral "rhetoric" [GBH+16, SD14] surrounding political issues such as presidential debates [Mar09]. Traditionally, moral politics and rhetoric were examined using discourse analytic methods such as critical narrative analysis [Rym95, SM14]. For instance, Rymes [Rym95] used critical narrative analysis to examine how at-risk youths assert their moral stance and "moral agency" towards violence through narrative and grammatical techniques. Though discourse analyses give thorough attention to how language and discourse elements are used to convey an individual's or group's moral stance, they are often based on smaller quantities of text data collected in specific social contexts. With these limitations in mind, we aim to leverage a larger-scale corpora of text data that covers social topics to investigate the narrative and linguistic features grounded in language on individuals' moral values and stances towards contemporary social issues.

One of the first empirical studies related to stance classification dealt with perspective identification. Lin, Wilson, Wiebe, and Hauptman [LWWH06] leveraged Bitter-Lemons' articles on Palestine-Israel conflict to automatically detect people's perspective regarding that issue. Hoover and colleagues [HPWY+20] used linear SVM (Support Vector Machines) to classify 35,108 tweets into "moral sentiments" (positive or negative) towards each of the five moral dimensions listed in the MFD. Similarly, Somasundaran and Wiebe [SW09] used a lexicon-based approach to identify arguments and sentiment in texts (on abortion, creationism, gun rights, and gay rights), and used these two features to classify stance. Both studies [HPWY+20, SW09] asserted that sentiment is a reliable indicator of an individual's perspective towards a social issue. Anand and colleagues [AWA+11] leveraged word-level features such as n-grams and syntactic dependencies to predict stance in debates. Li and Caragea [LC19] leveraged an existing sentiment lexicon [HL04] to predict stance by incorporating the lexicon words in the attention layer of a bidirectional LSTM.

Mohammad and colleagues [MKS+16] introduced the SemEval 2016 shared task for stance

detection (they had an inter-annotator agreement of 73.1%). Using Twitter as a source, they released a baseline model and dataset for analyzing stance in user-generated texts. Elfardy and Diab [ED16] analyzed the SemEval dataset by leveraging perspective detection, where they used frame and semantic analysis as well as textual information such as sentiment and Linguistic Inquiry and Word Count (LIWC) as features to predict stance. Recent work by Zhang and colleagues [ZYL+20] used the SemEval dataset to train a bidirectional LSTM that incorporates semantic and emotional valences as additional features to predict the stance of tweets. Aldayel and Magdy [AM19] leveraged social (media) network properties, such as a user interactions and connections, to study stance. Popat and colleagues [PMYW19] leveraged BERT (Bidirectional Encoder Representations from Transformers) to study stance, and showed the efficiency of their approach as it increased the state-of-the-art around 2-3%.

While prior research on stance detection has thoroughly investigated different linguistic and non-linguistic features to categorize stance, there is limited work on leveraging moral values as indicators of stance in text data. Baly and colleagues [BKA+18] used MFD as one of seven features to predict factuality of news articles (whether an article is unbiased or biased, fake or real). Johnson, Lee, and Goldwasser [JLG17] conceptualized morality as a frame of reference that politicians take to express their stance towards six political issues on Twitter. Ferreira and Vlachos [FV19] used moral values from the MFD as one of several lexical features to train a multi-label stance classifier that predicts either a presence or absence of stance in a tweet. Their classifier with moral values incorporated yielded a 12% higher performance compared to the baseline model. Prior research has also shown that moral values can be observed in language through the notion of stance [BKA+18, Jas99, RSD19]. However, in-depth knowledge about the relationship between morality and stance is still underexplored. Hence, in this study we further investigate the effect that moral values may have on opinion formation and expression, i.e., stance related to six social issues.

In this chapter, we investigate the impact(s) that moral values (care, authority, fairness, purity, loyalty) may have on opinion formation and expression, i.e., stances (in favor or against), related to six different social issues. For this purpose, we first study the impact of using Moral Foundations Dictionary (MFDO and MFDE) as an additional feature in predicting stance using both classical feature-based and deep learning machine learning

models. Furthermore, we study the topics and aspects related to each moral value and examine the correlation between morality and stance.

3.3 Moral Foundations Lexicon Expansion

The Moral Foundations Theory (MFT) categorizes human behavior into five basic principles that characterize opposing values (virtues and vices) as shown in Table 3.1. To enable the measurement of this theory based on text data, the Moral Foundations Dictionary (MFD) was developed and published [GHK+13, GHN09]. In the original MFD (called MFDO in this chapter), there is a sixth "miscellaneous" category, which is a collection of morally relevant words that were not yet mapped to any of the other categories. The MFDO associates 324 unique indicator terms (words) with the virtues and vices from the MFT. This lexical resource is highly valuable as it implements a theory. At the same time, it is limited in several ways: First, the number of entries is small and therefore might not capture all (variations of) terms indicative of morality in text data. This can lead to limited results, which may become part of our presumably valid knowledge about human morality. This problem can be mitigated through quality-controlled lexicon expansion, as presented in this chapter.

Second, we do not know based on what texts the MFDO was built, and even if we knew, these texts might be different from the ones to which researchers want to apply the MFDO. In NLP, this problem is known as domain adaptation. Several solutions to this problem have been developed [DI07, GBB11, SS07]. Given that the MFT aims to measure basic principles of human behavior, one could aim to build a generally valid, i.e., robust and validated resources with broad term coverage, which can then be used as is or further be adapted to domains, contexts, and culture. We chose the second strategy as it results in an improved general resource for others (and us) to use, and present our solution to this problem in this chapter.

In addition, the entries in the MFDO are not syntactically disambiguated, which can also limit the results, e.g., by capturing false positives. For example, one entry in the MFDO is "safe," which represents the virtue of care. In a text, "safe" can occur as a noun, which is probably not the intended meaning, or as an adjective, which is more likely to be the

intended meaning. This problem can be solved by adding the part of speech that represents the intended sense to each dictionary entry. We solve this problem as well. The outlined limitations of the MFDO in terms of size, scope, and syntactic ambiguity can lead to flawed analysis results. We fixed these issues as described in the remainder of this section and tested the benefit of this work as described in the next section (Method).

To expand the lexicon, we first sorted the words from the "miscellaneous" category (which we named "general") into virtues and vices. Next, we manually annotated each lexicon entry with one or more best fitting parts of speech (POS). We then manually added variations of the original words and senses, such as grammatical inflections, to the lexicon. All variations were added to the same category as the original root word. This expansion resulted in 1,085 words over 12 categories. We then added synonyms, antonyms, and (direct) hypernyms of all original entries automatically by using WordNet [Fel10]; a word graph of broad scope and general applicability. To evaluate and adjust the new additions, we trained two human annotators to analyze every word entry for its POS and morality category assignment. Their initial intercoder-agreement was 65% (Kappa). After that, we went through all entries again, resolved annotation disagreements, and removed the words that the annotators found not suitable for any predefined category. In MFDO, some words occurred in multiple categories. In our expanded lexicon, we made the word to category assignment exclusive by assigning each redundant entry to only the best fitting category. To justify these assignments, we asked the human annotators to study each applicable term and choose the most suitable dimension for the words by considering their common meaning. Finally, we expanded nouns with their plural or singular form, adjectives with comparatives and superlative, and lemmatized the verbs (following the MPQA subjectivity lexicon [WWC05]). Overall, our enhanced lexicon (MFDE) consists of 4,636 syntactically disambiguated, exhaustively expanded, and carefully pruned entries. Is this work worth the effort? To answer this question, we designed and ran experiments as described in the next sections. Our Enhanced Morality Lexicon can be accessed and downloaded at `https://doi.org/10.13012/B2IDB-3805242_V1`.

3.4 Data

We used two public benchmark datasets that were previously annotated for morality (Baltimore) and stance. The Baltimore data contains tweets related to the street violence that took place in Baltimore during the Freddie Gray protests (04/12/2015 to 05/08/2015). This dataset has been used to study if the rate of moral in tweets can assist in predicting violent protests [MHL+17]. From 19 million tweets that were collected, the authors of the original paper removed those tweets for which the geolocation was not the same as the cities where protests related to the death of Freddie Gray took place. Next, they had human annotators code 5,000 tweets for moral content based on the MFT. The annotated tweets were then used to train a deep neural network-based model (RNN and LSTMs) to predict moral values from tweets; resulting in 89.01% accuracy. To get the dataset, we ran the tweet IDs through the Twitter API and were able to extract 3,793 of the tweets (around 75.8% of the original tweets) for which human labels were available.

The stance dataset was made available for SemEval 2016 [MKS+16]. Using Twitter as a source, this dataset contains 4,870 tweets on six topics: abortion, atheism, climate change, feminism, Donald Trump, and Hillary Clinton. Tweets were hand-coded for stance, with the options being in favor, against, and none. The SemEval competition contained two tasks: Task A) was traditional supervised classification (on five topics mentioned above excluding Donald Trump), where 70% of the annotated data was used for training and the rest for testing. The highest accuracy (68.98%) was achieved by the baseline model, which used SVM and n-grams. Nineteen teams participated, and the best performing team achieved an overall accuracy (F-score) of 67.82% by using two RNN classifiers. Overall, about nine teams used some form of word embedding approaches, while some other teams leveraged publicly available lexicons (e.g., for sentiment, hashtags, and emotion), and Twitter specific features. For Task B, tweets on Donald Trump (a topic not used in Task A) were used. The highest F-score for Task B was 56.28% with nine teams participating. For our study, we combined the test and training sets from task A, and added the tweets on Donald Trump, resulting in a total of 4,870 tweets in our stance dataset.

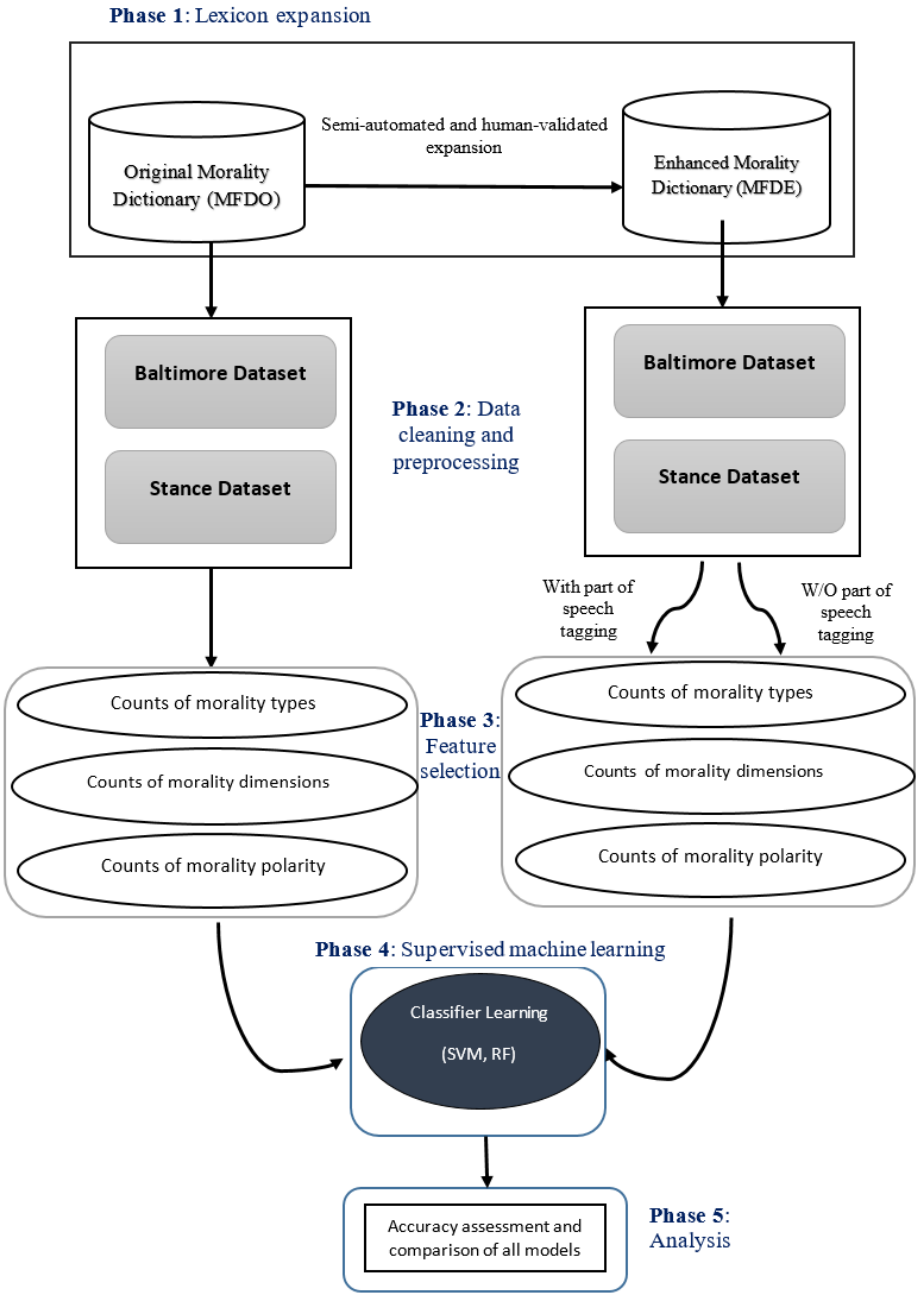


Figure 3.1: Task 1: Experimental design and workflow of the classic machine learning approach

3.5 Task 1: Leveraging Morality to Enhance the Prediction of Stance

To analyze the impact the morality lexicons have on predicting social effects, we built upon previous work in this domain. We assessed the performance of the lexicon and its expansion as features for both traditional feature-based and deep learning machine learning models. To test their impact on measuring social effects, we first created baseline models and then added the original and enhanced MFD to the baseline to test if morality is a useful feature and if the learning with MFDE outperforms MFDO.

3.5.1 Data Preprocessing

Tweets are noisy in that they do not follow conventional spelling schemes, and therefore require extensive data cleaning and preprocessing. To prepare our datasets for analysis, we removed all URLs, mentions (usernames), hashtag symbols, punctuation, and numbers from the tweets. We then expanded contracted words by automatically converting them to their assumed intended form (e.g., "I've" to "I have"). Finally, we lower-cased all words.

3.5.2 Classic Machine Learning

Figure 3.1 shows the overall experimental design used for this approach.

Feature Selection

We use morality words as additional attributes on top of the baseline models. We consider three types of counting to aggregate morality words per tweet: morality type count, morality dimension count, and morality polarity count. Morality dimension count represents the number of words per tweet that match any of the five morality dimensions plus the general category, resulting in six attributes (each horizontal row in Table 3.1).

Morality type count represents the number of words per tweet that match words in the vice or virtue category of each morality dimension (each box in the last two columns of Table

3.1). Using the MFDO results in 11 additional attributes, and the MFDE in 12 (since we divided the general category into vice and virtue).

Morality polarity count represents the number of words per tweet that match any virtue or vice category, regardless of the morality dimension (each of the last two columns in Table 3.1), which results in two additional attributes. We then test each counting approach with four feature sets, which are all subsequently explained: baseline (no morality feature), original morality, enhanced morality with POS, and enhanced morality without POS.

1) Baseline Model (BM): We replicated the baseline method from the SemEval competition, from which we re-used the stance detection dataset. In the original SemEval competition, the best performing model was the baseline, which only used word-level features, namely n-grams [MKS+16]. To re-create that model, we divided the dataset into its original sub-topics (feminism, climate change, atheism, Hillary Clinton, and abortion) and created one model for each sub-topic. We then replicated the unigram bag-of-words approach. To reduce the redundancy of the features, unlike in the original model, we removed stop words as well as words that appeared in less than 5% and more than 99% of the tweets. For the Baltimore dataset, we created a simple baseline by extracting unigrams from the dataset and using the counts of words to create feature vectors. We found that different numbers of tweets returned through the Twitter API and that a lack of transparency in the original models, such as preprocessing steps and metrics, limited our ability to reproduce the original works.

2) Original Morality Model (OM): The MFDO consists of five dimensions that are further divided into virtue and vice, and a sixth "miscellaneous" dimension. To aggregate the number of words per tweet, we used three types of counting, as explained earlier. For the morality dimension, we added 6 attributes on top of the baseline (OM6); for the morality types, we added 11 attributes (OM11); and for morality polarity, we added two attributes to the baseline model (OM2).

3) Enhanced Morality Model with POS (EM): We used the Python *NLTK* library to tokenize the tweets and tag each token with a POS [BKL09]. We then used all matches between the texts and the MFDE if they agreed in POS as features. Finally, we aggregated the extracted words using the three counting methods explained above.

4) Enhanced Morality Model without POS (EMNP): To test the impact not only of dictionary expansion in size but also of word sense disambiguation based on syntax, we built a set of models where any word from tweets that matched the MFDE was considered regardless of its POS. This model results in a higher number of words in the BOW than the EM model, since the grammatical agreement restriction was lifted from string matching. Again, we aggregated the extracted words using three count methods.

Classification

We used Support Vector Machine (SVM) and Random Forest (RF) as classification algorithms, as implemented in Python Scikit-learn package [PVG+11]. For the stance dataset, we replicated the approach from the original SemEval task; i.e., we used a 70%-30% split for training and testing. For the Baltimore dataset, we conducted 5-fold cross-validation. To test the performance of our models, we (1) built the baseline model by using the full set of unigrams (BOW), (2) added attributes created from MFDO to the baseline model, and (3) added attributes created from MFDE with POS and (4) without POS to the baseline model for each of the two datasets. For each model, we tested the previously explained counting options (morality dimension, type, and polarity). For assessing prediction accuracy, we used the standard metrics of overall accuracy, precision, recall, and F-score. Due to page limitation, we only report accuracy of the models (Table 3.2).

3.5.3 Deep Learning Models

We further investigated the usefulness of lexicons through a recurrent neural network (RNN) with bidirectional long short-term memory (LSTM) [HS97]. The advantage of LSTM compared to other RNNs is its ability to consider the whole context, since it is capable of bridging long time lags between inputs. To implement the models, we used Keras [C+18]. For the stance dataset, we used a 70%-30% split for training and testing, and for the morality dataset, we used 5-fold cross-validation. Baseline LSTM: To create the embedding layer, we leveraged the 200-dimensional word embedding from GloVe Twitter trained on two billion tweets [PSM14]. The embedding layer was followed by a Bidirectional LSTM of size 100, a

hidden layer with Sigmoid activation function, and an output layer with Softmax activation function. We further used Adam [KB14] to optimize the parameters and cross-entropy as the loss function.

Enhanced LSTM with Morality Lexicon: To create the enhanced model, we first created the embedding layers of the lexicon words for (1) the MFDO (OM), (2) the MFDE with POS (EM), and (3) the MFDE without POS (EMNP). Moreover, we first found the words that intersected between the lexicon and datasets and then created the embedding layers using the 200-dimensional GloVe Twitter [PSM14] without considering the morality dimensions, type, or polarity. After that, we concatenated the output of the baseline Bidirectional LSTM (as explained above) with the embedding of the morality words to build three types of models: (1) OM, (2) EM, and (3) EMNP. After concatenating the LSTM output and lexicon embedding, we used a hidden layer with Sigmoid activation function and an output layer with Softmax activation function. We further used Adam [KB14] to optimize the parameters and cross-entropy as the loss function.

One challenge in implementing neural network models is finding the best number of layers and settings (because there is no standard way of building the models). Since we are comparing different models, we found it difficult to choose a common set of numbers as the best hyperparameters, e.g., neurons, for both baseline and enhanced models. While we found that one hidden layer worked best for our models, to increase transparency we report the performance of our models with two sets of neuron sizes: 150 and 100. Table 3.3 shows the output of the LSTM models.

3.6 Task 2: Investigating the Correlation Between Morality and Stance

3.6.1 Morality Across Social Issues

To extract morality from the tweet corpus, we first preprocessed the data by converting all words to lower case, removing usernames and URLs, symbols, numbers, punctuation, and additional whitespace, and truncated repetitions of the same letters to two consecutive

occurrences. We then used *NLTK* [BKL09] to tokenize the tweets and tag each token with its respective POS. Next, searched the preprocessed texts for the terms listed in the MFDE. If a term in text and its POS coincided with a lexicon entry and its POS, we considered the term for our analysis and labeled the word with its respective morality type per tweet.

We then clustered the tweets based on social issues and analyzed the differences and similarities between these social issues with respect to their average moral values (Figure 3.2). We further grouped the tweets based on their labeled stances to investigate the relationship between stance, social issues, and morality (Figures 3.3).

3.6.2 Extracting Aspects Based on Morality

When discussing an issue or topic, people mostly refer to various aspects of that issue to better position their opinion with respect to their stance. For instance, to discuss the topic of "abortion," tweets expressing favor towards this topic may discuss *women's right* while tweets against it may talk about *intention to harm*. Based on this assertion, we investigated the potential connection between aspects related to each social issue and morality. According to [GL10, L$^+$10, ZHJU21, ZHJJU21], nouns are key factors for representing aspects or topics in texts. Following this finding, we extracted the top 50 nouns, including hashtags, from each social issue with respect to the $tf - idf$ score (Eq. 2.1) of the nouns.

To identify the aspects related to each morality type, we first clustered the tweets based on morality types (resulting in 12 clusters for each social issue) and averaged the $tf - idf$ score of each extracted aspect (noun) across these clusters. We then selected the top five aspects (if applicable) per morality type and social issue. Figure 3.4 visualizes the top aspects across morality types and social issues. In addition, Table 3.4 and Table 3.5 list the top aspects.

