Socioculturally Aware Language Technologies

Chan Young Park

CMU-LTI-24-017

August 2024

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Yulia Tsvetkov, Chair
David R. Mortensen
Maarten Sap
David Jurgens (University of Michigan)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
In Language and Information Technology.

Keywords: natural language processing, large language models, social context, cultural context, social norms, community values, community moderation, social biases

Dedicated to my grandmother and Jamie

Abstract

Despite recent advancements in natural language processing (NLP), current models often overlook the critical influence of social context on language. This thesis develops methods to incorporate sociocultural context into NLP models, enhancing their performance, fairness, and generalizability. By focusing on cultural, community, and personal contexts, this thesis aims to enrich NLP models with a deeper understanding of the social dynamics that shape language. The thesis introduces novel approaches for measuring and operationalizing social context, as well as for integrating this information into model architectures and datasets. Through empirical studies across various NLP tasks, this work demonstrates the effectiveness of the proposed methods in improving model performance and addressing social biases. Ultimately, this research contributes to the development of more robust, inclusive, and human-centered language technologies.

Contents

Chapter 1

Introduction

Recent years have witnessed a remarkable surge in the capabilities of language models. Fueled by the scaling of datasets, input lengths, and model complexity, these models [25, 209] have demonstrated unprecedented capabilities in understanding and generating human language, and achieved impressive results on a wide range of NLP tasks, ranging from machine translation to sentiment analysis [12, 15, 18, 25, 99, 112, 209, 226].

Despite these advancements, current NLP models face a critical limitation: their failure to incorporate the rich social contexts that profoundly shape language. The absence of social context in NLP models leads to significant shortcomings, including misinterpretations, biases, and a lack of generalizability. Language is inherently social, and its meaning is deeply embedded in the social contexts in which it is used. When these contexts are ignored, models can produce outputs that are contextually inappropriate or biased, failing to grasp the subtleties that human language often carries [77, 101]. For instance, a model trained solely on text data may struggle to accurately interpret culturally specific idioms or community-specific slang, leading to errors that can undermine its reliability in real-world applications [119, 227]. These limitations suggest that current NLP systems, while powerful, remain incomplete in their ability to fully understand and generate human language as it is naturally used.

Among various social factors, cultural, community, and personal factors are one of the most influential and essential factors in understanding of language, yet are often overlooked in current NLP systems. Cultural context plays a crucial role in how language is used and interpreted, influencing everything from sentiment analysis to cross-lingual transfer [132, 175]. For example, the same word or phrase can carry vastly different connotations depending on the cultural background of the speaker and listener. Similarly, community context, defined by the norms and values specific to particular groups, shapes language use in ways that are critical to understanding meaning within those groups [32, 34]. Personal context, which includes individual attributes like social identity, personal experiences and beliefs, further complicates the task of accurately modeling and understanding anguage [175]. Ignoring these dimensions can lead to oversimplifications and perpetuate biases within NLP systems, highlighting the need for more sophisticated approaches that integrate these diverse contexts.

This thesis aims to bridge this gap by developing NLP models that are deeply grounded in sociocultural context. By focusing on cultural, community, and personal contexts, we seek to create more accurate and inclusive language technologies that can operate effectively across

diverse contexts and populations. To achieve this goal, we introduce new models, datasets, and methodologies that explicitly capture and integrate social context into the NLP pipeline.

Along the three types of social contexts, this work focuses on two key areas. First, we develop methods for measuring and operationalizing social context. This includes identifying and quantifying relevant social contexts that can be directly integrated into NLP models or can provide valuable insights into the underlying mechanisms of language use and develop more effective models. Second, we explore ways to integrate social context directly into NLP models, enhancing their ability to understand and generate language that is contextually appropriate and socially sensitive. This involves developing model architectures that can effectively leverage social information, as well as creating datasets that are annotated with rich social context.

Ultimately, this thesis strives to push the boundaries of NLP by creating language technologies that are more socially aware and contextually grounded. By addressing the limitations of current models and developing new approaches, we aim to contribute to the development of NLP systems that are better aligned with the complexities and nuanced nature of human language and society.

1.1 Thesis Statement

This thesis aims to advance NLP by integrating sociocultural contexts into model development and analysis, addressing the limitations in performance, generalizability, and fairness resulting from current models' neglect of social context. By focusing on three primary sociocultural contexts—Cultural, Community, and Personal—the thesis presents novel methodologies and datasets that enable NLP models to adapt to diverse social norms and cultural backgrounds. It introduces new computational methods, model architectures, and analysis algorithms that incorporate sociocultural awareness, demonstrating the significant impact of social context on language understanding. Additionally, the thesis explores how state-of-the-art NLP models can be leveraged to uncover new social meanings, thereby enhancing our comprehension of language's role within various sociocultural environments. Through these contributions, this work seeks to make NLP more inclusive, accurate, and socially aware, pushing the boundaries of traditional language technologies.

1.2 Thesis Overview

This thesis aims to explore the question, "How can sociocultural context enhance NLP systems?" by presenting concrete examples of how social context can be integrated into NLP models to make them more socially aware and effective. The thesis is organized into three main parts: cultural context and language, community values and language, and personal context and language. Each section addresses specific aspects of sociocultural awareness in NLP:

Part I: Cultural Context and Language In this part, I explore the significance of cultural context in understanding language and demonstrate how it can be leveraged to improve NLP tasks such as cross-lingual transfer.

- Chapter 2. Public Sentiment on Culture and Language Adoption: This chapter explores how NLP systems can operationalize cultural context, particularly cultural public sentiment. By capturing and analyzing public sentiment, we illustrate how it contributes to a deeper understanding of social phenomena such as language borrowing. The chapter highlights the importance of considering social context when studying language change and provides empirical evidences of its role in influencing language borrowing.
- Chapter 3. Cultural Similarity for Cross-Lingual Transfer: This chapter investigates the use of cultural context to enhance the performance of multilingual NLP models, particularly in cross-lingual transfer tasks. The research emphasizes the importance of cultural similarity between languages as a critical factor in selecting training data. We demonstrate that cultural context provides valuable insights that go beyond what can be captured by semantic similarity alone, thereby contributing to more effective cross-lingual model performance.

Part II: Community Values and Language Part II focuses on the interplay between community values and language. I first introduce methods for uncovering the implicit values of communities, then discuss how these community-specific values and norms can be used to develop more effective NLP models for community moderation.

- Chapter 4. Uncovering Community Values Through Social Interactions: Community norms and values play a pivotal role in shaping both behavior and language within a group. This chapter introduces a framework for identifying and understanding these implicit norms and values based on the language and recognition signals used within the community. The insights gained from this analysis can provide valuable guidance on how community norms may evolve over time and offer practical support for community moderators in maintaining the cohesion and healthy environment for the community.
- Chapter 5. Community Norms and Community Moderation: This chapter examines how NLP systems can be improved by incorporating community context, with a particular focus on community moderation. We propose a novel task of detecting community-sensitive norm violations and introduce new datasets and models designed specifically for this purpose. The findings demonstrate the critical importance of understanding and integrating community context into moderation systems, both quantitatively and qualitatively, by showing how communities have their own ways of enforcing rules and maintaining their standards.

Part III: Personal Context and Language The final part of the thesis explores the impact of individual contexts, such as gender and age, on language use and explores how these factors can be leveraged to create more adaptive and effective NLP systems.

- Chapter 6. Identifying Social Biases Using Individuals' Social Attributes: This chapter focuses on identifying social biases in language by considering individual contexts, particularly social attributes. We present a method that matches individuals with others who are comparable except for the target attribute, enabling a controlled analysis of the attribute's impact. As a case study, we apply this matching algorithm to measure and analyze so-

cial biases present in the descriptions of the LGBT community in Wikipedia biographies, providing concrete examples of how social biases manifest in language.

- Chapter 7. Socially Aware Empowering Text Detection: In this chapter, we highlight the significance and potential of incorporating personal context into text classification tasks. We introduce a new dataset designed for the important yet underexplored task of empowerment detection, analyzing how various aspects of social context influence this task. Through extensive modeling, we demonstrate that integrating social context into text classification can significantly enhance the effectiveness of empowerment detection.

The thesis concludes with a discussion of the ethical implications of this work (Chapter 8) and a summary of the key contributions made throughout the research (Chapter 9). These final chapters reflect on the broader impact of integrating sociocultural awareness into NLP and suggest directions for future research.

Part I

Cultural Context and Language

Chapter 2

Public Sentiment on Culture and Language Adoption

Language and culture are intricately intertwined, with culture playing a crucial role in shaping the way language evolves and is used. In this first chapter, I investigate the importance of cultural context in understanding language, aligned with the two main themes of this thesis: (1) how NLP can be used to measure and operationalize social context to deepen our understanding of the relationship between society and language, and (2) how these operationalized social contexts can be leveraged to advance NLP systems.

Understanding the dynamics of language change is a central focus in linguistics, and culture is an indispensable factor in this process. Culture determines the words and expressions present in a language, and language, in turn, adapts according to the cultural and cognitive frameworks of its speakers. This chapter introduces methods to measure cultural context, such as the public sentiment of culture A towards culture B, and explores its relationship with language change, particularly word borrowing. We illustrate that advanced language technologies, such as word connotation classifiers, combined with historical corpora, can effectively measure public sentiment towards different cultures, providing quantitative evidence and insights into how such sentiments influence language adoption and change over time.

2.1 Introduction

Lexical borrowing is the phenomenon of transferring linguistic constructions from a "donor" language to a "recipient" language as a result of contacts between communities speaking those languages [205]. Borrowed words (or *loanwords*) are found in nearly all languages; they account for 10–70% of the vocabulary, making up the "overwhelming majority" of language [168]. Naturally, many studies in historical linguistics are aimed at understanding this universal phenomenon and at gaining meaningful insights into how languages have changed and evolved [91, 92, 221].

Prior work has claimed that the susceptibility of a language to borrowing is determined by both *linguistic* and *socio-political* factors [169, 205]. Linguistic considerations include filling a lexical gap for a concept that has no word yet in the recipient language, e.g., new information or technology, or foreign cultural terms [90]. Socio-political factors reflect the socio-economic

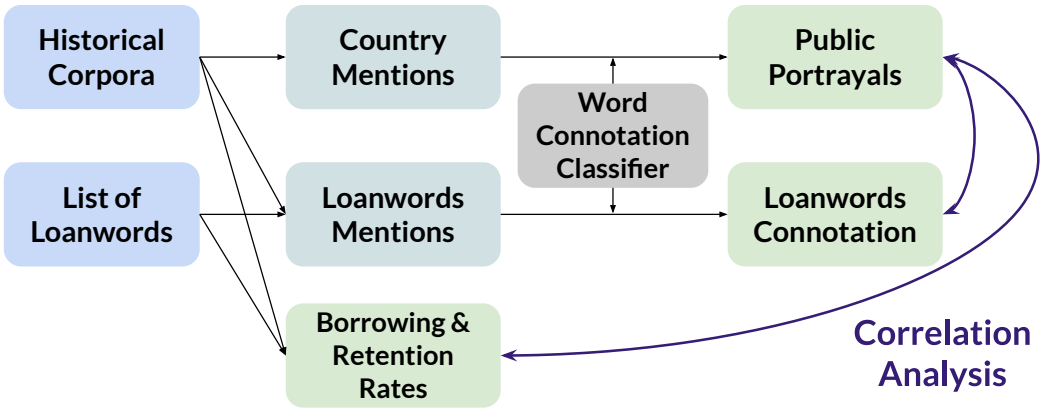


Figure 2.1: Overview of the framework. With two types of resources (historical corpora and lists of loanwords) and word connotation classifiers, we analyze the relationships between public portrayals and lexical borrowing.

and political status quo of contact communities [57]. For example, Krouglov [123] showed how changes in Russian-Ukrainian relations have historically influenced the linguistic structure and behavior of the Russian and Ukrainian languages. Also, McMahon [147] found that when loanwords enter from less prestigious to more prestigious language, they often connote derogatory meanings.

However, as pointed out by Appel and Muysken [7, p. 174], although such literature combines a number of case studies, it lacks generality because loanword lists are typically collected manually and therefore have limited scalability. In this work we develop computational tools and collect resources for studying the phenomenon of lexical borrowing *at scale*. We introduce a computational framework including (1) classifiers that identify the connotations that loanwords carry in context [70, 165]; and (2) classifiers that automatically detect the changes in public portrayals of contact communities throughout long periods of time. We automatically construct lists of loanwords, training data for connotation classifiers, and historical corpora.[1] Our method does not rely on any language-specific techniques, which makes it generalizable to any language of interest . We then use these resources to investigate the correlations between borrowing and socio-political relations.

We focus on two research questions:

RQ1 Is there a *systematic* correlation between socio-political relations in contact communities and the rate of borrowing?

RQ2 Are there correlations between such socio-political relations and the semantic orientation that loanwords acquire in borrowing?

We conduct a large-scale analysis to investigate these research questions using language resources that we collected in Hebrew, Korean, and Russian alongside donor languages for them (Arabic, English, Japanese, and French). Figure 2.1 outlines the structure of our paper. We propose a new method to quantify socio-political relations between countries by employing word connotation classifiers, which predict whether a word implies a positive or negative connotation

[1] All these resources (and our code) are publicly available.

in a given context (§2.3). We connect the measured socio-political relations with the borrowing rate and frequency of loanwords. Finally, we measure the connotations of loanwords and examine how they are connected to political relations and how they change over time (§6.5).

2.2 Background

2.2.1 Country and Language Pairs

Before we dive into the methodology and experiments, we begin by describing the specific target languages and their corresponding donor languages we focus on in this study. Additionally, we describe the nations affiliated with each recipient and donor language, which we categorize as donor and recipient countries. The countries we have chosen have a single dominant official language that is spoken natively by more than half of the population.[2]

Additionally, we provide concise descriptions of the relationships between the donor and recipient countries. Our primary objective is to select pairs of target and donor languages that have a well-documented history of interaction, resulting in a significant presence of loanwords and linguistic exchange between them. Ideally, one of the donor languages should exhibit a distinctively positive or negative political relationship in contrast to the other , thereby facilitating clear expectations for the findings from experimental results.

Figure 2.2 illustrates notable historical events involving the recipient and donor countries which indirectly connote the relationships between them. Notably, in this work we limit our analyses to time periods after 1920, as data collected before this point may not be reliable (e.g., due to scarcity of language resources or poor OCR quality) and may not be representative of the overall linguistic landscape and prevailing sentiment.

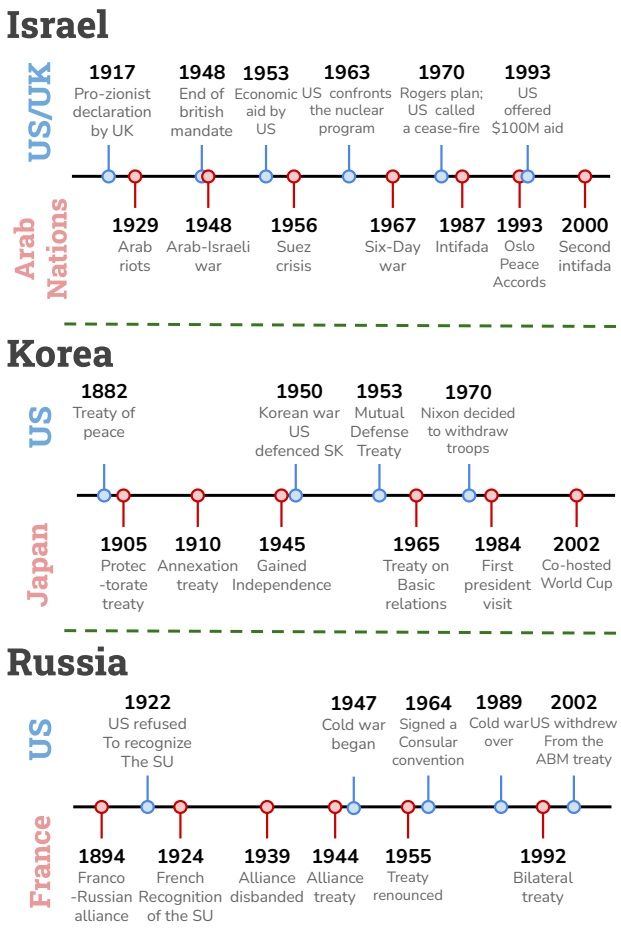


Figure 2.2: Major historical events involving Israel, Korea, and Russia, and their corresponding linguistic donor communities.

[2]Although we make the simplifying assumption that certain countries represent certain languages (e.g., Israel for Hebrew) to simplify our analyses, we acknowledge the fact that there are wider language communities outside of those countries and speakers of other languages within these countries.

Korean In the case of the Korean language, South Korea is identified as the primary recipient country.[3] We chose two donor languages for Korean: English and Japanese. We consider the United States as the primary donor country for English, given the relatively lower degree of linguistic and social interaction with other English-speaking countries. For Japanese, Japan is designated as the donor country.

This choice is underpinned by the historical context of Korea being colonized by the Empire of Japan from 1910 to 1945, following years of conflict and war. It is widely acknowledged that the Korean populace exhibits negative sentiments toward Japan, persisting even after the colonization period, although Japan is often perceived as a more advanced and powerful country.

In contrast, the United States was initially regarded as a foreign adversary before the end of the colonization period. However, the United States played a pivotal role in helping South Korea establish a modern state after gaining independence and notably became a key ally during the Korean war. Consequently, the United States has become South Korea's most influential ally. As a result, the South Korean populace generally maintains a favorable view of the United States and American culture, with South Korea being one of the most pro-American nations worldwide.

Hebrew For Hebrew, we designate Israel as the primary country associated with the recipient language. We consider two donor languages for Hebrew: English and Arabic. Unlike the Korean case, we incorporate multiple donor countries for each language. For English, both the United Kingdom and the United States are designated as donor countries during different time periods. The United Kingdom is associated with English until 1950, after which the United States exclusively represents the language. In contrast, Arabic is spoken across numerous countries. Therefore, we extend our association of Arabic to the entire Arab-speaking world.

In general, the Israeli populace consistently shows negative sentiments toward Arab nations since their establishment, but concurrently maintains more positive relationships with the United States due to substantial economic aid and a long-term alliance.

Russian For Russian, we designate Russia and the Soviet Union as the primary countries associated with the recipient language. We focus on two donor languages for Russian: English and French. The United States is the primary donor country associated with English, while France represents French. While the relationship between the United States and Russia has experienced fluctuations, such as during the Cold War, France and Russia have generally maintained a cooperative relationship. Unlike the previous two cases where English and the United States were associated with relatively more positive sentiments, in the case of Russian, we anticipate that English may be perceived as relatively less positive and prestigious than French, at least during some time periods .

[3]The majority of Korean text we can find online primarily reflects the South Korean context, although North Korea also uses the same language.

2.3 Methodology

The primary goal of this work is to develop a comprehensive framework for a large-scale investigation of lexical borrowing, with a specific emphasis on elucidating the intricate interplay between socio-political variables and lexical borrowing. Our work tries to extend the scope of analysis, overcoming the limitations associated with studies that rely on a limited number of illustrative examples.

An overview of the proposed framework is described in Figure 2.1. To address our research questions, we first develop a computational model that quantifies word connotations, i.e., the implicit sentiments associated with words (here, country mentions and loanwords) within a context. This model forms a foundational component of our analyses. Subsequently, we apply this model to extensive historical corpora, enabling us to gauge the prevailing sentiment exhibited in the text, specifically examining how donor countries and loanwords are portrayed at specific points in time.

In tandem with word connotation analyses, we employ several metrics, including borrowing rates of loanwords, derived from a curated list of loanwords and historical corpora. These metrics are designed to capture different aspects of lexical borrowing. By leveraging these quantitative measures, we provide correlation analyses that investigate the relationship between socio-political factors and lexical borrowing. We now describe in detail the methodology, resources, preprocessing, and computation involved.

2.3.1 Word Connotation Classifier

A standard approach to building a classifier for a specific application is finetuning a language model such as RoBERTa [139] with human-annotated data. Compared with general-purpose generative models such as GPT-3, finetuned models are usually more efficient, as they require fewer parameters, and more effective, as they are specifically trained with application-specific data . Ideally, to obtain a model that works well for the target data, we would want to minimize the difference between the training data and the test data of the specific application. However, in many cases, available annotated data sets are from different domains, which may result in lower performance .

On the other hand, generative language models have demonstrated remarkable capabilities in various tasks even in a zero-shot setting. For example, Koto et al. [122] showed that zero-shot multilingual sentiment classifiers can outperform models fine-tuned on English sentiment datasets, showing opportunities of using LLMs to eliminate the need for training data. Unfortunately, the most advanced models (e.g., GPT-4) are neither transparent (a problem for reproducibility) nor cheap to use. In our case, if we were to use GPT-4 for all country mentions in historical texts, it would have cost more than $5000+ for inferring word connotation for 20M+ sentences.

In order to leverage the benefits of each approach (good zero-shot performance of generative language models and efficient fine-tuned classifiers), we propose a new framework that consists of two steps: (1) use large generative language models with prompts for the target application to construct in-domain training data, and (2) train a RoBERTa-based word connotation classifier using the constructed data.

Training Data Construction with Generative Language Models We use three prompts to generate labels for word connotation classifiers. For example, one of the prompts we use is "{TEXT}. IS THE IMPLIED SENTIMENT TOWARDS {WORD} IN THE GIVEN TEXT POSITIVE, NEUTRAL OR NEGATIVE?". The language models generate answers, which we then categorize into one of the three labels based on the frequency of specific words that correspond to each label. Once the responses have been mapped to a label, we take the majority vote of the three prompts. Any ties, or samples where more than one response cannot be mapped to a label (e.g., "need more context to answer the question"), are ignored.

We employed GPT-3.5 to curate our classifier's training dataset, which comprises two distinct categories. In the first category, we used sentences that reference donor countries in historical texts as the input text in the prompt, with the donor country serving as the target word. For the second category, we used sentences containing loanwords in historical texts as the input in the prompt, with the loanword as the target word. To construct the first category, we randomly selected 3,000 sentences for each donor country across three languages, totaling 18,000 samples. For the second category, we employed a similar approach, randomly selecting 4,500 sentences for each of the six donor-recipient language pairs, which amounted to 27,000 training samples. We then combined these two subsets, resulting in a training dataset comprising 45,000 samples, which was subsequently used to train our classifier.

It's worth highlighting that our approach involves presenting a sentence to the language model and then ask its prediction for the word, making our labels inherently context-specific. Consequently, these labels are taken to indicate the connotation of the word within the context of the sentence, providing a nuanced perspective that differs from more abstract lexicon-based word connotation frameworks, which are primarily focused on aggregated sentiment. This contextualized approach not only enhances the robustness and comprehensiveness of our analysis but also enables us to conduct fine-grained temporal sentiment analysis, as we have access to the timestamp of each example.

RoBERTa-Based Classifier In line with the instructions provided in the data construction prompt, we have framed the word connotation classification task as a three-class text classification challenge. In this context, our model is tasked with analyzing input text and categorizing it into one of three sentiment categories: negative, neutral, or positive. To prepare a classifier for this specific task, we used the dataset we constructed using GPT-3.5 and followed the standard training procedure employed for RoBERTa-based classifiers.

The input format for our classifier is as follows: "{text}</s>the sentiment expressed towards {word}," with the </s> token indicating the end of a sentence. We intentionally chose this format over a more conventional structure , like "{text}</s>{word}," based on our observation that many pretrained language models trained on a task of predicting the subsequent sentence. Consequently, we aimed to ensure that the second part of our input constitutes a complete sentence. Our choice was reinforced by preliminary results, which demonstrated improved performance compared to using only the word as input.

2.3.2 Data and Resources

We now describe how we collected the list of loanwords and historical texts, which we used to study the relationship between political relations and language borrowing at large.

Lists of Loanwords

Previous research has attempted to automatically identify loanwords by comparing the phonemic similarity between words in source and target languages that are translation equivalents. However, these studies have demonstrated the significant challenges in robustly identifying loanwords due to various factors. For instance, two languages may sound similar, yet both could borrow the same word from a third language. In contrast, prior linguistic studies on loanwords have often relied on a limited set of expert-curated loanwords. However, these lists are not generalizable (i.e., collecting one list for one language pair does not provide useful resources for other language pairs) and often limited to the scope of specific phenomena the studies are investigating.

In our work, we adopt a hybrid approach by leveraging loanwords curated by Wiktionary users. Wiktionary hosts a collection of borrowed words for various language pairs. For example, the Category:Hebrew_terms_borrowed_from_English on English Wiktionary contains a list of over 200 English loanwords in Hebrew. Similar lists can be found in other languages on Wiktionary as well (e.g., English loanwords in Hebrew on Hebrew Wiktionary). For each donor–recipient language pair, we merge these two loanword lists from English and the target languages' Wiktionary.

By incorporating Wiktionary into our framework, we ensure that the lists remain continually updated and maintained by a large user community. Furthermore, this framework can be easily expanded to include other languages, thanks to the extensive coverage of Wiktionary pages across a wide range of languages beyond the three languages we analyze in this study. To further improve our lists' coverage, we supplemented the loanword datasets collected from Wiktionary with loanwords manually curated in previous studies on lexical borrowing in Hebrew and Korean.