3.6.3 Significance Testing

Given that our variables are categorical in nature, chi-square (χ^2) tests of associations were performed on each of the six social issues. We examined a series of correlations between

Experiments		Stance Dataset												Baltimore	
		Abortion		Atheism		Climate		Clinton		Feminist		Trump			
		SVM	*RF*	*SVM*	*RF*	*SVM*	*RF*	*SVM*	*RF*	*SVM*	*RF*	*SVM*	*RF*	*SVM*	*RF*
Baseline	**BM**	66.42	62.5	69.54	64.54	61.76	**68.23**	60.81	60.13	58.94	60.7	51.17	45.07	85.2	83.91
Morality Types	**OM11**	66.42	62.5	**71.81**	65	**63.52**	**67.05**	61.14	57.77	**61.05**	**59.29**	50.7	49.29	85.18	84.12
	EM12	**67.85**	62.85	71.36	62.72	**63.52**	60.58	**64.18**	58.78	57.19	57.19	51.64	47.88	**85.6**	**84.73**
	EMNP12	66.07	**63.21**	71.36	**66.81**	62.35	62.94	62.38	**61.48**	58.94	**59.29**	**52.58**	**52.58**	85.31	84.12
Morality Dimension	**OM6**	**68.21**	**63.57**	70.45	**69.09**	**62.35**	64.7	59.79	58.1	59.29	60.7	51.17	46.94	85.31	**84.73**
	EM6	**68.21**	62.5	**71.36**	66.81	60.58	64.11	62.83	57.43	58.94	58.59	52.58	**53.99**	**85.71**	84.44
	EMNP6	**68.21**	62.5	70.45	60	60	**66.47**	**64.52**	**59.12**	**60**	**62.45**	**54.92**	50.7	85.55	84.1
Morality Polarity	**OM2**	67.14	63.21	69.09	**69.54**	**62.94**	65.29	62.83	57.09	58.24	56.84	52.58	**50.23**	85.31	84.99
	EM2	**67.85**	**64.28**	**72.27**	66.81	**62.94**	61.17	**63.17**	58.78	57.19	**61.05**	50.7	43.19	**85.6**	**85.31**
	EMNP2	67.14	63.92	71.81	64.54	61.17	**67.05**	**63.17**	**60.13**	**59.29**	56.49	**53.52**	49.29	85.49	84.84

Table 3.2: Result of predicting stance (first 12 columns) and morality (last two columns) with SVM and RF for stance and Baltimore datasets (Accuracy) (the highest performance per set of experiments (OM, EM, and EMNP — each half column) in bold, the highest accuracy per each model (each column) in gray)

stance and (1) different morality dimensions (number of words per tweet that match the five morality dimensions plus general category, as shown in Table 3.1), (2) different morality types (number of words per tweet that match words in the vice or virtue category of each morality dimension), and (3) morality polarities (number of words per tweet that match any virtue or vice category regardless of the morality dimension or type). We set our confidence level for statistical significance for considering any pairs of correlations to 95% ($p = 0.05$).

3.7 Results

3.7.1 Task 1: Leveraging Morality to Enhance the Prediction of Stance

Table 3.2 and 3.3 show the results of predicting stance and morality. In both tables, the highest performance for each set of experiments (OM, EM, and EMNP) is marked with bold text, and gray cells indicate the highest accuracy per model (per column).

The results for the classic machine learning models are shown in Table 3.2. For the Baltimore dataset (originally annotated for morality, the last two columns in Table 3.2), using a simple set of basic unigram features and classic machine learning models resulted in a baseline accuracy of 85.20% for SVM. Adding the simplest morality model (OM11) led to a small decrease (about 0.02%) with SVM. For the RF model, adding OM11 increased the

#Neurons in Hidden Layer	Stance Dataset							Baltimore
	Experiments	Abortion	Atheism	Climate	Clinton	Feminist	Trump	
N = 150	BM	62.5	68.181	67.647	58.445	57.192	51.643	84.2391
	(1) OM	**68.214**	68.636	65.882	56.081	**57.894**	50.704	85.504
	(2) EM	67.5	72.272	**70**	**63.851**	**57.894**	50.234	**86.163**
	(3) EMNP	65.714	**73.181**	68.823	57.432	57.543	**54.929**	84.634
N = 100	BM	65.714	65.454	**70.588**	59.121	**58.596**	51.173	84.845
	(1) OM	64.642	66.363	69.411	**60.472**	56.842	51.643	85.9
	(2) EM	**67.142**	70.909	69.411	59.797	54.385	**53.521**	**86.612**
	(3) EMNP	64.642	**71.363**	67.647	56.756	58.245	49.765	83.58

Table 3.3: Result of predicting stance (first 7 columns) and morality (last column) with LSTM model for stance and Baltimore datasets (Accuracy) (the highest performance per set of experiments (OM, EM, and EMNP — each half column) in bold, the highest accuracy per each model (each column) in gray)

performance by about 0.21%. Adding information about morality-relevant words in more sophisticated ways (EMs and EMNPs) increased accuracy for both RF and SVM. As shown in Table 3.2, the best result for RF was achieved using EM2 (85.31%), and for SVM using EM6 (85.71%). For the stance datasets, the results are shown in the first 12 columns of Table 3.2. Depending on the sub-topic, our baseline accuracy ranged from 45.07% (RF, Trump, stances hardest to predict) to 69.54% (SVM, atheism, stances easiest to predict). As observed for the Baltimore data, adding lexical morality features to stance increased accuracy across our baseline in all but one case (Climate, RF).

The results for the LSTM model for both datasets are shown in Table 3.3. As mentioned before, we used two sets of neuron sizes for the hidden layer. For the Baltimore dataset, using the MFDE achieved better performance in both implemented models. The highest accuracy was obtained by the enhanced LSTM model that used EM, 86.61% (N=100). For the stance dataset, adding morality embedding to the output of LSTM (baseline) resulted in an outperformance of the baseline in 83.33% of cases (10 out of 12).

Does using morality as a lexical feature improve prediction accuracy for the selected NLP tasks? Comparing the baseline to any models that include morality, we conclude that adding morality as a lexical feature increases accuracy in 13 out of 14 cases (93%) for feature-based learning (considering RF and SVM models for each topic) and in 12 out of 14 cases (85.7%) for deep learning (considering experiments with two sets of neurons for each topic). This finding suggests that using morality as a feature is helpful for standard

NLP tasks — and possibly other tasks as well, which would need to be explored in future work. Does expanding the MFDO pay off? We find that for feature-based learning (Table 3.2), in 29 out of 42 cases (69.05%), the accuracy with any MFDE feature outperforms the models with MFDO features; in 21.43% of the cases, MFDO outperforms MFDE; and in 9.52% of the cases, both versions of the dictionary lead to equal results. For the LSTM, 9 out of 14 models (64.28%) had better performance when using MFDE, while 14.28% of models (2 models) worked better with MFDO (Table 3.3). From that result, we conclude that lexicon expansion is worthwhile as it improves prediction accuracy in the majority of our experiments, especially for feature-based learning.

Does disambiguating word sense in the MFDO via POS pay off? Based on the results in Table 3.2 and 3.3, we found that the syntactic disambiguation of lexicon entries leads to only minor quantitative improvements. We believe that the usefulness of POS tags can be further tested with other types of user-generated data that follow more conventional grammatical rules. Beyond what we measured in this chapter, this additional layer of information might further boost the quality of the data.

Based on the results of all implemented models, highlighted in Table 3.2 and 3.3, we found that using MFDE results in higher performance than other models (MFDO and BM).

3.7.2 Task 2: Investigating the Correlation Between Morality and Stance

Morality Across Social Issues

To understand how moral foundations are manifested in tweets, we identified the occurrence of 12 morality types in each of the six social issues by applying MFDE to the tweets (§3.6.1). Figure 3.2 visualizes the average morality across six social issues. We observe that social issues entail a distinctive distribution of morality types. While feminism contains the highest morality value of fairness, atheism contains the highest morality value of purity. Moreover, harm and authority are more prevalent in the tweets related to the topic of abortion, and authority is among the top morality types in all social issues.

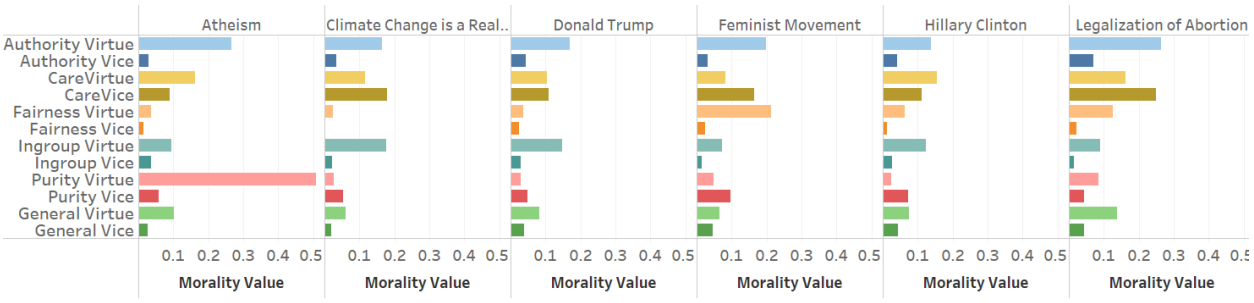


Figure 3.2: Average morality values across each social issue

Furthermore, tweets mentioning Donald Trump have higher values of loyalty and authority than tweets mentioning Hillary Clinton. On the other hand, tweets mentioning Hillary Clinton have higher morality values of care than the ones discussing Trump. In addition, the topic of climate change features higher values of loyalty and is the only social issue in our dataset that has no morality value of cheating. This finding may have resulted from sparsity in our dataset and invites a more detailed investigation in the future with additional datasets to confirm or reject this finding.

As shown in Figure 3.3, we also investigated similarities and differences between morality types with respect to stance and social issues. Similar to the previous analysis, the results show different profiles of morality types when considering stance. For instance, discussions against the topic of climate change do not indicate care but have high values of harm. Similarly, discussions against the topic of abortion and feminism demonstrate higher values of harm than discussions in favor of this topic. Tweets in favor of feminism also contain higher values of fairness than those against it. Moreover, discussions against the topic of Hillary Clinton project higher values of harm, and those written in favor of her consist of higher values of care and fairness. In contrast, tweets written in favor of the topic of Donald Trump consist of higher values of harm and care as well as authority and loyalty.

Our analysis highlights the importance of considering morality when studying stance. The differences that we observed in various viewpoints (stances) demonstrate linguistic differences in discourse and can assist in analyzing cultural values and biases in society.

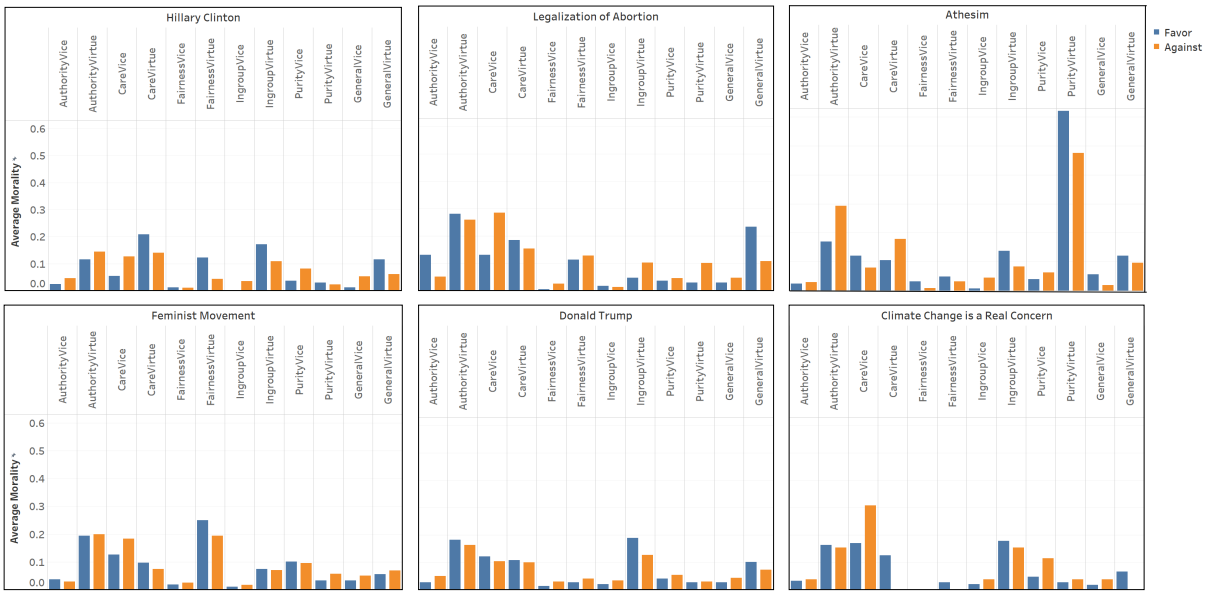


Figure 3.3: Average morality values of each social issue with respect to stance

Extracting Aspects Based on Morality

To further explore the connections between morality and aspects, we extracted aspects of the discussions for each morality type and social issue. Figure 3.4 visualizes the top five topics and their connected moral dimensions using a word network graph. In this network, nodes represent (12) morality types as well as top extracted aspects (listed in Table 3.4 and Table 3.5). The connection between a term and a morality type is represented by the edges in our network. The weight of the edge represents the average $tf - idf$ value of the aspects, while the colors of the edges represent social issues and stances.

Our results show that tweets written against the topic of atheism reference quotes and verses from the Bible and other holy books. For instance, for the word "lamb" in care (Tables 3.4 and 3.5), people bring quotes such as *I am washed and cleansed by the blood of the Lamb -Rev. 1:5; 7:14*[1], and for "acts" tweets include verses such as *Jesus commands you to follow Acts 2:38-39 to be saved*. On the other hand, tweets in favor of atheism discuss "country" and its rules and "establishment clauses"; i.e., *The establishment clause sets our country apart and prevents the radical religious zealots from taking charge*. Moreover, our findings

[1] *Italicized texts* in this section represent parts of the contexts.

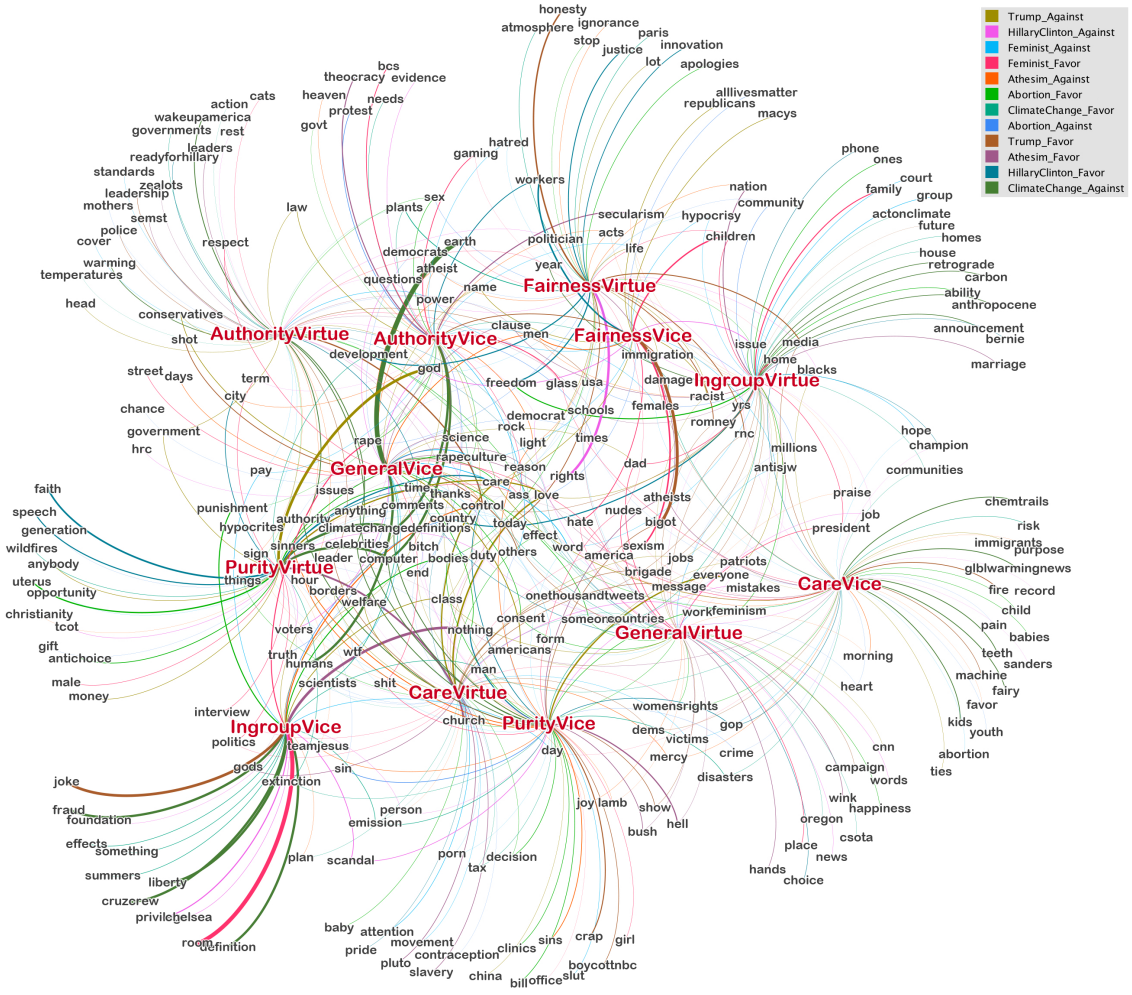


Figure 3.4: Network of top aspects and their connection to the 12 morality types (color=stance + aspect, thickness of the edge= weight of the word)

show that the topic of atheism is associated with words related to purity and sanctity. Based on the MFT, this dimension is inspired by the notion of living "in an elevated, less carnal, more noble way" [GHK+13].

Furthermore, context analysis of the tweets against the idea of climate change shows that they include aspects such as *tooth* "fairy," "fraud," and *flawed* "computer" *models*, indicating that climate change may have been perceived as a *hoax*. Those concerned about climate change include aspects such as "disaster," "government," "extinction," and "wildfire," showing concerns about *human and species extinction*. In addition, we found that people in this group discuss events such as "Paris climate change" and "climate summit of

the Americas (csota)" and show their support and interest regarding the events and topics discussed. Moreover, discussions of climate change feature both high loyalty and authority. Based on the MFT, loyalty is active and high "anytime people feel that it's one for all, and all for one" [GHK+13]. Also, aspects such as "countries," "homes," and "communities" are frequent in tweets related to loyalty. Authority, on the other hand, refers to "virtues of leadership and followership, including deference to legitimate authority and respect for traditions" [GHK+13]. The usage of aspects such as "government," "authority," and "wake-upamerica" may have resulted in the higher ratio of authority in this topic.

Additionally, tweets labeled as against the topic of Hillary Clinton discuss aspects such as "scandal," her daughter "Chelsea," her role in Clinton's "foundation," and some African-American ("black") voters who were *not supporting her in the election*. Tweets written in favor of her contain aspects such as "justice," "pride," and "faith." Also, in terms of stance, tweets labeled as in favor of the topic of Hillary Clinton used fewer negative and vulgar words related to harm and more positive words related to authority and care. On the other hand, tweets in favor of Donald Trump include more aspects related to loyalty, highlighting the likelihood of commitment to this figure. Moreover, the results show that aspects such as "racist," "China," and "border" are more prevalent in discussions against him while those in favor include aspects such as "Trump brigade," "RNC," and "honesty."

For the topics of feminism and abortion, we observe a high ratio of fairness. Also, the results show that tweets labeled as against feminism and abortion are concerned with lack of fairness compared to those in favor of them. Furthermore, tweets showing support for feminism discuss *breaking the* "glass" *ceiling* and the fact that they *refuse to accept that there is an unbreakable glass ceiling*. For the topic of abortion, we observe more aspects related to both care and harm. For instance, tweets against the topic of abortion talk about "pain," and "love," while those in favor of it use aspects such as "care," "babies," and "women's right."

	Atheism	Climate Change	Hillary Clinton	Feminist	Donald Trump	Abortion
CareVirtue	mercy, lamb, joy, sign, person		duty, rights, voters, work, berniesanders	attention, movement, victims, porn, USA	love, welfare, care, borders, jobs	love, truth, humans, end, contraception
CareVice	morning, America, mercy, favor, purpose	fairy, teeth, kids, glblwarmingnew, chemtrails	duty, champion, sanders, record, issue	victims, hope, word, females, abortion	blacks, immigrants, ties, care, patriots	times, pain, crime,youth, heart
FairnessVirtue	acts, nation, ignorance, atheist, light		rights, media, blacks, questions, yrs	god, USA, hypocrisy, reason, hatred	care, life, immigration, lot, racist	rape, sex, community, control, democrats
FairnessVice	men, acts, nation, ignorance		issue, god	power, hypocrisy, rapeculture, antisjw, sexism	racist, macys, name, thanks, republicans	millions, alllivesmatter, love
IngroupVirtue	nation, men, secularism, acts, praise	announcement, carbon, retrograde, year, anthropocene	politician, work, yrs, job, mistakes	group, court, hope, hypocrisy, word	patriots, love, jobs, life, everyone	country, family, community, USA, control
IngroupVice	control, plan	liberty, Cruzcrew, fraud, definition	privilege, scandal, Chelsea, foundation, person	pay, day	class, borders, welfare	country
AuthorityVirtue	power, hour, sinners, control, sign	celebrities, authority, anything, wakeupamerica, warming	duty, politician, questions, democrat, rest	power, standards, cover, pay, hypocrites	law, conservatives, government, head, god	mothers, control, government, men, USA
AuthorityVice	light, acts, city, atheist, heaven	climatechangedefinitions, computer	evidence, comments, issues, bitch, term	others, hatred, pay, females, life	law, class, borders, welfare, name	protest, times, end, rape, sex
PurityVirtue	sinners, hour, control, teamjesus, christianity		god, duty, voters, tcot	god, word, opportunity, today, wtf	god, love. government, anybody, money	sin, man, freedom, gift, generation
PurityVice	sinners, sins, hour, light, teamjesus	climatechangedefinitions, computer, celebrities, authority, anything	scandal, mistakes, person, bitch, comments	slut, day, porn, word, today	patriots, everyone, China, thanks, end	sin, everyone, others, crime, control
GeneralVirtue	joy, lamb, science, morning, praise		democrat, job, news, words, yrs	antisjw, form, USA, work, rapeculture	class, racist, today, CNN, Americans	heart, everyone, millions, others, control
GeneralVice	men, control, days, atheist, power	earth, humans	voters, questions, anything, hrc, god	word, god, wtf, form, rapeculture	Americans, today, conservatives, dad, name	truth, times, end, democrats, chance

Table 3.4: Top 5 terms for against stance

	Atheism	Climate Change	Hillary Clinton	Feminist	Donald Trump	Abortion
CareVirtue	slavery, love, gods, church, county	tax, countries, plan, scientists, emission	care, pride, love, country, control	work, politics, womensrights, interview, sexism	brigade, onethousandtweets, message, today, dems	care, effect, baby, decision, things
CareVice	atheists, love, hate, nothing	risk, damage, disasters, man, communities	GOP, home	glass, hate, sexism, day, ass	fire, machine, media, dems, immigration	care, effect, babies, child, womensrights
FairnessVirtue	atheists	plants, Paris, development, atmosphere	rights, justice, development, freedom, innovation	children, nudes, earth, gaming, home	honesty, today, RNC, Romney, media	apologies, freedom, bodies, stop, issue
FairnessVice			workers	children, sexism, hate	bigot	
IngroupVirtue	marriage, clause, nation, country, schools	countries, homes, communities, house, actonclimate	home, country, phone, champion, Bernie	home, family, dad, children, president	brigade, onethousandtweets, message, control, future	freedom, ones, Issue, ability, feminism
IngroupVice	nothing	effects, emission, summers, man, something		room, rape	joke	punishment, bodies, control
AuthorityVirtue	leaders, country, clause, zealots, schools	governments, temperatures, science, plants, action	time, control, readyforhillary, rock, love	respect, glass, year, cats, anyone	leader, leadership, control, police, USA	control, effect, care, sex, things
AuthorityVice	secularism, theocracy	needs, plants, damage	workers	glass, bcs, gaming, ass	immigration	rapeculture, punishment, govt, rape, sex
PurityVirtue	nothing, gods, reason, church, hate	development, extinction, wildfires	faith, care, speech, city, thanks	politics, ass, term, male, interview		church, uterus, rapeculture, antichoice, time
PurityVice	hell, bush, everyone, atheists	emission, disasters, countries, extinction, tax	nothing, control, GOP	shit, nudes, girl, everyone, office	crap, bigot, show, leader, boycottnbc	bill, decision, feminism, bodies, clinics
GeneralVirtue	hands, wink, hell, Bush, schools	reason, csota, oregon, countries, science	choice, time, rock, love, work	nudes, dad, president, place, hate	show, today, campaign, RNC, Romney	consent, happiness, someone, womensrights, church
GeneralVice	love, clause, reason, science	scientists		shit, ass, respect, America, street	shot, others	nothing, rapeculture, consent, someone, care

Table 3.5: Top 5 terms for in favor stance

Significance Testing

Chi-square procedures yield significant associations between stance and morality dimensions, types, and polarities in three social issues: abortion ($n = 711$), atheism ($n = 588$), and Hillary Clinton ($n = 728$). As a result of this observation, we excluded the other three social issues (climate change, Donald Trump, and feminism) from the results presented in Table 3.6. Our analysis shows that the highest number of associations are for the topic of abortion ($n = 711$), with significant relationships found between stance and the morality types of harm, subversion, purity, and general-virtue, as well as stance and the morality dimensions of purity and general morality (as shown in Table 3.6). A similar number of associations were also found on the topic of Hillary Clinton, with stance having significant relationships with morality types of harm, fairness, and betrayal, as well as the morality dimension of fairness and both morality polarities (all virtues and vices). Stances on the topic of atheism have significant relationships with the morality types of purity and general-vice. To further test for the strengths of association between these pairs, *Lambda* tests for association between nominal variables were conducted. We did not find any significance for the *Lambda* tests. With these results in mind, we believe that morality types and dimensions can be considered features that contribute to predicting stance, but they may not be the only variables with full explanatory power.

3.8 Discussion and Conclusion

In this chapter, we first investigated the usefulness of leveraging morality as an NLP feature for predicting two selected social effects (morality and stance) in Task 1. In addition, we showed how investments in the quality and general nature of lexical auxiliary tools and the rigorous evaluation of these investments improve the predictability of these social effects, thereby reducing biases in algorithmic solutions. This work matters, as personal values and social effects (which are often measured as the aggregation of personal values) are abstract and complex constructs, and their measurement requires researchers to find reliable and robust ways to operationalize them. The validity of such research hinges on the trustworthiness

		Abortion (N=711)	**Atheism** (N=588)	**Clinton** (N=728)
Morality Types	CareVirtue (Care)	-	-	-
	CareVice (Harm)	X2(3) = 10.99, p = 0.012**	-	X2(2) = 6.16, p = 0.046*
	Authority Virtue (Authority)	-	-	-
	AuthorityVice (Subversion)	X2(2) = 11.04, p = 0.004***	-	
	FairnessVirtue (Fairness)	-	-	X2(4) = 13.45, p = 0.009**
	FairnessVice (Cheating)	-	-	-
	IngroupVirtue (Loyalty)	-	-	-
	IngroupVice (Betrayal)	-	-	X2(1) = 5.93, p = 0.015**
	PurityVirtue (Purity)	X2(2) = 10.57, p = 0.005***	X2(3) = 8.06, p = 0.045*	-
	PurityVice (Degradation)	-	-	
	GeneralVirtue	X2(3) = 15.11, p = 0.002***	-	-
	GeneralVice	-	X2(1) = 5.08, p = 0.024*	-
Morality Dimension	Care (Care/Harm)	-	-	-
	Authority (Authority/Subversion)	-	-	-
	Fairness (Fairness/Cheating)	X2(2) = 6.84, p = 0.033*	-	-
	Purity (Purity/Degradation)	-	-	X2(4) = 13.49, p = 0.009**
	Ingroup (Loyalty/Betrayal)	-	-	-
	General (GeneralVirtue/GeneralVice)	X2(3) = 0.04, p = 0.045*	-	-
Morality Polarity	Virtue (All virtues)	-	-	X2(5) = 12.93, p = 0.024*
	Vice (All Vices)	-	-	X2(4) = 22.52, p = 0.000****

Table 3.6: Result of significance tests using chi-square (χ^2) ($p = 0.05$)

of our methods for capturing these effects in digital traces of human behavior. Hence, our work is based on the assumption that people's personal values, which might be impacted by their cultural contexts, are reflected in their language use [Bat00, MM85, Tri89], and that we can capture these values in user-generated text data.

Enhancing lexicons is expensive, as it requires trained human coders to assess each entry and its meta-data (in our case, category assignment and part of speech). This might help increase the reliability of social computing research, but does this effort make a difference for improving the accuracy of NLP tasks? In order to answer this question, we evaluated the usefulness of using no lexicon, a basic lexicon, and an enhanced lexicon for capturing morality in text data to measure two different social effects (morality and stance) based on public benchmark datasets. We found that using the lexicons we tested, namely the Moral Foundations Dictionary, does increase prediction accuracy in the majority of cases, especially when used for feature-based machine learning. Moreover, we found that the semi-automated

and human-validated verification and advancement of this lexical resource led to measurable improvements in capturing social effects in text data.

In Task 2, we performed a theory-driven and vocabulary-controlled detection of moral foundations from text data to expand the knowledge about stance analysis. Moral foundations can capture the influence of personal values and cultural differences on polarized or controversial discussions. Using a standard stance dataset, we applied MFDE to analyze people's discussions on six distinct social issues [RSD19, RD19]. The main objective of this task is to expand the scope of stance analysis by examining the narratives on either side of the topic in greater depth.

Our first research question in Task 2 asks what basic morality types are contained in tweets about social issues. To answer this question, we identified basic morality types contained in the sample of tweets we obtained on six different social issues: abortion, atheism, climate change, Hilary Clinton, Donald Trump, and feminism. Our results (Figure 3.2 and Figure 3.3) show that various social issues have distinctive distributions of morality types. While some project more authority-related words (Donald Trump), others consist of words related to care and purity (abortion and feminism). This finding suggests that social issues and individuals' stances on them are not morally equivalent. Additionally, our findings slightly confirm prior studies that found liberals' (e.g. Hillary Clinton) moral judgments more aligned with fairness and care [HWBS14, SM21]. This finding is further exemplified in Stewart and Morris [SM21]'s study that found that liberals' tended to exhibit "individualizing" moral foundations such as fairness and harm/care, while conservatives embraced group-based, "binding" foundations such as ingroup and authority. For the issue of climate change, some prominent morality types we observe such as authority, loyalty, and harm are also found in other studies [DMBA16, WAS16] as important moral concerns regarding environmental conservation. However, our finding that climate change discussions score low on fairness and cheating is in contrast with the existing evidence that fairness is a salient indicator of attitude towards climate change [DMBA16]. One possible reason for this difference may be that prior studies often included political orientation as either a moderating [WAS16] or mediating variable [DT12, VM16] of moral types and stances found in the discussion of climate change. However, due to the sparsity in our dataset, more detailed investigation is

needed to confirm or reject these findings.

Our second research question in Task 2 focuses on examining different characteristics associated with each of the five morality dimensions plus a general category, given there are two separate stances on each social issue. Referring to the Results section above, we find that each social issue and stance has a distinctive lexical profile with different aspects representing the prominent discussions surrounding it (Figure 3.3, and Tables 3.4 and 3.5). For instance, tweets in favor of atheism contain more discussions on purity and authority, while tweets in support of Donald Trump are more related to loyalty and authority. It is interesting to note that while tweets in support of atheism, Donald Trump, and Hillary Clinton have different lexical profiles, they allude to the importance of group-based moral values (e.g. purity and authority) that are essential in the discussions of religion [GH10] and politics [SM21]. In general, we see more instances of negative polarities (i.e. vice) in most morality dimensions–namely care, authority, and ingroup–a finding that was also supported by chi-square analyses. We assume that the discussions surrounding controversial issues on social media, especially on Twitter, are often polarized to either positive or negative sentiments [FY15, NBGRM15]. Specifically, for several controversial issues in our analysis such as those related to politics (i.e., Hillary Clinton and Donald Trump) or those related to climate change, abortion, and atheism, which all have longstanding debates, we expect to see distinctive characteristics in the discussions on opposing sides of each social issue.

Our third research question in Task 2 focuses on the correlations between morality and stances of each tweet according to the social issues. Furthermore, our chi-square tests of associations (Table 3.6) find variances in the number of statistically significant relationships on morality and stance across different social topics. For instance, we find six significant correlations on the topic of abortion, six correlations on Hillary Clinton, two correlations on atheism, and none for Donald Trump, climate change, and feminism. Among the significant associations we found, stance is most correlated with the *vice* morality type on various dimensions such as harm, subversion, general-vice, and betrayal. We performed a post-hoc analysis using a *Lambda* test to find direction of associations, but it did not yield significant relationships. We hypothesize that more words that fall under the *vice* spectrum would correlate with the increase in against stance, but the results do not reflect this correlation.

There are several reasons that we should perform further analyses to explicate all correlation pairs. Firstly, each social topic has a different sample of tweets, and some topics may be more controversial than others (i.e., the topic of abortion compared to Donald Trump). In the future, we hope to perform further aspect analyses to examine cross-cutting communication as well as the conceptual complexities within each social issue.

Our work has several limitations. For deep learning models, while using the enhanced morality lexicon yielded better overall accuracy, we still need to investigate more parameters and settings to find the most robust models. We plan to investigate these settings in the future. Moreover, the benchmark data we used were too small for this purpose. In addition, we only worked with tweets, which is just one out of many types of user-generated text data. The robustness of our evaluation might be further improved by working with texts from other genres and of higher formality, such as debates, congressional speeches, product reviews, and news articles. Moreover, our work follows the standard model of stances as a binary problem (in favor or against), but on certain issues there might be more than these two points of views. Additionally, negation detection was not considered in this study, an exclusion that might influence the amount of moral loading for contrasting polarities. Finally, we recognize the constraints and sparsity of our data sample, focusing only on US-related social issues and English-only language use. We recognize prior studies that have found that moral values and stances towards social issues significantly vary across cultures ([NE15]'s study with Swedish individuals found prominent concerns with fairness and harm; [KKY12]'s study with Korean individuals found emphasis on purity values). In the future, we hope to expand the study of morality and stance towards more issues from a more diverse set of social contexts and cultures.

Acknowledgments

This research is sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-17-2-0196. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the

U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. We appreciate Jessica Schmiederer, Tiffany Lu, Craig Evans and Max McKittrick for their great help in expanding and evaluating the morality lexicon. We also thank Ming Jiang, Chieh-Li (Julian) Chin, Shubhanshu Mishra, and Aseel Addawood for their assistance.

CHAPTER 4

ANALYZING PEOPLE'S SOCIAL INTERACTIONS USING SOCIAL AND PERSONAL VALUES

Contents of this chapter is based on the following papers (contributions are listed in §1.3):
Aref*, S., Dinh*, L., Rezapour*, R., & Diesner, J. (2020). Multilevel structural evaluation of signed directed social networks based on balance theory. *Scientific Reports*, 10(1), 1-12. (*authors have equal contributions)
Dinh*, L., Rezapour*, R., Jiang, L., & Diesner, J. (In preparation). Structural balance in signed digraphs: considering transitivity to measure balance in graphs constructed by using different link signing methods.(*authors have equal contributions)

4.1 Introduction

Social interactions can be depicted as sequences of formal or informal dynamic exchanges through which people influence (or are influenced by) other people [Mar07]. An interaction can be defined as "a process by which people act and react to those around them" [GSAK16]. The strength and pattern of these interactions are formed by people's social statuses, ties, and roles [Kle13, Gum64]. To better understand communities and groups, it is key to realize how people interact with each other and if these interactions are stable or impacted by any social or personal phenomena [DNMLPK12].

Real-world social and communication networks are composed of complex and continually evolving interactions among social agents. Analyzing this data allows for exploration of the structure and dynamics of relationships among social entities, incorporation of observations based on social science theories, and empirical testing of existing theories, among other uses. Researchers have leveraged social networks to analyze communication and interaction patterns in complex systems [New03, AB02, SGGN+17]. Moreover, social networks are

capable of indicating how people are connected with each other and what types of ties are connecting them in their networks [New03, AB02]. One existing theory that addresses the study of social interactions is that of structural balance [Hei46, CH56]. This theory has been widely used to explain local-level social dynamics that emerge within and among triads (three connected nodes forming a triangle), potentially causing ripples throughout networks and leading to network-wide effects.

With its root in social and cognitive psychology [Hei46], balance theory explains how different configurations of positive and negative relationships between pairs of nodes may impact the amount of tension in a closed triad (three nodes with an edge between every pair of nodes). This tension would be absent if the triad has an even number (0 or 2) of negative edges [CH56]. Applying this premise to real-world situations, the following four adages scope out the balanced configurations at the triadic level based on edge signs: *my friend's friend is my friend (+++); my friend's enemy is my enemy (+- -); my enemy's friend is my enemy (-+-); and my enemy's enemy is my friend (- -+).* The measurement of balance has more recently been expanded from the triad level (micro-level) to the subgroup level (meso-level) by partitioning nodes into two groups or "plus-sets" [DM96, DL67], such that the number of positive edges between groups and negative edges within groups is decreased to some extent. The measurement of balance could also be expanded to the network level (macro-level) using the *line index of balance* [Har59, Fla70, Zas87, AW18, AW19], which equals the minimum number of positive edges between groups and negative edges within groups across all possible ways to partition the nodes into two groups. Looking at the prior body of research on balance, we conclude that for a more comprehensive analysis of interactions across networks, balance should be assessed at multiple levels of the network, namely the micro-, meso-, and macro-level. In [ADRD20], we proposed a new methodological framework to "link micro and macro levels" [Gra77] for analyzing signed social networks.

Structural balance has primarily been studied for undirected signed networks [WF94, LHK10a, CHN$^+$14, AW18] and not directed signed networks [Rap83, LS69, She71]. Using undirected network models for balance assessment could be justified when the modeled relationships and interactions are truly undirected, such as collaborations [AN20], or inherently reciprocal, such as bi-lateral alliances [AW19]. However, many real-world relationships are

intrinsically directed, such as social preferences [LS52, New61], and not necessarily reciprocated, such as a friendship [ARPS16]. Therefore, disregarding directionality [FIA11] when it does apply can jeopardize the validity of network measures [SM13], including balance assessment [LHK10a, FIA11].

In this chapter, we propose to study people's real-world interactions in signed and directed networks by employing a multilevel structural balance analysis [ADRD20]. Moreover, we measure balance at three fundamental levels of analysis, namely at the (1) triad (micro), (2) subgroup (meso), and (3) network (macro) levels. We aim to answer the following research questions:

- **RQ1**: How can structural balance analysis at the micro-, meso-, and macro-levels help to measure people's interactions in signed and directed networks?

- **RQ2**: What insights can we gain from a multilevel evaluation of balance?

For this purpose, we extended the previous analysis of networks by leveraging people's generated texts to extract network signs and communication types. In recent years, social media platforms, online forums, and communication channels such as emails added new mediums for people to interact with each other via chatting, messaging, posting images or links, and sharing content [VDP13]. These generated data are a rich resource of information that can be used to extract and analyze personal-level as well organizational- and societal-level information in networks [DNMLPK12]. Moreover, language, in written or verbal form, can be used as a tool to influence others, reinforce behaviors or beliefs, and interact with other people [Die19]. Variation in the uses of language and signals in people's everyday communications can convey the purposes and forms of the interactions [Die19, IP10]. In addition, what is stated in the text is closely related to the speaker's attitude and culture, as well as the larger social context in which the discussion is occurring. Therefore, user-generated texts can provide insights about a community and the stability of the interactions within it.

Following these intuitions, in addition to the previous research questions, we aim to explore the following research question:

- **RQ3**: What insights can we gain about people's interactions in a network by extracting different properties and linguistic cues from user-generated texts?

To answer the proposed questions, in this work, we leveraged methods and concepts from natural language processing to extract two types of edge signs–moral values (virtue or vice) and sentiment (positive or negative)–from the exchanged (user-generated) text data between people. We used sentiment and moral values, as they can represent a person's emotional and moral status in their network [Die19, IP10, GHK+13]. For this purpose, we used two large user-generated email corpora, namely the Enron email dataset[1] and the Avocado Research email collection [OWKG15]. Emails, as one form of user-generated datasets, have gained popularity in social network society since they provide access to real-world communication data in an electronic form and are promising resources for research on human interactions [WHAT04, DFC05].

Our analysis of balance in the Enron and Avocado datasets showed that balance ratios vary across different measurements of social relations. With the average micro-level balance ratio being 81.7% for morality and 69.5% for sentiment, one can conclude that people's moral states are more balanced than their emotional states in communities, especially through tension (i.e., Enron's bankruptcy). This result confirms that texts indeed entail information about the social contexts and people's values, which can help in analysis of interactions. Moreover, balance ratios were high (70% and above) across the networks, a commonality that offers an empirical validation of balance theory. Furthermore, our multi-level balance analysis displayed that, while each layer manifests different characteristics of stability in networks, tension can be observed across different levels.

4.2 Literature Review

Social network analysis has gained attention in recent years, since it provides unique opportunities for studying people and their real-world interactions [FF13]. This analysis has been used extensively to study the relation between entities, types of ties, and inferring positive

[1] `https://www.cs.cmu.edu/~./enron/`

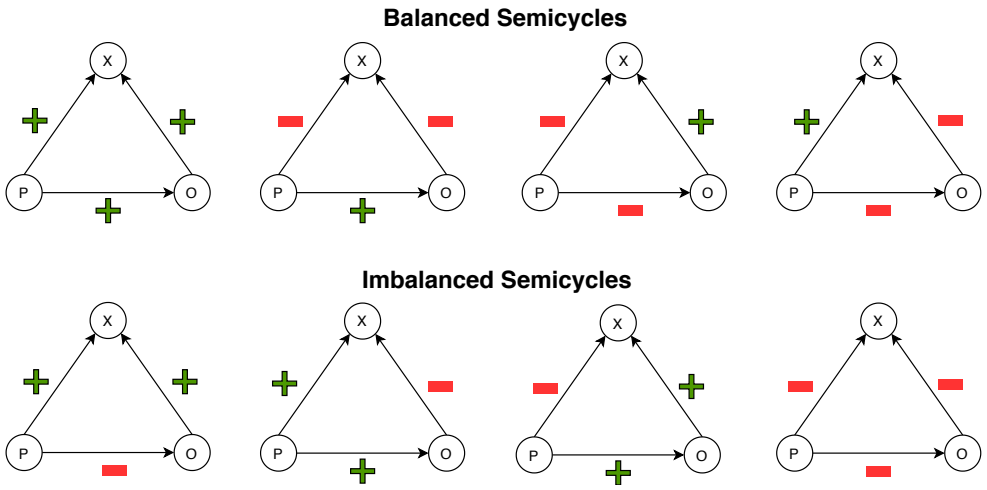


Figure 4.1: Balanced and imbalanced semicycles

or negative relations among people [Kle13, LHK10b]. One of the foundational theories in social network analysis is structural balance theory, proposed by Heider [Hei46], in which the relationship between individuals, groups, or objects are explained in a network setting [WF94]. Heider's original structural balance theory [Hei46] has its roots in cognitive psychology, which links (im)balanced structures in a network to cognitive dissonance [Fes62] and formulates the formation of the relationships between individuals in a triad. Based on Heider, the relationship between three entities in a triad is 'balanced' if P, O, and X all have positive feelings toward one another and is 'imbalanced' if one of the two entities (forming a dyad) doesn't like the other one. Heider proposed that balance co-exists with symmetry of relations [Hei58], such that P liking O also implies O liking P. Heider [Hei46] argued that the imbalanced state results in tension, and individuals in a network strive to move towards balance over time if they feel disturbance.