In the final step, we refined the lists by only retaining words that unambiguously qualify as common loanwords by filtering out the following categories: 1) words that were designated as loanwords in both donor languages, as they typically result from borrowing from a third language, 2) named entities (e.g., names of cities, individuals, organizations, etc.), 3) stop words,[4] 4) extremely short words (those with fewer than 4 characters in Hebrew and Russian, and fewer than 2 characters in Korean—these thresholds were empirically determined through manual inspection), and 5) words that appeared fewer than three times in our historical text corpus (detailed in the following section). To ensure the accuracy of the final lists of loanwords in each language, native speakers of Hebrew, Korean, and Russian with a background in linguistics conducted manual verification.

Historical Texts

In order to analyze when loanwords entered the recipient language and how they are used over time, we need a corpus of historical texts with creation dates. Unfortunately, there is no single

[4]Defined following https://pypi.org/project/stopwordsiso/

	Wik.	Manual	Final
Hebrew ← Arabic	208	542	314
Hebrew ← English	413	282	321
Korean ← Japanese	1694	124	506
Korean ← English	1036	242	497
Russian ← French	934	-	633
Russian ← English	745	-	363

Table 2.1: The number of loanwords found in Wiktionary (**Wik.**) and in manual sources (**Manual**), and the final number after filtering.

comprehensive dataset with such texts for all three of our target languages. Hence we sourced each language's texts differently. Our focus primarily centered on nonfiction texts, with the exclusion of literary, religious, and encyclopedic texts, as we expect these to be less relevant to the study of loanwords across communities speaking different languages.

Hebrew For Hebrew, we leveraged Google Books to gather historical texts. Google Books is an extensive digital library comprising a diverse range of books and printed materials, digitized through optical character recognition (OCR) technology. The Google Books API allows us to retrieve book excerpts containing specific queries, along with metadata such as the book's title, author, and publication date. We adopted the published year of the books as timestamps for the text snippets, covering a period ranging from the 16th century to the present. However, due to the unreliable OCR results in the earliest books, we opted to include only those published after 1920. Additionally, we excluded a substantial number of religious books, as they often lack accurate timestamps (e.g., religious texts from centuries ago may still be published in the 2000s) and are likely less related to the focus of our study on political relations.

It is worth noting that our approach for collecting historical corpora from Google Books is not limited to any specific language, such as Hebrew in our case, and can be applied to any language for which Google Books has data.

Korean The Yonsei 20th Century Corpus [186] comprises approximately 150 million words drawn from a variety of 20th-century texts. In this corpus, timestamps of the texts are binned into decades and do not provide the exact publication date or year. They also provide category information of the sources of the texts (e.g., newspapers, magazines, literature). We chose to exclude literary texts, which are more likely to be fictional, as well as encyclopedic texts.

Russian The Russian National Corpus (RNC) [1] is an extensive corpus of contemporary Russian, encompassing over 2 billion tokens. It covers a period ranging from the mid-18th century to the early 21st century and includes a diverse categories of content, such as fiction, essays, news articles, public speeches, and more. Similar to Korean, we excluded fictional texts such as novels, tales, scenarios, religious texts, and translations based on category information available in the dataset.

Data Collection through Search Interface All three resources we used for collecting historical texts do not grant full access to the entire dataset but instead provide a search interface that allows users to input search terms and retrieve texts containing those query terms. Consequently, we employed extensive query words to maximize the retrieval of texts and to construct the comprehensive historical corpora. To ensure the broadest coverage of our data while ensuring the inclusion of texts containing loanwords, we utilized a word list comprising the final loanword lists and the most frequently used 200,000 words selected by word frequency from Wikipedia.[5]

Following the collection of texts using the search interface, we performed postprocessing on the gathered historical corpora. This involved segmenting texts into sentences and removing duplicate sentences originating from the same source. Additionally, we excluded sentences with fewer than five words to ensure that each sentence in our corpus contained sufficient context. The final historical corpus we constructed consists of approximately 10.2 million, 5.6 million, and 5.7 million unique sentences for Hebrew, Korean, and Russian, respectively. It is important to note that our analysis is confined to the constructed historical corpora and may not represent spoken language.

Translation The word connotation classifier we developed is primarily optimized for English and may not perform as effectively in other languages. This limitation arises from the LLMs' limited capacity to generate training examples in languages other than English reliably. Therefore, we employed off-the-shelf translation models from the HuggingFace platform to translate the historical corpora we collected into English. Specifically, for Hebrew, Korean, and Russian, we used the tiedeman/opus-mt-he-en, Helsinki-NLP/opus-mt-ko-en, and Helsinki-NLP/opus-mt-ru-en models, respectively. Following the translation process, native speakers of each language conducted manual reviews of sampled texts to ensure that the overall translation quality and the sentiment expressed in the texts were preserved accurately after translation.

Lemmatization In our analysis, we focus on the implicit connotations that specific words carry within different sentences provided as context. To analyze trends for a given word, it is crucial to identify the sentences in which those words appear. This process was relatively straightforward before the translation because the search interfaces explicitly returned all texts containing the queried word. However, after translation, there is no guarantee that a direct translation of the word used in the search query will still be present in the translated sentence. As a solution, we employed lemmatization for all translated sentences in our historical corpora using SpaCy's en_core_web_md model. Then, we searched for the English translations of target words (i.e., loanwords and country names in our case) with the lemmatized sentences. When the possible English translation of target words are not found in the translated sentence, we exclude those sentences from word connotation analysis. It's worth noting that we used the lemmatized form exclusively for matching sentences to words but used the original forms for word connotation classification.

[5]Word frequencies were counted based on the Wikipedia dump version from March 2022.

2.3.3 Evaluation

In order to measure how well generative language models in the zero-shot setting and the fine-tuned models work, we collected human annotated data for word connotations in three languages.

Human Annotation We asked two native speakers of each language to assess the writer's sentiment towards a particular word in a given text. Annotators were asked to annotate whether a word carries either positive, neutral, or negative connotation within a given sentence. For the sentences, we used the samples from our historical corpora, and specifically chose sentences that mention one of the target countries. Participants were given the option to skip instances that were ambiguous in sentiment or not suitable for annotation (e.g., too short or not in their language). After excluding any skipped examples, we obtained 124, 42, and 97 annotated instances for Korean, Hebrew, and Russian, respectively. Tthe annotator agreement measured using Cohen's Kappa for Hebrew, Korean, and Russian was 0.48, 0.41, and 0.38, respectively.

Experiment Settings We used the RoBERTa-base checkpoint from the HuggingFace transformer repository[6] as a pretrained model for finetuning experiments. We train one classifier each to infer word connotations of loanwords and countries in historical corpora. As described in §2.3.1, we collected silver labels for 27,000 and 18,000 examples for loanwords and countries using GPT-3. We divided each dataset into training, development, and test sets in an 8:1:1 ratio, and used them to train a separate classifier. We conducted a hyperparameter search for each model to find the best model using the development data, and used the best model for evaluation.

Baselines We evaluate our trained classifier using human-annotated data by comparing it with two baselines, Majority and 165. The majority baseline predicts the most frequent class, which is neutral in our dataset, for all inputs. Park et al. [165] is a logistic regression model introduced in Park et al. [165] which uses contextualized word embeddings of a target word from a large language model to predict its word connotations. We trained the model with our classifier training data. Park et al. [165] is considered the most recent contextual word connotation classification model among existing works.

Additionally, we evaluate two GPT-3 variants, GPT-3 Curie and GPT-3 Davinci, to determine their performance on word connotation classification, which is crucial as we employ these models to construct a synthetic training dataset for our classifier. The comparison between GPT-3 Davinci and GPT-3 Curie helps us understand whether using a larger model (GPT-3 Davinci) is essential by comparing it to the performance of a relatively smaller model (GPT-3 Curie). It is important to note that our classifier was trained with data generated by GPT-3-davinci, so we do not anticipate it to outperform GPT-3-davinci. Instead, our evaluation offers insights into the extent of performance drop when transitioning from the large, computationally intensive language model to a more efficient, task-specific model for inference on the entire dataset.

Classification Evaluation Results Table 2.2 presents the performance measured by the macro F1 score for both the baselines and our proposed RoBERTa-based classifier. The evaluation

[6] https://huggingface.co/roberta-base

	All	Kor	Heb	Rus
Majority	27.4	29.4	22.2	26.5
Park et al. [165]	31.4	33.6	19.4	30.2
GPT-3 Curie	26.7	20.0	37.6	28.4
GPT-3 Davinci	62.5	63.9	61.7	59.1
Finetuned RoBERTa (ours)	58.0	60.9	53.6	53.2

Table 2.2: **Macro F1 scores of word connotation classifiers**. Our student model, trained on synthetic data generated by GPT-3 Davinci, significantly outperforms other baselines and performs closely to GPT-3 Davinci, while being significantly more efficient and cost-effective.

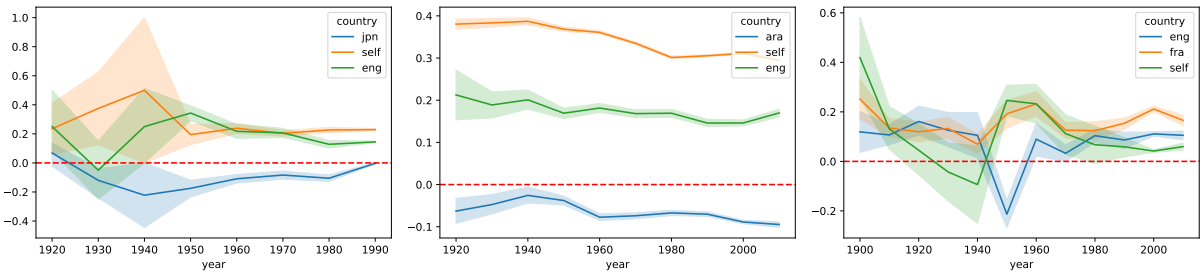


Figure 2.3: Average sentiment expressed towards the countries in Korean(left), Hebrew (middle), and Russian (right).

results reveal three key findings: First, our RoBERTa-based classifier outperforms the two baseline models significantly. Second, our model, trained with data generated by GPT-3, almost approaches the effectiveness of the original GPT-3 Davinci model. This result validates the use of our fine-tuned classifier as a cost-effective alternative to GPT-3 models for processing large corpora, thus reducing API cost from using GPT-3 without substantial compromises in performance. Lastly, the noticeable performance gap between GPT-3 Curie and GPT-3 Davinci shows the impact of model size on their ability to infer word connotations. This observation highlights the necessity of employing larger models in the data generation process.

2.3.4 Results

We applied our trained RoBERTa classifier to mentions of countries and loanwords in our historical texts to measure general public sentiments expressed towards countries and loanwords in the texts.

Country Sentiment

We applied the trained classifier to all mentions of donor countries in our historical texts. Additionally, we employed the classifier to evaluate self-mentions, such as references to Israel in the Hebrew corpus, and compared their sentiment with that of donor countries. Some of these mentions were too brief and may not have provided sufficient context for the model to accurately determine sentiment. To address this issue, we filtered out sentences with fewer than five words.

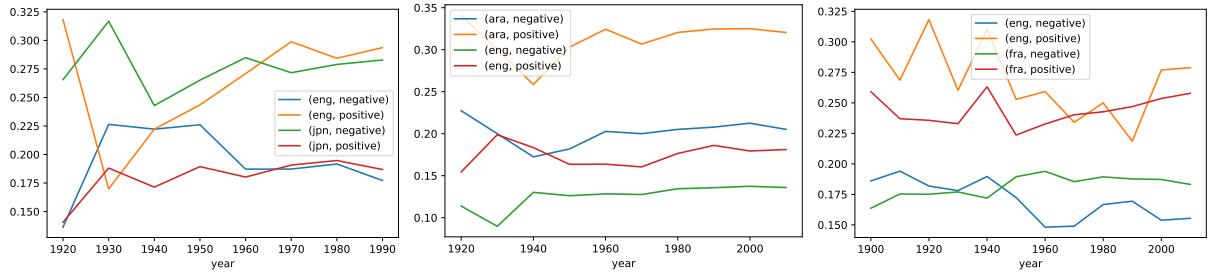


Figure 2.4: Average sentiment expressed towards the loanwords in Korean(left), Hebrew (middle), and Russian (right).

As a result, our dataset consisted of 849,430 country mentions in Hebrew (Israel: 649,111, Arab Nations: 123,028, US and UK: 77,291), 58,620 country mentions in Korean (Korea: 21,232, Japan: 20,846, US: 16,542), and 55,319 country mentions in Russian (Russia: 26,829, France: 12,975, US: 15,515).

Subsequently, we assigned numerical values to the class labels inferred by our classifier: -1 for negative, 0 for neutral, and 1 for positive sentiment. We calculated the average sentiment score for all mentions of a given country, grouped by decade, to analyze sentiment changes over time.

Figure 2.3 shows the measured sentiment toward the donor countries in each of the three recipient languages. We also depict, for each recipient language, the average sentiment toward mentions of its own country. The results qualitatively show patterns that are aligned with historical events:

- **Korean.** In 1930, the sentiment towards the US was mostly neutral, but it dramatically improved from 1940 to 1950, when the US became a powerful ally after World War II. On the other hand, the sentiment towards Japan in Korean historical records was consistently negative from 1930 to 1990, reaching its lowest point in 1940 during the harshest period of Japanese oppression during the colonial era. However, the sentiment gradually improved over time and became almost neutral by 1990.
- **Hebrew.** Compared to the other two languages, Hebrew demonstrates the most stable trend in country sentiment. The sentiment towards Arab nations remained consistently negative throughout the entire period, becoming even more negative after 1940.
- **Russian.** The most striking trend in Russian is the significant decline in sentiment towards the US in 1950, which coincides with the start of the Cold War. The sentiment towards the US improved consistently from 1960 onwards. Additionally, the sentiment towards France was always more positive than the sentiment towards the US, except for 1920 and 1940, which is in contrast to the other two languages where the US had a more positive sentiment than the other country.

Loanwords Sentiment

We followed the same procedure to keep the filtered loanword mentions and assign numerical values to the sentiment implied by the word in context. We then calculated the average senti-

ment score of each loanword across the entire timespan in our data. Based on this score, we categorized the loanwords as positive, neutral, or negative. Words that had average scores more than half a standard deviation above the mean were labeled as positive, while those that were half a standard deviation below the mean were labeled as negative. Finally, we determined the average proportion of positive and negative loanwords in all mentions at a given time period. This calculation was done using the full loanwords data, not just the filtered mentions, as it does not depend on the classifier. Figure 2.4 shows the measured loanwords sentiment of language pairs in three languages. We find that:

- **Korean.** In the Korean data, we observe that the sentiment of loanwords borrowed from English shows a clear distinction over time. Positive sentiment is relatively stable, while negative sentiment fluctuates, particularly during the 1930s and 1950s, possibly reflecting historical events such as Japanese occupation and the Korean War. The trend suggests that negative sentiment peaks correspond with periods of political tension, highlighting how socio-political circumstances can influence the sentiment of borrowed words.
- **Hebrew.** The Hebrew data presents an interesting pattern, where both positive and negative sentiments show relatively stable trends over the years. The positive sentiment towards English loanwords remains consistently higher compared to the negative sentiment, especially from the 1970s onwards. This could be attributed to the growing influence of Western culture and technology during this period, leading to a more favorable perception of English loanwords. On the other hand, Arabic loanwords exhibit a different trend, where positive sentiment towards Arab nations to be consistently high except around 1940, possibly reflecting changing socio-political relations between Hebrew speakers and Arabic-speaking communities.
- **Russian.** In the Russian data, there is a notable fluctuation in the sentiment of loanwords. English loanwords show a distinct peak in positive sentiment in the early 20th century, followed by a decline during the Soviet era, likely due to the political climate and the Cold War. However, towards the late 20th and early 21st centuries, positive sentiment towards English loanwords rises again, indicating a renewed openness to Western influence post-Soviet Union. The sentiment towards French loanwords, while more stable, generally trends upwards, suggesting a longstanding cultural appreciation for French language and culture, which has persisted despite political changes.

2.4 Analysis: Socio-Political Relations and Lexical Borrowing

In this section, we use the measured word connotation of loanwords and the sentiment towards nations to address our two original research questions concerning the relationship between socio-political dynamics and lexical borrowing. We focus on two key metrics: the level of interaction between countries and the public sentiment they share. First, we explain how we measure the interaction between countries using historical data, and then we conduct a regression analysis to examine how socio-political factors influence language borrowing, including the rate at which words are borrowed and the sentiment associated with these borrowed terms.

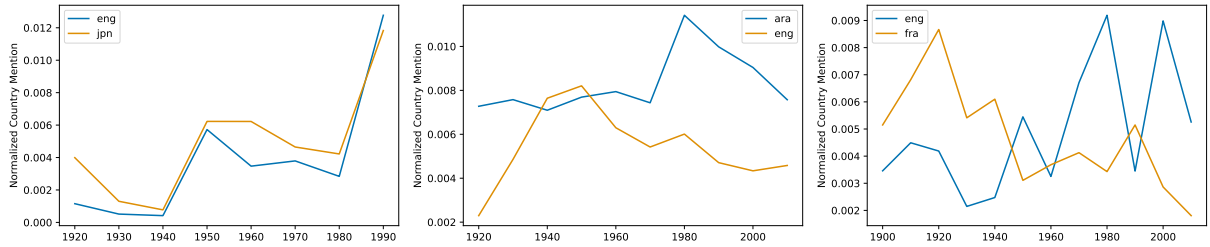


Figure 2.5: Normalized country mention rate in Korean (left), Hebrew (middle), and Russian (right).

2.4.1 Measuring Interaction Between Countries

To quantify the intensity of (geo-political as well as linguistic) interaction between two countries, we defined two frequency-based metrics: country mention rate and borrowing rate.

Country Mention Rate *Country mention rate* is the number of mentions of countries in our historical corpus, normalized by the corpus size. We assume that if a country was frequently mentioned in publicly available texts written in language L, then the country and the corresponding speakers of L had intensive interaction with the mentioned country. This approach, however, has limitations, such as the possibility that mentions of a country might be driven by specific events rather than sustained interaction, and that the context of mentions (e.g., negative or neutral) is not considered in this metric.

Figure 2.5 shows the country mention rates of respective target countries in Hebrew, Korean, and Russian. We observe distinct patterns across languages:

In Hebrew, the data shows that mentions of the US/UK were higher than mentions of Arab nations during the 1940s–1950s. However, starting in the 1960s, mentions of Arab nations began to surpass those of the US/UK, coinciding with the escalation of the Arab-Israeli conflict. This trend suggests a shift in focus towards regional geopolitical dynamics during this period, reflecting the increased importance of interactions, both conflictual and diplomatic, between Israel and its neighboring Arab countries.

In Korean texts, since Korea's liberation in 1950, Japan was more frequently mentioned than US except for 1950 and 1990. This trend might be rather surprising as one could expect the US, which had helped with the nation's liberation in 1950, might have more interaction than Japan after 1950. However, our results show that even after the colonization period, Japan and Korea had a significant amount of interaction. We note that there are significantly fewer country mentions during the 1920–1940s, which can be attributed to the fact that the Korean language was forbidden in schools and in official documents during the colonization period.

In Russian, the data shows that France had significantly more mentions than the US until the 1980s, after which mentions of the US steadily increased, eventually surpassing those of France. This shift likely reflects the changing geopolitical landscape, particularly during the Cold War, where the US became increasingly central to Soviet foreign relations and domestic discourse. The increase in US mentions coincides with the deterioration of US-Soviet relations and the eventual dissolution of the Soviet Union.

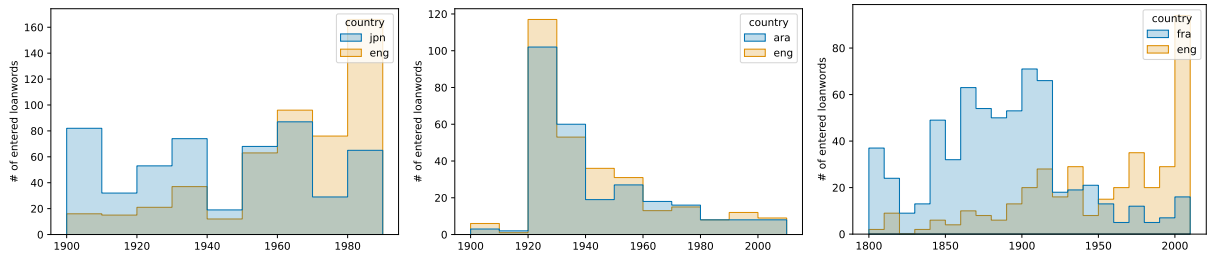


Figure 2.6: Normalized borrowing rate of Korean (left), Hebrew (middle), and Russian (right).

Overall, these results align with historical expectations: the patterns of country mentions reflect the geopolitical realities of the times, suggesting that our country mention rate metric is a valid indicator of the level of interaction between countries.

Borrowing Rate Figure 2.6 shows the borrowing rate of each donor-recipient language pair normalized by the total number of new words entered in the target language. The total number of words entered in each time frame was calculated using the control group – all words in target language's bilingual dictionaries. Thus the y-axis value denotes the portion of loanwords from each language.

The results show how borrowing rate is closely related to the relationship between nations. For example, in Korean, the crossing point of two graphs in 1950 coincides with an important historical event – the Korean war which began on 1950. This war, after the Japanese colonial period, was when the country with the biggest influence was changed from Japan to the US. Since 1940, the number of loanwords adopted from English continuously increased. For Hebrew, while borrowing from Arabic consistently constitutes a significant portion of loanwords, the borrowing rate of English is significantly higher than that of Arabic.

In contrast, in Russian, until 1990, the borrowing rate of French was always higher than that of English. However, after 1990, the end of Cold War, there are slightly more English loanwords than French loanwords.

2.4.2 Socio-political Relations and Lexical Borrowing

Regression analysis To determine the extent to which the metrics we have established for socio-political relations affect borrowing rates, we performed a regression analysis. We provide three features as inputs (x): country mention , country sentiment, and year. These inputs are standardized and were used to predict the borrowing rate. In total, there were 64 data points from three languages to fit the model.

Socio-political Relations and Borrowing Rates We investigated the relationship between socio-political relations and the frequency of language borrowing. Table 2.3 shows the coefficients for the input features of the trained model. The results indicate a significant correlation between the frequency of mention and the rate of borrowing ($p < 0.01$). This finding is consistent with prior research suggesting that higher visibility or salience of a country in dis-

	country mention	country sentiment	year
borrowing rate	**0.074***	0.008	0.015
loanwords sentiment (pos)	-0.010	**0.040**†	-0.011
loanwords sentiment (neg)	**-0.029***	0.001	0.005

Table 2.3: Regression coeffcients of trained regression models. (*: $p < 0.01$, †: $p < 0.1$.)

course—measured by mentions—can be associated with increased linguistic influence, including borrowing of lexical items.

However, it is important to note that this relationship does not necessarily imply a direct causal link. While increased interaction and discussion of a country might contribute to greater borrowing from its language, the direction of causality is not definitive. Other factors, such as the underlying political power of the donor country, could drive both increased mentions and borrowing. Thus, the observed correlation should be interpreted with caution, considering the complex socio-political factors that may influence these dynamics.

Socio-political Relations and Loanwords Sentiment We examined our second research question on the connection between socio-political relations and the semantic orientation (sentiment) of borrowed words. We used the same regression analysis method as in the previous section, the only difference being the dependent variable (y). This time, the sentiment of loanwords, measured as the proportion of positive and negative words in Section 5.2, was used as the target feature.

Table 2.3 shows that the positive sentiment of loanwords is positively correlated with country sentiment, while the negative sentiment of loanwords is negatively correlated with the country mentions. This suggests that when a country is viewed positively in the socio-political context, the words borrowed from that country's language tend to carry a positive connotation. Conversely, frequent mentions of a country, which might indicate heightened interaction or visibility, are associated with borrowing words that have a negative sentiment. This could be because increased mention may often occur in contexts of conflict or criticism, leading to the adoption of terms with negative connotations. Our results corroborate prior work that found during periods of political tension, languages often borrow words with negative connotations from the languages of opposing groups, reflecting the strained relations.

2.5 Conclusion

In this work, we proposed methods to operationalize political relationships between nations and lexical borrowing. We introduced Public Mentions and Public Portrayals which represent how much interaction and how positive portrayal they have with a donor country. For lexical borrowing, we quantified borrowing rate (i.e. volume of loanwords entering the language), use

rate (i.e. how frequently the loanwords are used), and loanwords connotation by their adopted time and by used time. Our analysis on Hebrew, Korean, and Russian showed that the proposed political relation-related metrics were well aligned with various historical events and the relationship with their major language donor countries at the time. We further investigated how socio-political relations between contact communities are related with lexical borrowing behavior. Our results show that there is a high correlation between how donor countries are portrayed in public and how positive their loanwords are, and also between how frequently donor countries are mentioned and how frequently their loanwords are used. Our work is the first attempt to quantitatively study the intersection of lexical borrowing and political factors at a large-scale using advanced NLP methods. We publicly release the trained models and codes and the entire pipeline of our method is language-agnostic, thus can be extended to any other languages with adequate resources. [163]

Chapter 3

Cultural Similarity for Cross-Lingual Transfer

Having explored the impact of cultural context on language change in Chapter 2, we now turn our attention to how cultural context can be leveraged to enhance NLP systems. One of the key challenges in NLP is improving the performance of models across different languages, particularly in low-resource settings. Cross-lingual transfer learning, where models trained on one language are adapted to another, offers a promising solution. This chapter examines the role of cultural similarity in this process, proposing that understanding the relationships between cultures can guide the selection of training datasets for cross-lingual NLP systems, especially for pragmatics-related tasks. By proposing metrics that can operationalize similarity between cultures and languages, we demonstrate how they can be applied to improve cross-lingual transfer, enhancing model performance and enabling more effective and culturally aware NLP applications. This chapter presents the findings of feature analyses and extensive experiments that validate the proposed methods, shedding light on the importance of cultural context in cross-lingual NLP.