In addition to the condition of symmetry, transitivity also plays a vital role in explaining the formation of ties within triads. Heider pointed to transitivity as a prerequisite for balance [Hei46] and posited that "three positive relations may be considered psychologically transitive", in that "P tends to like X if $P\mathscr{R}O$ and $O\mathscr{R}X$ are valid at the same time" ($\mathscr{R}$ represents positive relation between two nodes). Thus, showing transitivity as a necessary condition for stability [HM75] and balance [HL78, KH06] in a social network. In other studies, Davis, Holland, and Leinhardt [Dav79], and Stix [Sti74] found transitivity as an

important property in balance.

Cartwright and Harary (1956) generalized Heider's balance theory and proposed the context of signed and directed networks, in which relationships between pairs of nodes were presented as either a (+) or (−) [CH56]. They further extended the concept of balance and called a triad structurally balanced if the product of the signs of its edges is positive or, in other words, if every cycle has an even number of negative ties (Figure 4.1). They later pointed to the necessity of studying balance in a network beyond triads. Davis [DL67], and Dorian and Mrvar [DM96] studied balance on the meso-level by analyzing the conditions for separating the graph into two or more subsets of nodes, where each positive edge would link two nodes of the same subset and a negative edge would link nodes from different subsets. Davis [DL67] introduced the concept of weak balance theory, in which triads with all negative signs are permitted and clusterable. Following this line, researchers used the notion of the line index of balance [Har59, Fla70, Zas87, AW18, AW19, AMW18] to study the global balance in networks. In their research, Facceti and colleagues [FIA11] analyzed balance in large undirected networks on global (macro) level and found that real-world large networks are extremely balanced.

Structural balance has been used for studying a variety of research problems in real-world networks, such as team formation and clusterablity of individuals in networks [DFFL13, KSZ$^+$20], polarization [XCJ15, LCL16], community detection [FYWL16, DAE$^+$16, MG20, Est19], and stability of communication [DE15, Fea67], to name a few. Leskovec and colleagues [LHK10c] analyzed how "interplay between positive and negative relationships affects the structure of on-line social networks" in large signed directed networks using both balance theory and status theory and found that status theory provide more accurate predictions in directed networks compared to balance theory (which found to be more accurate in undirected networks). Teixeira and colleagues [TSF17] studied the relations between individuals over time and found that balance in triads tend to increase and tension tend to minimize between the individuals (depending on the initial distribution of the signs). Tang and colleagues [TLK12] developed a framework to predict various types of social relationships across heterogeneous networks using social theories such as structural balance and status theory and showed that leveraging network information can result in high accu-

racy for inferring particular relationships, e.g., manager-subordinate. Rawlings and Friedkin [RF17] tested balance theory via a sentiment conversion process in the Urban Communes Data Set (UCDS). Moreover, the authors analyzed the relational tensions and changes in sentiment structures in the networks and found evidence for temporal reduction of balance violations. Askarisichani and colleagues [ALB+19] studied traders' affective relations in a financial institution using triads and found "strong propensity for stability in the 'classic' balanced states."

Following the previous work, we found structural balance a suitable model for analyzing social interactions and knowing whether if relationships are (or will be) stable over time or are (or will remain) polarized.

As mentioned earlier, structural balance has been primarily studied in undirected signed networks [WF94, LHK10a, CHN+14, AW18] as opposed to directed signed networks [Rap83, LS69, She71]. One important point here is that real-world relations and interactions are not always reciprocated [WF94], e.g., P may regard person O as a friend, but O may not see P as a friend. Therefore, considering the directionality of interactions is crucial to better investigate such networks. Prior research considered direction in balance analysis by either symmetrizing all the edges, following Heider's proposition [Hei58], or removing the unreciprocated edges [DE15]. Some scholars, on the other hand, refined the assessment of balance to the level of semicycles (containing directed ties) embedded in each triad [Fea64, Rob74, dN99] and analyzed the balance of triads with respect to the semicycles. Following this intuition, semicycles that are cyclic ($P \rightarrow O$, $O \rightarrow X$, $X \rightarrow P$) were found not suitable for balance analysis because (1) cycles contain limited information on the process of influence among relationships [RRP10, VDSVZ13], and (2) they are intransitive [Rob74, Blo15].

Furthermore, Holland and Leinhardt [HL78] developed triad census in which all possible combinations of directed ties between three nodes were presented (sixteen classes of MAN (Mutual, Asymmetric, Null) triads). Following the intuitions discussed in the triad census, four triad types (as show in figure 4.2) are driven by both transitivity and balance, which can be used to extend the theory of structural balance and integrate balance analysis and directed edges in triads. We show that these four configurations are relevant for our operationalization of balanced triads, with transitivity as a precondition of (micro-level) balance.

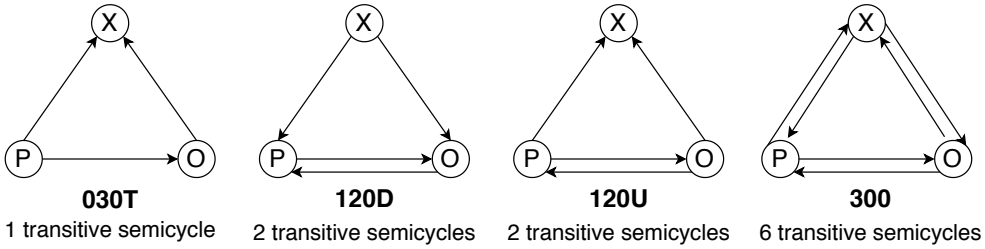


Figure 4.2: Triads in the triad census [HL70] with transitive semicycles. Signs of edges (not shown in the figure) can either be positive or negative. Triad types are labeled based on the number of mutual (first digit), asymmetric (second digit), and null (third digit) dyads, and an additional letter for direction (T:transitive, D:down, U:up). See [HL70] for more details about nomenclature for the triad census.

Moreover, while a number of methodologies has been developed to incorporate directionality into the calculation of balance, studies have mostly dismissed the direction of ties and thus analyzed balance for undirected networks as opposed to directed ones. In this chapter, we present a solution for calculating structural balance for signed and directed networks to better study social interactions and their stability and changes in social networks.

4.3 Notations and Basic Definitions

We denote a directed signed digraph as $G = (V, E, \sigma)$, where V and E are sets of vertices and directed edges, respectively, and σ is the sign function that maps edges to $\{-, +\}$, i.e.; $\{-1, +1\}$. A signed digraph G contains $|V| = n$ nodes and $|E| = m$ directed edges. The set E of directed edges contains m^- negative edges and m^+ positive edges.

A *triad* (T) in G is a set of three nodes with at least one directed edge between each two of them (could be in either direction) as shown in Figure 4.2. Given a triad, if there are 3 edges incident on its nodes such that for every pair of nodes there is one edge, then those three edges form a *semicycle* (S). A triad has at least one semicycle, but it can include more semicycles. In Figure 4.2, the leftmost triad has one semicycle while the rightmost triad has eight semicycles. If the binary relation $\mathscr{R}$ that defines edges $A\mathscr{R}B \leftrightarrow (A, B) \in E$ is transitive over the set of a semicycle's edges (i.e. $A\mathscr{R}B \;\&\; B\mathscr{R}C \rightarrow A\mathscr{R}C$), the semicycle is

called a *transitive semicycle*. A transitive semicycle is balanced (unbalanced) if and only if the product of the signs on its edges is positive (negative). A signed digraph is (completely) balanced if and only if its set of nodes can be partitioned into two groups, such that all positive edges are within each group and all negative edges are between the groups.

4.4 Analyzing Interactions using Multi-level Structural Balance

In this section, we discuss our proposed multilevel evaluation framework, which involves measuring balance at the micro-, meso-, and macro-level. The overall workflow of this evaluation is shown in Figure 1.3.

4.4.1 Micro-level Structural Balance

To evaluate balance in a signed network, the most common method is to quantify balance per triad [CH56, Rap83, Joh86, TW11]. This step is usually followed by adding up and comparing frequencies or ratios of balanced versus unbalanced triads, with the implicit assumption being that this aggregation represents a network's overall balance. The majority of studies do not consider edge directionality when calculating triadic balance. In real-world social networks with positive and negative relationships, ties are not necessarily reciprocated. For instance, A might perceive B as a friend, but B is neutral towards A, which can be formulated as $(A, B) \in E^+, (B, A) \notin E$ using a signed digraph notation. Another example would be A trusting B but B distrusting A, which can be formulated as $(A, B) \in E^+, (B, A) \in E^-$. Undirected signed networks are incapable of modeling such basic cases, leading to the exclusion of these situations from network models [FIA11] or the disregard of all unreciprocated edges for analysis [DE15, LHK10a]. This fundamental flaw is resolved by using signed digraphs, which results in a more flexible and comprehensive network model. Addressing this problem requires the consideration of edge directionality for measuring balance. Our unit of analysis for the micro-level evaluation of balance is a transitive semicycle. We only evaluate triads in which all semicycles are transitive (which we refer to as transitive triads).

As mentioned earlier, a *semicycle* S in signed directed T is a set of three directed edges that starts from a vertex V, follows the direction of edges, and does not return to the same vertex. In other words, S is *transitive* and *non-cyclic*. Based on the triad census [HL70], four types of triads are transitive; '030T', '120D', '120U', and '300' (illustrated in Fig. 4.1). For triad '300', we only consider its six transitive semicycles, and disregard its two cyclic semicycles. For triads '030T', '120D', and '120U', we consider all their semicycles since they are all transitive. Therefore, we have six permutations of P, O, X for '300', two permutations for '120D' and '120U' and one permutation for 030T.

We define T a *completely balanced* triad if and only if every transitive semicycle in T is balanced (positive). A transitive semicycle S is balanced (or positive) if it contains an even number of negative directed edges. Furthermore, we define T a *partially balanced* triad if it contains at least one negative semicycle. Finally, T is *completely imbalanced* if every transitive semicycle in T is imbalanced (negative).

For simplicity of notation, we define $T^{(i)}$ as the set of all transitive triads of type i, where $i = 1, 2, 3, 4$ refers to 030T, 120D, 120U, and 300, respectively. Algorithm 1 shows our step-by-step computation of balance in a signed directed network. After calculating balance $B^j_{T^{(i)}}$ for each triad j of type i, we compute the weighted balance ratio for the set of all transitive triads of type i ($B_{T^{(i)}}$). Finally, the *overall balance ratio of* G ($T(G)$) is calculated by averaging the balance ratio of all types i across a network.

Furthermore, evaluating balance solely at the micro-level is a common practice, but rests on the assumption that aggregating triad-level balance is sufficient to determine network-level balance. Also, measuring balance at the triad level does not consider how configurations within triads influence neighboring nodes and edges, as well as broader areas of the network. Based on prior literature, there are structural configurations beyond the triad, such as longer cycles, that contribute to balance of a network or lack thereof [Har59, DM96, Bon12, EB14, AW18, AW19]. These findings show that aggregating balance scores from the micro-level might not capture other structural features such as density. To mitigate the limitations resulting from a single-level evaluation, we leverage and apply complementary methods proposed by [AN20, ADRD20] to evaluate meso- and macro-level balance as parts of a comprehensive multilevel evaluation framework.

Algorithm 1 Computing triadic balance for a signed directed network

for $i = 1, 2, 3, 4$ **do**

  Consider set $T^{(i)}$

  $\triangleright$ Take element j of $T^{(i)}$, for $j = 1, \cdots, N_j$:

    Find the Semicycles and calculate: $B_k^{sign} := \prod_r$ sign of edge r

    Consider $S_{T^{(i)}}^{+,j} = \{$semicycle k, B_k^{sign} is $+\}$

    Consider $S_{T^{(i)}}^{-,j} = \{$semicycle k, B_k^{sign} is $-\}$

    Let $S_{T^{(i)}}^{j} = S_{T^{(i)}}^{+,j} \cup S_{T^{(i)}}^{-,j}$

  $\triangleright$ Define:

$$B_{T^{(i)}}^{j} := \frac{|S_{T^{(i)}}^{+,j}|}{|S_{T^{(i)}}^{-,j}|}, \quad \Big(\text{Note: } B_{T^{(i)}}^{j} \in [0,1]\Big)$$

  Let $\tilde{N}_{T^{(i)}} := \{T_j^{(i)} : B_{T^{(i)}}^{j} \neq 0\}$, $\tilde{Z}_{T^{(i)}} := \{T_j^{(i)} : B_{T^{(i)}}^{j} = 0\}$, where, $T^{(i)} = \tilde{N}_{T^{(i)}} \cup \tilde{Z}_{T^{(i)}}$

  Define:

$$B_{T^{(i)}} := \frac{|\tilde{N}_{T^{(i)}}|}{|T^{(i)}|}$$

end for

$$T(G) = \frac{1}{4} \sum_{i=1}^{4} B_{T^{(i)}}$$

[1]

4.4.2 Meso-level Structural Balance

Following the methodology proposed by [AN20, ADRD20], meso-level balance can be evaluated by partitioning the vertices of a network into two mutually antagonists but internally solidary subgroups [Zas87, Zas12, AMW18, AMW20]. Solidarity within subgroups is referred to as cohesiveness, and antagonism between subgroups as divisiveness within a network. An internally solidary subgroup means that there are only positive edges within a subgroup. Two internally solidary subgroups are mutually antagonistic when they are connected by only negative edges. This approach returns the minimum number of negative edges within subgroups and positive edges between subgroups across all possible ways of partitioning nodes into two subsets. The following example illustrates this approach: a balanced network that contains both positive and negative edges has an extreme amount of cohesiveness (because all edges within its two subgroups are positive) and an extreme amount of divisiveness (because all edges between its two subgroups are negative). We quantify cohesiveness and divisiveness through the deviation from this extreme case.

Using a signed digraph $G = (V, E, \sigma)$ as input, the set of vertices, V, can be partitioned based on $P = \{X, V \setminus X\}$ into the two subgroups X and $V \setminus X$. Given partition $P = \{X, V \setminus X\}$, edges that cross the subgroups are *external* edges that belong to $E_P^e = \{(i,j) \in E | i \in X, j \notin X \text{ or } i \notin X, j \in X\}$. Edges that do not cross the subgroups are *internal* edges that belong to $E_P^i = \{(i,j) \in E | i,j \in X \text{ or } i,j \notin X\}$. We measure cohesiveness (divisiveness) of a partition P by only looking at the signs of its internal (external) edges. We quantify the cohesiveness of a given partition P using the fraction of its positive internal edges to all internal edges $C(P) = |E_P^i \cap E^+|/|E_P^i|$. Similarly, we quantify the divisiveness of partition P as the fraction of its negative external edges to all external edges $D(P) = |E_P^e \cap E^-|/|E_P^e|$. We compute cohesiveness and divisiveness using P^*, which is the best fitting bi-partition of nodes, as explained further in the next subsection. This bi-partition is also connected to our proposed macro-level analysis. The proposed measures of cohesiveness and divisiveness are consistent with prior social networks literature, especially with the concepts of ranked clusterability [WF94], partitioning nodes via blockmodeling [DM96], in-group attraction and out-group repulsion mechanisms [STV20], and intra- and inter-group conflicts in small groups [CL13, CI08], as well as sociological literature on faultline theory [LM98].

4.4.3 Macro-level Structural Balance

The line index of balance, denoted as $L(G)$ and also referred to as the *frustration index* [Zas87, AW19] and *global balance* [FIA11], is defined as the minimum number of edges whose removal leads to balance. These edges can be thought of as sources of tension in this approach. While the historical roots of the frustration index go back to the 1950s [AR58, Har59], this approach only started to receive major attention in recent years [IRSA10, FIA11, AW18, AW19, AN20]. This might be due to the computational complexity of obtaining this index exactly, which is an NP-hard problem [HBN10]. While even approximating this measure has been difficult [IRSA10, FIA11], recent developments have enabled the exact and efficient computation of $L(G)$ for graphs with up to 10^5 edges [AMW18, AMW20]. This model for directed signed graphs is developed by [ADRD20] by leveraging the exact method and building on recently proposed optimization models [AMW20, AN20].

Frustration of an edge depends on how the edge resides with respect to the partition $P = \{X, V \setminus X\}$ that is applied to V. Positive edges with endpoints in different subsets and negative edges with endpoints in same subset are frustrated edges under P. The frustration index offers a top-down evaluation mechanism for assessing partial balance by providing an optimal partition P^*. The optimal partition P^* minimizes the number of frustrated edges and is therefore the best fitting partition of nodes into two mutually antagonistic and internally solidary groups. A simple normalization of $L(G)$ using a line index upper bound (which equals a half of the edge count, $m/2$ [AMW18]) leads to the normalized line index $F(G) = 1 - 2L(G)/m$ [AW18]. The normalized line index provides values within the unit interval such that large values represent higher partial balance and therefore higher consistency of a network with balance theory at the macro-level. We refer the readers to the Supplementary Information in [ADRD20] and [AMW20, AN20] for more details on this measure and the optimization model used for this computation.

We solve the optimization model that produces P^* using *Gurobi* solver [Gur20] (version 9.0) in *Python*. For large networks, we follow the two-step method presented in [AN20, ADRD20], which involves computing a lower bound for the frustration index before solving the optimization model.

4.5 Data

To collect data on communication networks, researchers have used different methods [BJK$^+$90], such as observations [New61], surveys [Sam68], and text analyses [Die15, CMB06]. To study people's interactions, we used two large datasets: the Enron and Avocado email datasets. The Enron email data comprise a large-scale, temporal dataset from a global, U.S.-based, former energy brokerage that went bankrupt in 2001. The communication (email) dataset of 158 employees was released in 2002 by the FERC [DFC05, Inv]. The original dataset went through various edits and modifications over the years. In this study, we used the latest release of the dataset from 2015[2]. The Enron dataset is of special importance in the social network community since it provides real-world organizational communication data over a

[2] `https://www.cs.cmu.edu/~./enron/`

span of 3.5 years.

The Avocado Research Email Collection [OWKG15] is provided by the Linguistic Data Consortium[3] and consists of emails among 279 accounts in a defunct information technology company referred to as "AvocadoIT", a pseudonym assigned for anonymity. The dataset consists of calendars, attachments, contacts, reports, and emails. In this study, we only used the email communications.

For both datasets, we preprocessed the emails and removed numbers, punctuation, times, and dates. In addition, we removed the email threads tagged as "Original Email", and only used the latest communication between the sender and receiver(s). We kept all forwarded emails tagged as "Forwarded Emails" for our analysis, since we believe that the senders found this information relevant. Furthermore, We removed all emails sent by list-serves as well as spam-like email addresses, e.g., "outlook.team@enron.com" and "all@avocadoit.com". We identified these email addresses by analyzing a random sample from both the Enron and Avocado data sets.

4.6 Network Construction and Edge Labeling

Real-world communication, verbal or nonverbal, written or visual, entails various types of explicit and implicit relationships, such as like versus dislike and trust versus distrust. To validate our proposed methods for calculating micro-, meso-, and macro-level balance in directed signed graphs, we constructed the interaction networks in two different social contexts: two business organizations (the Enron email dataset and the Avocado Research Email collection). We then leveraged natural language processing methods to extract two types of edge signs from the text data: moral values (virtue or vice) and sentiment (positive or negative). The next sections provide more details about the edge labeling as well as the constructed networks.

[3] `https://catalog.ldc.upenn.edu/LDC2015T03`

4.6.1 Edge Labeling Based on Morality and Sentiment

To label the interactions between people (email exchanges) with their valence, we leveraged two linguistic properties as the moral values and sentiment. This approach is based on the premise that people's language use can reflect their emotional, cultural, economic, and ideological states and backgrounds [Tri89]. Differences in people's feelings, opinions, and moral or personal values may be the sources of tension and conflict in relationships and groups. Therefore, extracting and analyzing these relationships from the language exchanged between network participants can help in better understanding the structure and balance in social networks, as well as the stability of interactions.

To capture moral values in our email data sets, we leveraged the Moral Foundations Theory (MFT) [GHK+13, GHN09]. As explained in Chapter 3, MFT can help capture people's spontaneous reactions and categorizes human behavior into five basic principles (fairness/cheating, care/harm, authority/subversion, loyalty/betrayal, and purity/degradation) that are characterized by opposing values (virtues and vices). The Moral Foundations Dictionary (MFD) enables the measurement of MFT based on text data by associating 324 words with virtues and vices from the MFT [GHN09, GHK+13]. To extract moral values from our email data, we used MFDE, our enhanced version of MFD[4] as developed, introduced, and validated in Chapter 3, §3.3 [RSD19, RD19]. As noted previously, compared to the original MFD, our enhanced lexicon consists of about 4,636 terms that were syntactically disambiguated and manually pruned and verified. In order to analyze balance and label edges with signs, we only considered the polarity of moral words (virtue or vice) and did not take the moral dimensions into consideration.

For the second language property, we leveraged sentiment analysis, a technique commonly used for understanding people's emotions, opinions, and affective states, to label the links (emails) with signs [PL+08]. The basic task with sentiment analysis is to identify the polarity of communication or discourse and to label pieces of text data as positive, negative, or neutral. To identify the sentiment of each email, we leveraged the Subjectivity Lexicon, a widely adopted and previously evaluated sentiment lexicon developed by Wiebe and Riloff

[4] `https://doi.org/10.13012/B2IDB-3805242_V1.1`

[WR05]. This lexicon contains a total of 8,222 syntactically disambiguated words that are tagged with negative, positive, or neutral polarity.

Furthermore, we domain-adopted both morality and sentiment lexicons to align them with the language of our email datasets. As an example, *"power"*, one of the words in our morality lexicon, is regularly used in Enron emails since this company was an energy broker. With no domain adaptation, this word would skew the results. To remove such words from the lexicons, we first selected the top salient words of each email dataset (separately) using their $tf - idf$ scores (Eq. 2.1). Next, we trained two human annotators to (1) remove overly common words (false positives) from the lexicons and (2) add relevant but missing words (false negatives) to the lexicons. Using the list of top words, the annotators checked if the extracted words already existed in the lexicons and whether their prior polarity and part of speech (POS) were appropriate given the context of the email datasets. If a word did not exist in the lexicons, and both annotators found it appropriate for the purpose of this study, the word was added to the respective lexicon. If the word was not found suitable, or it already existed in the lexicons, we removed the entry. Finally, if a word did exist in a lexicon, but both annotators agreed on changing the polarity or POS of the word, we modified the entry in the lexicons.

After preprocessing the emails from both datsets and domain-adapting the lexicons, we used *spaCy* [HM17], a *Python* library, to split the emails into sentences, tokenize the sentences into words, and tag each word with its respective POS. In order to assign an edge (email communication between node P to O) with virtue (+) or vice (-) morality, we counted the number of words for either morality polarity value (+, -) if a word and its POS matched an entry in MFDE, and then we tagged the sentence with the moral polarity that had the highest count. We used the same approach to label each edge with its sentiment valence. For any word that matched an entry in our domain-adapted Subjectivity Lexicon in surface form and POS, we logged a match, counted all matches per sentence and sentiment class (positive, negative, or neutral), and tagged each sentence with the majority class. We also checked each sentence for negation using the $NLTK$ package [LB02]. If a negation was found in a sentence, we flipped the morality or sentiment polarity to its opposite value; e.g., for morality, from virtue to vice. Finally, we aggregated the moral or sentiment polarity of

all sentences per email and normalized the score by the number of sentences per email.

After tagging morality and sentiment in both email datasets, we constructed four interactions data, aka directed edgelists, (Avocado Morality, Enron Morality, Avocado Sentiment, and Enron Sentiment), in which email addresses are nodes (senders are source nodes, and receivers are target nodes), emails sent from a node to another node are directed edges, morality or sentiment scores (normalized counts of each email) are the weights of each edge, and morality or sentiment polarity (+, -) are the signs of the edges. If an email did not contain any word that matched a lexicon entry, that email was not considered in the respective edgelists. Therefore, an edge could be present in the sentiment edgelists but not in the morality edgelists. Furthermore, we normalized the morality or sentiment scores (signs) of the edges between every two nodes if they had interacted more than one time and were connected with more than one edge (email exchange).

4.6.2 Edgelist Preparation

One challenge with the Enron dataset is that individuals may have more than one email address [DFC05]. For instance, an employee with initials *"K L"* was using the following email addresses:
kl@enron; k@enron.com;*k.l@enron.com*; *k_l@enron.com*; *ke.l@enron.com*; *ke_l@enron.com*; *k.l.l@enron.com*; *knn.l@enron.com*; *knn_l@enron.com*; *knn_l@enron.net*; *k@enron.com*; *kll@enron.com*; *l.k@enron.com*;*lk@enron.com*; *sssk.l@enron.com*.

After extracting the edge signs, we first converted the email addresses into actual names of the people in the Enron dataset [DFC05, DE15]. In order to disambiguate the email addresses, we leveraged the work by [DFC05], which includes the disambiguated names and email addresses of 558 employees of Enron. For the Avocado dataset, to maintain consistency with the Enron dataset, we only considered emails that were sent to or from corporate email addresses (emails ending in *@avocadoit.com*).

The number of nodes and edges of both Avocado and Enron datasets are shown in Table 4.1. The difference in the number of nodes and edges of the two lists is due to the availability of sentiment and morality words in the emails. In addition, Figure 4.3 visualizes the final

networks of Enron and Avocado with morality and sentiment as signs.

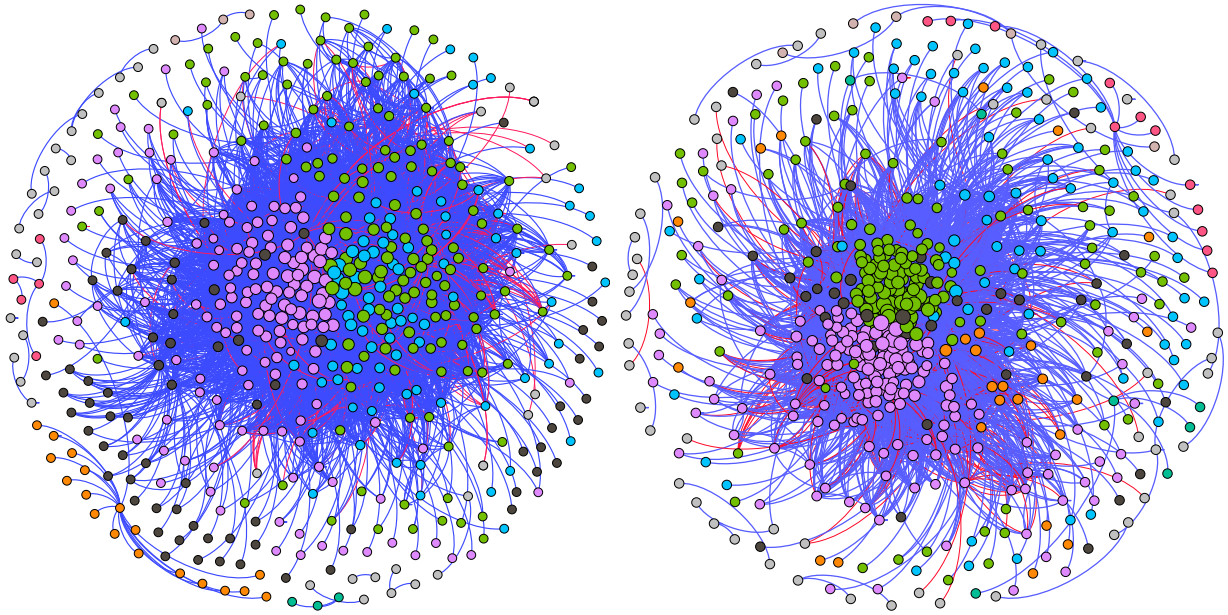
4.6.3 Balance Calculation

To analyze the micro-level balance, after cleaning the edge lists and disambiguating names and email addresses, we used *NetworkX*, a *Python* library, to remove self-loops, isolates, and pendants, as well as the edges with neutral (0) scores, as they have no impact on calculating balance. Table 4.1 shows the number of nodes and edges after preprocessing. Furthermore, we extracted instances of four transitive triads (030T, 120D, 120U, and 300) and analyzed balance within each triad with respect to their semicycles. Tables 4.2, 4.3, 4.4, and 4.5 show the final counts and ratios of completely balanced, partially balanced, and completely imbalanced transitive triads in each dataset.

To analyze balance in the meso- and macro-levels, we followed the steps explained in §4.4. Before analyzing the networks, we chose the largest component in each network, but we did not remove the pendants. Table 4.6 shows the divisiveness ($D(P^*)$) and cohesiveness ($C(P^*)$) scores for the meso-level and the number of frustrated edges ($L(G)$ and the normalized line index ($F(G)$) for the macro-level balance of all networks.

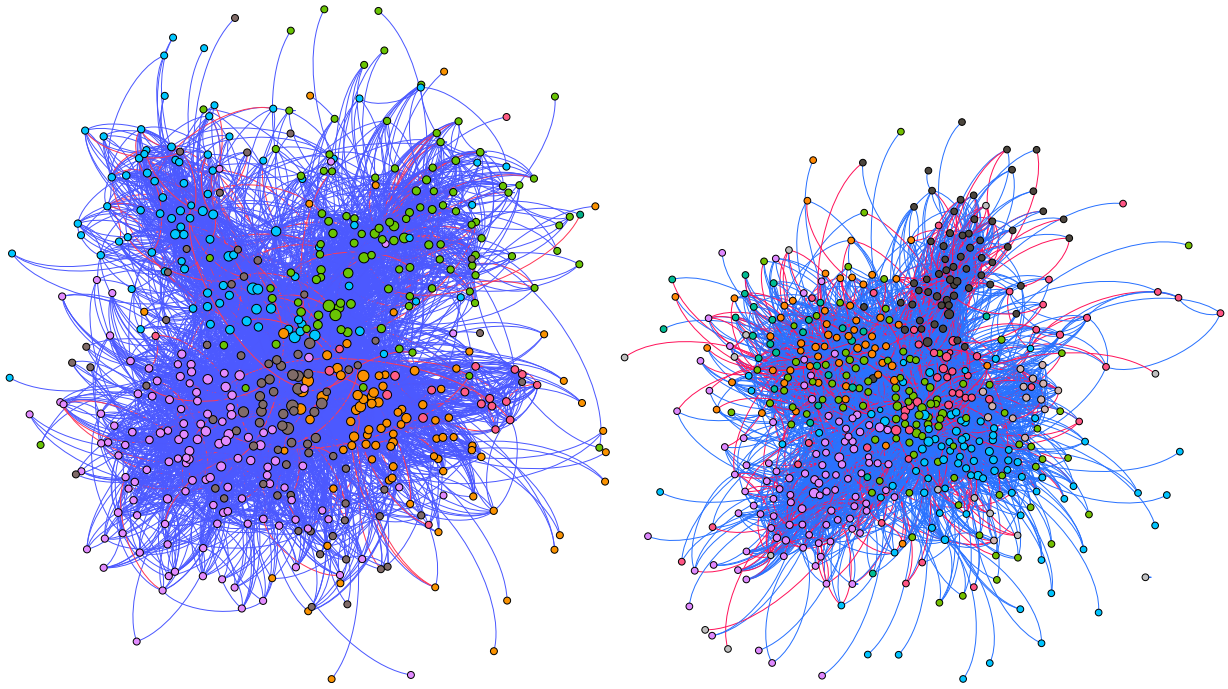
4.7 Results

Tables 4.2 and 4.3 present the results of the micro-level balance for the Enron networks. The morality network has an overall (micro) balance ratio of 92.37%. All four triad types have high balance ratios, ranging from 91.47% - 93.89%. The sentiment network has an overall balance ratio of 67.50%, with triad 300 having the highest balance ratio (69.94%) and triad 120U having the lowest balance ratio (64.36%). The prevalence of balanced triad 300s shows that balance is present in situations where individuals initiate and reciprocate email communication. One notable difference in triad 300 counts between the morality and sentiment networks is that there is higher partial balance in the sentiment network than in the morality network, where complete balance is higher. This difference indicates that, while three individuals are fully interacting and connected in terms of sending or



(a) Avocado Morality

(b) Avocado Sentiment



(c) Enron Morality

(d) Enron Sentiment

Figure 4.3: Enron and Avocado networks. Clustering based on Louvain modularity. Edge colors: purple=positive, red=negative

Network Measures	Enron		Avocado	
	Morality	**Sentiment**	**Morality**	**Sentiment**
# of nodes	517	518	557	557
# of edges	7605	7510	22479	23910
Transitivity	0.21	0.2	0.48	0.49
Degree Centralization	0.06	0.06	0.14	0.14
Density	0.03	0.03	0.07	0.07
Clustering Coefficient	0.44	0.44	0.49	0.47
# of Components	1	2	19	19
# of node in largest component	517	517	498	526
Average Path Length in Largest Component	2.46	2.47	1.39	1.32
# of node after removing pendants, isolates, and self-loops	494	491	395	402
# of edges after removing pendants, isolates, and self-loops	7432	7336	22085	23497

Table 4.1: Descriptive network measures of (1) Enron, and (2) Avocado networks

receiving emails, there may be inconsistencies with the sentiment exchanged, but not so much with morality. This finding is expected, since sentiment is on the surface and shows the spontaneous reactions or feelings of the people, while moral states of the people may remain more stable across times and events.

Enron's morality and sentiment networks have similar triadic profiles, in which triad 030T occurs most frequently and is often balanced (91.47% for morality, 67.46% for sentiment). In the context of this dataset, the 030T triad represents triples of individuals who are bounded by a certain "local hierarchy" - P sends an email to O, who then sends an email to X, and then P sends an email to X as well. Such behavior implies a hierarchy, where both P and O initiate communication with X, and X may be at a higher level of influence (consistent with the assumptions of the Ranked Clusters model; see [dN99]). High counts of balanced triads of type 030T also indicate a strong correlation between transitivity and balance at the triad level of the network.

High triad 030T counts also mean that there is lower reciprocity at the triad level. This

Enron_ Morality	**Type**	**Count**	**Completely Balanced**	**Partially Balanced**	**Completely Imbalanced**	**Balance Ratio** ($B_{T^{(i)}}$)
Transitive Triads	030T	4514	4129	0	385	91.47%
	120D	2390	2120	161	109	92.07%
	120U	3615	3244	167	204	92.04%
	300	3056	2696	339	21	93.89%
Total		13575	12189	667	719	$T(G) =$ 92.37%

Table 4.2: Balance counts with respect to morality in Enron network

Enron_ Sentiment	**Type**	**Count**	**Completely Balanced**	**Partially Balanced**	**Completely Imbalanced**	**Balance Ratio** ($B_{T^{(i)}}$)
Transitive Triads	030T	4238	2859	0	1379	67.46%
	120D	2384	1333	588	463	68.24%
	120U	3513	1775	972	766	64.36%
	300	3056	1312	1605	139	69.94%
Total		13191	7279	3165	2747	$T(G) =$ 67.50%

Table 4.3: Balance counts with respect to sentiment in Enron network

insight has implications for professional email communication and practices for companies in crisis, as we observe more instances of initiating emails to other individuals and less reciprocity (i.e., replying) in exchanging emails. Triad 300 represents complete and reciprocated interactions among three individuals, and these communications are carried out with less tension. In addition, we observe high counts of triads of type 120U, which indicate information reporting (120U, P and O reporting up to X), but not of type 120D, which indicate the act of passing down information. This finding suggests hierarchical information flow at Enron, where email communication is initiated by employees and sent to personnel at different levels in the organization.

Tables 4.4 and 4.5 show the results of the micro-level balance for the Avocado networks. The overall balance ratio for morality is 86.70%, with triad 030T having the lowest balance ratio (80.74%) and 300 having the highest balance ratio (93.47%). The overall balance ratio for sentiment is 82.47%, with the same profile triad 030T has the lowest balance ratio (76.22%), and triad 300 has the highest balance ratio (90.28%). In addition, triad 300 is the most frequently-occurring one in both Avocado networks.

Similar to the Enron networks, the Avocado networks contain substantially more counts of 120U than 120D. Recurring prominence of 120U triads in email communication networks may

Avocado_ Morality	**Type**	**Count**	**Completely Balanced**	**Partially Balanced**	**Completely Imbalanced**	**Balance Ratio** ($B_{T^{(i)}}$)
Transitive Triads	030T	8787	7095	0	1692	80.74%
	120D	14111	11627	882	1602	85.52%
	120U	26165	22257	1047	2861	87.06%
	300	124371	109528	13203	1640	93.47%
Total		173434	150507	15132	7795	$T(G) =$ 86.70%

Table 4.4: Balance counts with respect to morality in Avocado network

Avocado_ Sentiment	**Type**	**Count**	**Completely Balanced**	**Partially Balanced**	**Completely Imbalanced**	**Balance Ratio** ($B_{T^{(i)}}$)
Transitive Triads	030T	8577	6538	0	2039	76.22%
	120D	14276	10816	1408	2052	80.69%
	120U	28615	22802	1725	4088	82.69%
	300	144865	118673	23870	2322	90.28%
Total		196333	158829	27003	10501	$T(G) =$ 82.47%

Table 4.5: Balance counts with respect to sentiment in Avocado network

indicate the prevalence of information reporting. We observe more consistency in balance ratios of the Avocado networks compared to Enron, where balance difference is only 4.23% for Avocado and 24.87% for Enron. One reason that such inconsistencies are only in the Enron networks could be that this company underwent a series of crises that resulted in bankruptcy, which may have had profound effects on the sentiment of the emails.

The overall micro-level balance of Avocado's morality network (86.70%) is slightly lower than Enron's morality network (92.37%), possibly because Avocado's network size is three times larger, hence providing more opportunities to develop balance (or in this case imbalance) among triads. On the other hand, Avocado's sentiment network has a higher balance ratio (82.47%) than Enron's sentiment network (67.50%), indicating that there may be less tension in the emails exchanged between Avocado employees compared to those at Enron. Another difference between Avocado and Enron is that Avocado networks contain higher proportions of 300s triads (72% for morality; 74% for sentiment). In contrast to Enron networks, which contain mostly 030T triads, Avocado networks are more tightly-connected with frequent and reciprocated communications. With respect to triad counts, Enron's morality and sentiment networks have a similar total number of triads (13,575 and 13,191, respectively). Avocado's morality network has notably fewer triads than the sentiment network

Network	**# of Edges (m) (m^+, m^-)**	**Triad Level Balance** $T(G)$	**Subgroup Level Balance**		**Frustration Index** $L(G)$	**Network Level Balance** $F(G)$
			Cohesiveness $C(P^*)$	**Divisiveness** $D(P^*)$		
Enron Morality	7605 (7319,286)	0.92	0.96	0.6	285	0.92
Enron Sentiment	7509 (5948, 1561)	0.67	0.81	0.61	1477	0.61
Avocado Morality	22435 (21351, 1084)	0.87	0.95	0.86	1058	0.91
Avocado Sentiment	23874 (22281, 1593)	0.82	0.93	0.8	1570	0.87

Table 4.6: Multi-level Balance Results

(174,434 and 196,333, respectively), although the number of edges and nodes in these networks are comparable (Table 4.1. This difference in triad counts may be the result of more sentimental terms than morality terms in the email exchanges.

Overall, the results show that different linguistic properties can manifest distinct interactions characteristics.

Furthermore, Table 4.6 presents the multi-level structural balance across the four networks. Comparing the micro-level balance with the meso-level shows that the cohesiveness (intra-group solidarity) is high in all four networks compared to the micro-level balance. Divisiveness (inter-group antagonism), on the other hand, is low compared to the other measurements. These results indicate positive interactions and association between individuals in the same group.

Enron Morality has the highest micro- and macro-level balance, as well as the highest cohesiveness, which indicates that balance is present in this network across different levels. More specifically, while micro-level and cohesiveness measurements of Enron Morality are much higher than those of the other networks, its divisiveness is the lowest. This result is interesting as it shows that positive morality is predominant across the network in general within and between the two groups. This suggests the existence of one (almost-) cohesive community (with respect to moral values) rather than two divided subgroups for this particular network. Enron Sentiment, on the other hand, shows low divisiveness, although the micro-level balance is very low, and we expect to see two divided communities.

Regarding the macro-level balance, the results show that all except Enron Sentiment have

high proportions of balance. Low micro- and macro-level balances suggest that tension is present in the network. This is expected as Enron was going through bankruptcy, and more negative emotions were involved when interactions were taking place in this organization.

Both Avocado networks are showing similar characteristics, with high cohesiveness and macro-level compared to the other two measurements. This result indicates that, while some negative interactions may be present in the lower levels of the networks, this organization is almost stable with positive associations within the sub-groups.

Consistent with previous observations [ADRD20, AW18], we find that each level of balance may lead to distinct observations, and aggregating triads may not yield the same intuitions, especially in sparse networks. Moreover, aside from the Enron Morality network, all the networks have different micro- and macro-level balance measurements. For instance, the Avocado Morality network, with the density of 0.07 and clustering coefficient of 0.49, has the micro-level balance of 0.87, high cohesiveness of 0.95, and macro-level balance of 0.91. A similar trend is observed for the Avocado Sentiment network. The Enron Sentiment network, on the other hand, with the density of 0.03 and clustering coefficient of 0.44, has higher micro-level balance (0.67) than the macro-level (0.61). Overall, while there are cases in which the two measurements match, balance at the micro- and macro-levels is not generally the same property measured at different levels.

4.8 Discussion and Conclusion

In this chapter, we leveraged structural balance analysis to study people's interactions in signed directed networks. The overall purpose of this chapter is to explore the potential use value of user-generated texts for better investigating people's interactions and of different language properties for gaining insights about the structure of the networks. In addition, we utilized a theoretical framework for calculating balance on three different levels (micro-, meso-, and macro-levels) in signed directed networks. Prior work in this area has mainly examined structural balance in signed and indirect graphs. Our approach extends the current analysis of structural balance by incorporating directionality. For the purpose of this study, we leveraged the email exchanges in two large-scale organizations (Enron and AvocadoIT

email datasets) and extracted two language properties, namely sentiment and morality, to investigate the stability and structure of the interactions. Our rationale for testing our approach on different networks was to determine whether mechanisms of structural balance and transitivity hold true across diverse social contexts.

Our findings showed that the degree to which a network was balanced was strongly impacted by choices of measuring social relations. When the direction of edges was taken into account, along with sign consistency, we expected that the overall balance ratio might be different than it would be in findings where only sign consistency was considered [DE15, LHK10a]. Choices of edge type may also have an effect on the overall balance. Our findings showed that utilizing different language properties to construct networks captured distinct characteristics, as reflected in the different balance ratios across morality and sentiment. While balance ratios for both edge types were about 70% and above (balance higher than imbalance), we found that networks labeled with morality as the edge type had the highest balance ratios, while networks labeled with sentiment as the edge types were notably lower.