3.1 Introduction

Hofstede et al. [100] defined culture as the collective mind which "distinguishes the members of one group of people from another." Cultural idiosyncrasies affect and shape people's beliefs and behaviors. Linguists have particularly focused on the relationship between culture and language, revealing in qualitative case studies how cultural differences are manifested as linguistic variations [198].

Quantifying cross-cultural similarities from linguistic patterns has largely been unexplored in NLP, with the exception of studies that focused on cross-cultural differences in word usage [81, 134]. In this work, we aim to quantify cross-cultural similarity, focusing on *semantic* and *pragmatic* differences across languages.[1] We devise a new distance measure between languages based on linguistic proxies of culture. We hypothesize that it can be used to select transfer languages and improve cross-lingual transfer learning, specifically in pragmatically-motivated tasks such as sentiment analysis, since expressions of subtle sentiment or emotion—such as

*The first three authors contributed equally.

subjective well-being [200], anger [162], or irony [113]—have been shown to vary significantly by culture.

We focus on three distinct aspects in the intersection of language and culture, and propose features to operationalize them. First, every language and culture rely on different levels of *context in communication*. Western European languages are generally considered low-context languages, whereas Korean and Japanese are considered high-context languages [88]. Second, similar cultures construct and construe *figurative language* similarly [29, 215]. Finally, *emotion semantics* is similar between languages that are culturally-related [105]. For example, in Persian, 'grief' and 'regret' are expressed with the same word whereas 'grief' is co-lexified with 'anxiety' in Dargwa. Therefore, Persian speakers may perceive 'grief' as more similar to 'regret,' while Dargwa speakers may associate the concept with 'anxiety.'

We validate the proposed features qualitatively, and also quantitatively by an extrinsic evaluation method. We first analyze each linguistic feature to confirm that they capture the intended cultural patterns. We find that the results corroborate the existing work in sociolinguistics and linguistic anthropology. Next, as a practical application of our features, we use them to rank transfer languages for cross-lingual transfer learning. Lin et al. [135] have shown that selecting the right set of transfer languages with syntactic and semantic language-level features can significantly boost the performance of cross-lingual models. We incorporate our features into Lin et al. [135]'s ranking model to evaluate the new cultural features' utility in selecting better transfer languages. Experimental results show that incorporating the features improves the performance for cross-lingual sentiment analysis, but not for dependency parsing. These results support our hypothesis that cultural features are more helpful when the cross-lingual task is driven by pragmatic knowledge. [2]

3.2 Cultural Similarity Features

We propose three language-level features that quantify the cultural similarities across languages.

Language Context-level Ratio A language's *context-level* reflects the extent to which the language leaves the identity of entities and predicates to context. For example, an English sentence *Did you eat lunch?* explicitly indicates the pronoun *you*, whereas the equivalent Korean sentence 점심 먹었니? (= *Did eat lunch?*) omits the pronoun. Context-level is considered one of the distinctive attributes of a language's pragmatics in linguistics and communication studies, and if two languages have similar levels of context, their speakers are more likely to be from similar cultures [155].

The language context-level ratio (LCR) feature approximates this linguistic quality. We compute the pronoun- and verb-token ratio, $\texttt{ptr}(l_k)$ and $\texttt{vtr}(l_k)$ for each language l_k, using part-

[1] In linguistics, *pragmatics* has both a broad and a narrow sense. Narrowly, the term refers to formal pragmatics. In the broad sense, which we employ in this paper, pragmatics refers to contextual factors in language use. We are particularly concerned with cross-cultural pragmatics and finding quantifiable linguistic measures that correspond to aspects of cultural context. These measures are not the cultural characteristics that would be identified by anthropological linguists themselves but are rather intended to be measurable correlates of these characteristics.

[2] Code and data are publicly available at https://github.com/hwijeen/langrank.

of-speech tagging results. We first run language-specific POS-taggers over a large mono-lingual corpus for each language. Next, we compute `ptr` as the ratio of count of pronouns in the corpus to the count of all tokens. `vtr` is obtained likewise with verb tokens. Low `ptr`, `vtr` values may indicate that a language leaves the identity of entities and predicates, respectively, to context. We then compare these values between the *target language* l_{tg} and *transfer language* l_{tf}, which leads to the following definition of `LCR`:

$$\texttt{LCR-pron}(l_{tf}, l_{tg}) = \texttt{ptr}(l_{tg})/\texttt{ptr}(l_{tf})$$
$$\texttt{LCR-verb}(l_{tf}, l_{tg}) = \texttt{vtr}(l_{tg})/\texttt{vtr}(l_{tf})$$

Literal Translation Quality Similar cultures tend to share similar figurative expressions, including idiomatic multiword expressions (MWEs) and metaphors [125, 126]. For example, *like father like son* in English can be translated word-by-word into a similar idiom *tel père tel fils* in French. However, in Japanese, a similar idiom 蛙の子は蛙 (*Kaeru no ko wa kaeru*) "A frog's child is a frog." cannot be literally translated.

Literal translation quality (`LTQ`) feature quantifies how well a given language pair's MWEs are preserved in literal (word-by-word) translation, using a bilingual dictionary. A well-curated list of MWEs is not available for the majority of languages. We thus follow an automatic extraction approach of MWEs [211]. First, a variant of pointwise mutual information, PMI[3] [43] is used to extract noisy lists of top-scoring n-grams from two large monolingual corpora from different domains, and intersecting the lists filters out domain-specific n-grams and retains the language-specific top-k MWEs. Then, a bilingual dictionary between l_{tf} and l_{tg} and a parallel corpus between the pair are used. [3] For each n-gram in l_{tg}'s MWEs, we search for its literal translations extracted using the dictionary in parallel sentences containing the n-gram. For any word in the n-gram, if there is a translation in the parallel sentence, we consider this as hit, otherwise as miss. And we calculate *hit ratio* as $\frac{hit}{(hit+miss)}$ for each n-gram found in the parallel corpus. Finally, we average the hit ratios of all n-grams and z-normalize over the transfer languages to obtain $\texttt{LTQ}(l_{tf}, l_{tg})$.

Emotion Semantics Distance Emotion semantic distance (`ESD`) measures how similarly emotions are lexicalized across languages. This is inspired by Jackson et al. [105] who used colexification patterns (i.e., when different concepts are expressed using the same lexical item) to capture the semantic similarity of languages. However, colexification patterns require human annotation, and existing annotations may not be comprehensive. We extend Jackson et al. [105]'s method by using cross-lingual word embeddings.

We define `ESD` as the average distance of emotion word vectors in transfer and target languages, after aligning word embeddings into the same space. More specifically, we use 24 emotion concepts defined in Jackson et al. [105] and use bilingual dictionaries to expand each concept into every other language (e.g., *love* and *proud* to *Liebe* and *stolz* in German). We then remove the emotion word pairs from the bilingual dictionaries, and use the remaining pairs to align word

[3] While dictionaries and parallel corpora are not available for many languages, they are easier to obtain than the task-specific annotations of MWEs.

embeddings of source into the space of target languages. We hypothesize that if words correspond to the same emotion concept in different languages (e.g., *proud* and *stolz*) have similar meaning, they should be aligned to the same point despite the lack of supervision. However, because each language possesses different emotion semantics, emotions are scattered into different positions. We thus define ESD as the average cosine distance between languages:

$$\texttt{ESD}(l_{tf}, l_{tg}) = \sum_{e \in E} \cos(\mathbf{v}_{tf,e}, \mathbf{v}_{tg,e})/|E|$$

where E is the set of emotion concepts and $\mathbf{v}_{tf,e}$ is the aligned emotion word vector of language l_{tf}.

3.3 Feature Analysis

In this section, we evaluate the proposed pragmatically-motivated features intrinsically. Throughout the analyses, we use 16 languages listed in Figure 3.4 which are later used for extrinsic evaluation (§3.5).

Implementation Details We used multilingual word tokenizers from NLTK and RDR POS Tagger [158] for most of the languages except for Arabic, Chinese, Japanese, and Korean, where we used PyArabic, Jieba, Kytea, and Mecab, respectively. For monolingual corpora, we used the news-crawl 1M corpora from Leipzig [84] for both LCR and LTQ. We used bilingual dictionaries from Choe et al. [36] and TED talks corpora [176] for both parallel corpora and an additional monolingual corpus for LTQ. We focused on bigrams and trigrams and set k, the number of extracted MWEs, to 500. We followed Lample et al. [130] to generate the supervised cross-lingual word embeddings for ESD.

LCR and Language Context-level ptr approximates how often discourse entities are indexed with pronouns rather than left conjecturable from context. Similarly, vtr estimates the rate at which predicates appear explicitly as verbs. In order to examine to which extent these features reflect context-levels, we plot languages on a two-dimensional plane where the x-axis indicates ptr and the y-axis indicates vtr in Figure 3.1.

The plot reveals a clear pattern of context-levels in different languages. Low-context languages such as German and English [88] possess the largest values of ptr. On the other extreme are located Korean and Japanese with low ptr, which are representative of high-context languages. One thing to notice is the isolated location of Turkish with a high vtr. This is morphosyntactically plausible as a lot of information is expressed by the affixation to verbs in Turkish.

LTQ and MWEs LTQ uses n-grams with high PMI scores as proxies for figurative language MWE (PMI MWEs). We evaluate the quality of selected MWEs and the resulting LTQ by comparing with human-curated list of figurative language MWE (gold MWEs) that are available in some languages. We collected gold MWEs in multiple languages from Wiktionary[4]. We dis-

[4]For example, https://en.wiktionary.org/wiki/Category:English_idioms

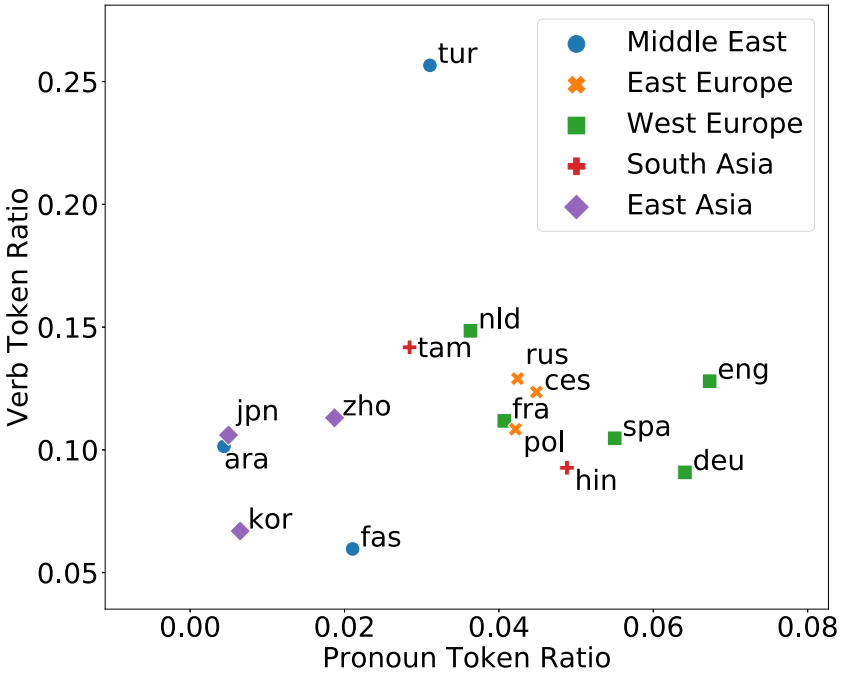


Figure 3.1: Plot of languages in `ptr` and `vtr` plane. Languages are color-coded according to the cultural areas defined in Siegel [198].

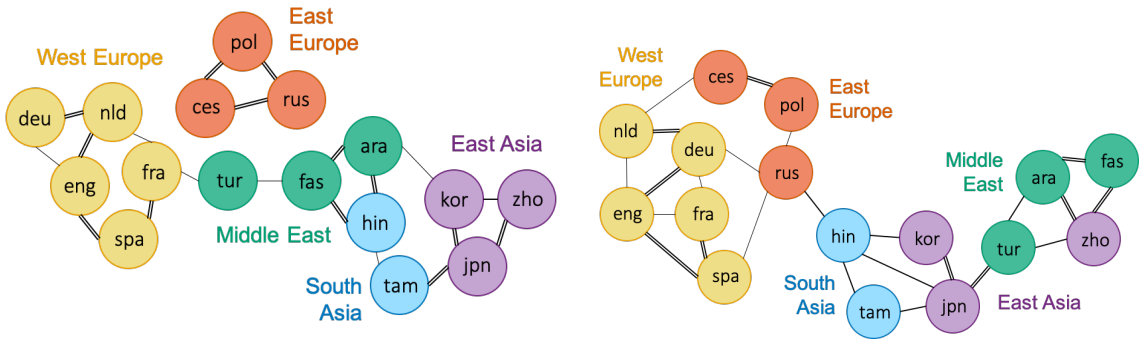
carded languages with less than 2,000 phrases on the list, resulting in four languages (English, French, German, Spanish) for analysis.

First, we check how many PMI MWEs are actually in the gold MWEs. Out of the top-500 PMI bigrams and trigrams, 19.0% of bigrams and 3.8% of trigrams are included in the gold MWE list (averaged over four languages). For example, the trigrams in the PMI MWEs, *keep an eye* and *take into account*, are considered to be in the gold MWEs as *keep an eye peeled* and *take into account* are in the list. The seemingly low percentages are reasonable, regarding that the PMI scores are designed to extract collocations patterns rather than figurative languages themselves.

Secondly, to validate using PMI MWEs as proxies, we compare the `LTQ` of PMI MWEs with the `LTQ` using gold MWEs. Specifically, we obtained the `LTQ` scores of each language pair with target languages limited to the four European languages mentioned above. Then for each target language, we measured Pearson correlation coefficient between the two `LTQ` scores based on the two MWE lists. The average coefficient was 0.92, which indicates a strong correlation between the two resulting `LTQ` scores, and thus justifies using PMI MWEs for all other languages.

ESD and Cultural Grouping We investigate what is carried by `ESD` by visualizing and looking at the nearest neighbors of emotion vectors.[5] Jackson et al. [105] used word colexification patterns to reveal that the same emotion concepts cluster with different emotions according to the language family they belong to. For instance, in Tai-Kadai languages, *hope* appears in the same

[5]A visualization demo of emotion vectors can be found at `https://bit.ly/emotion_vecs`.



(a) Network based on Emotion Semantics Distance.

(b) Network based on syntactic distance.

Figure 3.2: Network of languages color-coded by their cultural areas. An edge is added between the two languages if a language is ranked in the top-2 closest languages of the other language in terms of feature value.

cluster as *want* and *pity*, while *hope* associates with *good* and *love* in the Nakh-Daghestanian language family. Our results derived from ESD do not rely on colexification patterns, but also support this finding. The nearest neighbors of the Chinese word for *hope* was *want* and *pity*, while they were found as *love* and *joy* for *hope* in Arabic.

In Figure 3.2, we compare ESD to the syntactic distance between languages by constructing two networks of languages based on each feature. Figure 3.2a uses ESD as reference while Figure 3.2b uses the syntactic distance from the URIEL database [137]. Each node represents a language, color-coded by its cultural area. For each language, we sort the other languages according to the distance value. When a language is in the list of top-k closest languages, we draw an edge between the two. We set $k = 2$.

We see that languages in the same cultural areas tend to form more cohesive clusters in Figure 3.2a compared to Figure 3.2b. The portion of edges *within* the cultural areas is 76% for ESD while it is 59% for syntactic distance. These results indicate that ESD effectively extracts linguistic information that aligns well with the commonly shared perception of cultural areas.

Correlation with Geographical Distance Regarding the language clusters in Figure 3.2a, some may suspect that geographic distance can substitute the pragmatically-inspired features. For Chinese, Korean and Japanese are the closest languages by ESD, which can also be explained by their geographical proximity. Do our features add additional pragmatic information, or can they simply be replaced by geographical distance?

To verify this speculation, we evaluate Pearson's correlation coefficient of each pragmatic feature value with geographical distance from URIEL. The feature with the strongest correlation was ESD (r=0.4). The least correlated was LCR-verb (r=0.03), followed by LCR-pron (r=0.17) and LTQ (r=−0.31)[6]. The results suggest that the pragmatic features contain extra information that cannot be subsumed by geographic distance.

[6]When two languages are more similar, LTQ is higher whereas geographic distance is smaller.

3.4 Extrinsic Evaluation: Ranking Transfer Languages

To demonstrate the utility of our features, we apply them to a *transfer language ranking* task for cross-lingual transfer learning. We first present the overall task setting, including the datasets and models used for the two cross-lingual tasks. Next, we describe the transfer language ranking model and its evaluation metrics.

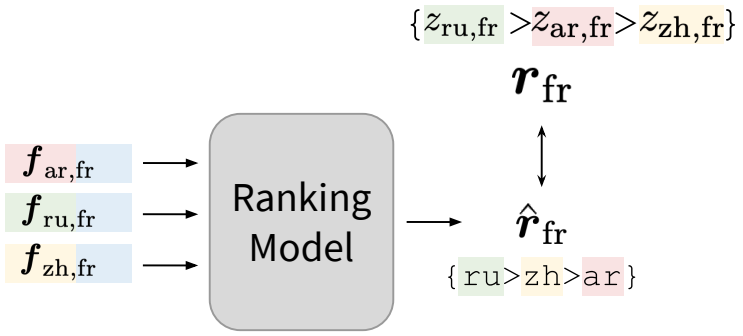


Figure 3.3: Illustration of transfer language ranking problem when the target language is French (fr) and there are three available transfer languages: Arabic (ar), Russian (ru), and Chinese (zh). The output ranking $\hat{\boldsymbol{r}}_{\text{fr}}$ is compared to the ground truth ranking $\boldsymbol{r}_{\text{fr}}$ which is determined by the zero-shot performance $\boldsymbol{z}$ of cross-lingual models.

Task Setting

We define our task as the *language ranking* problem: given the target language l_{tg}, we want to rank a set of n candidate transfer languages $\mathcal{L}_{tf}$=$\{l_{tf}^{(1)}, \ldots, l_{tf}^{(n)}\}$ by their usefulness when transferred to l_{tg}, which we refer to as *transferability* (illustrated in Figure 3.3). The effectiveness of cross-lingual transfer is often measured by evaluating the joint training or zero-shot transfer performance [193, 223]. In this work, we quantify the effectiveness as the zero-shot transfer performance, following Lin et al. [135]. Our goal is to train a model that ranks available transfer languages in $\mathcal{L}_{tf}$ by their transferability for a target language l_{tg}.

To train the ranking model, we first need to find the ground-truth transferability rankings, which operate as the model's training data. We evaluate the zero-shot performance $z_{tf,tg}$ by training a task-specific cross-lingual model solely with transfer language l_{tf} and testing on l_{tg}. After evaluating $z_{tf,tg}$ for each candidate transfer language in $\mathcal{L}_{tf}$, we obtain the optimal ranking of languages $\boldsymbol{r}_{tg}$ by sorting languages according to the measured $z_{tf,tg}$. Note that $\boldsymbol{r}_{tg}$ also depends on downstream task.

Next, we train the language ranking model. The ranking model predicts the transfer ranking of candidate languages. Each source, target pair (l_{tf}, l_{tg}) is represented as a vector of language features $\boldsymbol{f}_{tf,tg}$, which may include phonological similarity, typological similarity, word-overlap to name a few. The ranking model takes $\boldsymbol{f}_{tf,tg}$ of every l_{tf} as input, and predicts the transferability ranking $\widehat{\boldsymbol{r}}_{\boldsymbol{tg}}$. Using $\boldsymbol{r}_{tg}$ from the previous step as training data, the model learns to find optimal transfer languages based on $\boldsymbol{f}_{tf,tg}$. The trained model can either be used to select the optimal

set of transfer languages, or to decide which language to additionally annotate during the data creation process.

Task & Dataset

We apply the proposed features to train a ranking model for two distinctive tasks: multilingual sentiment analysis (SA) and multilingual dependency parsing (DEP). The tasks are chosen based on our hypothesis that high-order information such as pragmatics would assist sentiment analysis while it may be less significant for dependency parsing, where lower-order information such as syntax is relatively stressed.

SA As there is no single sentiment analysis dataset covering a wide variety of languages, we collected various review datasets from different sources. [7] All samples are labeled as either positive or negative. In case of datasets rated with a five-point Likert scale, we mapped 1–2 to negative and 4–5 to positive. We settled on a dataset consist of 16 languages categorized into five distinct cultural groups: West Europe, East Europe, East Asia, South Asia, and Middle East (Figure 3.4).

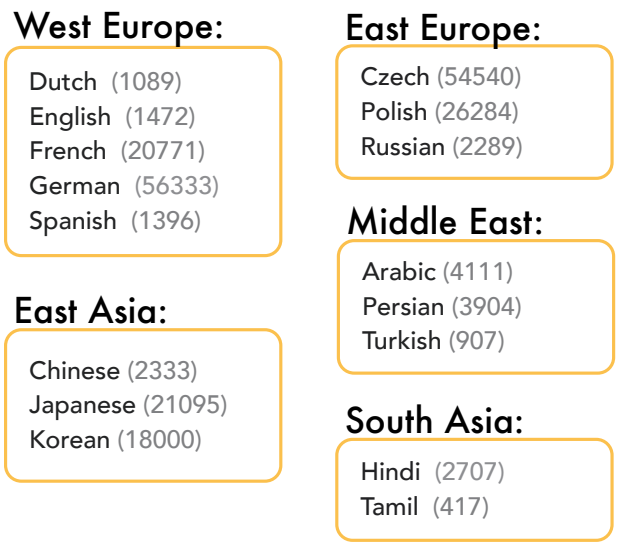


Figure 3.4: Languages used throughout the experiments are grouped by their cultural areas [198]. The numbers indicate the size of each dataset.

DEP To compare the effectiveness of the proposed features on syntax-focused tasks, we chose datasets of the same set of 16 languages from Universal Dependencies v2.2 [159].

Task-Specific Cross-Lingual Models

SA Multilingual BERT (mBERT) [51], a multilingual extension of BERT pretrained with 104 different languages, has shown strong results in various text classification tasks in cross-lingual settings [133, 204, 224]. We use mBERT to conduct zero-shot cross-lingual transfer and to extract optimal transfer language rankings: fine-tune mBERT on transfer language data and test it on target language data. The performance is measured by the macro F1 score on the test set.

DEP We adopt the setting from Ahmad et al. [2] to perform cross-lingual zero-shot transfer. We train deep biaffine attentional graph-based models [58] which achieved state-of-the-art

[7] Note that the difference in domain and label distribution of data can also affect the transferability, and a related discussion is in §3.5

performance in dependency parsing for many languages. The performance is evaluated using labeled attachment scores (LAS).

Ranking Model & Evaluation

Ranking Model For the language ranking model, we employ gradient boosted decision trees, LightGBM [115], which is one of the state-of-the-art models for ranking tasks.

Ranking Evaluation Metric We evaluate the ranking models' performance with two standard metrics for ranking tasks: Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain at position p (NDCG@p) [106]. While MAP assumes a binary concept of relevance, NDCG is a more fine-grained measure that reflects the ranking positions. The *relevant* languages for computing MAP are defined as the top-k languages in terms of zero-shot performance in the downstream task. In our experiments, we set k to 3 for MAP. Similarly, we use NDCG@3.

We train and evaluate the model using leave-one-out cross-validation: where one language is set aside as the test language while other languages are used to train the ranking model. Among the training languages, each language is posited in turn as the *target* language while others are the *transfer* languages.

3.5 Experiments

Baselines

LANGRANK LANGRANK [135] uses 13 features to train the ranking model: The dataset size in transfer language (`tf_size`), target language (`tg_size`), and the ratio between the two (`ratio_size`); Type-token-ratio (`ttr`) which measures lexical diversity and `word_overlap` for lexical similarity between a pair of languages; various distances between a language pair from the URIEL database (geographic `geo`, genetic `gen`, inventory `inv`, syntactic `syn`, phonological `phon` and featural `feat`).

MTVEC Malaviya et al. [143] proposed to learn a language representation while training a neural machine translation (NMT) system in a simliar fashion to Johnson et al. [109]. During training, a language token is prepended to the source sentence and the learned token's embedding becomes the language vector. Bjerva et al. [17] has shown that such language representations contain various types of linguistic information ranging from word order to typological information. We used the one released by Malaviya et al. [143] which has the dimension of 512.

Individual Feature Contribution

We first look into whether the proposed features are helpful in ranking transfer languages for sentiment analysis and dependency parsing (Table 3.1). We add all three features (PRAG) to the two baseline features (LANGRANK, MTVEC) and compare the performance in the two tasks.