The patterns of structural balance that we discovered across the networks offer implications for existing communication and organizational networks literature. First, we found that email communication is highly positive in both morality and sentiment. In addition, communication flow was upwards through a hierarchy in the form of information-reporting behavior. One implication of this finding is that the observed communication patterns can provide insights into an organization's formal hierarchy and shed light on the types of influences (e.g., organizational status) that exist to maintain balance in the network. A methodological implication of our findings is that preprocessing text data for network construction impacts balance assessment results. For the sentiment results specifically, overall balance ratios decreased after negation handling and domain adaptation of the applied lexicon. Thus, balance measures may also depend on the researcher's choices about network data preprocessing. This work further expanded research on the impact of human choices about extracting relational data from text data [Die15, DC10].

Second, we observed that choices about constructing and aggregating social network data may impact balance ratios. For the Enron and Avocado email communication networks, we made an informed choice to normalize all communications between two people (Tables 4.2,

4.3, 4.4. We performed additional analyses on email datasets and found that choosing the first instance of email communication between two people results in different balance ratios (77.3% for Avocado-morality, 73.5% for Avocado-sentiment, 86.7% for Enron-morality, 61.2% for Enron-sentiment) than those from considering the last instance of email communication between the same people (76.7% for Avocado-morality, 64.6% for Avocado-sentiment, 86.7% for Enron-morality, 60.0% for Enron-sentiment). These results highlight the recurrent problem of constructing static networks from temporal network data, where researchers must make decisions on either aggregating or disregarding instances. These solutions may result in biasing the overall balance ratio of a network. To address this issue, incorporating temporal data (if applicable) into balance analysis will ensure a more comprehensive analysis of networks, since it would enable an examination of how networks gravitate towards balance throughout time [DE15, UH13].

Furthermore, multi-level analysis showed that balance in the micro-level does not necessarily translate to balance in the meso- or macro-levels. Our results (Table 4.6) showed that each level of networks presents different characteristics and profiles, and, therefore, it is necessary to perform independent analysis to gain better insights about people's interactions as well as the overall structure and stability of the networks. In the Avocado dataset, we observed relatively high values of balance across the three levels using both morality and sentiment. However, for the Enron dataset, while the balance for morality was high, the sentiment in all three levels was low compared to other networks. These findings provide evidence that tension can be observed at all levels in times of crisis.

Analyzing signs of the triads in the micro-level showed that both the Enron and Avocado networks contain higher proportions of positive edges, and as a result higher proportions of positive transitive semicycles. Our findings are consistent with prior work [LHK10a] in which the majority semicycles in three real-world social networks were all found positive (70% to 87%). Moreover, we found that all-negative semicycles are rare (about 0.5% in Enron networks, and 0.03% in Avocado networks), suggesting that it is not common to engage in chains of negative emails.

Our analysis has several limitations. First, we did not consider the temporal characteristics of the interactions in two datasets in our study. We plan to extend this work to analyze

multilevel balance over time and investigate temporal changes in balance and stability in networks. In addition, we only considered one type of communication channel (emails). Further analysis is needed to research a wider range of social networks, which include different temporal and linguistic properties.

Acknowledgment

We would like to thank the graduate students in the iSchool Network Analysis class who helped us with the annotation tasks.

CHAPTER 5

CONCLUSION AND FUTURE DIRECTION

5.1 Revisiting the Proposed Research Questions

This thesis builds on previous work in the areas of natural language processing and computational social science that focuses on developing robust methods for analyzing user-generated texts to understand people while considering their socio-cultural settings. More specifically, in this work we explored three specific and novel research questions, and we investigated different types of user-generated texts to (1) extract and analyze the impact of information products on people by looking at reviews, (2) leverage moral values to better understand polarized viewpoints in tweets, and (3) analyze people's interactions in social networks by utilizing expressed emotions and moral states in emails. The preceding chapters presented the methodologies and research design that were employed to answer the proposed research questions of this dissertation, as well as their results. This chapter revisits these questions and provides a discussion about the findings and contributions of each study.

- **RQ1: How Can We Leverage User-Generated Reviews to Analyze the Impact of Information Products on People's Behavior and Cognition?** *Detecting the Impact of Issue-Focused Documentary Films on People Using Reviews*

  In *Chapter 2*, we first analyzed user-generated reviews to investigate whether users simply indicate their opinion and sentiment regarding a product, or whether they provide any discussion regarding the influence or impact of this product on their knowledge, behavior, or emotions. We then used methods from natural language processing and machine learning to extract textual indicators or representations of different types of impact from texts. More specifically, we developed a theoretically grounded and data-

driven classification schema for impact and generated an annotated corpus of reviews in which each sentence was labeled with one type of impact. We then used our new impact corpus to develop a prediction model to detect different types of impact that documentaries have on people. The results of our analysis showed that documentary films, as information products, are capable of changing peoples' perceptions and cognitions about various types of social issues, e.g., climate change, sugar consumption, and women's roles in communities. In addition, our findings showed that impact is not uniform across all types of information products and can be associated with the messages that these products try to convey. For instance, some films could change the behavior of the users, while others had more influence on people's awareness and cognition. Overall, our results confirm that user-generated reviews may contain information beyond just sentiment, which can provide insights about users' opinions, socio-cultural information, and emotional states.

- **RQ2: How Can We Utilize Personal Values to Study Social Effects?** *Investigating the Effect of Moral Values on Stance Prediction and Analysis in Tweets*

  In *Chapter 3*, we investigated the relationship between principles of human morality and the expression of stances in user-generated text data, namely tweets. This work was based on the intuition that leveraging the social context of the speakers, such as personal beliefs, biases, and values or societal and political environments, is helpful when analyzing user-generated texts. To extract textual indicators or representations of moral values from tweets and analyze their effect on the measurement of stance, we first developed an extended version of the Moral Foundations Dictionary using a quality-controlled, human-in-the-loop process. We then used our enhanced lexicon to develop both feature-based and deep learning classification models to test the usefulness of moral foundations in predicting stance. Using the enhanced lexicon led to measurable improvements in prediction accuracy of stance analysis. After showing the usefulness of moral foundations as a feature for prediction, we performed a detailed analysis of the correlation between different types of moral foundations and social issues (topics) in our dataset, e.g., abortion, climate change, or atheism. Our results showed

that different topics and stances have distinctive distributions of morality types. In addition, we found that each social issue has a distinctive lexical profile, some containing more topics related to purity and authority and others showing more loyalty, for instance. The correlation tests also showed variances in the numbers of statistically significant relationships of morality and stance across different social topics. For instance, we observed that stance is more well-correlated with the *vice* morality type on various dimensions, such as harm, subversion, betrayal, and general-vice. Overall, through this work, we introduced and operationalized morality as a feature for natural language-processing tasks and showed that leveraging socio-cultural settings can result in better understanding of the human language.

- **RQ3: How Can Social and Personal Values as Textual Properties Facilitate Studying People's Social Interactions?** *Exploring Social Networks and their Stability by Extracting Sentiment and Moral Values from Email Communications*

  In *Chapter 4*, we analyzed people's interactions in social networks by first extracting two different linguistic properties, namely emotional and moral states, from user-generated texts (emails). We then leveraged multi-level structural balance analysis to explore individuals' interactions in their network. In this work, we developed a new methodology that advances the structural balance theory by including direction in the analysis of balance. Moreover, structural balance theory has been (primarily) analyzed in undirected graphs, but in real-world networks, relationships are not always reciprocated; individual A may perceive individual B as a friend, but B may not have the same perception of A. Our extended version of structural balance addresses this shortcoming. The results of our analysis showed that each linguistic property and each level of analysis provides unique and different insights about people's interactions, as well as the stability of the networks. For instance, we found that a person's moral status stays balanced through the network even in the presence of tension, while the sentiment networks demonstrated tension across the interactions. Furthermore, our analysis advances research in the area of network analysis by extending the theory and incorporating direction into the analysis of structural balance. We also leveraged

natural language processing to infer two distinct social aspects from texts and used those aspects to analyze interactions and balance in social networks.

Overall, our studies broaden our understanding of the impact of information products on individuals' everyday lives and people's interactions with online communities. In addition, we show that studying language can decode social relations as well as the social roles, dynamics, and structures of different people, and can contribute to the development of socially aware NLP models and a better understanding of real-world communication.

As mentioned in Chapters 2, 3, and 4, for each study, we made specific data and methodological choices. Our selections have several limitations. For instance, when studying the impact of documentary films, we only focused on specific domains and genres of films, and when studying stance, we solely considered Twitter data. Our results show that our computational models and approaches can yield compelling outcomes in some cases, as well as unsatisfactory performances in others. Future work is needed to expand the methodologies and frameworks proposed in this thesis to better explore and explain user-generated text data.

5.2 Future Directions and Research

This thesis opens up several research avenues that are worth further pursuit. Here, we discuss a few possible avenues of investigation.

Impact Analysis at a Larger Scale

In *Chapter 2*, we focused on identifying the impact of documentary films on people's behaviors, cognitions, and emotions. First, future work can extend this analysis to study other types of information products [Rez20] and films, such as motion pictures, and compare different types of observed impact. In addition, in this work, we only focused on micro-level impact. However, since people are the building blocks of groups, communities, and societies, it may be insightful to investigate impact on a larger scale to find if information products

can have any effect on the meso-level (e.g., norms or rules in workplaces or organizations) and macro-level (e.g., social structures and legislations).

In addition, as shown in *Chapter 2*, information products are capable of affecting people's attitudes, behaviors, and cognitions. However, little is known about the correlation between various socio-cultural settings such as personality and culture, and about the impact of these products. Leveraging computational models and causal analysis can lead to a better understanding of these relationships in a more systematic manner. Future analyses can examine how people with different demographics, personality types, cultures, and value systems perceive or are affected by different types of information products. Overall, the following questions can deliver more insights about the impact of information products:

- What methods, frameworks, and taxonomies can be used to study and analyze the broader impact (meso- and macro-level) of different types of information products, e.g., (funded) research and (e-)books?

- Can leveraging causal inference extend the analysis of impact? How are impact and people's socio-cultural information correlated?

User-Generated Texts, Social Contexts, and Socio-Cultural Information

In *Chapter 3*, we studied the usefulness of user-informed features, namely moral values, in analyzing stance and polarized viewpoints in tweets. Our work followed the standard model of stance analysis as a binary problem (in favor or against), but on certain issues there might be more than these two points of view. Additionally, we recognize the constraints of our data sample, as we focused only on US-related social issues and English-only language use from one medium (Twitter). In the future, we hope to expand the study of morality and stance to encompass more issues from a more diverse set of social contexts and effects.

Furthermore, automated communication tools are becoming a cornerstone of online communication. However, to be effective, these tools must consider various aspects of natural language, which are informed by social factors and peoples' socio-cultural contexts. Knowing and understanding different aspects of human language not only increases the performance

of these models, but also improves the models' communication efficiency and minimizes discrimination against marginalized communities. In addition, knowing these characteristics can facilitate better investigations of problems such as misinformation or disinformation on online platforms and detecting malicious behaviors. Moreover, studying language and its context can decode social relations as well as the social roles, dynamics, and structures of different groups and people. Current language processing models do not take social and personal differences into consideration when analyzing users' text data. In their recent work, Hovy and Yang [HYS21] discuss the importance of different social factors in analyzing user-generated texts and provide a taxonomy of social factors to include in NLP systems in order to enhance natural language understanding and analysis. With the vast amount of texts available online, there is a unique opportunity to study these social and personal dimensions in language. As an example, we can focus on the following questions to develop socially aware, human-centered NLP models for social good:

- What ethical frameworks can be used when analyzing user-generated texts while leveraging social contexts and user-informed features?

- How can systems and NLP models leverage socio-cultural information such as personality, moral values, and culture to better understand and analyze affects, sentiment, sarcasm, emotion, and stance in user-generated texts?

- How can users' social constructs be utilized to better understand online phenomena such as (mis)information perception, controversy, and social movements in different communities and cultures?

- How can online platforms leverage NLP models to maintain and facilitate cross-cutting communications in diverse and polarized communities?

Social Interactions and Socio-Cultural Information

In *Chapter 4*, we studied social interaction in a specific network setting by extracting two linguistic properties from texts. Obviously, considering different networks as well as different language properties can provide more insights into the interactions in a more general setting.

In addition, future research can look into the impact of various personal factors on people's interactions. For example, in organization settings, people's statuses and roles can change, moving upward or downward and vertically or horizontally, within their community. It is worth further pursuit to analyze how these changes affect interactions. In addition, previous studies found evidence that men tend to dominate online interactions and can be more aggressive in their communication than women [HO12]. Moreover, previous work suggests that women offer more social support, while men try to display their knowledge and maintain their social statuses. Future work in analyzing social interactions can examine these phenomena to further investigate the following questions:

- How do changes in social statuses and roles impact people's interactions and, as a result, the balance in their networks?

- How do demographic differences (e.g., gender) affect the balance and communication flow in social networks?

- What personal or social factors can be utilized to better understand social interactions?

5.3 Final Statement

In this dissertation, we showed that language is a rich source of information that can be used to extract various properties such as opinion, sentiment, and stance. We showed that language, in the form of user-generated content, is a powerful means though which we can study people's values, cultures, and beliefs and utilize them to better understand people's real-world communications and behaviors. We introduced and explored various methodologies and concepts to measure properties such as social impact, stances and polarized view points, moral values, emotional states and sentiment, and social interactions in user-generated contents.

While there have been tremendous achievements and improvements in the area of NLP and the study of user-generated contents, we are still in the early stages of fully leveraging social contexts and personal factors to analyze humans' language and online texts while

considering ethical frameworks. We hope that this thesis paves the way for such opportunities and provides researchers in the field of computational social science with potential venues of exploration in the future.

REFERENCES

[80117] *Iso/iec/ieee international standard - systems and software engineering–vocabulary*, ISO/IEC/IEEE 24765:2017(E) (2017), 1–541.

[AB02] Réka Albert and Albert-László Barabási, *Statistical mechanics of complex networks*, Reviews of modern physics **74** (2002), no. 1, 47.

[ADRD20] Samin Aref, Ly Dinh, Rezvaneh Rezapour, and Jana Diesner, *Multilevel structural evaluation of signed directed social networks based on balance theory*, Scientific Reports **10** (2020), no. 1, 1–12.

[ALB+19] Omid Askarisichani, Jacqueline Ng Lane, Francesco Bullo, Noah E Friedkin, Ambuj K Singh, and Brian Uzzi, *Structural balance emerges and explains performance in risky decision-making*, Nature communications **10** (2019), no. 1, 1–10.

[AM19] Abeer Aldayel and Walid Magdy, *Your stance is exposed! analysing possible factors for stance detection on social media*, Proceedings of the ACM on Human-Computer Interaction **3** (2019), no. CSCW, 1–20.

[AMW18] Samin Aref, Andrew J Mason, and Mark C Wilson, *Computing the line index of balance using integer programming optimisation*, Optimization Problems in Graph Theory (Boris Goldengorin, ed.), Springer, 2018, pp. 65–84.

[AMW20] Samin Aref, Andrew J. Mason, and Mark C. Wilson, *A modeling and computational study of the frustration index in signed networks*, Networks **75** (2020), no. 1, 95–110.

[AN20] Samin Aref and Zachary Neal, *Detecting coalitions by optimally partitioning signed networks of political collaboration*, Scientific Reports **10** (2020), no. 1, 1–10.

[Apa14] Apache Software Foundation, *openNLP Natural Language Processing Library*, 2014, http://opennlp.apache.org/.

[AR58] Robert P. Abelson and Milton J. Rosenberg, *Symbolic psycho-logic: A model of attitudinal cognition*, Behavioral Science **3** (1958), no. 1, 1–13 (en).

[ARAD17] Aseel Addawood, Rezvaneh Rezapour, Omid Abdar, and Jana Diesner, *Telling apart tweets associated with controversial versus non-controversial topics*, Proceedings of the Second Workshop on NLP and Computational Social Science, 2017, pp. 32–41.

[ARPS16] Abdullah Almaatouq, Laura Radaelli, Alex Pentland, and Erez Shmueli, *Are you your friends' friend? poor perception of friendship ties limits the ability to promote behavioral change*, PloS one **11** (2016), no. 3.

[AW18] Samin Aref and Mark C. Wilson, *Measuring partial balance in signed networks*, Journal of Complex Networks **6** (2018), no. 4, 566–595.

[AW19] Samin Aref and Mark C Wilson, *Balance and frustration in signed networks*, Journal of Complex Networks **7** (2019), no. 2, 163–189.

[AWA+11] Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor, *Cats rule and dogs drool!: Classifying stance in online debate*, Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis, Association for Computational Linguistics, 2011, pp. 1–9.

[BAB+11] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky, *Exposure to opposing views on social media can increase political polarization*, Proceedings of the National Academy of Sciences **115** (2018-09-11), no. 37, 9216–9221.

[Bat00] Gregory Bateson, *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology*, University of Chicago Press, 2000.

[BC12] Danah Boyd and Kate Crawford, *Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon*, Information, communication & society **15** (2012), no. 5, 662–679.

[BCD17] Adrian Benton, Glen Coppersmith, and Mark Dredze, *Ethical research protocols for social media health research*, Proceedings of the First ACL Workshop on Ethics in Natural Language Processing (Valencia, Spain), Association for Computational Linguistics, April 2017, pp. 94–102.

[BCZ+16] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai, *Man is to computer programmer as woman is to homemaker? debiasing word embeddings*, arXiv preprint arXiv:1607.06520 (2016).

[Bec01] Henk A Becker, *Social impact assessment*, European Journal of Operational Research **128** (2001), no. 2, 311–321.

[BES14] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen, *Gender identity and lexical variation in social media*, Journal of Sociolinguistics **18** (2014), no. 2, 135–160.

[BF88] Douglas Biber and Edward Finegan, *Adverbial stance types in english*, Discourse processes **11** (1988), no. 1, 1–34.

[BHNS16] Johanna Blakley, Grace Huang, Sheena Nahm, and Heesung Shin, *Changing appetites & changing minds: Measuring the impact of" food, inc."*, The USC Annenberg Norman Lear Center, nd (2016).

[BJK+90] H Russell Bernard, Eugene C Johnsen, Peter D Killworth, Christopher McCarty, Gene A Shelley, and Scott Robinson, *Comparing four different methods for measuring personal social networks*, Social networks **12** (1990), no. 3, 179–215.

[BK20] Emily M Bender and Alexander Koller, *Climbing towards nlu: On meaning, form, and understanding in the age of data*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5185–5198.

[BKA+18] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov, *Predicting factuality of reporting and bias of news media sources*, arXiv preprint arXiv:1810.01765 (2018).

[BKL09] Steven Bird, Ewan Klein, and Edward Loper, *Natural language processing with python: analyzing text with the natural language toolkit*, " O'Reilly Media, Inc.", 2009.

[BL08] Diana Barrett and Sheila Leddy, *Assessing creative media's social impact*, The Fledgling Fund (2008).

[Blo15] Per Block, *Reciprocity, transitivity, and the mysterious three-cycle*, Social Networks **40** (2015), 163–173.

[BMP21] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria, *Investigating gender bias in bert*, Cognitive Computation (2021), 1–11.

[BMR+20] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., *Language models are few-shot learners*, arXiv preprint arXiv:2005.14165 (2020).

[Bon12] Phillip Bonacich, *Introduction to mathematical sociology*, Princeton University Press, Princeton, 2012 (eng).

[BRS02] Ruth Berman, Hrafnhildur Ragnarsdóttir, and Sven Strömqvist, *Discourse stance: Written and spoken language*, Written Language & Literacy **5** (2002), no. 2, 255–289.

[Bru02] Amy Bruckman, *Studying the amateur artist: A perspective on disguising data collected in human subjects research on the internet*, Ethics and Information Technology **4** (2002), no. 3, 217–231.

[BWP+15] Ryan L Boyd, Steven R Wilson, James W Pennebaker, Michal Kosinski, David J Stillwell, and Rada Mihalcea, *Values in words: Using language to evaluate and understand personal values*, Ninth International AAAI Conference on Web and Social Media, 2015.

[C+18] François Chollet et al., *Keras: The python deep learning library*, Astrophysics Source Code Library (2018).

[CA11] J Clark and B Abrash, *Social justice documentary: Designing for impact*, Center for Social Media (2011).

[CBDC19] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury, *Who is the" human" in human-centered machine learning: The case of predicting mental health from social media*, Proceedings of the ACM on Human-Computer Interaction **3** (2019), no. CSCW, 1–32.

[CBHK02] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer, *Smote: synthetic minority over-sampling technique*, Journal of artificial intelligence research **16** (2002), 321–357.

[CD14] Caty Borum Chattoo and Angelica Das, *Assessing the social impact of issues-focused documentaries: Research methods & future considerations*, Center for Media and Social Impact, School of Communication at American University (2014).

[CH56] Dorwin Cartwright and Frank Harary, *Structural balance: a generalization of Heider's theory*, Psychological Review **63** (1956), no. 5, 277–293.

[CHN+14] Kai-Yang Chiang, Cho-Jui Hsieh, Nagarajan Natarajan, Inderjit S Dhillon, and Ambuj Tewari, *Prediction and clustering in signed networks: a local to global perspective*, The Journal of Machine Learning Research **15** (2014), no. 1, 1177–1213.

[CI08] Taya R Cohen and Chester A Insko, *War and peace: Possible approaches to reducing intergroup conflict*, Perspectives on Psychological Science **3** (2008), no. 2, 87–93.