	SA		DEP	
	MAP	NDCG	MAP	NDCG
LANGRANK	71.3	86.5	**63.0**	**82.2**
LANGRANK+PRAG	**76.0**	**90.9**	61.7	80.5
- LCR	75.0	88.3	60.3	79.6
- LTQ	72.4	89.3	63.1*	81.3*
- ESD	77.7*	92.1*	58.2	78.5
MTVEC	71.1	89.5	43.0	69.7
MTVEC+PRAG	**74.3**	**90.8**	**49.7**	**74.8**
- LCR	72.9	90.1	54.1*	76.3*
- LTQ	71.2	89.0	53.0*	78.6*
- ESD	73.1	90.7	45.3	73.9

Table 3.1: Evaluation results of our features (PRAG) added to each baseline. The higher scores are **boldfaced**. Rows in gray indicate ablation studies.
* is marked when improvements are made compared to LANGRANK+PRAG, MTVEC+PRAG respectively.

Results show that our features improve both baselines in SA, implying that the pragmatic information captured by our features is helpful for discerning the subtle differences in sentiment among languages.

In the case of DEP, including our features brings inconsistent results to performance. The features help the performance of MTVEC while they deteriorate the performance of LANGRANK. Although some performance increase was observed when applied to MTVEC, the performance of MTVEC in DEP remains extremely poor. These conflicting trends suggest that pragmatic information is not crucial to less pragmatically-driven tasks, represented as dependency parsing in our case.

The low performance of MTVEC in DEP is noticeable as MTVEC is generally believed to contain a significant amount of syntactic information, with much higher dimensionality than LANGRANK. It also suggests the limitation of using distributional representations as language features; their lack of interpretability makes it difficult to control the kinds of information used in a model.

We additionally conduct ablation studies by removing each feature from the +PRAG models to examine each feature's contribution. The SA results show that LCR and LTQ significantly contribute to overall improvements achieved by adding our features, while ESD turns out to be less helpful. Sometimes, removing ESD resulted in a better performance. In contrast, the results of DEP show that ESD consistently made a significant contribution, and LCR and LTQ were not useful. The results imply that the emotion semantics information of languages is surprisingly not useful in sentiment analysis, but more so in dependency parsing.

Group-wise Contribution

The previous experiment suggests that the same pragmatic information can be helpful to different extents depending on the downstream task. We further investigate to what extent each kind of information is useful to each task by conducting group-wise comparisons. To this end,

	SA		DEP	
	MAP	NDCG	MAP	NDCG
Pretrain-specific	39.0	55.5	-	-
Data-specific	68.0	85.4	37.2	55.0
Typology	44.9	60.7	**58.0**	**79.8**
Geography	24.9	55.0	32.3	65.1
Orthography	34.2	56.6	35.5	60.5
Pragmatic	**73.0**	**88.0**	46.5	71.8

Table 3.2: Ranking performance using each feature group as input to the ranking model.

we group the features into five categories: Pretrain-specific, Data-specific, Typology, Geography, Orthography, and Pragmatic. Pretrain-specific features cover factors that may be related to the performance of pretrained language models used in our task-specific cross-lingual models. Specifically, we used the size of the Wikipedia training corpus of each language used in training mBERT.[8] Note that we do not measure this feature group's performance on DEP as no pretrained language model was used in DEP. Data-specific features include `tf_size`, `tg_size`, and `ratio_size`. Typological features include `geo`, `syn`, `feat`, `phon`, and `inv` distances. Geography includes `geo` distance in isolation. Orthographic feature is the `word_overlap` between languages. Finally, the Pragmatic group consists of `ttr` and the three proposed features, `LCR`, `LTQ`, and `ESD`. `ttr` is included in Pragmatic as Richards [182] have suggested that it encodes a significant amount of cultural information.

Table 3.2 reports the performance of ranking models trained with the respective feature category. Interestingly, the two tasks showed significantly different results; the Pragmatic group showed the best performance in SA while the Typology group outperformed all other groups in DEP. This again confirms that the features indicating cross-lingual transferability differ depending on the target task. Although the Pretrain-specific features were more predictive than the Geography and Orthography features it was not as helpful as the Pragmatic features.

Controlling for Dataset Size

The performance of cross-lingual transfer depends not only on the cultural similarity between transfer and target languages but also on other factors, including dataset size and label distributions. Although our model already accounts for the dataset size to some extent by including `tf_size` as input, we conduct a more rigorous experiment to better understand the importance of cultural similarity in language selection. Specifically, we control the data size by down-sampling all SA data to match both the size and label distribution of the second smallest Turkish dataset.[9] We then trained two ranking models equipped with different sets of features: LANGRANK and LANGRANK+PRAG.

In terms of languages, we focus on a setting where Turkish is the target and Arabic, Japanese and Korean are the transfer languages. This is a particularly interesting set of languages because

[8] https://meta.wikimedia.org/wiki/List_of_Wikipedias

[9] The size of the smallest language (Tamil; 417 samples) was too small to train an effective model.

the source languages are similar/dissimilar to Turkish in different aspects; Korean and Japanese are typologically similar to Turkish, yet in cultural terms, Arabic is more similar to Turkish.

In this controlled setting, the ground-truth ranking reveals that the optimal transfer language among the three is Arabic, followed by Korean and Japanese. It indicates the important role of cultural resemblance in sentiment analysis which encapsulates the rich historical relationship shared between Arabic- and Turkish-speaking communities. LANGRANK+PRAG chose Arabic as the best transfer language, suggesting that the imposed cultural similarity information from the features helped the ranking model learn the cultural tie between the two languages. On the other hand, LANGRANK ranked Japanese the highest over Arabic, possibly because the provided features mainly focus on typological similarity over cultural similarity.

3.6 Related Work

Quantifying Cross-cultural Similarity A few recent work in psycholinguistics and NLP have aimed to measure cultural differences, mainly from word-level semantics. Lin et al. [134] suggested a cross-lingual word alignment method that preserves the cultural, social context of words. They derive cross-cultural similarity from the embeddings of a bilingual lexicon in the shared representation space. Thompson et al. [206] computed similarity by comparing the nearest neighborhood of words in different languages, showing that words in some domains (e.g., time, quantity) exhibit higher cross-lingual alignment than other domains (e.g., politics, food, emotions). Jackson et al. [105] represented each language as a network of emotion concepts derived from their colexification patterns and measured the similarity between networks.

Auxiliary Language Selection in Cross-lingual tasks There has been active work on leveraging multiple languages to improve cross-lingual systems [5, 156]. Adapting auxiliary language datasets to the target language task can be practiced through either language-selection or data-selection. Previous work on language-selection mostly relied on leveraging syntactic or semantic resemblance between languages (e.g. ngram overlap) to choose the best transfer languages [217, 233]. Our approach extends this line of work by leveraging cross-cultural pragmatics, an aspect that has been unexplored by prior work.

3.7 Conclusions and Future Work

In this work, we propose three pragmatically-inspired features that capture cross-cultural similarities that arise as linguistic patterns: language context-level ratio, literal translation quality, and emotion semantic distance. Our feature analyses validate these features as effective proxies for cross-cultural similarity, with practical applications demonstrated in selecting the best transfer language for cross-lingual transfer in pragmatically-driven tasks like sentiment analysis.

Looking ahead, these features open new avenues for research by providing a provisional quantitative cross-linguistic typology of pragmatics in language. This typology can serve as a stand-in for cross-cultural similarity in studies and raises intriguing questions about what other

quantitative features of language may correlate with cultural and pragmatic differences. Additionally, fine-tuning pretrained models to downstream tasks has become standard practice in NLP, but the learning dynamics of these models remain largely obscure. Our proposed features offer a valuable tool for probing models to evaluate their knowledge of cross-cultural pragmatics. Exploring how different pretraining tasks influence the learning of pragmatic knowledge presents an exciting direction for future research. Our work not only introduces innovative features for analyzing cross-cultural pragmatics but also lays the groundwork for further exploration into the intersection of language, culture, and machine learning models.

Part II

Community Values and Language

Chapter 4

Uncovering Community Values Through Social Interactions

Within a single culture, various communities may emerge, each with its own distinct values and norms. Understanding these community-specific contexts is essential for numerous NLP tasks, particularly in content moderation, where the goal is to determine whether a given comment should be moderated (i.e., removed). As prior research has shown, different online communities adhere to unique rules and norms, making community context critical in this task. However, existing NLP approaches often overlook this community-specific context, focusing instead on developing general classifiers, such as toxicity detectors, that capture a broad sense of acceptability.

In this chapter, we introduce a framework for measuring the implicit norms and values within communities, laying the groundwork for improving NLP systems by incorporating community context. We explore how advanced language models can simulate community preferences, answering questions like, "Would a community prefer a more or less formal comment?" This approach is inspired by social science literature on social norms and illustrates the potential for combining NLP with social theory to conduct large-scale, insightful analyses of community behavior.

4.1 Introduction

Social norms—the perceived, informal, and mostly unwritten rules that govern acceptable behaviors within a community—are foundational to understanding the dynamics of social interactions and shaping the community's identity [212]. Social values, in turn, are the deeper ideals and principles that a community aspires to uphold, guiding the creation and enforcement of these norms [145]. Social norms and values emerge organically through the interplay of behaviors [16] and are difficult to grasp without gaining experience of the community firsthand. This complexity poses challenges for new users to assimilate [129] and makes it difficult for automatic community moderation systems, which lack the nuanced understanding of in-community human moderators [164].

Previous studies have focused on a small subset of norms outlined by explicit rules, known as

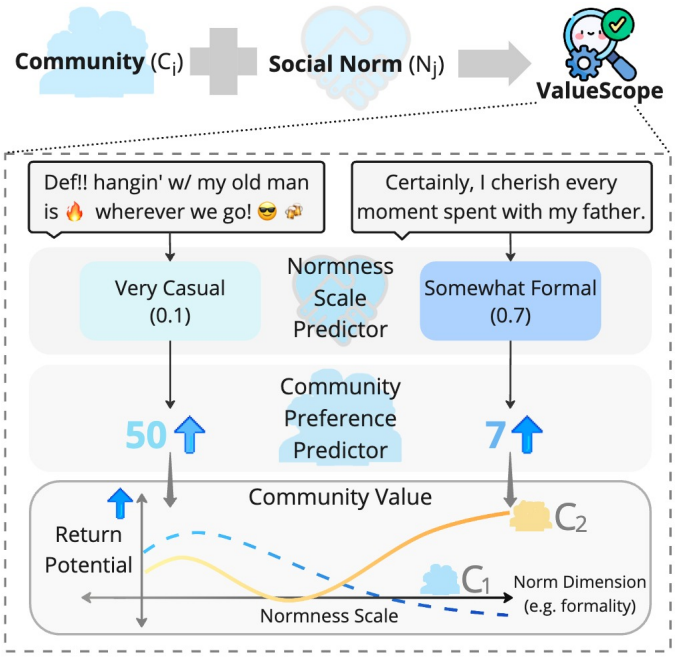


Figure 4.1: **The VALUESCOPE framework**. We characterize a comment along a norm dimension (e.g., formality), outputting the *normness scale* (e.g., a very casual comment has a formality scale of 0.1). Then, we predict the *return potential*, reflecting community preference (e.g., the number of upvotes). Finally, we plot the return potential against the normness scale using the Return Potential Model (RPM) to visualize community values.

active norms, to examine active moderation and governance [34, 73, 157, 164]. However, most social norms remain *implicit*, subtly revealed through social interactions and reinforced by the community, presenting significant challenges for computational modeling. Most current methods either rely on qualitative analysis and case studies [32, 114, 197] or analyze lexical variations, which offer limited explanatory power and generalizability [202]. Consequently, we ask (RQ1): *How can we identify and measure implicit social norms ingrained in community interactions?* We posit that social norms should not be categorical but understood on a spectrum, reflecting the diversity of human behavior and social groups [103], thereby defining the notion of *normness scale*—the degree of conformity to a norm dimension inspired by Labovitz and Hagedorn [127].

To answer RQ1, we draw inspiration from social science, particularly the **Return Potential Model** [RPM; 103], which views norms as dynamic elements shaped by interactions. We propose a theoretically-grounded computational framework—**VALUESCOPE** (Figure 4.1)—to quantify behaviors along social norm dimensions and investigate the interplay of normness scale and community preference to study the formation and evolution of *values*. This leads to our second research question (RQ2): *Can we predict the change in community norms based on observed normative behaviors?* To address this question, we extend VALUESCOPE along the temporal axis to capture the shifts in community norms. We examine whether the magnitude and variance of community preferences can help predict future changes in norms.

VALUESCOPE offers a scalable framework applicable to diverse online communities and

norm dimensions, facilitating large-scale analysis of social norm dynamics. Our contributions include:

1. We introduce VALUESCOPE —a theoretically-grounded framework based on the Return Potential Model (RPM)—to analyze social norms and values within online communities.
2. To operationalize the framework, we develop an innovative modeling pipeline consisting of a **Normness Scale Predictor** to measure the scale of social norms in text and a **Community Preference Predictor** to quantify community reactions to these variations. We also introduce novel evaluation methods to validate both individual components and the pipeline holistically.
3. Our work offers new insights into the dynamics of social norms, especially how they evolve over time. These findings have important scientific and practical implications for social scientists and community moderators, helping them identify norms that are likely to change and enabling proactive intervention.

4.2 Related Works

Social Science Literature on Social Norms A *community* represents a collective of individuals united by shared interests [222] that develop unique norms, linguistic practices, and identities, cultivating specific in-group languages and norms over time [59, 60, 61, 85]. To analyze these norms, Jackson [103] introduced the Return Potential Model (RPM), viewing social norms as dynamic processes influenced by community members' (dis)approval of behaviors [104]. While previous studies have applied RPM through qualitative methods in areas like communication and leadership [83, 95, 160, 208], our work diverges as we computationally analyze implicit norms and values in online communities at scale, focusing on the interplay between community preference and behaviors.

Norms in Online Communities Computational studies have examined linguistic norms and semantic changes in online communities [32, 45, 49, 50, 94, 118, 142, 202]. However, these often focus narrowly on language use and neologisms, neglecting the wider spectrum of community values influenced by feedback. While some research has addressed explicit governance [34, 73, 164] or qualitatively studied implicit norms [114, 197], our approach fills the gap by (1) focusing on a range of implicit norms (e.g., formality and sarcasm) automatically selected through a generalizable norm induction process, and (2) analyzing collective community preference over behaviors along the selected norm dimensions to capture a comprehensive spectrum of community values, which can provide more fine-grained and objective measurement for alignment [14, 76].

4.3 Methodology

We introduce VALUESCOPE—a theoretically-grounded framework to model social norms and values in online communities (§4.3.1). This framework is operationalized through a modeling pipeline consisting of a Normness Scale Predictor (§4.3.2) and a Community Preference Predictor (§4.3.3) to capture two interwoven dimensions of community values.

4.3.1 The VALUESCOPE Framework

Theoretical Background Community members acquire social adeptness by learning unwritten rules, or implicit norms, that govern appropriate actions under various conditions, with feedback from others to guide their behaviors [41, 230]. The Return Potential Model [103, RPM] quantifies these norms by mapping the *return potential*—expected (dis)approval—across different behaviors. Individuals in a community adjust their actions based on the learned mental model of return potential. We propose **VALUESCOPE**, a computational framework that adapts RPM to analyze the expected community preference to behaviors with varying *normness scales* (i.e., conforming to a norm dimension to different extents), offering scalable insights into community values.

Problem Definition Let $\mathcal{C}$ be communities, $\mathcal{A}$ be comments, and $\mathcal{D}$ be norm dimensions (e.g., sarcasm). For an arbitrary community $c \in \mathcal{C}$ and norm dimension $d \in \mathcal{D}$, VALUESCOPE measures the *normness scale* Φ via the Normness Scale Predictor, $\Phi_d : \mathcal{A} \to \mathbb{R}$, and the *community preference* Ψ via the Community Preference Predictor, $\Psi_c : \mathcal{A} \to \mathbb{R}$, of all N comments in c: $\mathcal{A}_c$.[1] For an arbitrary range of normness scales $\Phi_d^i := [\phi_d', \phi_d'')$ (e.g., "somewhat sarcastic"), we take the set of comments $\mathcal{A}_{c,d}^i := \{a_i | \Phi_d(a_i) \in \Phi_d^i\}$ with normness scales in the given range, and let $N_{c,d}^i := ||\mathcal{A}_{c,d}^i||$ be the number of comments in this subset. We compute the community preference of these comments:

$$\Psi_{c,d}^i := \Psi_c(\mathcal{A}_{c,d}^i) = \{\psi_1, \ldots, \psi_{N_{c,d}^i} | \psi_i = \Psi_c(a_i), a_i \in \mathcal{A}_{c,d}^i\},$$

and the estimated community preference of the given normness scale range: $\widehat{\psi_{c,d}^i} = \frac{1}{N_{c,d}^i} \sum_{j=1}^{N_{c,d}^i} \psi_j$.

Finally, we obtain $(\Phi_d^i, \widehat{\psi_{c,d}^i})$ as one point on the return potential curve[2] representing community preferences for comments of varying normness scales. For instance, we later show that `r/askscience` strongly prefers "very supportive" comments compared to its spin-off `r/shittyaskscience` (§4.5).

Differing from the social-science RPM theory, our work proposes *bidirectional continuous normness dimensions* to capture behaviors at both ends of a spectrum, such as identifying both rude and polite comments rather than just measuring politeness. This bidirectionality broadens the representational span of our analysis, empirically reduces cases where a comment is orthogonal to the norm dimension, and leads to easier generalization.

Interpreting VALUESCOPE Via VALUESCOPE, we quantitatively observe a number of features of the RPM model proposed in social science literature [103, 136, 160], enhancing our understanding of community values. Specifically, we use the **point of maximum return**—the highest point on the RPM curve—to locate the ideal normative behavior one should follow to maximize community preference, and the **potential return difference**—total positive feedback minus total negative feedback—to discover norm regulation strategies; i.e., whether the community tends to use reward or punishment to guide the formation and adaptation of its values.

[1] Here, we define Φ_d and Ψ_c to be the vanilla normness scale and community preference; to obtain the distilled scores derived in §4.3.2 and §4.3.3, we simply take the delta between two comments (a_i, a_i') to get $\nabla\Phi_d : (\mathcal{A} \times \mathcal{A}) \to \mathbb{R} = \Phi_d(a_i') - \Phi_d(a_i)$ and $\nabla\Psi_c : (\mathcal{A} \times \mathcal{A}) \to \mathbb{R} = \Psi_c(a_i') - \Psi_c(a_i)$.

[2] Alternatively, $(\nabla\Phi_d^i, \Delta\widehat{\psi_{c,d}^i})$ for the distilled RPM plot.

4.3.2 Normness Scale Predictor (NSP)

The Normness Scale Predictor (NSP) quantifies the extent to which a comment exhibits a specified social norm and is decomposed into two stages: normness measurement and normness distillation.

Normness Measurement The measurement module should map a comment to a numerical score that represents the scale of normness in the comment. We describe the challenges we tackle to construct a robust norms measurement pipeline. First, the intricacy and complexity of social norms make them extremely difficult to learn using a small regression model with limited expressive power and scarce data. Yet, it is not ideal either to use an LLM to score the comments directly; although LLMs can perform tasks with few labeled data, they are computationally expensive or rely on external APIs, posing security risks [86]. To address this, we reformulate the regression task into a binary classification task inspired by Lee and Vajjala [131]. Instead of assigning a numerical normness label to a comment, the model only learns the relative normness of comments. Then, we obtain numerical normness scales using win-rates and mathematically show that this reformulation is equivalent to a regression task given that we are only interested in relative differences in normness scales.

The second challenge is the lack of labeled data; to the best of our knowledge, there is no oracle dataset with normness scale labels. To this end, we automatically label comment pairs in terms of their *relative normness scale* using an LLM with high utility [231] to train a student model [179, 203]. To summarize, we operationalize the NSP via training a *lightweight binary classifier* using high-quality synthetic labels and evaluate both the synthetic labels and the trained classifier with human annotations.

Normness Distillation The normness distillation stage addresses two key challenges. First, unlike survey-based social science studies, our approach observes normative behaviors *post-hoc*, lacking the opportunity to explore "alternative behaviors." We attempt to recreate the "hypothetical conditions" proposed in Jackson [103], in which the individual considers alternative options to maximize return [230]. We achieve this with a **community language simulation** module, which generate comments identical to the original, except for controlled variations in one norm dimension. We then apply the normness measurement module to quantify the normness scales of the transformed comments. E.g., for an original comment, "*ty!*," we generate "*thank you*" by varying formality, and obtain formality scales of 0.2 and 0.4, respectively.

Second, the unconstrained nature of language brings a myriad of potential confounding factors biasing the predictions of the NSP, such as content variations and personal linguistic habits. By varying only one norm dimension and comparing the original and rewritten comments, the norm distillation stage aims to mitigate these confounding factors. In the above example, comparing "*ty!*" and "*thank you*" eliminates gratitude as a potential confounder for formality. We use a series of filters to ensure the quality of the generated text, including fluency and content preservation, and evaluate with annotations from in-community members.

4.3.3 Community Preference Predictor (CPP)

The Community Preference Predictor (CPP) estimates community reactions to comments, thereby serving as an indicator of prevailing community norms that govern behavior within online com-

munities. Similar to the NSP, the CPP also consists of a measurement stage and a distillation stage.

Community Preference Measurement The measurement stage of the CPP focuses on estimating community preference, which is quantified using net preference scores computed as the number of upvotes minus the number of downvotes of each comment. Unlike the NSP, which requires synthetic labeling, the CPP leverages real-world data for training. To capture the nuances of community approval, the CPP accounts for various contextual factors—post titles and time metadata—in addition to the comments as inputs, and outputs the predicted net community preference score.

Community Preference Distillation Is a comment receiving more upvotes because of its timing, its content, or because the amount of sarcasm is just right? To answer such questions, the distillation stage of the CPP aims to isolate the effects of specific norm dimensions on community reactions by calculating the difference in predicted preference between the original comment and its rewrite (which vary only in one norm dimension), and comparing it with the change in normness. Returning to the "*ty!*" and "*thank you*" example (§4.3.2), the CPP uses identical contextual information and produces community preference scores of 2 and 5; thus, a preference increase of 3 can be attributed to a formality increase of 0.2. Overall, this approach addresses confounders such as temporal dynamics and content differences, by constraining variations to a single norm dimension and comparing the preference predictions with the original comments.

4.4 Experiments

We outline our data curation process (§4.4.1) and describe experiments done to thoroughly validate the Normness Scale Predictor (§4.4.2) and the Community Preference Predictor (§4.4.3).

4.4.1 Datasets

We obtain data from the Reddit Dump via Academic Torrents, which includes posts, comments, and their metadata. Our analysis primarily focuses on first-order comments directly responding to posts from the time period 2019 to 2023.

Inductive Norm Identification Given the flexibility of VALUESCOPE, we can select any norm dimensions that describe the comments (aka behaviors) in the community. We employ an inductive norm identification process to surface the overarching norms in Reddit communities to use in our experiments as a proof of concept. First, we assume familiarity of GPT-4 with the top 5,000 subreddits [52], and instruct it to categorize them into 30 broad thematic topical groups such as finance or politics. Then, we identify the prominent norm dimensions within each category; for instance, the politics subreddits often consist of *argumentative* discussions. Consultations with subreddit experts help prioritize the six most significant norms based on their prevalence and relevance: Politeness, Supportiveness, Sarcasm, Humor, Formality, and Verbosity.

Subreddit Selection We select the subreddit topics of gender, politics, finance, and science based on their relevance and on prior work discussing their norms [62, 96, 98, 177]. For each topic, we take the most representative subreddits out of the top 5,000 SFW (safe-for-work) subreddits based on the size of the subreddit. For the gender topical group, we have `r/askmen`,

r/askwomen and r/asktransgender; for the politics topical group, we have r/republican, r/demcorats and r/libertarian. For the science topical groups, we select r/askscience, its spinoff subreddit r/shittyaskscience which was created to mock r/askscience, and a more open variant r/asksciencediscussion that discusses topics *in* science and *related* to science, such as academia [98]. Lastly, for the finance-related topics, we selected the most popular three subreddits from the top 5,000: wallstreetbets, stocks, pennystocks, and additionally consider r/wallstreetbetsnew, which is the spinoff subreddit of r/wallstreetbets. Table 4.1 summarizes the topics, subreddits, and dataset sizes examined in this study.

Topic	Subreddit	Raw Data	Synthetic Data
Gender	r/askmen	4.56M	1.08M
	r/askwomen	2.13M	1.21M
	r/asktransgender	1.61M	1.01M
Politics	r/libertarian	3.66M	1.00M
	r/democrats	534K	922K
	r/republican	502K	1.01M
Science	r/askscience	426K	1.23M
	r/shittyaskscience	185K	761K
	r/asksciencediscussion	141K	1.10M
Finance	r/stocks	3.51M	1.05M
	r/pennystocks	1.23M	1.04M
	r/wallstreetbets	49.3M	864K
	r/wallstreetbetsnew	655K	784K

Table 4.1: Selected online communities (subreddits) across various topics. For each subreddit, we show the number of existing comments within the community ("Raw Data") and the number of synthetic comments remaining after applying filters to ensure the quality of the simulated comments ("Synthetic Data").

4.4.2 Normness Scale Predictor (NSP)

Normness Measurement

Data Preprocessing Each topical group and norm dimension except for the verbosity dimension [3] has a dedicated classifier model, enabling comparisons across similar subreddits. Normness measurement relies on synthetic labels generated through stratified sampling and automatic labeling. During the sampling stage, comments are rated on a 5-point Likert scale by GPT-3.5 [27] to gauge normness. Then, 10 comments are sampled per scale point per subreddit, resulting in 150 comments per topic (200 for finance with 4 subreddits included). From these, 1,250 comment pairs are randomly selected to create binary synthetic labels using GPT-4 We train DeBERTa-base [93] with the synthetic labels for each of the 4 topic groups and 5 norm dimensions.

[3] Instead of training a verbosity scale classifier, we use the number of characters to intuitively measure verbosity. Using the character counts, we compute winrates in the range [0-1] to match the scale of other dimensions.

	Gender		Politics		Science		Finance	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
GPT-4	81.8	82.4	73.7	75.2	80.8	82.1	81.3	82.2
Val.	84.4	84.3	83.2	83.1	88.6	88.5	83.9	83.8
Test	78.1	77.9	74.3	74.2	83.0	83.0	78.1	78.1

Table 4.2: **Normness Scale Predictor evaluation** for each topic, averaged across norm dimensions: accuracy (Acc.) and F1 scores (F1). We evaluate GPT-4 generated labels against human annotations (GPT-4) and the binary classifier predictions against a held-out set with synthetic labels (Val.) and human annotations (Test).

Evaluation To evaluate the quality of GPT-4 generated training labels and the NSP models, we curate a high-quality human annotation set of 450 samples for each norm dimension, where each sample is annotated by 3 annotators with an average inter-annotator agreement, measured by Fleiss's kappa, of 0.58. Quantitative evaluation results of the GPT-4 generated synthetic labels and trained classifiers using human annotations and a held-out validation split of the synthetic labels are reported in Table 4.2, validating the quality of the NSP models.

Community Language Simulation

The norm distillation stage of NSP employs a **community language simulation** module to synthesize comments and control for norm variations.

Data Generation To simulate community language, we instruct Llama-3-8B-Instruct [209] to perform linguistic style transfer while preserving the original content and context. The model takes post titles and comment content as input and generates five variations of each comment representing different scales of normness, such as: "`Very Toxic`," "`Somewhat Toxic`," "`Neutral`," "`Somewhat Supportive`," "`Very Supportive`" for the Toxic–Supportive dimension.

Data Processing We sample 50K comments per subreddit[4] to use as the seed comments for community language simulation. To ensure the synthetic data quality, we apply preprocessing, lexical, fluency, and content preservation filters (Figure 4.2) inspired by prior works in style transfer evaluation [23, 150], removing 33% of the synthetic comments.

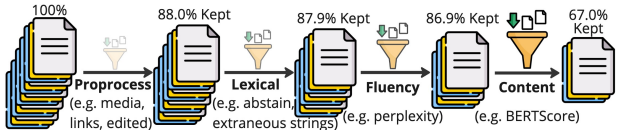


Figure 4.2: **Data filtering pipeline**, including preprocessing, lexical, fluency, and content preservation filters to ensure data quality, keeps 67% data after filtering.

[4]The data is sampled from the subset *not* used to train the community preference predictor, which ensures that the trained CPP model does not perform any inference on its training data in the community preference distillation stage.

Metric	Cont. Sim.	Fluency	Authorship	Holistic
Threshold	roughly similar	somewhat fluent	human-written	suitable
Original	86.0	94.0	81.0	91.0
Synthetic		95.9	50.0	71.3

Table 4.3: **Human evaluation results** of community language simulation. Numbers indicate the % of original/synthetic comments rated at/above the threshold.

Evaluation Three expert annotators familiar with each topical group evaluated 5 original–synthetic comment pairs per subreddit, resulting in 195 annotated samples. The annotators assessed (1) content similarity of the pair, (2) fluency, (3) authorship (LLM or human), and (4) overall quality (i.e., whether the comment is suitable to be posted in the subreddit) of each comment. Table 4.3 shows that synthetic data fluently preserves content, and is of good overall quality. Expert annotators *failed* to identify synthetic data as machine-generated 50% of the time. Moreover, postmortem interviews revealed that being "politically correct" is a strong identifier for machine-ness, and authorship is indistinguishable in science and finance topics. Overall, these results validate the quality of the filtered data.

4.4.3 Community Preference Predictor (CPP)

Data Preprocessing We take all first-level comments and their associated up-/down-vote counts. We filter out samples that are deleted, edited, created after 1 day of the post creation time, or created within 1 day of the data scraping time as they would skew the true preference.
Models CPP is fine-tuned on the DialogRPT model—a dialog response ranking GPT-2 based model trained on 133M data from Reddit [80]. Initializing CPP with DialogRPT weights enhances its understanding of general dialogue dynamics and community preferences. We train a distinct CPP model for each selected subreddit; the fine-tuning process customizes the model to better predict the preference habits of the specific community.
Baselines We investigate the effect of contextual data with 4 input format variants: **comment only**, **comment+post**, **comment+post+timestamp**, and **comment+post+timestamp+author**.
Evaluation Following Gao et al. [80], model performance is evaluated using binary accuracy: whether the relative relations between the predictions and ground truth labels of comment pairs align. We found that including contextual information such as the post title and time of the post significantly improves the accuracy, while adding author information only helps in certain subreddits such as `r/libertarian`. The most performant setup, **comment+post+timestamp**, achieved an accuracy of 73.9% ($\pm$4.1), suggesting reliable prediction performance.

4.5 Results

Using the validated NSP and CPP, we explore prevailing norms and values of online communities by modeling return potentials and analyzing the *point of maximum return* (PMR) and *potential return difference* (PRD) to corroborate our findings with existing work on similar communities

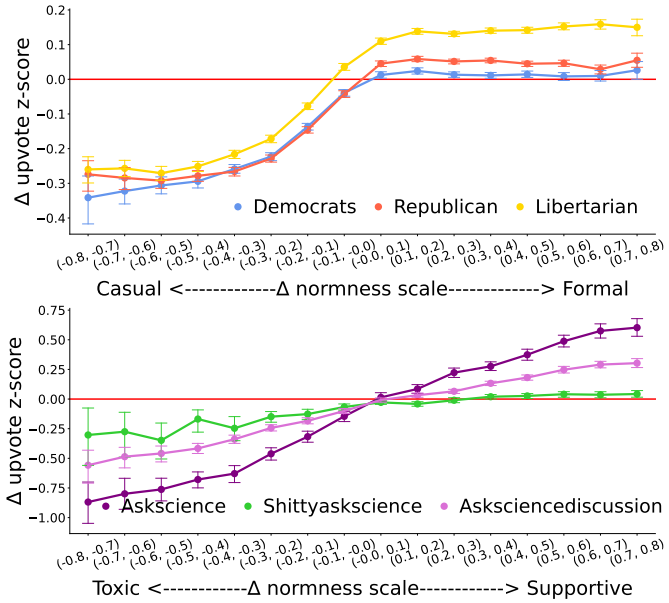


Figure 4.3: Estimated return potential over normness scales. Formality preferences in politics subreddits (top) and supportiveness preferences in science subreddits both corroborate prior findings about the communities.

and then uncover additional insights at scale.

Return Potential Modeling (RPM) Our RPM results demonstrate how a community's preferences varies with the scale of normness. We highlight two key RPM plots—formality preferences in politics subreddits and supportiveness preferences in science subreddits—to validate ValueScope in Figure 4.3.

In the politics subreddits, community preference for formal to neutral comments is nearly invariant, but as comments become progressively more casual, there is a steep decrease in preference across all subreddits. These patterns align with community rules that encourage more formal interactions (e.g., "quality control" and "no disinformation") and denounce casual behaviors (e.g., "no trolling" and "no spamming"). Higher preferences toward formal comments in `r/libertarian` is consistent with its strict guidelines encouraging detailed explanations and references to policies.

The RPM results of science subreddits show a general disapproval for toxic behaviors that gradually changes to approval as the comments become supportive. `r/askscience` and `r/askscicencediscussion`—subreddits designated for scientific discussion with guidelines discouraging offensive language and encouraging helpful answers—show a stronger preference for supportive comments than `r/shittyaskscience`, which is a parody created to mock `r/askscience` [98]. Overall, ValueScope effectively surfaces community norms shaped by guidelines and core premises.

What Are the Ideal Norm Behaviors? The point of maximum return (PMR) signifies the behaviors most favored by each community. Figure 4.4 illustrates the PMR for the top 5 subreddits

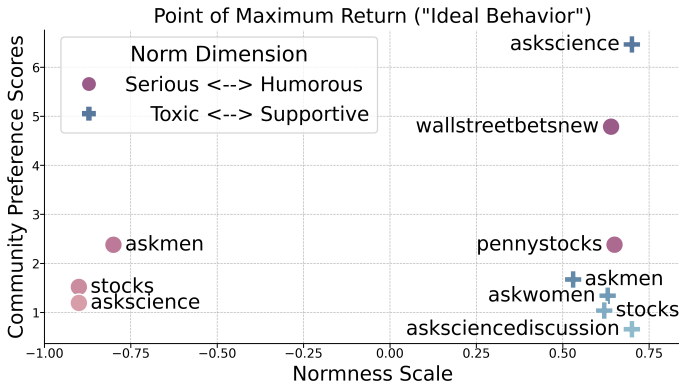


Figure 4.4: **PMR of the top five subreddits for Serious–Humorous and Toxic–Supportive**. The point of maximum return on an RPM curve describes the "ideal" behavior that would maximize community preference. For instance, these results show that `r/askscience` strongly prefers supportive comments.

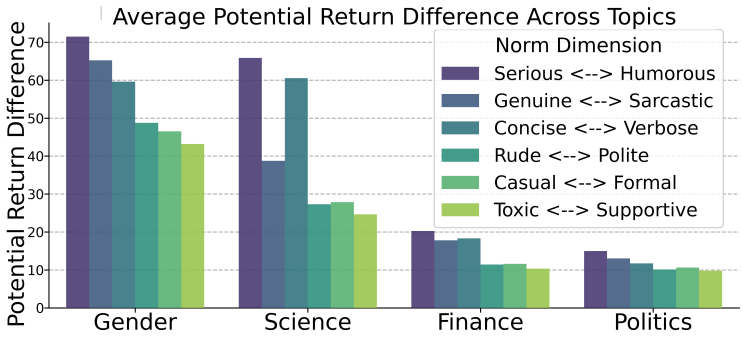


Figure 4.5: **PRD across topical groups**, reflecting the feedback strategy used by the community to regulate certain norms. All studied communities tend to use positive feedback: the gender related subreddits extensively reward behaviors aligned with their values, while the politics subreddits reward much more conservatively.

across humor and supportiveness dimensions. For instance, `r/askscience` prefers supportive comments, as discussed above, and serious comments, which is in line with its explicit community rules (e.g., "memes or jokes are not allowed") and implicit rules identified in prior work; e.g., "no personal anecdotes" [34]. Additionally, all subreddits show a preference for supportiveness over toxicity to varying degrees, which aligns with Redditquette, which are informal values held by most redditors [73].

Inferring Norm Regulation Strategies Potential return differences (PRD) in Figure 4.5 reveal how much communities emphasize rewards (PRD$>$0) or punishments (PRD$<$0) to enforce norms. All communities significantly favor positive reinforcement, indicating a generally supportive atmosphere [103], echoing calls for positivity in Redditquette [73]. Moreover, punitive measures are ineffective in maintaining prosocial communities [48, 154, 197].

Feedback intensity distinctly varies across topics. Gender-related subreddits extensively reward behaviors aligned with their values, suggesting a strong preference for promoting norms

	NI-only		NI+CR		
	$\mathbf{c}_{NI}$	$\mathbf{R}^2$	$\mathbf{c}_{NI}$	$\mathbf{c}_{CR}$	$\mathbf{R}^2$
Politeness	0.26	0.17	0.16	-0.14	0.23
Supportiveness	0.16	0.04	0.05	-0.13	0.10
Sarcasm	0.42	0.13	0.45	-0.13	0.14
Humor	0.50	0.27	0.50	-0.13	0.28
Formality	0.40	0.17	0.27	-0.07	0.18
Verbosity	2.57	0.09	2.57	-0.35	0.09

Table 4.4: **Coefficients of NI and CR, and R^2** of two linear regression models (NI-only and NI+CR).

that enhance inclusivity and respect. In contrast, politics subreddits are more conservative with rewards, possibly due to explicit rules against "disproportionate upvoting," "brigading," and "upvote spam," which aim to prevent bias. These regulations may contribute to more measured rewards. Lastly, PRD variations across norm dimensions reveal which normative behaviors are most regulated. The serious–humorous, genuine–sarcastic and concise–verbose dimensions witness the most intense regulation in all topical groups, suggesting the importance of tone and authenticity of interactions in cultivating social identity [24].

Findings in this section validate VALUESCOPE and, more importantly, allude to the impact of moderation on social norms and potential applications of VALUESCOPE: if undesirable behaviors are detected to rise, moderation strategies should be updated to maintain healthy community norms.

4.6 Analysis

To address RQ2—*Can we predict the change in norms based on observed normative behaviors?*—we study the fluidity and stability of social norms and its implications using VALUESCOPE and social science theories, specifically norm intensity and crystallization [103, 160], then analyze their temporal changes in the context of external events and internal community conflicts.

Norm Crystallization Social norms are constantly evolving. Understanding such changes and their predictive features can help community moderators respond effectively. Jackson [103] introduces the concepts of *norm intensity* (NI) and *crystallization* (CR). NI measures the magnitude of community (dis)approval of behaviors at a given normness scale, indicating how strongly the community cares about the norm, while CR represents the level of consensus on the preference.

Taking the year 2021 as a cutoff, we test the predictive power of NI and CR on upcoming temporal changes ($TC := \Delta NI$) with a linear regression model. We use results from VALUESCOPE predictions and follow implementation defined in Linnan et al. [136]. Our results in Table 4.4 show that NI and $NI + CR$ are both significant predictors of TC, while adding CR increases the coefficient of determination R^2 significantly. Additionally, higher norm intensity

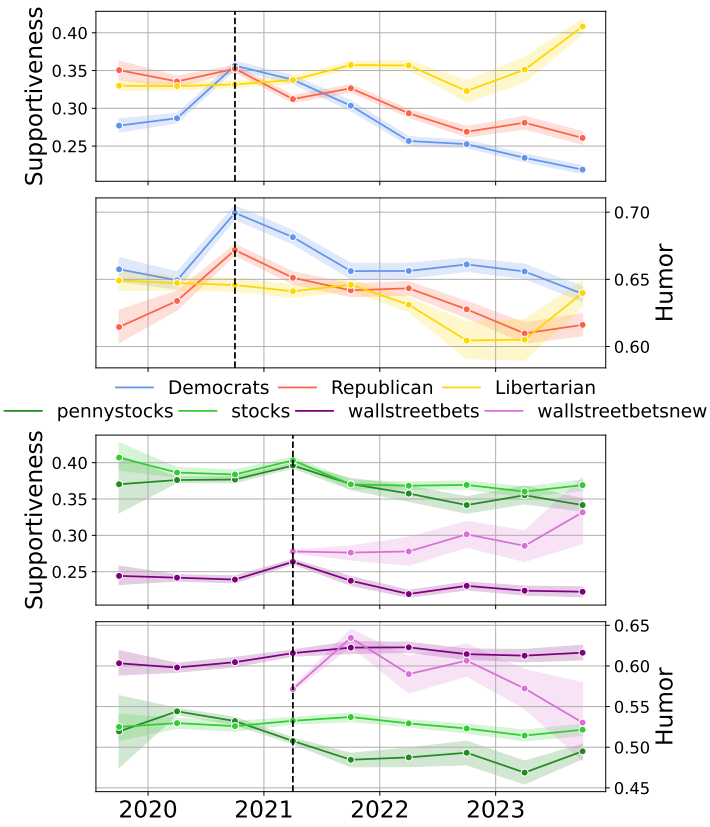


Figure 4.6: **Temporal changes in average norm intensity** for politics and finance subreddits. Comments were binned by 6 month intervals based on their posting date. For instance, a point for 2020.25 represents the average norm intensity of comments posted from January to June 2020. The vertical lines mark two events: the U.S. presidential election and the creation of `r/wallstreetbetsnew`, highlighting changes before and after these events.

and less crystallization (i.e., community members have strong opinions but less agreement) are correlated with larger shifts in norm intensity. Our findings support Jackson [104]'s hypothesis that these volatile instances are more likely to generate conflicts and trigger changes in norms. This demonstrates VALUESCOPE's potential to help moderators identify and proactively address norms likely to change by setting explicit community rules.

Temporal Change in Norm Intensity We further investigate how NI changes over time, particularly in relation to external events. Figure 4.6 shows NI of the humor and supportiveness dimensions from 2019-2023 in politics and finance subreddits.

For politics, a significant event during this period is the 2020 U.S. presidential election, represented by the vertical line in the plot (corresponding to July-December 2020). Our results reveal highly similar patterns of norm shifts in `r/republican` and `r/democrats`, characterized by a steep increase of community preference of humor and supportiveness during the election period. Following this peak, both dimensions experienced a continuous decline until 2023. On the other hand, `r/libertarian` bears a notable increase in supportiveness over time and was not

community shift / norm dimension	politeness	supportiveness	sarcasm	humor	formality
r/wallstreetbets → r/wallstreetbetsnew (925.6)	-0.003	0.013	0.003	0.005	0.018
r/wallstreetbets → r/stocks (2157.6)	0.084	0.092	-0.044	-0.062	0.131
r/wallstreetbets → r/pennystocks (1052.0)	0.091	0.094	-0.023	-0.084	0.063
r/askwomen → r/askmen (717.4)	-0.015	-0.022	0.026	0.036	0.004
r/republican → r/democrats (223.8)	0.026	0.016	0.036	0.018	-0.008

Table 4.5: User behavior shifts in select subreddit transition pairs. Gray cells indicate changes that are insignificant ($p > 0.05$); red and green cells represent significant negative and positive changes.

impacted as much by the election. These results suggest that external events, such as elections, could potentially shape the overall norms in online communities.

For finance subreddits, a notable event was the creation of r/wallstreetbetsnew—a spinoff from r/wallstreetbets—in 2021 by members dissatisfied with the culture of r/wallstreetbets in an attempt to create a less toxic environment focused on serious trading strategies on risky stocks.[5] Among the finance subreddits, our results show that the NI of r/wallstreetbetsnew starts diverging from r/wallstreetbets and begins to resemble the NI of r/stocks and r/pennystocks, becoming more supportive and less humorous over time. This finding aligns with Zhang et al. [228] in showing that new communities establish their own identities and norms over time. Additionally, after the creation of r/wallstreetbetsnew, the NI of r/wallstreetbets also shifts, becoming less supportive and more humorous. This suggests that the culture of the original community may be influenced when some members leave to form a new spinoff community as explored below.

Community Norm Adaptation by Users Social norms can influence the behavior of community members [146], so we examine how individual users modify their language and interaction styles based on the subreddit they are participating in. We define user-level norm behavior in a community as the average NI of comments left by the specific user in that community. For related subreddits with shared users, we compute the change in normative behavior of these users when they switch from subreddit A to subreddit B using a paired two-tailed t-test (Table 4.5).

Our results reveal significant variability in user normative behaviors between the selected subreddit pairs. For example, users in r/wallstreetbets, known for its usage of profane jargon and aggressive trading strategies [96], significantly modify their behaviors in r/stocks and r/pennystocks, but adapt much less in the spinoff subreddit r/wallstreetbetsnew. Additionally, user behaviors tend to remain consistent in identity-related subreddits (e.g., r/askwomen, r/askmen) or those with competing relationships (r/republican, r/democrats). These findings highlight the context-specific nature of community norm adaptations by users. We also observe that users are more likely to change their formality to fit different subreddit contexts than other dimensions, such as humor, indicating that certain norms are more malleable and adaptable than others.

[5] As one user noted: "The moderators in the original r/wallstreetbets are driving the narrative away from $GME and $AMC and the vibe is very negative/toxic over there" (paraphrased from a subreddit post in r/wallstreetbetsnew).

Different extents to which users adapt their language to the audience suggest that digital identities are fluid and context-dependent. This can inform the development of tailored moderation tools to align with the behavioral norms of specific communities, potentially improving user experience and engagement on a more fine-grained level.

4.7 Conclusion and Future Work

We introduced VALUESCOPE, a novel framework based on the RPM theory from social science, to quantify social norms and values at scale. We comprehensively validated the effectiveness of VALUESCOPE to assess the normness of behaviors and predict community preferences while controlling for confounders. VALUESCOPE enables numerous quantitative analyses, including predicting norm shifts and contextualizing temporal changes with external events, providing a deeper understanding of social norm dynamics in online communities.

Our work contributes a robust and generalizable method that can be easily extended to various norms and communities. It opens up many exciting possibilities for applications and future research:

Computational Modeling Applications Our framework can enhance community moderation tools by integrating theoretically grounded insights, such as maximum return potential, to refine toxicity detectors. It can also guide generation models to produce contextually appropriate responses specialized to each community's unique norms.

Applications for Social Scientists Our method empowers the development of new hypotheses about social norms, by providing social scientists with enhanced tools to explore how norms form and influence social interactions within communities.

Support Tools for Communities VALUESCOPE can enhance community management by enabling moderators to monitor and address norm shifts in real-time. It can help transform widely accepted but informal norms into explicit rules, clarifying guidelines and easing new member integration. This approach is applicable in various settings (e.g., workplaces) where it can guide individuals on appropriate cultural expressions, improving their integration and acceptance. Platform developers can use this method to refine community recommendation engines, aligning users with groups that match their preferences and values, thereby enhancing user engagement and community growth.

Ethical Considerations

We use publicly accessible LLMs to conduct this research, which includes generating more toxic versions of comments. In our investigation to understand the implicit norms of online communities, our experiments inevitably produced toxic content to measure how communities react to toxicity. However, we believe the benefits of our research outweigh the risks, as community moderators and platform developers can use our framework to understand the implicit norms in various communities, especially in response to toxic content, and self-assess and monitor their culture. The generated toxic content was only used to compute aggregated metrics to identify high-level patterns, and it will not be released to the public. To ensure reproducibility while

protecting the rights of Reddit users, we will only release the IDs of the comments used in our analysis. Using these provided IDs, practitioners will need to independently fetch the comments from the publicly accessible Reddit Dump.

Chapter 5

Community Norms and Community Moderation

Building on the previous chapter's exploration of implicit community norms and values, we now shift our focus to the explicit rules that govern community moderation. While reward systems, such as likes and dislikes, reflect implicit community norms, explicit community rules define what behaviors are acceptable or unacceptable within a community. Community moderation is typically guided by these rules, which vary from one community to another. In this chapter, we propose several models that incorporate community context and rules into the content moderation process, improving the ability to detect and explain norm violations. To enable these models, we introduce a new dataset, NormVio, which includes instances of community moderation along with contextual information about the communities and their rules. Our analysis of the model results provides further insights into the distinct ways communities operate and apply different rules in moderation, highlighting the need for community-aware approaches in NLP.

5.1 Introduction

Online communities establish their own norms of what is acceptable behavior [46, 108, 178]. These norms run the gamut from *no hate speech* or *no personal attacks* to more idiosyncratic expectations of *content formatting* and *content sharing* [34, 74]. Community moderators are responsible for identifying and removing rule-breaking content, regardless of whether users violate rules intentionally or unintentionally due to unfamiliarity with community norms.

Moderators of online communities often face a tough challenge of triaging the massive flow of content [56, 120, 121]; for example, over 2 billion comments were posted to Reddit in just 2020.[1] Moderators have looked to technology to help support their role, using regex-based tools like Automoderator to flag potentially rule-breaking comments [107]. Prior work has aimed to assist by developing machine learning techniques to recognize unacceptable content—yet these have focused on only the most socially-harmful violations, such as hate speech. Furthermore, the rules moderators enforce vary widely both in their formulation and interpretation across communities, making a one-size-fits-all approach increasingly brittle. Since successful moderation

[1] https://backlinko.com/reddit-users#reddit-statistics

relies on fine-grained understanding of a given community's norms, we present a new dataset and models for community-specific, contextualized norm violation detection for over twenty types of norms.

We introduce a new approach to context-sensitive automated content moderation that explicitly encodes community norms. Using a new dataset of 51K conversations across 3.2K communities, we show that the most commonly-studied norm violation behavior in NLP, hate speech, corresponds to a small minority of cases in which moderators intervene in practice. We then create multiple models to detect when moderators intervene and *why* they intervene, adapting to the norms and rules of a community.

Our paper offers the following four contributions towards advancing the future of NLP in community and context-specific moderation. First, in a large scale analysis of rule and moderation behavior, we show that subreddits vary considerably in their rules, with only some common themes. However, in practice, most rules are not enforced and, further, the enforcement of some types of rules, e.g., *incivility*, is highly varied across communities. Second, we introduce a new dataset, NORMVIO, of 51K conversations across 3.2K subreddits and map the 25K rules from these communities into nine categories of context-specific unacceptable behavior, including five types of incivility. Third, we introduce a new series of models aimed at detecting and explaining rule-violating behavior based on norms and rules of the community. Our approach enables not only identifying that conversation in a particular community (with particular rules) is likely to violate a rule, but also *which rule*. We demonstrate the effectiveness of these models, showing our best model attains an F1 of 78.64 across all rule types, a 50% improvement over context-insensitive baselines. Finally, we perform an in-depth analysis of how much conversation context and community-sensitivity affects predictability. Our work points towards key challenges in detecting particular rule violations, while providing high accuracy in others, which can allow moderators to quickly intervene.