[CJK04] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz, *Special issue on learning from imbalanced data sets*, ACM Sigkdd Explorations Newsletter **6** (2004), no. 1, 1–6.

[CL13] Eean R Crawford and Jeffery A LePine, *A configural theory of team processes: Accounting for the structure of taskwork and teamwork*, Academy of Management Review **38** (2013), no. 1, 32–48.

[CMB06] Aron Culotta, Andrew McCallum, and Jonathan Betz, *Integrating probabilistic extraction models and data mining to discover relations and patterns in text*, Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (USA), HLT-NAACL '06, Association for Computational Linguistics, 2006, p. 296–303.

[CMD06] Hang Cui, Vibhu Mittal, and Mayur Datar, *Comparative experiments on sentiment classification for online product reviews*, AAAI, vol. 6, 2006, p. 30.

[CPCO20] Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice J O'Toole, *Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?*, IEEE Transactions on Biometrics, Behavior, and Identity Science (2020).

[Cra17] Kate Crawford, *The trouble with bias*, Conference on Neural Information Processing Systems, invited speaker, 2017.

[CRF+11] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini, *Political polarization on twitter*, Fifth international AAAI conference on weblogs and social media, 2011.

[CS92] Herbert H Clark and Michael F Schober, *Asking questions and influencing answers.*

[DA07] Ann Devitt and Khurshid Ahmad, *Sentiment polarity identification in financial news: A cohesion-based approach*, Proceedings of the 45th annual meeting of the association of computational linguistics, 2007, pp. 984–991.

[DAE+16] Hongzhong Deng, Peter Abell, Ofer Engel, Jun Wu, and Yuejin Tan, *The influence of structural balance and homophily/heterophobia on the adjustment of random complete signed networks*, Social Networks **44** (2016), 190–201.

[DAPGD11] Jorge Carrillo De Albornoz, Laura Plaza, Pablo Gervás, and Alberto Díaz, *A joint model of feature mining and sentiment analysis for product review rating*, European conference on information retrieval, Springer, 2011, pp. 55–66.

[Dav79] James A Davis, *The Davis Holland Leinhardt studies: An overview*, Perspectives on social network research, Elsevier, 1979, pp. 51–62.

[DC10] Jana Diesner and Kathleen M Carley, *Extraktion relationaler daten aus texten*, Handbuch Netzwerkforschung, Springer, 2010, pp. 507–521.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805 (2018).

[DE15] Jana Diesner and Craig S Evans, *Little bad concerns: Using sentiment analysis to assess structural balance in communication networks*, 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2015, pp. 342–348.

[DFC05] Jana Diesner, Terrill L Frantz, and Kathleen M Carley, *Communication networks from the enron email corpus "it's always about the people. enron is no different"*, Computational & Mathematical Organization Theory **11** (2005), no. 3, 201–228.

[DFFL13] Lúcia Drummond, Rosa Figueiredo, Yuri Frota, and Mário Levorato, *Efficient solution of the correlation clustering problem: An application to structural balance*, On the Move to Meaningful Internet Systems: OTM 2013 Workshops (Berlin, Heidelberg) (Yan Tang Demey and Hervé Panetto, eds.), Springer Berlin Heidelberg, 2013, pp. 674–683.

[DI07] Hal Daume III, *Frustratingly easy domain adaptation*, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 256–263.

[Die15] Jana Diesner, *Words and networks: How reliable are network data constructed from text data?*, Roles, Trust, and Reputation in Social Media Knowledge Markets, Springer, 2015, pp. 81–89.

[Die19] David K Diehl, *Language and interaction: applying sociolinguistics to social network analysis*, Quality & Quantity **53** (2019), no. 2, 757–774.

[DJH+16] Morteza Dehghani, Kate Johnson, Joe Hoover, Eyal Sagi, Justin Garten, Niki Jitendra Parmar, Stephen Vaisey, Rumen Iliev, and Jesse Graham, *Purity homophily in social networks.*, Journal of Experimental Psychology: General **145** (2016), no. 3, 366.

[DL67] James A Davis and Samuel Leinhardt, *The structure of positive interpersonal relations in small groups*, Sociological Theories in Progress **2** (1967), 218–251.

[DM96] Patrick Doreian and Andrej Mrvar, *A partitioning approach to structural balance*, Social networks **18** (1996), no. 2, 149–168.

[DMBA16] Janis L Dickinson, Poppy McLeod, Robert Bloomfield, and Shorna Allred, *Which moral foundations predict willingness to make lifestyle changes to avert climate change in the usa?*, PloS One **11** (2016), no. 10, e0163852.

[DMK11] Jean Decety, Kalina J Michalska, and Katherine D Kinzler, *The contribution of emotion and cognition to moral sensitivity: a neurodevelopmental study*, Cerebral cortex **22** (2011), no. 1, 209–220.

[dN99] Wouter de Nooy, *The sign of affection: Balance-theoretic models and incomplete signed digraphs*, Social Networks **21** (1999), no. 3, 269–286.

[DNMLPK12] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg, *Echoes of power: Language effects and power differences in social interaction*, Proceedings of the 21st international conference on World Wide Web, 2012, pp. 699–708.

[DNMSJ+13] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts, *A computational approach to politeness with application to social factors*, arXiv preprint arXiv:1306.6078 (2013).

[DR15] Jana Diesner and Rezvaneh Rezapour, *Social computing for impact assessment of social change projects*, International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, Springer, 2015, pp. 34–43.

[DRJ16] Jana Diesner, Rezvaneh Rezapour, and Ming Jiang, *Assessing public awareness of social justice documentary films based on news coverage versus social media*, IConference 2016 Proceedings (2016).

[DRJD20] Ly Dinh, Rezvaneh Rezapour, Lan Jiang, and Jana Diesner, *Structural balance in signed digraphs: Considering transitivity to measure balance in graphs constructed by using different link signing methods*, arXiv preprint arXiv:2006.02565 (2020).

[DSSG14] Morteza Dehghani, Kenji Sagae, Sonya Sachdeva, and Jonathan Gratch, *Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the "ground zero mosque"*, Journal of Information Technology & Politics **11** (2014), no. 1, 1–14.

[DT12] Sharon L Dawson and Graham A Tyson, *Will morality or political ideology determine attitudes to climate change*, Australian Community Psychologist **24** (2012), no. 2, 8–25.

[DTMWF19] Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández, *You shall know a user by the company it keeps: Dynamic representations for social media users in nlp*, arXiv preprint arXiv:1909.00412 (2019).

[EB14] Ernesto Estrada and Michele Benzi, *Walk-based measure of balance in signed networks: Detecting lack of balance in social networks*, Physical Review E **90** (2014), no. 4, 1–10.

[Eck12] Penelope Eckert, *Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation*, Annual review of Anthropology **41** (2012), 87–100.

[ED16] Heba Elfardy and Mona Diab, *CU-GWU perspective at SemEval-2016 task 6: Ideological stance detection in informal text*, Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (San Diego, California), Association for Computational Linguistics, June 2016, pp. 434–439.

[ELS17] Benjamin Edelman, Michael Luca, and Dan Svirsky, *Racial discrimination in the sharing economy: Evidence from a field experiment*, American Economic Journal: Applied Economics **9** (2017), no. 2, 1–22.

[Est19] Ernesto Estrada, *Rethinking structural balance in signed social networks*, Discrete Applied Mathematics **268** (2019), 70–90.

[FCUPP16] Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preoţiuc-Pietro, *An empirical exploration of moral foundations theory in partisan news sources*, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 3730–3736.

[Fea64] Norman T Feather, *A structural balance model of communication effect.*, Psychological Review **71** (1964), no. 4, 291.

[Fea67] _____, *A structural balance approach to the analysis of communication effects*, Advances in experimental social psychology, vol. 3, Elsevier, 1967, pp. 99–165.

[Fel10] Christiane Fellbaum, *Wordnet*, Theory and applications of ontology: computer applications, Springer, 2010, pp. 231–243.

[Fes62] Leon Festinger, *A theory of cognitive dissonance*, vol. 2, Stanford university press, 1962.

[FF13] Diane Felmlee and Robert Faris, *Interaction in social networks*, Handbook of social psychology, Springer, 2013, pp. 439–464.

[FIA11] Giuseppe Facchetti, Giovanni Iacono, and Claudio Altafini, *Computing global structural balance in large-scale signed social networks*, Proceedings of the National Academy of Sciences **108** (2011), no. 52, 20953–20958 (en).

[Fin14] Patricia Finneran, *Documentary impact: Social change through storytelling*, StoryMatters and HotDocs (2014), 3–8.

[Fis93] Susan T Fiske, *Controlling other people: The impact of power on stereotyping.*, American psychologist **48** (1993), no. 6, 621.

[Fla70] Claude Flament, *Équilibre d'un graphe: quelques résultats algébriques*, Mathématiques et Sciences Humaines **8** (1970), 5–10.

[Fle20] Lucie Flek, *Returning the N to NLP: Towards contextually personalized classification models*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Online), Association for Computational Linguistics, July 2020, pp. 7828–7838.

[FV19] William Ferreira and Andreas Vlachos, *Incorporating label dependencies in multilabel stance detection*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 6350–6354.

[FY15] Emilio Ferrara and Zeyao Yang, *Measuring emotional contagion in social media*, PloS one **10** (2015), no. 11, e0142390.

[FYWL16] Lei Fang, Qun Yang, Jiawen Wang, and Weihua Lei, *Signed network label propagation algorithm with structural balance degree for community detection*, International Conference on Smart Homes and Health Telematics, Springer, 2016, pp. 427–435.

[GBB11] Xavier Glorot, Antoine Bordes, and Yoshua Bengio, *Domain adaptation for large-scale sentiment classification: A deep learning approach*, Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 513–520.

[GBH+16] Justin Garten, Reihane Boghrati, Joe Hoover, Kate M Johnson, and Morteza Dehghani, *Morality between the lines: Detecting moral sentiment in text*, Proceedings of IJCAI 2016 workshop on Computational Modeling of Attitudes, 2016.

[Geb20] Timnit Gebru, *Race and gender*, The Oxford handbook of ethics of aI (2020), 251–269.

[GH10] Jesse Graham and Jonathan Haidt, *Beyond beliefs: Religions bind individuals into moral communities*, Personality and Social Psychology Review **14** (2010), no. 1, 140–150.

[GHJ+18] Justin Garten, Joe Hoover, Kate M Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani, *Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis*, Behavior research methods **50** (2018), no. 1, 344–361.

[GHK+13] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto, *Moral foundations theory: The pragmatic validity of moral pluralism*, Advances in experimental social psychology, vol. 47, Elsevier, 2013, pp. 55–130.

[GHN09] Jesse Graham, Jonathan Haidt, and Brian A Nosek, *Liberals and conservatives rely on different sets of moral foundations.*, Journal of personality and social psychology **96** (2009), no. 5, 1029.

[GI10] Anindya Ghose and Panagiotis G Ipeirotis, *Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics*, IEEE Transactions on Knowledge and Data Engineering **23** (2010), no. 10, 1498–1512.

[GK09] Eric Gilbert and Karrie Karahalios, *Predicting tie strength with social media*, Proceedings of the SIGCHI conference on human factors in computing systems, 2009, pp. 211–220.

[GL10] Vishal Gupta and Gurpreet Singh Lehal, *A survey of text summarization extractive techniques*, Journal of emerging technologies in web intelligence **2** (2010), no. 3, 258–268.

[GP13] Dan Green and Mayur Patel, *Deepening engagement for lasting impact: A framework for measuring media performance and results*, John S. and James L. Knight Foundation and Bill & Melinda Gates Foundation (2013).

[Gra77] Mark S. Granovetter, *The strength of weak ties*, Social Networks (Samuel Leinhardt, ed.), Academic Press, 1977, pp. 347 – 367.

[GS03] Douglas A Gentile and Arturo Sesma, *Developmental approaches to understanding media effects on individuals*, Media violence and children (2003), 19–37.

[GS13] Justin Grimmer and Brandon M Stewart, *Text as data: The promise and pitfalls of automatic content analysis methods for political texts*, Political analysis **21** (2013), no. 3, 267–297.

[GSAK16] Cornelia Gerdenitsch, Tabea E Scheel, Julia Andorfer, and Christian Korunka, *Coworking spaces: A source of social support for independent professionals*, Frontiers in psychology **7** (2016), 581.

[Gum64] John J Gumperz, *Linguistic and social interaction in two communities 1*, American anthropologist **66** (1964), no. 6_PART2, 137–153.

[Gur20] Gurobi Optimization Inc., *Gurobi optimizer reference manual*, 2020, url: `gurobi.com/documentation/9.0/refman/index.html` date accessed 1 April 2020.

[GWGK19] Raul Vicente Garcia, Lukasz Wandzik, Louisa Grabner, and Joerg Krueger, *The harms of demographic bias in deep face recognition research*, 2019 International Conference on Biometrics (ICB), IEEE, 2019, pp. 1–6.

[Har59] Frank Harary, *On the measurement of structural balance*, Behavioral Science **4** (1959), no. 4, 316–323.

[HBF20] Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari, *"you sound just like your father" commercial machine translation systems include stylistic biases*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Online), Association for Computational Linguistics, July 2020, pp. 1686–1690.

[HBN10] Falk Hüffner, Nadja Betzler, and Rolf Niedermeier, *Separator-based data reduction for signed graph balancing*, Journal of Combinatorial Optimization **20** (2010), no. 4, 335–360.

[HBR19] Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces, *Aspect-based sentiment analysis using bert*, Proceedings of the 22nd Nordic Conference on Computational Linguistics, 2019, pp. 187–196.

[Hei44] Fritz Heider, *Social perception and phenomenal causality*, Psychological Review **51** (1944), no. 6, 358–378.

[Hei46] _____, *Attitudes and cognitive organization*, The Journal of psychology **21** (1946), no. 1, 107–112.

[Hei58] F Heider, *Psychological theory of attribution: Thee psychology of interpersonal relation*, 1958.

[HF19] Christoph Hube and Besnik Fetahu, *Neural based statement classification for biased language*, Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, ACM, 2019, pp. 195–203.

[HFH+09] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten, *The weka data mining software: an update*, ACM SIGKDD explorations newsletter **11** (2009), no. 1, 10–18.

[HG07] Jonathan Haidt and Jesse Graham, *When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize*, Social Justice Research **20** (2007), no. 1, 98–116.

[HJ04] Jonathan Haidt and Craig Joseph, *Intuitive ethics: How innately prepared intuitions generate culturally variable virtues*, Daedalus **133** (2004), no. 4, 55–66.

[HJ+07] Jonathan Haidt, Craig Joseph, et al., *The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules*, The innate mind **3** (2007), 367–391.

[HL70] Paul W Holland and Samuel Leinhardt, *A method for detecting structure in sociometric data*, American Journal of Sociology (1970), 492–513.

[HL78] _____, *An omnibus test for social structure using triads*, Sociological Methods & Research **7** (1978), no. 2, 227–256.

[HL04] Minqing Hu and Bing Liu, *Mining and summarizing customer reviews*, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2004, pp. 168–177.

[HM75] Maureen Hallinan and David D McFarland, *Higher order stability conditions in mathematical models of sociometric or cognitive structure*, Journal of Mathematical Sociology **4** (1975), no. 1, 131–148.

[HM17] Matthew Honnibal and Ines Montani, *spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing*, To appear **7** (2017).

[HO12] Libby Hemphill and Jahna Otterbacher, *Learning the lingo? gender, prestige and linguistic adaptation in review communities*, Proceedings of the ACM 2012 conference on computer supported cooperative work, 2012, pp. 305–314.

[HPWY+20] Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al., *Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment*, Social Psychological and Personality Science **11** (2020), no. 8, 1057–1071.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber, *Long short-term memory*, Neural computation **9** (1997), no. 8, 1735–1780.

[HS16] Dirk Hovy and Shannon L Spruit, *The social impact of natural language processing*, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2, 2016, pp. 591–598.

[HWBS14] Wilhelm Hofmann, Daniel C Wisneski, Mark J Brandt, and Linda J Skitka, *Morality in everyday life*, Science **345** (2014), no. 6202, 1340–1343.

[HYS21] Dirk Hovy, Diyi Yang, and CODA Tech Square, *The importance of modeling social factors of language: Theory and practice*, NAACL, 2021.

[Inv] FERC Western Energy Markets-Enron Investigation, *Pa02-2.(nd). retrieved october 18, 2004.*

[IP10] Molly E Ireland and James W Pennebaker, *Language style matching in writing: Synchrony in essays, correspondence, and poetry.*, Journal of personality and social psychology **99** (2010), no. 3, 549.

[IRSA10] Giovanni Iacono, Fahimeh Ramezani, Nicola Soranzo, and Claudio Altafini, *Determining the distance to monotonicity of a biological network: a graph-theoretical approach*, IET Systems Biology **4** (2010), no. 3, 223–235.

[Jas99] Katarzyna Jaszczolt, *Discourse, beliefs, and intentions: Semantic defaults and propositional attitude ascription.*

[JCT+21] Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea, *How good is NLP? a sober look at NLP tasks through the lens of social impact*, Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (Online), Association for Computational Linguistics, August 2021, pp. 3099–3113.

[JJ11] S John and L James, *Impact: A practical guide for evaluating community information projects*, Knight Foundation (2011).

[JLG17] Kristen Johnson, I-Ta Lee, and Dan Goldwasser, *Ideological phrase indicators for classification of political discourse framing on twitter*, Proceedings of the Second Workshop on NLP and Computational Social Science, 2017, pp. 90–99.

[Joh86] Eugene C Johnsen, *Structure and process: Agreement models for friendship formation*, Social Networks **8** (1986), no. 3, 257–306.

[JWOH20] Hamed Jelodar, Yongli Wang, Rita Orji, and Shucheng Huang, *Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach*, IEEE Journal of Biomedical and Health Informatics **24** (2020), no. 10, 2733–2742.

[KB14] Diederik P Kingma and Jimmy Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980 (2014).

[KBQ+19] Ahmet Baki Kocaballi, Shlomo Berkovsky, Juan C Quiroz, Liliana Laranjo, Huong Ly Tong, Dana Rezazadegan, Agustina Briatore, and Enrico Coiera, *The personalization of conversational agents in health care: systematic review*, Journal of medical Internet research **21** (2019), no. 11, e15360.

[KH06] David Krackhardt and Mark S Handcock, *Heider vs simmel: Emergent features in dynamic structures*, ICML Workshop on Statistical Network Analysis, Springer, 2006, pp. 14–27.

[KJ11] Beth Karlin and John Johnson, *Measuring impact: The importance of evaluation for documentary film campaigns*, M/C Journal **14** (2011), no. 6.

[KK+11] John Kania, Mark Kramer, et al., *Collective impact*, 2011.

[KKY12] Kisok R Kim, Je-Sang Kang, and Seongyi Yun, *Moral intuitions and political orientation: Similarities and differences between south korea and the united states*, Psychological Reports **111** (2012), no. 1, 173–185.

[Kle13] Jon Kleinberg, *Analysis of large-scale social and information networks*, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **371** (2013), no. 1987, 20120378.

[KPCP06] Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti, *Automatically assessing review helpfulness*, Proceedings of the 2006 Conference on empirical methods in natural language processing, Association for Computational Linguistics, 2006, pp. 423–430.

[KS16] Rishemjit Kaur and Kazutoshi Sasahara, *Quantifying moral foundations from various topics on twitter conversations*, 2016 IEEE International Conference on Big Data (Big Data), IEEE, 2016, pp. 2505–2512.

[KSZ+20] Ioannis Kouvatis, Konstantinos Semertzidis, Maria Zerva, Evaggelia Pitoura, and Panayiotis Tsaparas, *Forming compatible teams in signed networks*, arXiv preprint arXiv:2001.03128 (2020).

[L+10] Bing Liu et al., *Sentiment analysis and subjectivity.*, Handbook of natural language processing **2** (2010), no. 2010, 627–666.

[Lak95] George Lakoff, *Metaphor, morality, and politics, or, why conservatives have left liberals in the dust*, Social Research (1995), 177–213.

[Lat81] Bibb Latané, *The psychology of social impact.*, American psychologist **36** (1981), no. 4, 343.

[LB02] Edward Loper and Steven Bird, *Nltk: the natural language toolkit*, arXiv preprint cs/0205028 (2002).

[LC19] Yingjie Li and Cornelia Caragea, *Multi-task stance detection with sentiment and stance lexicons*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (Hong Kong, China), Association for Computational Linguistics, November 2019, pp. 6299–6305.

[LCL16] Duan-Shin Lee, Cheng-Shang Chang, and Ying Liu, *Consensus and polarization of binary opinions in structurally balanced networks*, IEEE Transactions on Computational Social Systems **3** (2016), no. 4, 141–150.

[Lei04] Anthony A Leiserowitz, *Day after tomorrow: study of climate change risk perception*, Environment: Science and Policy for Sustainable Development **46** (2004), no. 9, 22–39.

[LHK10a] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg, *Signed networks in social media*, Proceedings of the SIGCHI conference on human factors in computing systems, 2010, pp. 1361–1370.

[LHK10b] Jure Leskovec, Daniel Huttenlocher, and Jon M. Kleinberg, *Predicting positive and negative links in online social networks*, Proceedings of the 19th international conference on World wide web, ACM, 2010, pp. 641–650.

[LHK10c] ———, *Signed networks in social media*, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (New York, NY, USA) (Elizabeth D. Mynatt, Don Schoner, Geraldine Fitzpatrick, Scott E. Hudson, W. Keith Edwards, and Tom Rodden, eds.), CHI '10, ACM, 2010, pp. 1361–1370.

[LHPW+18] Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji, *Acquiring background knowledge to improve moral value prediction*, 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2018, pp. 552–559.