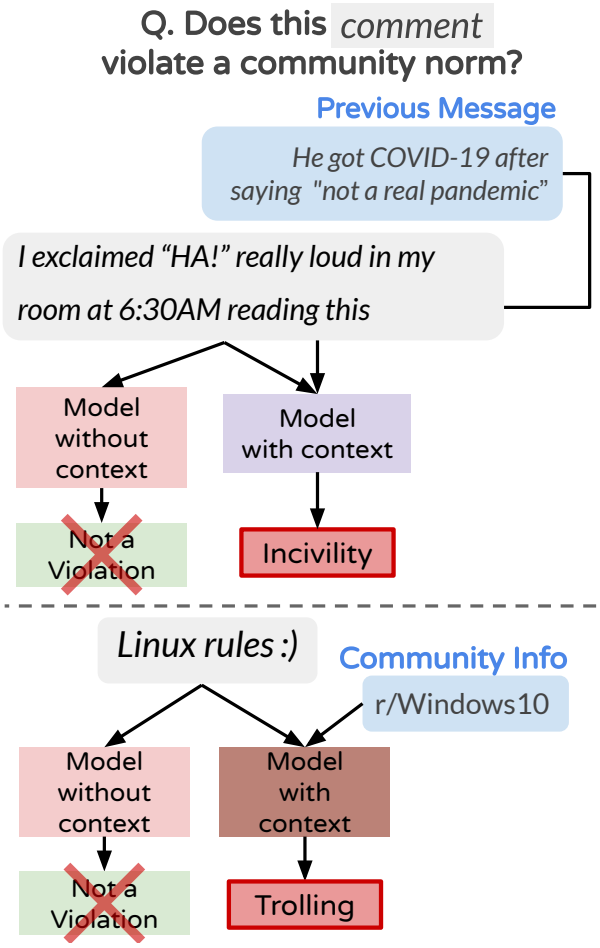


Figure 5.1: Two example comments that were moderated due to violating community norms. The examples highlight the importance of contexts (i.e. conversation history and community information) in detecting community norm violation.

[1] All example comments used in this paper are lightly paraphrased to preserve privacy.

More generally, our work provides a clear next step for NLP to look beyond one-size-fits-all methods for detecting incivility to developing holistic, context-sensitive approaches that better suit the needs of moderators and their communities.

5.2 NORMVIO Dataset

Prior work has created datasets used to detect single types of norm violations in social media messages (e.g. incivility, hate speech or hostility) [79, 219]. However, these datasets typically focus on isolated texts and do not provide prior conversational context or community-specific details.

In order to detect representative types of norm violations and account for context, we construct a new dataset—NORMVIO—a collection of 52K English conversation threads on Reddit. NORMVIOincludes comments removed for violating a variety of community norms beyond the traditional hate speech and incivility, such as spamming or violating community format/topics. Furthermore, NORMVIOprovides additional context beyond the norm-violating comment itself with (a) the entire conversation thread (i.e., the original post and prior comments) and (b) the subreddit (i.e., community) in which the comment was posted.

Data Collection We collected our initial data via the Reddit API, which provides list of moderators and their comments for each subreddit. For each of the top 100K most popular subreddits,[2] we identified the most recent 500 comments from each moderator and retrieved comments that moderators posted in response to a removed comment (henceforth, *moderation comments*).

Moderation comments often provide useful signals for inferring which community norm was violated. From the full set of moderation comments, we selected those that contain a phrase explicitly stating the rule number (e.g. "this comment violates Rule 2") or the exact text of one of its subreddit's rules (e.g. "don't be rude").

We then fetch the entire conversation thread for this set of moderation comments: the original post and all parent comments prior to the moderator's comment. We also fetched the norm-violating comment that was removed by moderators, by searching archived comments via the Pushshift API [11].[3]

The final dataset is comprised of 20K conversations that have the last comment removed by one of the moderators of the community. Following the approach in Chang and Danescu-Niculescu-Mizil [35], we include 32K paired unmoderated conversations as a control set. Each moderated conversation is matched with up to two unmoderated conversations from the same post and with most similar conversation lengths as the target moderated conversation.

Ethical Considerations for Protecting User Privacy Our dataset focuses, in part, on comments that moderators have viewed as objectionable and therefore removed. While these moderated comments are still publicly available, their use requires additional ethical reflection and

[2]Ranked by number of subscribers as of April 2021

[3]We were unable to retrieve an additional 21K removed norm-violating comments, which were unavailable in the PushShift archive. We still include these corresponding conversations in our data release as they can be useful in the task of forecasting future norm violations.

precautions to preserve the dignity and privacy of users [210]. Moderated comments offer significant benefit to the study of supporting moderators and authorities in their goals of having supportive technologies that match their community's norms. At the same time, users who made those comments may object to having them included in a dataset [75]. Therefore, we take additional measures to ensure that user privacy is protected, especially for the deleted comments.

We use Reddit data through Pushshift [11], an archive that has been widely used in NLP and related fields since its first release in 2015 [54, 97, 117, 188, *among many others*]. Pushshift's collection policy explicitly states that it conforms to Reddit's rules and user agreement with regards to data collection. In releasing our dataset, we provide only the associated identifiers of comments but *not* their textual content. Practitioners will need to independently fetch the texts from Pushshift by using the provided comment IDs. Releasing only IDs ensures that any users who request their data to be removed in Pushshift will also have it removed in our dataset. Additionally, in our dataset we anonymize individual usernames and personal identifiers of posters and moderators. Finally, along with our data release, we provide guidelines to the users who wish to delete their comments from the Pushshift dump.

Classification of Community Norms Moderator comments as well as rules defined in each subreddit are free-form and diverse, and it is not trivial to map the rule/comment to a specific community norm it refers to. In order to study norm violations, we thus first train classifiers that given a rule description label it with a type of norm it violates.

We follow Fiesler et al. [74]'s qualitative analysis of 1K subreddits, that identified main categories of rules through annotating 3,789 rules from the subreddits.[4] We then use the annotations from [74] to fine-tune a BERT-based binary classifier for each rule type.[5] Table 5.1 shows the list of the resulting 21 categories of community norms and the performance of our classifiers evaluated using macro F1 scores with stratified 10-fold cross validation.

Rule Types	**F1**	**Rule Types**	**F1**
Advertising	71.0	NSFW	88.2
Moderation Enforcement	87.0	Off-topic	63.5
Copyright/Piracy	70.6	Personal Army	43.2
Doxxing	75.4	Personality	81.9
Format	73.5	Politics	85.7
Harassment	67.9	Reddiquette	83.2
Hate Speech	84.2	Reposting	81.4
Images	65.1	Spam	86.9
Outside Content	68.0	Spoilers	76.7
Low-Quality Content	45.6	Trolling	96.0
		Voting	85.6
AVERAGE	75.3		

Table 5.1: Macro F1 of classifying the diverse sets of rules across subreddits to rule violation types.

We use the final models to map 183K rules from the top 100K subreddits to their corresponding rule types. Table 5.2 shows the examples of labeled community rules randomly sampled from our data. Finally, we classify mod-

[4]Out of 24 categories, we exclude the ones describing the tone of rules (whether a rule is "Prescriptive" or "Restrictive") and one (Behavior/Content) that is extremely broad, covering over 90% of coded rules.

[5]Binary classifiers were used since each community rule can be categorized with multiple types. We used the default hyperparameters suggested in the Transformers library and trained each model for 20 epochs.

erators' explanations of the rule-violating comments in NORMVIO. Because we only kept moderators' comments that mention a rule number or a rule's exact text, we can determine which rule was violated by the conversation.[6] Using our binary classifiers on rule text, we can now infer the type of norm that was violated by the moderated (removed) comment.

Although the 21 types are well suited for fine-grained analysis of rules on Reddit, they might leave insufficient number of examples per type which can make it more challenging to computationally model them. We define relatively more coarse-grained nine types and map the 21 types with the nine types as shown in Table 5.2. We designed these types to reflect our interest in text-based analysis of abusive language. We kept five different subcategories of uncivil comments (general incivility, trolling, harassment, hate speech, spam) while aggregating Voting, Reddiquette, and Moderation Enforcement into a broad "Meta-rules" category. In the remainder of this paper, we only use the coarse-level norm violation types.

Incivility: {Personality}	*"Be civil"*
Harassment: {Harassment, Doxxing}	*"Don't harass others"*
Spam:{Spam, Reposting, Copyright}	*"No excessive posting"*
Format: {Format, Images, Links}	*"Use the correct tags"*
Content:{Low-quality Content, NSFW, Spoilers}	*"No low-quality posts"*
Off-topic :{Off-topic, Politics}	*"Only relevant posts"*
Hate speech:{Hatespeech}	*"No racism, sexism"*
Trolling:{Trolling, Personal Army}	*"No trolls or bots"*
Meta-rules:{Voting, Moderation Enforcement, Reddiquette}	*"No Downvoting"*

Table 5.2: The mapping between coarse- and fine-grained rule types and examples.

Ultimately, each moderated comment in NORMVIOhas the following information: (1) its subreddit, (2) its conversation thread, (3) the community-specific rule violated, and (4) the coarse- and fine-grained rule types that were violated. To maximize user privacy, all comments are provided as IDs, the content for which can be retrieved through the Reddit and PushShift APIs.

Analysis of Community Norm Violations We analyze the types of rules and comments comprising NORMVIOwith a focus on what kinds of rules are established by communities, what kinds of rules are violated in practice, and when in conversations these rules are violated.

The results in Figure 5.2 show that the rule types are evenly distributed over rules (left) while the actual violations (right) are relatively more focused on abusive language rule types such as Incivility and Harassment. A large proportion of all rules in our dataset fall under the Format and Content categories, suggesting that there is a diverse set of community norms, beyond regulating incivility, needed to operate healthy online communities. Critically, while the majority of efforts

[6]Any data collection procedure that relies on user-generated labels has the risk to absorb human biases. In our setting too, there is a risk of moderator biases to be incorporated when we match moderation comments to rules and violation types. However, in pilot work examining moderator comments with explicit rule violations and those where we had to infer the rule(s), we found a near-identical distribution of violation types.

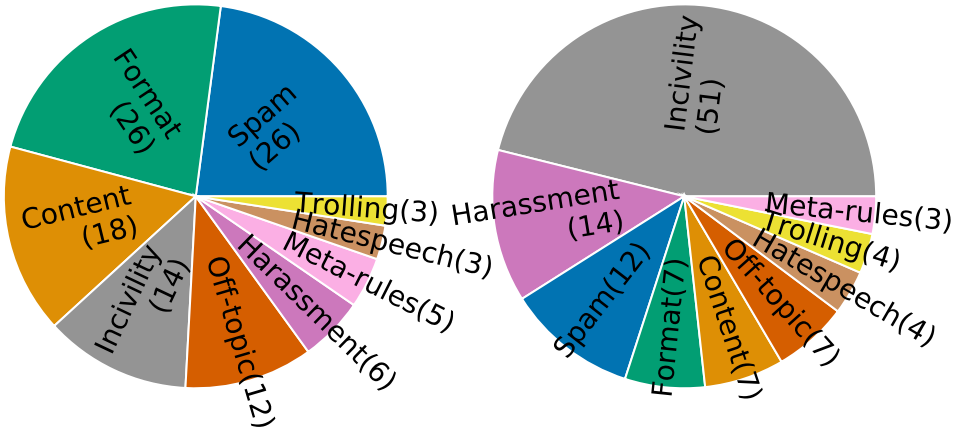


Figure 5.2: % of rule types of rules (left) and comments violating those rules (right) in NORMVIO.

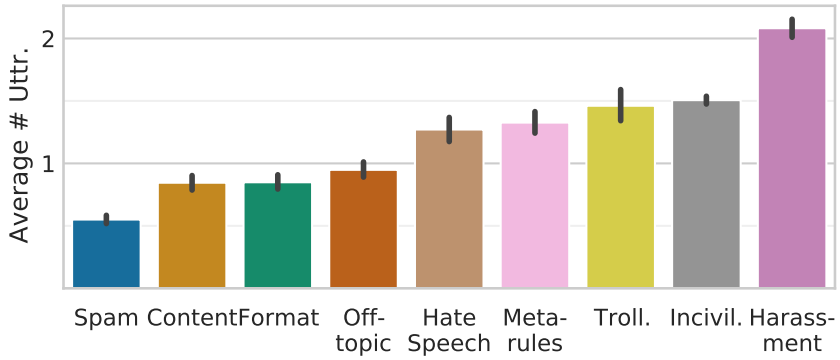


Figure 5.3: Average number of utterances between the original post and the moderated comments.

on identifying abusive language in the NLP community have been focused on hate speech, more subtle types of incivility are significantly more prevalent in removed comments, which are also harder to detect [20, 71, 111]. Moreover, only 55% of removed comments are violations of Incivility and Hate Speech rules, again highlighting the importance of understanding the spectrum of community norms in designing automated moderation assistance systems.

Figure 5.3 shows the average number of utterances from the original post to the norm-violating removed comment. Overall, violations related to abusive language such as Harassment, Incivility, and Trolling occur *later* in conversations than comments removed for other reasons (e.g. Spam and Format). This timing has implications for the "forecastability" of violation types. For example, the average conversation length within the Spam category is about 0.5 which indicates that half of the violations happen in the original post or a reply to it, making it impractical trying to forecast such violations.

Even though Hate Speech and Harassment are both related to abusive language, comments removed due to Harassment occur after more interactions. We hypothesize this is because ha-

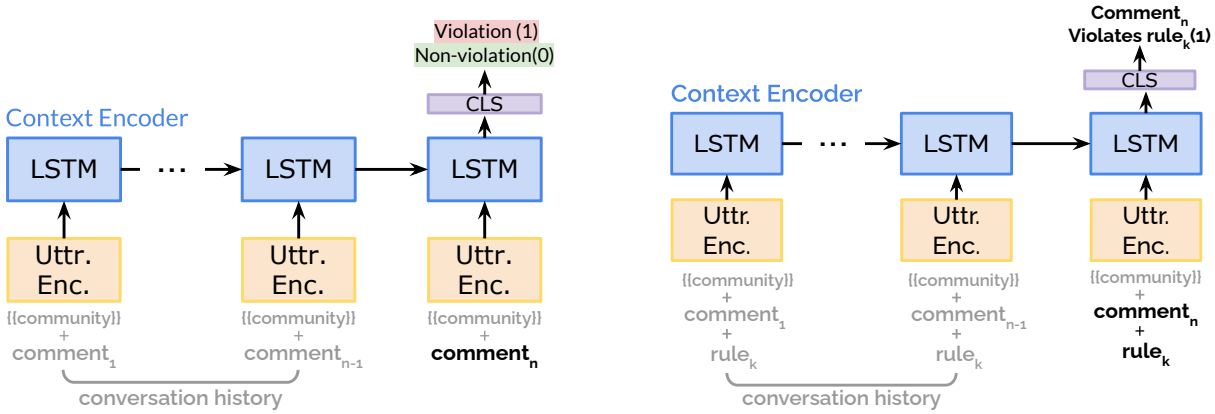


Figure 5.4: Structure of the baselines of the two proposed tasks: detecting norm violation (left) and explaining norm violation (right). Inputs in gray (conversation history and community information) are optional context.

rassment and trolling are intentionally expressed in less overt forms to delay the moderators' intervention. These findings illustrate that with a more representative set of community rules and a larger-scale dataset, NORMVIOfacilitates deeper understanding of community norm violation behaviors and provides guidance on more urgent tasks our field should be focusing on for a practical impact.

5.3 Detecting and Explaining Community Norm Violations

With NORMVIO, we can now train models for detecting contextualized, fine-grained community norm violations. We present two tasks: (1) Detecting community norm violations, and (2) Explaining community norm violations. The former identifies coarse categories of norm violations detailed in §6.3, and the latter is aimed at identifying specific local community rules being violated, to facilitate moderation transparency. For each task, we compare model variants without or with varying types of incorporated context, including conversation history and community information (e.g. subreddit name).

Detecting Community Norm Violations In this task we assume a set of pre-defined categories of norm violations. For each category, we train a binary classifier to detect violations, since the categories are not mutually exclusive.

As shown in Figure 5.4, we encode a conversational context of arbitrary length along with community rules. Following Chang and Danescu-Niculescu-Mizil [35], we use a uni-directional LSTM context encoder. The utterance encoder is initialized with a pretrained BERT model, with each classifier is then fine-tuned using training data specific to each rule type. The last hidden state from the last comment is fed into the classifier. The flexibility of this design allows for both retroactive detection after violations occur (the focus of this work) as well as proactive prediction of future rule violations.

We experiment with four model variants with different input contexts:

- COMMENT: Only the final comment.
- +HISTORY: Past conversation history and the final comment.
- +COMMUNITY: Community information and the final comment. We concatenated the subreddit name in front of the comment (e.g. "r/AskReddit ask anything!").[7]
- +HISTORY+COMMUNITY: Conversation history and community information.

Explaining Community Rule Violations In addition to categorizing rule violations by type (type-based), we develop a model that leverages the specific community rule text to identify violations in context. This text-based model facilitates explanations of rule violations, and improves transparency [110]. Such a system could lighten moderators' workload through highlighting why they might moderate a comment, enable more productive interventions, and improve the relationship between community members and moderators.

Similar to the violation category detection task, we construct binary classifiers that detect violations given conversational and community context. However, as shown in Figure 5.4, the full input and training procedure are different; we include the community's verbatim rule description as a model input. The rule text is appended to the input comment with a special token ([SEP]) added between the comment and the rule to leverage pretrained language models' ability to infer relationships between two sentences. Since the precise formulation of the target rule is given as an input, we no longer need to train one model per rule type; we train one universal model with all available training data.

NORMVIOcontains information about which rules are violated in each removed comment, and we use these rule-comment pairs as positive examples. If a comment is tagged for violating more than one rule, we include all comment-rule pairs as positive examples. We construct negative training examples using matched unmoderated conversations from NORMVIO(described in §6.3) by adding the text of the violated rule to the corresponding unmoderated conversation.

To guide the model in better discriminating rules, we construct additional negative examples by mapping each removed comment with an randomly chosen incorrect rule from the same subreddit (e.g. "Here's my referral code! [SEP] No Politics").

Similarly, we experiment with three model variants with different input contexts:

- +RULE: Only the final comment and a rule text.
- +RULE+HISTORY: Past conversation history, the final comment, and a rule text.
- +RULE+HISTORY+COMMUNITY: Both conversation and community history, the final comment, and a rule text.

The main advantage of the text-based model is in its interpretability and generalizability. Since the model now looks at the community-specific rule texts, the system can provide more meaningful feedback to moderators and users. For example, instead of saying "potential hate speech detected", now the model can be more informative in notifying users that "the comment has breached our community's Rule 2: No Racial Slurs". Moreover, since the model takes free-form rules as input, it can generalize to unseen rules and novel rule types.

[7]Note that the model variants without conversation history do not use a context encoder at all and thus have a smaller number of trainable parameters.

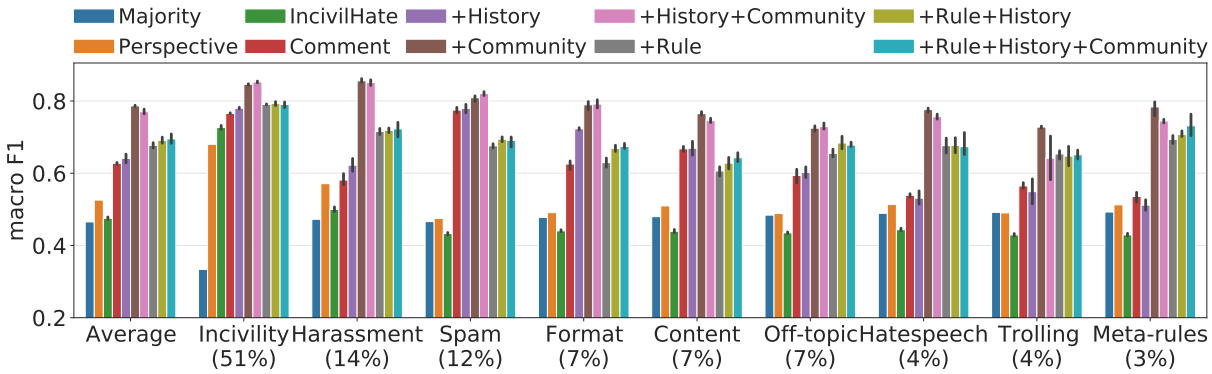


Figure 5.5: Average and breakdown of Macro F1 scores of the baselines and the model variants. Error bars indicate 95% confidence interval and the types are sorted by their violation frequency (percentage below x-axis labels).

5.4 Experiments

Baselines In addition to the seven model variants in §5.3, we consider three baselines that represent current common approaches:

- MAJORITY: Majority class baseline.
- PERSPECTIVE: Perspective API's toxicity score of the final comment to make a binary decision. For each rule type, a threshold value was tuned to maximize development set F1 score.
- INCIVILHATE: We train a model using just the incivility and hate speech violations from NORMVIO. The test set predictions from the trained model was evaluated over different rule types.

Training Details We perform an 80-10-10 train/dev/test random split of moderated comments in NORMVIO and then appended paired unmoderated comments into the same split. The resulting number of examples of train/dev/test split was 41667, 5214, and 5131, respectively. We ran training for five different random seeds and report the average scores of multiple runs except for MAJORITY and PERSPECTIVE baselines.

The base utterance encoder is a pretrained Conversational BERT model. Each model was trained for 10 epochs with an early stopping patience of 5, and with Adam optimizer with a learning rate of 1e-5. We used a batch size of 32 for models that do not leverage past conversations and 8 for the ones that use comment history. We used 2 layers of GRUs with a hidden size of 768 for the context encoder and 2 linear layers for the final classifier.

Evaluation We used macro F1 to evaluate all models. For models in §5.3, at test time we cannot assume that we know which rule will be violated in a given conversation. We thus create multiple comment-rule pairs for each comment in the test set by matching it with each community rule. Out of the resulting pairs, we mark the pairs that were observed in the original test set as positive, and the remaining pairs are marked as negative. We refer to these negative pairs added to the test set of models explaining rule violations as *augmented pairs*. Note that the test

sets of models in §5.3 are now different from the text sets in §5.3 and the F1 scores of two tasks are not directly comparable.

Experiment Results Information from the social context of a comment substantially improves performance (Figure 5.5). Compared to current approaches for inferring toxicity, all type-based violation detection model performed significantly better—even for rule violation categories those approaches are tailored for. While PERSPECTIVE and INCIVILHATE performed better in Incivility and almost comparable with COMMENT and +HISTORY, adding community information still resulted in a significant improvement of +8.0 absolute increase in F1.

Across all rule violation types, adding the context about community significantly improved the performance, often resulting in the highest performing models when added. Adding conversation history showed mixed results. +HISTORY showed improvements over COMMENT whereas +HISTORY+COMMUNITY was not necessarily better than +COMMUNITY. Models with conversation history tend to perform worse on scarce violation types such as Meta-rules and Trolling; we speculate that this decreased performance is due to the increased number of parameters from adding context encoder layer to process conversation history and future work with more examples of these violations may substantially improve performance. This result for history greatly expands an analysis by Pavlopoulos et al. [167] that found minimal performance gain when adding a single prior comment to identify toxicity; while we too find minimal improvement for Incivility and Harassment norms, adding history *does* improve the recognition for other norm violations (e.g., Format and Content) indicating that prior context can be useful.

While the results of text-based violation detection models (+RULE, +RULE+HISTORY, +RULE+HISTORY+COMMUNITY) and type-based models are not directly comparable due to the augmented pairs, they were evaluated over the same set of comments so the numbers can provide a general sense of text-based model performance. An interesting distinction between the two detection tasks is in how much additional context helps. In type-based models, adding context made significant improvements in all or in some cases. However, with text-based models, the performance was relatively more uniform and additional context did not contribute as much. This result suggests that providing full text of rules may help resolve certain ambiguous comments and thus the model rely less on the additional context.

5.5 Analysis

How many violations do current systems miss? In part due to their targeted focus, the PERSPECTIVE and INCIVILHATE baseline models miss a substantial proportion of the total norm violations. Figure 5.6 shows the confusion matrices of the violation detection task, where labels are aggregated over all violation types to test how many violations overall are not captured by these systems. The results

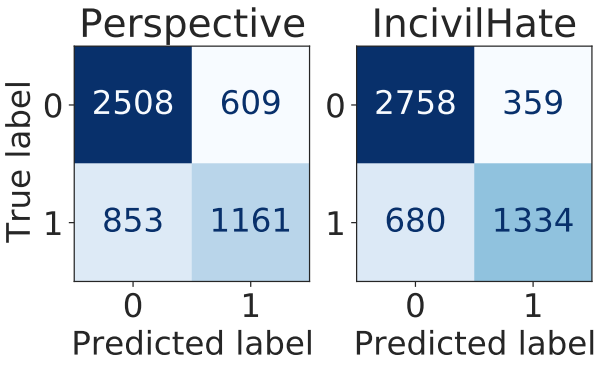


Figure 5.6: Confusion matrices of two baselines over the norm violation detection task.

show that PERSPECTIVE and INCIVILHATE fail to recognize 42% and 34% of all violations, respectively. Moderators on platforms like Reddit must triage huge numbers of comments daily and this points to a clear gap between current practice (represented by the baselines) and indicates what moderators act on in practice.

How does community information help? We observed that adding community information provides the most significant improvements in Harassment in Figure 5.5. We now look into the Harassment type to understand more about how did the additional community information actually help to improve the performance.

What kinds of errors are corrected by adding community context? By comparing confusion matrices of COMMENT and +COMMUNITY (Figure 5.7), we find that +COMMUNITY has fewer false positives. Out of 154 false positives from the COMMENT model that were corrected in the +COMMUNITY model, 106 (69%) were Incivility violations. Consider the following example:

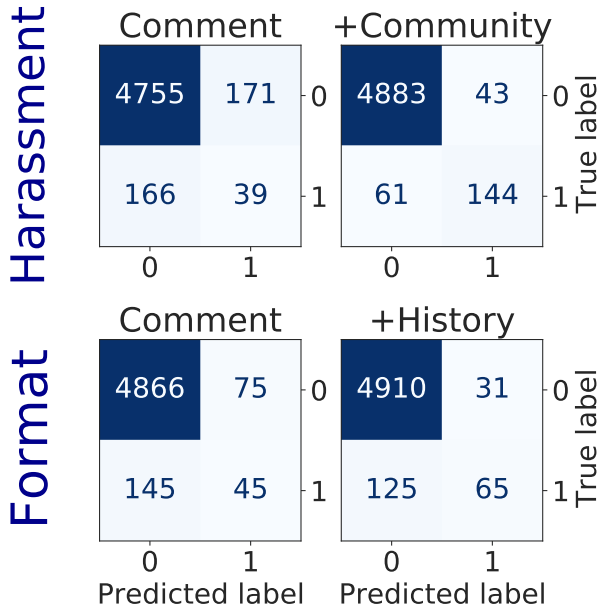


Figure 5.7: Confusion matrices of COMMENT and +COMMUNITY for the Harassment detection task (top), and of COMMENT and +HISTORY for the Format violation detection task (bottom).

Comment:
That game's already dead to 99% of the world a few weeks later, get over it you stupid idiot.
Moderator Comment:
Your comment has been removed for Rule 2. Be civil and respectful. Do not attack or harass other users or engage in hate-speech.
Paired Rule: Rule 2: Be civil and respectful.
Violation: Incivility
Community: r/classicwow

The final comment in this example could be considered as both a Incivility and Harassment violation and COMMENT model labels it as Harassment. Although the moderator refers to the community's Incivility rule, the rule mentions "do not attack or *harass* other users", which makes it clear that this example falls into both categories. However, the +COMMUNITY model labels this comment as Incivility and not Harassment. We speculate that the +COMMUNITY model learns about what rules exist in each community; *r/classicwow* has 8 rules and none of them are about Harassment, so moderators refer to the Incivility rule when moderating Harassment violations. In other words, depending on the community and their available community rules, the same comments can be moderated as either incivility or harassment violation. Therefore, providing the community information can help the model disambiguate this decision and ground

its moderator support in the norms of the community.

How does conversation history help? Likewise, for the conversation history context, the largest gain was achieved in the Format type. In Figure 5.7, we compare confusion matrices of COMMENT and +HISTORY. The result again shows that additional context can help the model in reducing the false positive rate.

Among the corrected false positives, the most prevalent type mistaken for Format was Spam. One example of such case is given below:

Comment: *UPDATE: I found it! here you go if you need it_LINK_*
Violation: Spam (Piracy)
Moderator Comment:
See Rule 1: No Merchandise / Spam
Previous Message:
Does anyone know where to buy this?

If we only consider the final comment, there are two possible explanations for which rule was violated: 1) Format: the outside link does not follow the community guideline 2) Spam: self-promotion / promoting specific merchandise is banned. However, the previous message makes it clear that the author had just posted about a product and then made a self-reply with a link to buy the product. With this information, model can disambiguate this situation and choose the right violation type.

5.6 Related Work

Community Norms and Rules Many studies have investigated how online conversations are moderated and how each community has different norms to ensure a safe environment for discussions [4, 34, 107, 108, 110, 178]. Fiesler et al. [74] conduct an analysis over the rules of Reddit communities and define 24 types of the rules. They provide a thorough and large-scale analysis over how the rules are phrased and how rules are different across subreddits. We adopt their rule categorization and extend it to code actual rule violations.

Chandrasekharan et al. [34] also studied removed comments on Reddit to understand what types of rules exist on Reddit by clustering the moderator comments and investigated how they are governed. However, their dataset provides limited context of moderated comments, whereas we focus on providing a dataset that has enough context and also explicit violation type that can be leveraged in modeling rule violation.

Context in Detecting Online Abuse Most of the existing datasets for abusive language detection implicitly assumes that comments may be judged independently taken out of context. Pavlopoulos et al. [167] challenged this assumption and examined if context matters in toxic language detection. While they found a significant number of human annotation labels were changed when context is additionally given, they could not find evidence that context actually improves the performance of classifiers. Our work also examines the importance of context, but

we do not limit our scope to toxic language detection and investigate a broader set of community norm violation ranging from formatting issues to trolling.

Beyond Incivility and Hate Speech Jurgens et al. [111] claims "abusive behavior online falls along a spectrum, and current approaches focus only on a narrow range" and urges to expand the scope of problems in online abuse. Most work on online conversation has been focused on certain types of rule violation such as incivility and toxic language [e.g., 4, 35, 229]. In this work, we focus on a broader concept of *community norm violation* and provide a new dataset and tasks to facilitate future research in this direction.

5.7 Conclusions and Future Work

Online communities establish their own norms for what is acceptable behavior. However, current NLP methods for identifying unacceptable behavior have largely overlooked the context in which comments are made, and, moreover, have focused on a relatively small set of unacceptable behaviors such as incivility. In this work, we introduce a new dataset, NORMVIO, of 51K conversations grounded with community-specific judgements of which rule is violated. Using this data, we develop new models for detecting context-sensitive rule violations, demonstrating that across nine categories of rules, by incorporating community and conversation history as context, our best model provides a nearly 50% improvement over context-insensitive baselines; further, we show that using our models, we can *explain* which rule is violated, providing a key assistive technology for helping moderators identify content not appropriate to their specific community and better communicate to users why. Our work enables a critical new direction for NLP to develop holistic, context-sensitive approaches that support the needs of moderators and communities.

Part III

Personal Context and Language

Chapter 6

Identifying Social Biases Using Individuals' Social Attributes

The last part of this thesis addresses the role of personal context in NLP, following our exploration of cultural and community contexts. Personal context encompasses the unique social attributes and experiences of individuals, which can significantly influence language use and interpretation. Understanding and modeling personal context is crucial for conducting more controlled analyses and improving text classification. In this chapter, we focus on identifying social biases in corpora by leveraging individuals' social attributes, such as those found in Wikipedia biographies. We propose a methodology for matching social attributes to create controlled comparisons, enabling the analysis of how different social attributes, such as gender, race, or sexual orientation, influence the portrayal of individuals in text. Specifically, we use the method to analyze Wikipedia biographies to uncover content biases related to the LGBT community, demonstrating the utility of personal social context in revealing systematic social biases.

6.1 Introduction

Alan Turing was prosecuted for being gay in 1952 and underwent a hormonal injection. His English Wikipedia page describes this situation as "He *accepted* the option of injections of what was then called stilboestrol." which suggests that Turing had little control over the situation. In contrast, the Russian version describes the same situation as " Учёный предпочёл инъекции стильбэстрола (The scientist *preferred* stilbestrol injections)" which implies that he actively made the decision and even implies positive sentiment towards the injections. These variations in phrasing on Wikipedia could suggest that biases and stereotypes about LGBT individuals manifest differently in English, and Russian-speaking cultures, violating the platform's "Neutral Point of View" policy.

In this work, we present computational methods for analyzing biases in multilingual narrative text, which can help identify social stereotypes and reduce bias. Previous studies have used Contextual Affective Analysis (CAA) to analyze affective dimensions of power, agency, and sentiment to reveal biases in movie scripts and newspaper articles [68, 181, 190]. This analysis relies on connotation frames, lexicons of verbs annotated to elicit implications, which until now

have only existed for English. While machine translation of lexicons is possible [152, 180], no in-language evaluations of connotation translatability or extensive analysis in other languages have been conducted.

Our goal is to measure the power, agency, and sentiment of people in multilingual text, particularly in English, Spanish, and Russian. This involves conducting in-language (in Russian text, are LGBT people portrayed as more powerful than non-LGBT people?) and cross-language analysis (is the power differential between LGBT people and non-LGBT people greater in English or in Russian?) by crowdsourcing annotations of connotation frames in the three languages. The researchers use this data to develop multilingual CAA classifiers for power, agency, and sentiment. The methodology is tested by analyzing how members of the LGBT community are portrayed on Wikipedia in different languages, revealing that LGBT biographies have more negative connotations in Russian, while English and Spanish pages are generally neutral or positive. These findings align with perceptions of LGBT people in English, Spanish, and Russian-speaking countries [78].

6.2 Matching Methodology

We introduce a method for identifying a *comparison* biography page for each page that corresponds to a target attribute, ensuring that the comparison page closely matches the target page in all known attributes except for the target attribute.[1]

The concept stems from modifying observational data to mimic the conditions of a randomized trial. Researchers create treatment and control groups from the observational data, ensuring that the distribution of all covariates, except for the target one, is as similar as possible between the two groups. [184].[2] By comparing the constructed treatment and control groups, researchers can separate the effects of the target attribute from other confounding variables. Matching is also increasingly being recognized in language analysis. [33, 37, 63, 116, 183]. Here, our target attribute is gender or race as perceived by editors and readers. Our goal is to create corpora that balance other characteristics, such as age, occupation, and nationality, which could influence how articles are written.

Given a set of target articles $\mathcal{T}$ (e.g., all biographies about women), our objective is to create a set of comparison articles $\mathcal{C}$ from a pool of candidates $\mathcal{A}$ (e.g., all biographies about men), such that $\mathcal{C}$ has a similar distribution of covariate as $\mathcal{T}$ for all attributes except the target one. We build $\mathcal{C}$ using a greedy matching approach. For each $t \in \mathcal{T}$, we identify $c_{best} \in \mathcal{A}$ that most closely matches t and include c_{best} in $\mathcal{C}$. For instance, if t is about an American female actress, c_{best} might be about an American male actor. To find c_{best}, we utilize the category metadata associated with each article. For example, the page for Steve Jobs includes categories such as "Pixar people", "Directors of Apple Inc.", "American people of German descent", etc. While articles may not always be categorized correctly or with consistent detail, this metadata allows us to focus on covariates likely to influence how the article is written. Individuals may have relevant traits that are not listed on their Wikipedia page, but if no editor assigned a category related to those traits,

[1] All code and data is available at https://github.com/anjalief/wikipedia_bias_public

[2] Instead of treatment/control, we use target/comparison to clarify that this work does not involve actual "treatment".

we have no reason to believe that editors were aware of them or that they influenced the edits. We describe six methods for identifying $c_{best} \in A$. $CAT(c)$ indicates the set of categories associated with c.

NUMBER We choose c_{best} as the article with the highest number of categories in common with t, as this intuitively represents the best match.

$$c_{best} = arg\,max_{c_i} |CAT(c_i) \cap CAT(t)|$$

PERCENT NUMBER tends to favor articles with more categories. For instance, a candidate c_i with 30 categories is more likely to share more categories with t than a candidate c_j with only 10 categories. However, this does not necessarily mean that c_i has more traits in common with t—it might simply indicate that the article is more thoroughly written. To mitigate this bias, we normalize the number of shared categories by the total number of categories in the candidate c_i:

$$c_{best} = arg\,max_{c_i} |CAT(c_i) \cap CAT(t)| \frac{1}{|CAT(c_i)|}$$

TF-IDF Both of the previous methods assume that all categories are equally significant, which oversimplifies the situation. A candidate c_i sharing the category "American short story writers" with t is likely a better match than one sharing "Living People." To address this, we use TF-IDF weighting to give more weight to rarer categories [187]. We represent each $c_i \in \mathcal{A}$ as a sparse category vector, where each element is a product of the category's frequency in c_i ($\frac{1}{|CAT(c_i)|}$ if the category is in $CAT(c_i)$, 0 otherwise) and the inverse frequency of the category, which down-weights broad, common categories. We then select c_{best} as the c_i with the highest cosine similarity between its vector and a similarly constructed vector for t.

PROPENSITY For each article, we calculate a propensity score, estimating the probability that the article contains the target attribute [184, 185], using a logistic regression classifier trained on one-hot encoded category features. We then choose c_{best} as the article with the closest propensity score to t's. While propensity matching is not ideal in our context, first, because it was designed for lower-dimensional covariates and can fail with high-dimensional data, and second, because it doesn't always produce meaningful matched pairs, which hinders manual examination of matches [183], we include it as a baseline due to its popularity in controlling for confounding variables.

TF-IDF PROPENSITY We develop an additional propensity score model, using TF-IDF weighted category vectors as features instead of one-hot encoded vectors.

PIVOT-SLOPE TF-IDF Both TF-IDF and PERCENT incorporate the term $\frac{1}{|CAT(c_i)|}$ to normalize for differences in the number of categories across articles. However, information retrieval research suggests that this over-correction can cause the algorithm to favor articles with fewer categories [199]. To address this, we adopt pivot-slope normalization [199] and normalize TF-IDF terms with an adjusted value: $(1.0 - slope) * pivot + slope * |CAT(c_i)|$. This method requires setting the slope and pivot, which control the adjustment's strength. Following Singhal et al. [199], we set the pivot to the average number of categories across all articles and tune the slope using a development set. Tuning the slope is crucial, as adjusting this parameter affects the selected matches. PIVOT-SLOPE TF-IDF is our novel proposed approach.

In practice, it is unlikely to find close matches for every target article, as some characteristics of people in the target corpus may not be shared by anyone in the comparison corpus. To address this, we discard "weak matches": for direct matching methods, pairs with fewer than 2 categories in common, and for propensity matching methods, pairs whose difference in propensity scores

is more than 1 standard deviation away from the mean difference.

6.3 Crowdsourcing Contextualized Connotation Frames

We collected a corpus of multilingual connotation frames in English, Spanish, and Russian. These annotations ask annotators to evaluate the power, agency, and sentiment of the agent and theme of verbs in complete contexts drawn from newspaper articles. The connotation frames differ from existing lexicons in that they are collected in languages other than English and use complete contexts. Because these affective dimensions can be difficult to define, we took numerous steps to ensure annotations would be of high quality.

We collected a corpus of multilingual connotation frames in English, Spanish, and Russian by extracting frequent verbs and contexts representative of each verb's most common usage from a News Crawl corpus [10], which included tuples with at least one human subject or object using the *noun.person* category of WordNet [67]. Annotators from the United States, Russia, and several South American countries were asked to provide judgments on the power, agency, and sentiment of the verbs in these contexts. The judgments were aggregated, and connotations were labeled as positive, neutral, or negative. For each of the three target languages, we collected power, agency, and sentiment annotations for 300 verbs in three contexts each (900 instances). For each instance, we collected judgements from three annotators, leading to $32,400$ total annotations.

6.4 Classification of Connotations

Method We aim to develop a methodology for analyzing the portrayal of people in different languages. To achieve this, we train a supervised classifier on contextualized multilingual annotations to predict a connotation frame label for unseen verbs and contexts. Specifically, we use pre-trained cross-lingual language models to obtain language-agnostic feature representations, allowing us to combine different languages in the training and test data. The multilingual embeddings of verbs in-context are extracted using XLM [40], a state-of-the-art pre-trained model. A logistic regression model with sample weighting is used as the primary classifier, which is trained on frozen contextual embeddings. Finally, the trained classifier is used to predict connotation frame labels for verbs provided with their context in a target language corpus.

Evaluation We evaluate our model on the contextualized multilingual annotation data. Table 6.1 reports macro F1 scores for single-language evaluations, where we train on the training set of one language ("Src") and evaluate on the test set of a second language ("Tgt"). Unsurprisingly, training and testing on the same language achieves better performance than training on one language and testing on another. While the cross-lingual model works well in certain cases, in most cases transferring languages yields a substantial decrease in performance. These results offer further evidence of the importance of in-language connotations.

We also explore using in-language and out-of-language data in combination. This experimentation is based on the hypothesis that while connotations differ in different languages, there

Tgt	Src	$\text{Sent}_{\text{subj}}$	Sent_{obj}	Pow.	Agen.
EN	EN	**43.4***	43.0	**41.1**	**48.2***
	ES	38.1	43.4	29.5	43.4
	RU	41.1	**44.3**	40.1	41.4
ES	EN	38.9	36.6	24.5	31.3
	ES	**49.5***	**51.2***	**43.6***	**43.6***
	RU	39.0	42.2	34.0	38.9
RU	EN	43.6	49.2	36.4	44.5
	ES	37.2	49.3	38.2	42.7
	RU	**46.4***	**54.9***	**45.3***	**49.9***

Table 6.1: Macro F1 score of classifiers trained and evaluated with different target and source languages.

Tgt	Src	S_{subj}	S_{obj}	Pow.	Agen.
EN	EN	43.4	43.0	41.1	48.2
	+ES	44.8	**45.2***	40.5	49.7
	+RU	**46.5***	43.2	**41.8**	49.9
	+ES+RU	45.0	44.3	41.7	**50.0***
ES	ES	49.5	51.2	**43.6**	43.6
	+EN	50.4	51.6	36.4	45.5
	+RU	51.0	**55.0***	42.1	**45.6***
	+EN+RU	**51.8***	54.8	40.8	44.9
RU	RU	46.4	54.9	45.3	49.9
	+EN	45.6	55.7	44.1	50.9
	+ES	46.0	**59.2***	42.1	49.8
	+EN+ES	**47.7**	53.7	**46.9***	**51.7***

Table 6.2: Macro F1 score of classifiers, where in-language training data is augmented with training data from other languages.

may be enough overlap for the cross-lingual model to learn useful signals from out-of-language data. Table 6.2 shows results. For all connotation dimensions, the best performing augmented models outperform the non-augmented models.

6.5 Multilingual Affective Analysis of LGBT People

Finally, we demonstrate how the new corpus of annotations and the cross-lingual model facilitate multilingual analysis by examining portrayals of LGBT people on Wikipedia. Little prior computational work has studied narratives about the LGBT community, likely due to data scarcity [149]. To facilitate analysis, we collect a new corpus, titled LGBTBIO, which contains Wikipedia biography pages of LGBT people and provide our analysis.

LGBTBIO Corpus We collected a multilingual corpus of 1,340 Wikipedia biography pages for people in the LGBT community using Wikidata properties and lists of LGBT people from Wikipedia. This corpus allows us to analyze how the same person is portrayed in different languages. However, we cannot draw conclusions from this corpus alone, because we need to control for overall language differences. For example, if we find that LGBT people have higher power in English pages than in Russian pages, we cannot determine if this difference occurs because LGBT people are actually described differently or because English verbs tend to have more positive power connotations than Russian verbs. Therefore, we built a control corpus by matching each LGBT person with a non-LGBT person who has similar characteristics using a matching method we developed in Field et al. [69]. We constructed weighted TF-IDF vectors from each person's associated categories to identify matches, and selected the control person as the one who has the most similar category vector as the LGBT person. Together, the 1,340 pages

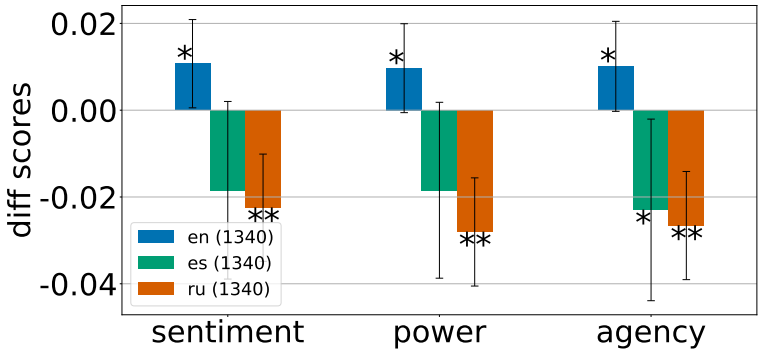


Figure 6.1: Average differences in affective scores in narratives about LGBT people vs. matched control people across languages. Asterisks indicate scores are statistically different from zero ($*$:p<0.05, $**$:p<0.01). Numbers in the legend indicate the number of biographies in each group.

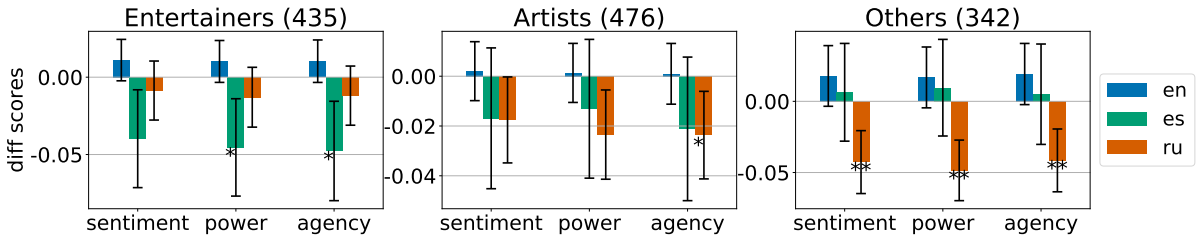


Figure 6.2: Average sentiment/power/agency diff scores for narratives of occupation subgroups.

about LGBT people and their matched controls constitute the corpus used for the analysis of how the same person is portrayed in different languages.

Contextual Affective Analysis of Narratives Describing LGBT People For each language, we used the best performing model from Table 6.2 to predict contextualized verb annotations for all sentences that contain a target person's name or pronoun as the subject of a verb. We then mapped these verb scores to entities following Field et al. [68]. We report *diff score* as the difference between sentiment/power/agency scores in each each article about an LGBT person and its matched control, averaged across all pairs (e.g. "average treatment effect"); a positive score means the articles about LGBT people had a higher affect connotation in aggregate.

Figure 6.1 shows diff scores across the entire corpus. The results reveal that English articles had significantly positive connotations, while Russian articles had significantly negative connotations. Spanish articles also had negative connotations, but only agency was statistically significant. These differences could be influenced by global perceptions of LGBT individuals, as studies have shown that social acceptance of LGBT individuals varies across countries [78].

In order to further examine possible cultural differences and in recognition that sexual orientation does not reflect an individual's entire identity, we divide our corpus according to nationality, birth year, and occupation and test if additional social theories manifest in our data. In this proposal document, we only discuss the occupation dimension, while additional results will be available in the main thesis.

We focus on the two most common occupations identified in our corpus (Entertainer and Artist) in order to ensure sufficient sample size. Survey and behavioral studies have suggested that LGBT people are perceived as better suited to some occupations than others [38, 207]. In Figure 6.2, we see little difference in how LGBT people of different occupations are portrayed on Wikipedia in English. However, in Spanish, articles about entertainers have significantly negative connotations, but articles about people of other occupations do not. Conversely, in Russian, while entertainers have near-neutral connotations, people of other occupations, such as politicians, scientists, and activists, are portrayed with significantly negative connotations. While more investigation is needed, our data offers evidence suggesting that perceptions about occupational stereotypes of LGBT people may differ across cultures.

6.6 Related Work

Our work follows on a series of prior work: Rashkin et al. [181] introduced sentiment connotation frames, Sap et al. [190] extended them to power and agency, and Field et al. [68] introduced the CAA framework. Connotation frames have been used to analyze films, newspaper articles, and online stories [6, 68, 181, 190]. Rashkin et al. [180] extend connotation frames to other languages through mapped embeddings, but they do not conduct evaluations against in-language annotations nor provide multilingual annotations.

Our work is generally consistent with existing literature on cross-cultural biases and online biographies. Dong et al. [55] show that perceptions of social roles differ across cultures, while De-Arteaga et al. [47] reveal gender bias in online biographies. Other work has examined biases in Wikipedia. Wagner et al. [216] show that portrayals of men and women differ across languages, and Callahan and Herring [28] reveal systematic cultural biases, particularly in biography pages.

Several studies in social science literature have analyzed biases and their effects on the LGBTQIA+ community, for example, examining mental health [3] microaggressions [8], and sociopolitical involvement [89]. With a few exceptions [53, 65, 149, 192], biased language about or against the LGBTQIA+ community has not been examined and analyzed extensively in automated analyses. The closest study to ours is an examination of gender, race, and LGBT portrayals in 700 popular films [201].

6.7 Conclusions and Future Work

Our work provides methodology and datasets that extend the capabilities of affective analysis to multilingual settings. While we focus on Wikipedia, our methodology could be used to conduct analyses in any English, Russian, and Spanish narrative text, which can aid writers in obtaining a neutral point of view and provide insight into social stereotypes, especially when used in combination with other methods. This framework supports the investigation of a wide range of research questions, and offers multiple avenues for future work such as improving the multilingual model, expansion to additional languages, investigation of Wikipedia edit histories, and the incorporation of additional connotations and existing linguistic databases.