[LM98] Dora C Lau and J Keith Murnighan, *Demographic diversity and faultlines: The compositional dynamics of organizational groups*, Academy of Management Review **23** (1998), no. 2, 325–340.

[LPA+09] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al., *Social science. computational social science.*, Science (New York, NY) **323** (2009), no. 5915, 721–723.

[LS52] Thomas B Lemann and Richard L Solomon, *Group characteristics as revealed in sociometric patterns and personality ratings*, Sociometry **15** (1952), no. 1/2, 7–90.

[LS69] Claude Levi-Strauss, *The elementary structures of kinship*, no. 340, Beacon Press, 1969.

[LWWH06] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann, *Which side are you on?: identifying perspectives at the document and sentence levels*, Proceedings of the tenth conference on computational natural language learning, Association for Computational Linguistics, 2006, pp. 109–116.

[LZ12] Bing Liu and Lei Zhang, *A survey of opinion mining and sentiment analysis*, Mining text data, Springer, 2012, pp. 415–463.

[LZ16] Michael Luca and Georgios Zervas, *Fake it till you make it: Reputation, competition, and yelp review fraud*, Management Science **62** (2016), no. 12, 3412–3427.

[Mar07] Laila Naif Marouf, *Social networks and knowledge sharing in organizations: a case study*, Journal of knowledge management (2007).

[Mar09] Morgan Marietta, *The absolutist advantage: sacred rhetoric in contemporary presidential debate*, Political Communication **26** (2009), no. 4, 388–411.

[MG20] Megan Morrison and Michael Gabbay, *Community detectability and structural balance dynamics in signed networks*, Physical Review E **102** (2020), no. 1, 012304.

[MHL+17] Marlon Mooijman, Joseph Hoover, Ying Lin, Heng Ji, and Morteza Dehghani, *When protests turn violent: The roles of moralization and moral convergence*.

[MKS+16] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry, *Semeval-2016 task 6: Detecting stance in tweets*, Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 31–41.

[ML13] Julian McAuley and Jure Leskovec, *Hidden factors and hidden topics: understanding rating dimensions with review text*, Proceedings of the 7th ACM conference on Recommender systems, ACM, 2013, pp. 165–172.

[MLA+11] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and James Rosenquist, *Understanding the demographics of twitter users*, Proceedings of the International AAAI Conference on Web and Social Media, vol. 5, 2011.

[MM85] James Milroy and Lesley Milroy, *Linguistic change, social network and speaker innovation*, Journal of linguistics **21** (1985), no. 2, 339–384.

[MMSP14] Myriam D Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen, *Are they different? affect, feeling, emotion, sentiment, and opinion detection in text*, IEEE transactions on affective computing **5** (2014), no. 2, 101–111.

[MOPS18] Stefano Marchesin, Nicola Orio, Chiara Ponchia, and Gianmaria Silvello, *Thirty years of digital libraries research at the university of padua: The user side*, Digital Libraries and Multimedia Archives: 14th Italian Research Conference on Digital Libraries, IRCDL 2018, Udine, Italy, January 25-26, 2018, Proceedings, vol. 806, Springer, 2018, p. 42.

[MR11] Nicolaas Matthijs and Filip Radlinski, *Personalizing web search using long term browsing history*, Proceedings of the fourth ACM international conference on Web search and data mining, 2011, pp. 25–34.

[MR17] Michela Menegatti and Monica Rubini, *Gender bias and sexism in language*, Oxford Research Encyclopedia of Communication, 2017.

[MS21] Milad Moradi and Matthias Samwald, *Evaluating the robustness of neural language models to input perturbations*, arXiv preprint arXiv:2108.12237 (2021).

[MSK17] Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko, *Stance and sentiment in tweets*, ACM Transactions on Internet Technology (TOIT) **17** (2017), no. 3, 26.

[MWZ+19] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, *Model cards for model reporting*, Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 220–229.

[NA09] Matthew C Nisbet and Patricia Aufderheide, *Documentary film: Towards a research agenda on forms, functions, and impacts*, Mass Communication and Society **12** (2009), no. 4, 450–456.

[Nap14] Philip M Napoli, *Measuring media impact: An overview of the field*, Norman Lear Center Media Impact Project. https://learcenter.org/pdf/measuringmedia. pdf (2014).

[NBGRM15] Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler, *Tea party in the house: A hierarchical ideal point topic model and its application to republican legislators in the 112th congress*, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), vol. 1, 2015, pp. 1438–1448.

[NE15] Artur Nilsson and Arvid Erlandsson, *The moral foundations taxonomy: Structural validity and relation to political ideology in sweden*, Personality and Individual Differences **76** (2015), 28–32.

[Neu16] Kimberly A Neuendorf, *The content analysis guidebook*, Sage, 2016.

[New61] Theodore M. Newcomb, *The acquaintance process*, Aldine Publishing Co, Chicago, IL, USA, 1961.

[New03] Mark EJ Newman, *The structure and function of complex networks*, SIAM review **45** (2003), no. 2, 167–256.

[Nob18] Safiya Umoja Noble, *Algorithms of oppression: How search engines reinforce racism*, nyu Press, 2018.

[OWKG15] Douglas Oard, William Webber, David Kirsch, and Sergey Golitsynskiy, *Avocado research email collection*, Philadelphia: Linguistic Data Consortium (2015).

[PBJB15] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn, *The development and psychometric properties of liwc2015*, Tech. report, 2015.

[PJ20] Jiaxin Pei and David Jurgens, *Quantifying intimacy in language*, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 5307–5326.

[PL+08] Bo Pang, Lillian Lee, et al., *Opinion mining and sentiment analysis*, Foundations and Trends® in Information Retrieval **2** (2008), no. 1–2, 1–135.

[PMYW19] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum, *Stancy: Stance classification based on consistency cues*, arXiv preprint arXiv:1910.06048 (2019).

[PP10] Marco Pennacchiotti and Ana-Maria Popescu, *Detecting controversies in twitter: a first study*, Proceedings of the NAACL HLT 2010 workshop on computational linguistics in a world of social media, Association for Computational Linguistics, 2010, pp. 31–32.

[PP11] ———, *A machine learning approach to twitter user classification*, Proceedings of the International AAAI Conference on Web and Social Media, vol. 5, 2011.

[PRERSVC20] Cristina M Pulido, Laura Ruiz-Eugenio, Gisela Redondo-Sama, and Beatriz Villarejo-Carballido, *A new application of social impact in social media for overcoming fake news in health*, International journal of environmental research and public health **17** (2020), no. 7, 2430.

[PRT08] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini, *Discrimination-aware data mining*, Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 560–568.

[PS07] Luís Moniz Pereira and Ari Saptawijaya, *Modelling morality with prospective logic*, Portuguese Conference on Artificial Intelligence, Springer, 2007, pp. 99–111.

[PSE+15] Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman, *Automatic personality assessment through social media language.*, Journal of personality and social psychology **108** (2015), no. 6, 934.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning, *Glove: Global vectors for word representation*, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[PTGN10] Jason Priem, Dario Taraborelli, Paul Groth, and Cameron Neylon, *Altmetrics: A manifesto*.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., *Scikit-learn: Machine learning in python*, Journal of machine learning research **12** (2011), 2825–2830.

[Rap83] Anatol Rapoport, *Mathematical models in the social and behavioral sciences*, Wiley New York, 1983.

[RBF+20] Rezvaneh Rezapour, Jutta Bopp, Norman Fiedler, Diana Steffen, Andreas Witt, and Jana Diesner, *Beyond citations: Corpus-based methods for detecting the impact of research outcomes on society*, Proceedings of the 12th Language Resources and Evaluation Conference (Marseille, France), European Language Resources Association, May 2020, pp. 6777–6785 (English).

[RD17] Rezvaneh Rezapour and Jana Diesner, *Classification and detection of micro-level impact of issue-focused documentary films based on reviews*, Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, ACM, 2017, pp. 1419–1431.

[RD19] ———, *Expanded morality lexicon*, University of Illinois at Urbana-Champaign (2019), 10.13012/B2IDB--3805242_V1.1.

[RDD21] Rezvaneh Rezapour, Ly Dinh, and Jana Diesner, *Incorporating the measurement of moral foundations theory into analyzing stances on controversial topics*, Proceedings of the 32nd ACM Conference on Hypertext and Social Media, 2021, pp. 177–188.

[Rez20] Rezvaneh Rezapour, *Text mining for social good; context-aware measurement of social impact and effects using natural language processing*, Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing, 2020, pp. 141–146.

[RF17] Craig M Rawlings and Noah E Friedkin, *The structural balance theory of sentiment networks: Elaboration and test*, American Journal of Sociology **123** (2017), no. 2, 510–548.

[RKC06] Danny Roobaert, Grigoris Karakoulas, and Nitesh V Chawla, *Information gain, correlation and support vector machines*, Feature extraction, Springer, 2006, pp. 463–470.

[Rob74] Fred S Roberts, *Structural characterizations of stability of signed digraphs under pulse processes*, Graphs and Combinatorics, Springer, 1974, pp. 330–338.

[Ros12] Frank Rose, *The art of immersion: How the digital generation is remaking hollywood, madison avenue, and the way we tell stories*, WW Norton & Company, 2012.

[RRP10] Olaf N Rank, Garry L Robins, and Philippa E Pattison, *Structural logic of intraorganizational networks*, Organization Science **21** (2010), no. 3, 745–764.

[RSD19] Rezvaneh Rezapour, Saumil H Shah, and Jana Diesner, *Enhancing the measurement of social effects by capturing morality*, Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2019, pp. 35–45.

[RSW+20] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes, *Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing*, Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 33–44.

[RWAD17] Rezvaneh Rezapour, Lufan Wang, Omid Abdar, and Jana Diesner, *Identifying the overlap between election result and candidates' ranking based on hashtag-enhanced, lexicon-based sentiment analysis*, 2017 IEEE 11th International Conference on Semantic Computing (ICSC), IEEE, 2017, pp. 93–96.

[Rym95] Betsy Rymes, *The construction of moral agency in the narratives of high-school drop-outs*, Discourse & Society **6** (1995), no. 4, 495–516.

[S+10] Galit Shmueli et al., *To explain or to predict?*, Statistical science **25** (2010), no. 3, 289–310.

[Sam68] Samuel F Sampson, *A novitiate in a period of change: An experimental and case study of social relationships*, Cornell University, Ithaca, NY, USA, 1968.

[SCG+19] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith, *The risk of racial bias in hate speech detection*, Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 1668–1678.

[Sch04] Shalom H Schwartz, *Mapping and interpreting cultural differences around the world*, International studies in sociology and social anthropology (2004), 43–73.

[Sch10] Thomas Schwartz, *The friend of my enemy is my enemy, the enemy of my enemy is my friend: Axioms for structural balance and bi-polarity*, Mathematical Social Sciences **60** (2010), no. 1, 39–45.

[SCNP21] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng, *Societal biases in language generation: Progress and challenges*, arXiv preprint arXiv:2105.04054 (2021).

[SD14] Eyal Sagi and Morteza Dehghani, *Measuring moral rhetoric in text*, Social science computer review **32** (2014), no. 2, 132–144.

[SEK+13] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al., *Personality, gender, and age in the language of social media: The open-vocabulary approach*, PloS one **8** (2013), no. 9, e73791.

[SGGN+17] Hugo Saiz, Jesús Gómez-Gardeñes, Paloma Nuche, Andrea Girón, Yolanda Pueyo, and Concepción L Alados, *Evidence of structural balance in spatial ecological networks*, Ecography **40** (2017), no. 6, 733–741.

[SGQ+19] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi, *Social bias frames: Reasoning about social and power implications of language*, arXiv preprint arXiv:1911.03891 (2019).

[She71] Ronald G Sherwin, *Introduction to the graph theory and structural balance approaches to international relations*, Tech. report, University of Southern California Los Angeles, 1971.

[SM13] Jeffrey A Smith and James Moody, *Structural effects of network sampling coverage i: Nodes missing at random*, Social networks **35** (2013), no. 4, 652–668.

[SM14] Mariana Souto-Manning, *Critical narrative analysis: The interplay of critical discourse and narrative analyses*, International Journal of Qualitative Studies in Education **27** (2014), no. 2, 159–180.

[SM21] Brandon D Stewart and David SM Morris, *Moving morality beyond the ingroup: liberals and conservatives show differences on group-framed moral foundations and these differences mediate the relationships to perceived bias and threat*, Frontiers in Psychology **12** (2021).

[SPE+14] Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and H Andrew Schwartz, *Developing age and gender predictive lexica over social media*, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1146–1151.

[SS07] Sandeepkumar Satpal and Sunita Sarawagi, *Domain adaptation of conditional probability models via feature subsetting*, European Conference on Principles of Data Mining and Knowledge Discovery, Springer, 2007, pp. 224–235.

[SSS+16] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims, *Recommendations as treatments: Debiasing learning and evaluation*, international conference on machine learning, PMLR, 2016, pp. 1670–1679.

[SSW+17] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, *Fake news detection on social media: A data mining perspective*, ACM SIGKDD explorations newsletter **19** (2017), no. 1, 22–36.

[ST06] Dietram A Scheufele and David Tewksbury, *Framing, agenda setting, and priming: The evolution of three media effects models*, Journal of communication **57** (2006), no. 1, 9–20.

[Sti74] Allen H Stix, *An improved measure of structural balance*, Human Relations **27** (1974), no. 5, 439–455.

[STV20] Christoph Stadtfeld, Károly Takács, and András Vörös, *The emergence and stability of groups in social networks*, Social Networks **60** (2020), 129–145.

[STZ05] Xuehua Shen, Bin Tan, and ChengXiang Zhai, *Implicit user modeling for personalized search*, Proceedings of the 14th ACM international conference on Information and knowledge management, 2005, pp. 824–831.

[SVY14] Anya Schiffrin, Cristobal Vasquez, and Nawei Yang, *Measuring media impact.*

[SW09] Swapna Somasundaran and Janyce Wiebe, *Recognizing stances in online debates*, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1 (Stroudsburg, PA, USA), ACL '09, Association for Computational Linguistics, 2009, pp. 226–234.

[SW17] Anna Schmidt and Michael Wiegand, *A survey on hate speech detection using natural language processing*, Proceedings of the fifth international workshop on natural language processing for social media, 2017, pp. 1–10.

[Swe13] Latanya Sweeney, *Discrimination in online ad delivery*, Communications of the ACM **56** (2013), no. 5, 44–54.

[SWL19] Kai Shu, Suhang Wang, and Huan Liu, *Beyond news contents: The role of social context for fake news detection*, Proceedings of the twelfth ACM international conference on web search and data mining, 2019, pp. 312–320.

[SWM19] Yiting Shen, Steven R Wilson, and Rada Mihalcea, *Measuring personal values in cross-cultural user-generated content*, International Conference on Social Informatics, Springer, 2019, pp. 143–156.

[SZ15] Anya Schiffrin and Ethan Zuckerman, *Can we measure media impact? surveying the field*, 2015.

[TCH+20] Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle CM Belgrave, Daphne Ezer, Fanny Cachat van der Haert, Frank Mugisha, et al., *Ai for social good: unlocking the opportunity for positive impact*, Nature Communications **11** (2020), no. 1, 1–6.

[THLS13] Mike Thelwall, Stefanie Haustein, Vincent Larivière, and Cassidy R Sugimoto, *Do altmetrics work? twitter and ten other social web services*, PloS one **8** (2013), no. 5, e64841.

[TJB+21] Samson Tan, Shafiq Joty, Kathy Baxter, Araz Taeihagh, Gregory A Bennett, and Min-Yen Kan, *Reliability testing for natural language processing systems*, arXiv preprint arXiv:2105.02590 (2021).

[TLK12] Jie Tang, Tiancheng Lou, and Jon Kleinberg, *Inferring social ties across heterogenous networks*, Proceedings of the fifth ACM international conference on Web search and data mining, 2012, pp. 743–752.

[Tri89] Harry C Triandis, *The self and social behavior in differing cultural contexts.*, Psychological review **96** (1989), no. 3, 506.

[TSF17] Andreia Sofia Teixeira, Francisco C Santos, and Alexandre P Francisco, *Emergence of social balance in signed networks*, International Workshop on Complex Networks, Springer, 2017, pp. 185–192.

[Tur02] Peter D Turney, *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews*, Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 417–424.

[TW11] Evimaria Terzi and Marco Winkler, *A spectral algorithm for computing social balance*, Proceedings of International Workshop on Algorithms and Models for the Web-Graph (Alan Frieze, Paul Horn, and Paweł Prałat, eds.), WAW 2011, Springer, 2011, pp. 1–13.

[UH13] Shahadat Uddin and Liaquat Hossain, *Dyad and triad census analysis of crisis communication network*, Social Networking (2013).

[Van99] Frank Vanclay, *Social impact assessment*, Handbook of environmental impact assessment **1** (1999), 301–326.

[VDP13] José Van Dijck and Thomas Poell, *Understanding social media logic*, Media and communication **1** (2013), no. 1, 2–14.

[VDSVZ13] René Veenstra, Jan Kornelis Dijkstra, Christian Steglich, and Maarten HW Van Zalk, *Network–behavior dynamics*, Journal of Research on Adolescence **23** (2013), no. 3, 399–412.

[VHJC+13] Tracy Van Holt, Jeffrey C Johnson, Kathleen M Carley, James Brinkley, and Jana Diesner, *Rapid ethnographic assessment for cultural mapping*, Poetics **41** (2013), no. 4, 366–383.

[Vil01] Susan Villani, *Impact of media on children and adolescents: a 10-year review of the research*, Journal of the American Academy of child & adolescent psychiatry **40** (2001), no. 4, 392–401.

[VM16] Annukka Vainio and Jaana-Piia Mäkiniemi, *How are moral foundations associated with climate-friendly consumption?*, Journal of Agricultural and Environmental Ethics **29** (2016), no. 2, 265–283.

[Wal14] Hanna Wallach, *Big data, machine learning, and the social sciences: Fairness, accountability, and transparency*, NIPS Workshop on Fairness, Accountability, and Transparency in Machine Learning, 2014.

[WAS16] Christopher Wolsko, Hector Ariceaga, and Jesse Seiden, *Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors*, Journal of Experimental Social Psychology **65** (2016), 7–19.

[WF94] Stanley Wasserman and Katherine Faust, *Social network analysis: Methods and applications*, vol. 8, Cambridge university press, 1994.

[WHAT04] Fang Wu, Bernardo A Huberman, Lada A Adamic, and Joshua R Tyler, *Information flow in social groups*, Physica A: Statistical Mechanics and its Applications **337** (2004), no. 1-2, 327–335.

[Whi04] David Whiteman, *Out of the theaters and into the streets: A coalition model of the political impact of documentary film and video*, Political Communication, **21** (2004), no. 1, 51–69.

[WLZ10] Hongning Wang, Yue Lu, and Chengxiang Zhai, *Latent aspect rating analysis on review text data: a rating regression approach*, Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACm, 2010, pp. 783–792.

[WR05] Janyce Wiebe and Ellen Riloff, *Creating subjective and objective sentence classifiers from unannotated texts*, International conference on intelligent text processing and computational linguistics, Springer, 2005, pp. 486–497.

[WWC05] Janyce Wiebe, Theresa Wilson, and Claire Cardie, *Annotating expressions of opinions and emotions in language*, Language resources and evaluation **39** (2005), no. 2-3, 165–210.

[WWH05] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, *Recognizing contextual polarity in phrase-level sentiment analysis*, Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005.

[XCJ15] Weiguo Xia, Ming Cao, and Karl Henrik Johansson, *Structural balance and opinion separation in trust–mistrust social networks*, IEEE Transactions on Control of Network Systems **3** (2015), no. 1, 46–56.

[Zas87] Thomas Zaslavsky, *Balanced decompositions of a signed graph*, Journal of Combinatorial Theory, Series B **43** (1987), no. 1, 1–13.

[Zas12] ———, *A mathematical bibliography of signed and gain graphs and allied areas*, The Electronic Journal of Combinatorics, Dynamic Surveys in Combinatorics DS8 (2012), 1–340 (en), url: `www.combinatorics.org/ojs/index.php/eljc/article/view/DS8` date accessed 2015-03-29.

[ZBT+14] Natascha Zeitel-Bank, Ute Tat, et al., *Social media and its effects on individuals and social systems*, Journal Management, Knowledge, And Learning (2014).

[ZF20] Samira Zad and Mark Finlayson, *Systematic evaluation of a framework for unsupervised emotion recognition for narrative text*, Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events, 2020, pp. 26–37.

[ZHJJU21] Samira Zad, Maryam Heidari, H James Jr, and Ozlem Uzuner, *Emotion detection of textual data: An interdisciplinary survey*, 2021 IEEE World AI IoT Congress (AIIoT), IEEE, 2021, pp. 0255–0261.

[ZHJU21] Samira Zad, Maryam Heidari, James H Jones, and Ozlem Uzuner, *A survey on concept-level sentiment analysis techniques of textual data*, 2021 IEEE World AI IoT Congress (AIIoT), IEEE, 2021, pp. 0285–0291.

[ZJZ06] Li Zhuang, Feng Jing, and Xiao-Yan Zhu, *Movie review mining and summarization*, Proceedings of the 15th ACM international conference on Information and knowledge management, ACM, 2006, pp. 43–50.

[ZL09] Lili Zhao and Chunping Li, *Ontology based opinion mining for movie reviews*, International Conference on Knowledge Science, Engineering and Management, Springer, 2009, pp. 204–214.

[ZV06] Zhu Zhang and Balaji Varadarajan, *Utility scoring of product reviews*, Proceedings of the 15th ACM international conference on Information and knowledge management, ACM, 2006, pp. 51–57.

[ZYL+20] Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai, *Enhancing cross-target stance detection with transferable semantic-emotion knowledge*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3188–3197.