Chapter 7

Socially Aware Empowering Text Detection

The previous chapter demonstrated the importance of personal context in conducting controlled analyses and revealing social biases. In this final chapter, we explore how diverse social attributes can be directly leveraged to improve text classification tasks. For instance, determining whether a statement like ”go get it girl” is empowering or condescending may depend heavily on the gender of both the speaker and the recipient. In real-world communication, we naturally consider these personal contexts to interpret the true meaning of text. This chapter presents methods for modeling and incorporating diverse social contexts into text classification, without the need for exhaustive data covering all possible scenarios. We introduce the TalkUp dataset and demonstrate through various experiments how socially-aware models can significantly enhance the accuracy and relevance of text classification tasks, paving the way for more nuanced and contextually informed NLP systems.

7.1 Introduction

Empowerment – the act of supporting someone's ability to make their own decisions, create change, and improve their lives – is a goal in many social interactions. For instance, teachers aim to empower their students, social workers aim to empower their clients, and politicians aim to empower their supporters. A growing body of psychology and linguistics research shows how empowerment – and disempowerment – can impact people by increasing their sense of self-efficacy and self-esteem [31, 161].

Understanding how empowerment is conveyed in language becomes more important as language technologies are increasingly being used in interactive contexts like education [153], workplace communication [171, 173], and healthcare [141, 194]. Whether we are building dialogue agents for mental health support, supplementing children's education, or analyzing managers' feedback to their employees, language that empowers or disempowers the reader can have drastically different effects.

With a few exceptions [196, 232], prior NLP research has focused on flagging *harmful* text, but there has been much less investigation of what makes text *helpful*. Other works have studied related concepts like condescension [218] and implicit toxicity [21, 189, 213], and we build off of these to construct a dataset that *complements* those tasks.

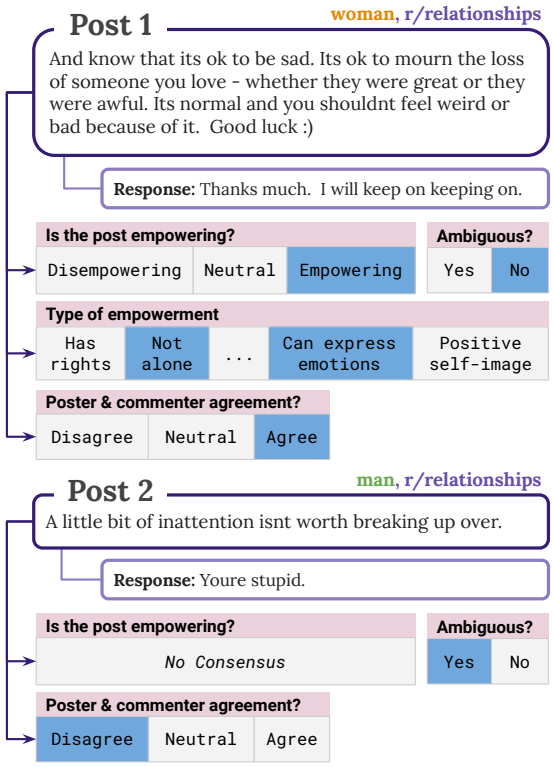


Figure 7.1: Two examples of annotated conversations in TalkUp. Post 1 is straightforwardly empowering, but Post 2 is inherently ambiguous and could either be interpreted as helpful advice or as a dismissive, belittling comment. Social context can also affect Post 2's implications: the post might elicit different reactions if it were written by a woman to a man or vice versa.

Consider the two examples of potentially empowering interactions in Figure 7.1. Empowerment exhibits the importance of social context in understanding the pragmatics of language: whether an exchange is interpreted as empowering or disempowering may depend on the participants' social roles and the power dynamics implied by their identities, including race, age, socioeconomic class, and many other social dimensions. Furthermore, empowerment cannot be easily detected with sentiment or emotion analyzers, since interactions with negative implicatures can be empowering (e.g., *you can quit!!!*), and messages that are positive on the surface can be disempowering (e.g., *you are so articulate for a girl!*) [72]. Modern language technologies do not model social context or deeper pragmatic phenomena, and thus are unable to capture or control for empowerment.

This work makes concrete steps towards understanding these linguistic phenomena by investigating the following research questions: **[RQ1]** What makes language empowering, and how is it manifested in language? **[RQ2]** Can empowerment be detected with computational approaches?

Our contributions are threefold: (1) We introduce the new task of empowerment detection, grounding it in linguistic and psychology literature. (2) We create TalkUp, a novel dataset of Reddit posts labeled for empowerment, the fine-grained type of empowerment felt by the reader, and the social relationships between posters and readers. (3) We analyze the data and demonstrate

how it can be used to train models that can capture empowering and disempowering language and to answer questions about human behavior.

Ultimately, TalkUp aims to assist future researchers in developing models that can detect, generate, and control for empowerment, and to facilitate broader exploration of pragmatics. We have by no means covered every possible social dimension, but by focusing on a few social factors in the simplified setting of two-turn dialogues, we hope that TalkUp's framework can make strides toward understanding language in more complex social interactions, such as conversations involving intersectionality as well as longer multi-turn dialogues.

7.2 Background

We discuss empowerment following its definitions in clinical psychology [31]. We find this most appropriate for studying language because clinical psychology practice is usually centered around dialogue between clinician and patient, and because it involves concrete implications about individuals rather than vague cultural phenomena. Thus, summarizing the different characteristics of empowerment described in psychology literature, we define empowering text as *text that supports the reader's rights, choices, self-fulfillment, or self-esteem.*

Incorporating empowerment in dialogue agents, mental health support chatbots, educational assistants, and other social-oriented NLP applications is clearly a desirable goal. However, empowerment is inherently challenging to operationalize for several reasons. First, it is a flexible term that describes a wide range of behaviors across many domains – empowerment in economics, for example, looks very different from empowerment in a therapy session [148]. We follow recent literature outside of NLP in trying to distill these varied interactions into a concrete definition. Second, empowerment is implicit: it is often read in between the lines rather than declared explicitly. Text might be empowering by reminding someone of their range of options to choose from, encouraging them to take action, asking for and valuing their opinion, or even validating their feelings [31]. Third, empowerment is heavily dependent on social context: whether or not a person is empowered depends on who is saying what to whom. We incorporate these consideration in our data collection process described next.

7.3 The TalkUp Dataset

We now discuss the construction process of the TalkUp dataset.

Annotation Scheme Our annotation task was shaped through multiple pilot studies, where we learned that context is useful for judging a post, annotators' free-response descriptions of social roles lack consistency, and posts are often inherently ambiguous. Based on these insights, the final task, which is illustrated in Figure 7.1, consists of three main parts:

(1) *Rating the post on an empowerment scale.* This scale has "empowering" on one end, "disempowering" one the other, and "neutral" in the middle. We define text to be empowering if it supports the reader's rights, choices, self-fulfillment, or self-esteem, and disempowering if it actively denies or discourages these things. Notably, posts that discuss an external topic without

making any implications about the conversants, such as a comment about a celebrity's lifestyle, are defined as neutral.

(2) *Selecting why a post is empowering or disempowering.* We adopt the 15 points from Chamberlin [31], with slight modifications to adapt them to written text, as *reasons why a post can be empowering* to a reader. If a post is empowering, it should imply one or more of these reasons (e.g. that the reader is capable of creating change), and if it is disempowering, it should imply the opposite (e.g. that the reader is not capable of creating change).

(3) *Selecting whether the poster and commenter have agreeing or disagreeing stances.* We define "agreeing" and "disagreeing" loosely in order to accommodate a wide range of social relationships: "agree" means that the poster and reader support the same point of view on a topic, whether it be politics, sports teams, or music preferences. "Disagree" means that they take opposing sides.

Data Source TalkUp consists of English Reddit posts from RtGender [214], a collection of 25M comments on posts from five different domains, each labeled with the genders of the commenter and the original poster. We take advantage of the fact that these conversations are already annotated for gender, which provides contextual information about who is speaking to whom and allows us to explore at least one dimension of social context. [1]

Though RtGender contains posts from several platforms, given our focus on *conversational* language, we specifically selected RtGender posts from Reddit because they were the most generalizable and contained natural-sounding conversations. We manually chose five subreddits, aiming to include (1) a diverse range of topics and user demographics, and (2) discussions that are personal rather than about external events unrelated to the conversants. The subreddits are listed in Table 7.1.

We filtered data from these subreddits to exclude posts or responses that exceeded 4 sentences in length or were shorter than 5 words. Additionally, we excluded posts with "Redditisms", and posts that were edited after they were initially posted (marked "EDIT:" by the original poster) and posts that began with quoted text (marked ">") were removed.

From pilot studies, we found that models can help to surface potentially empowering posts and help increase the yield of posts that were actually labeled as empowering by annotators. We trained a RoBERTa-based regression model with the data we collected from the pilot studies to predict the level of empowerment (0 for disempowering, 0.5 for neutral, 1 for empowering) in Reddit posts. We used this model to rank and select the top-k posts for annotation, and continually updated the model as we collected more data. To ensure we annotate a diverse range of posts, our final annotation task was done with half model-surfaced posts and half randomly-sampled posts.

Annotation on Amazon Mechanical Turk With 1k model-surfaced posts and 1k randomly-sampled posts spread evenly among the five subreddits, we collected annotations via Amazon Mechanical Turk (AMT). Each example was annotated by 3 different workers.

[1] We only consider men and women here due to the availability of data. We were not able to find any corpora that included nonbinary genders, but this is an important area for future work. Though we focus on gender, there are many other social variables that may impact empowerment, such as race and socioeconomic status.

	Size	#E	#D	#A	%W
TalkUp	2000	962	129	267	43
AskReddit	400	186	26	43	47
relationships	400	199	35	83	72
Fitness	400	193	28	64	14
teenagers	400	173	29	48	34
CasualConversation	400	211	11	29	50

Table 7.1: Data Statistics of TalkUp and breakdown of five subreddits in the data. #E: number of empowering examples, #D: number of disempowering examples, #A: number of ambiguous exmaples, %W: percentage of women posters in the data.

To ensure high quality annotations, we required annotators to have AMT's Masters Qualification,[2] a task approval rate of at least 95%, and a minimum of 100 prior tasks completed. Additionally, since our task requires English fluency, we limited annotators to those located in the US or Canada. Workers were compensated at $15/hour, and we calculated the reward per task based on the average time spent on each annotation in our pilot studies.

Following best practices to increase annotator diversity [30], we staggered batches of data to be released at different times of day over multiple days. After each batch was completed, we manually quality-checked the responses and computed each annotator's standard deviation. We discarded data from unreliable annotators, including those who straightlined through many annotations with the same answer, those who clearly had not read instructions, and those whose alignment scores were more than 2 standard deviations from the mean. Annotator alignment scores were calculated by dividing the number of disagreements by the number of agreements between their label and the majority vote. We subsequently released new batches to re-label data previously annotated by the identified unreliable annotators.

Data Statistics We combined the *maybe empowering* with the *empowering* label, and did the same for the *disempowering* labels. We then used majority voting to aggregate the three annotations into the final labels for empowerment, ambiguity, and stance for each post. When all three annotators disagreed on the empowerment label (i.e., one vote each for empowering, neutral, and disempowering), we marked it as *No Consensus* and considered it an ambiguous case. For reason labels, where annotators can mark more than one categories per example, we only kept the reason labels that were marked by at least two annotators.

Table 7.1 shows the overall size of our dataset and the distribution of labels, the number of ambiguous cases, and percentage of posts made by women across the entire dataset and also by different subreddits. We annotated 400 posts from 5 different subreddits resulting in a total of 2000 samples. Of these, 962 were labeled as empowering, 129 as disempowering, and 267 as ambiguous, with 642 being labeled as neutral. We note that 265 out of the 962 empowering cases had no final reason marked, indicating that there was no reason category annotators agreed on.

The inter-annotator agreement, Krippendorff's alpha, was 0.457, and the percentage agreement was 65.2%. These agreement scores are reasonable given the complexity and nuance of

[2]AMT grants the "Master Worker" qualification to highly reliable workers.

this task – we would neither expect nor want to have perfect annotator agreement because it is an inherently ambiguous problem even for humans, and there is often no objective "ground truth" on whether a text is empowering or not. Our agreement scores are comparable to those of other computational social science papers on tasks of similar nature, especially when concerning pragmatics. For example, our percentage agreement is higher than that of ElSherief et al. [64]'s dataset on latent hatred, and our agreement is similar to that of the Microaggression dataset [22].

7.4 Data Analysis

We present preliminary analyses of TalkUp. Empowerment is a nuanced phenomenon in pragmatics and deeper exploration of social and linguistic variables remains open for future work. The analyses we present here provide some initial, surface-level insights into what makes language empowering.

7.4.1 Characteristics of Empowering Language

We use the LIWC-22 software to compute LIWC features for all annotated posts [19]. These features measure the percentage of word overlap between the text and predefined lexicons that capture different social and psychological characteristics of language, such as prosocial words or words associated with positive tone. For a more concise and generalized analysis, some related features are combined into compound features: the *I* and *You* features are grouped into one feature *I+You*, *We* and *They* into *We+They*[3], and *male* and *female* into *gendered words*. We standardize LIWC feature scores using the mean and variance calculated from TalkUp's randomly sampled posts. Model-surfaced posts are excluded as they may not reflect the distribution of Reddit posts in the wild.

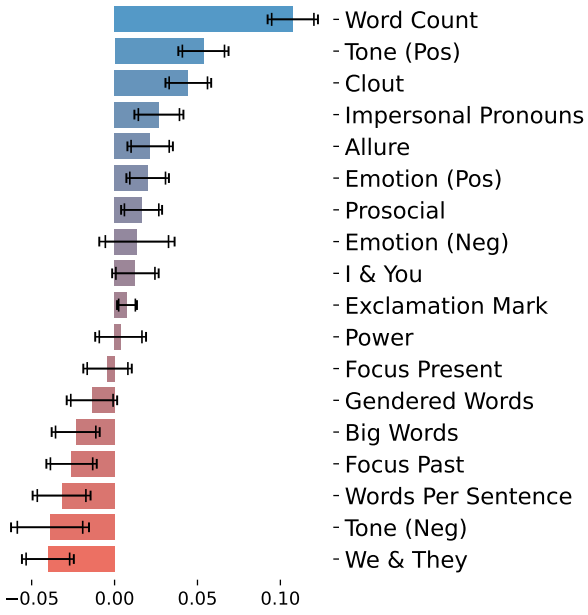


Figure 7.2: Weights of LIWC features with 90% and 95% confidence intervals assigned by linear regression model trained with TalkUp. All features except for Negative Emotion, Power, Focus Present have statistically significant weights ($p < 0.1$).

To understand how each of these features contributes to empowerment in language, we train a linear regression model to predict the likelihood of a post being empowering. Figure 7.2 shows the regression coefficients assigned to each feature. Looking at the positive coef-

[3]We combined I/you and we/they because they consistently followed the same patterns. This was also motivated by qualitative analysis: I/you occurred frequently in language that was addressed directly to the other conversant, which was common in empowering posts, whereas disempowering posts often dismissed a group vaguely without addressing the conversant as an individual, leading to greater use of we/they.

ficients reveals that empowerment is associated with lexical features like *clout*, *allure*, *prosocial words*, and *exclamation marks*. Meanwhile, disempowerment is associated with features that have negative coefficients, such as big words and words-per-sentence, which may indicate sentence complexity. We expand on a few of the most notable findings below.

Tone vs. Emotion. We find that the *tone* of language is more influential to empowerment than the *emotion* conveyed. Positive tone has a significantly higher coefficient than positive emotion; likewise, negative tone is highly associated with disempowerment, while negative emotion is not statistically significant. This suggests that the concept of empowerment is distinct from sentiment and cannot be captured by sentiment analysis models alone.

Power. Power is not a statistically significant feature in predicting empowerment. This corroborates the idea that empowerment is not the same as power – empowerment is a more nuanced and subtle concept that extends beyond power-related lexicons, relying more on the implications between the lines like the tone of the message.

Singular vs. Plural Pronouns. Interestingly, empowerment and disempowerment tend to use different types of pronouns. Singular pronouns (*I, you*) are positively associated with empowering language, while plural pronouns (*we, they*) are linked to disempowering language. Our manual inspections suggest one possible explanation: people who write empowering posts tend to speak directly to the listener, and also include elements of their own personal experience, hence the prevalence of *you* and *I* pronouns. Disempowering conversations are less personal and individualized, often making generalized assumptions or judgments about people.

7.4.2 Empowering Language by Gender

As a preliminary analysis of empowerment across one social dimension, we explore the differences in empowering posts written by men and women. First, we standardize the LIWC feature values for men and women's empowering language over the entire dataset. We find that women's empowering language displays significantly higher levels of positive tone and positive emotions than men. Women also use more exclamation points, while men use more swear words. These findings align with prior works in sociolinguistics that have associated exclamation points with higher expressiveness and excitability [9, 87, 220], which is usually more socially acceptable for women. Meanwhile, men's use of strong or offensive language is linked with masculinity or aggressiveness, and is less socially accepted in women. Additionally, there are other features where women and men's empowering posts diverge – women use more present tense than men, and men are much less likely to use gendered words.

We then control for gender, comparing men's empowering language with all men's posts, and likewise for women. The results show that positive tone, positive emotions, and exclamation marks remain strongly correlated with empowering language even after accounting for gender. However, considering gender does impact the degree of positivity and the use of exclamation marks. Men's empowering language, when compared to men's average language, displays a greater increase in positive tone, positive emotions, and the use of exclamation marks compared to women's empowering language in relation to their average language. This suggests that men tend to exhibit a more pronounced shift towards positive and expressive language when expressing empowerment, whereas women's empowering language already aligns closely with their overall language patterns. Our findings highlight the complex interplay between language,

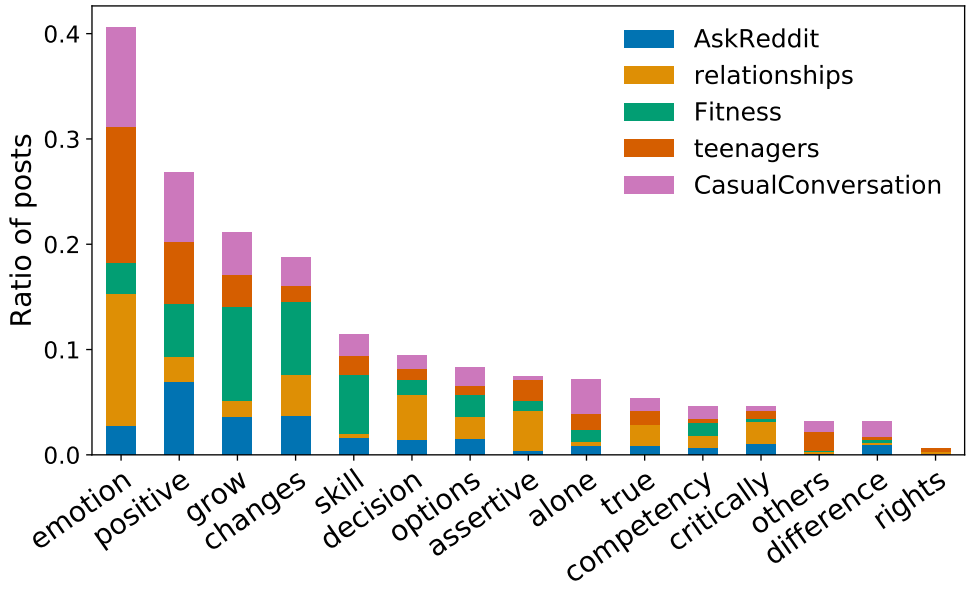


Figure 7.3: Distribution of empowering reasons. One post can have more than one empowering reason.

gender, and empowerment, motivating future research into the influence of social factors on communication of empowerment.

7.4.3 Reasons Why Posts Are Empowering

Figure 7.3 illustrates the distribution of reasons selected by at least two annotators for why a post was empowering/disempowering, broken down by subreddit. The most common reasons a post was considered empowering are encouraging expression of emotions (40.6%), supporting the reader's self-image (26.8%), and supporting the reader's ability to grow (21.1%) and change (18.8%).

Notably, there are significant differences in the reasons most commonly used in different subreddits. For example, the *teenagers* and *relationships* subreddits tend to empower users by promoting expression of emotions, while empowerment in *Fitness* was more focused on encouraging people to improve themselves and make changes. The unique distributions of reasons among different communities and topics of discussion suggests that empowerment serves diverse purposes and implies different meanings depending on the context. Future work could explore which techniques should be used to empower people in specific contexts, such as empowering clients in clinical psychology or students in educational settings, based on the desired interaction goals.

7.4.4 Empowerment and Poster-Commenter Alignment

While a commenter can take either an agree, neutral, or disagree stance with the poster, most empowering posts were in conversations where the poster and commenter *agreed* (79.6%). Like-

	RoBERTa		GPT-3	
Input Type	F1	Acc	F1	Acc
Post	63.5	77.7	36.9	59.7
+response	66.1	78.3	31.5	52.1
+context	65.5	77.9	37.5	64.2
+all	**67.1**	**78.4**	**38.2**	**67.1**

Table 7.2: Model performance of RoBERTa-based classifier fine-tuned on TalkUp and GPT-3 without fine-tuning. RoBERTA-+all is significantly better than RoBERTA-Post in terms of F1 (p¡0.1).

wise, most disempowering posts occurred when the poster and commenter *disagreed* (45.5%). Intuitively, this makes sense for the majority of cases – people often respond agreeably to empowerment and negatively to disempowerment.

Importantly, however, this is not always the case: empowering posts can sometimes have commenters who *disagree*, and disempowering posts can have commenters who *agree*. These cases often involve more complex pragmatics. Empowering posts that contain toxic positivity are frequently met with disagreement, and sometimes commenters will reject or minimize empowering compliments for the sake of politeness. Empowerment can also be met with antagonism from an ill-intentioned commenter, regardless of how genuine the original post may be. Disempowering posts that disparage a particular group might receive an agreeing comment from someone who also shares that view of the group. Additionally, posts that discuss heavy topics may cause a reader to feel disempowered, but they may still respond with an agreeing stance by supporting the original poster (such as by sympathizing with their struggles or validating their feelings). Overall, the empowering-disagree and disempowering-agree cases provide a rich corpus for studying implicature and interactions in social contexts.

7.4.5 Modeling Empowering Language

To explore how well empowerment can be captured by computational methods, we present empowerment detection experiments with two large language models: fine-tuned RoBERTa and zero-shot GPT-3.[4] We note that our goal here is not to build a state-of-the-art model, but to give a general picture of how well existing models work and to illustrate the usefulness of our dataset.

Fine-tuned RoBERTa. We assess how well empowerment can be identified by a pre-trained RoBERTa model [140] fine-tuned on TalkUp, and we conduct an ablation study to examine the importance of contextual information in helping the model classify a post as empowering, disempowering, or neutral. We test four model variants: post, +response (post and response), +context

[4]We opt to experiment with zero-shot rather than few-shot settings because fine-tuning a model of GPT-3's scale is impractical for most users, and because our preliminary experiments indicated that few-shot prompts resulted in lower performance than zero-shot. Although in-context examples often improve performance, there are cases in which few-shot underperforms zero-shot due to models becoming excessively fixated on the provided examples and struggling to generalize effectively. This phenomenon is documented in numerous previous studies (e.g. 66), and we consistently observed this in our case.

(post, posters' gender, subreddit), +all (post, response, context). We divide 1733 unambiguous samples from TalkUp into 60:20:20 for train:validation:test sets and select the model with best validation macro-f1 score.

Table 7.2 presents the average macro-f1 scores across 10 separate runs using different random seeds on the test set. The results show that additional context improves model performance.

Zero-Shot GPT-3. Additionally, we evaluate GPT-3 Davinci's [26] ability to detect empowerment using prompts. We design seven different prompts for each of the four combinations of post+context, and generate responses. While most of GPT-3's responses are single word (e.g. "empowering"), some are longer. To map GPT-3's responses to empowerment labels, we use a simple lexical counting method: if the generated text contains more empowering-related words (e.g. empowering, empowered, empower) than words related to other labels, it is classified as empowering. GPT-3's final classification for each post takes the majority vote over its responses to the seven prompts.

Our results indicate that GPT-3 performs poorly in zero-shot settings compared to RoBERTa-based classifiers fine-tuned on TalkUp. This reveals that even large language models cannot effectively capture empowering language, highlighting the importance of having a carefully annotated dataset of nuanced examples like TalkUp.

7.4.6 Ambiguity of Empowering Language

TalkUp contains 228 samples that either were labeled as "ambiguous" by at least two annotators, or were labeled "no consensus" because all three annotators marked different answers for the empowerment question. We qualitatively analyzed this subset of TalkUp, and we find that these ambiguous posts are not "bad data," but rather are linguistically interesting *because* they are ambiguous – they are examples of language that could reasonably be interpreted in several different ways.

For example, the post "*Maybe call a relative or friend who has a car? Youll figure it out. I wish you luck, kid.*" was unanimously labelled as "empowering" and "ambiguous" by annotators. This makes sense – the post overall seems to provide a helpful suggestion, but calling the responder "kid" could be interpreted in different ways (e.g. as an endearing nickname vs. a condescending title) depending on the social relationship between the poster and the responder. Notably, many of the posts with inherent ambiguity display *sarcasm*, such as the posts "i love you too?!" and "thats grimy as f*ck but sure you do that." Sarcasm, by design, disguises a negative message in positive words, and so a sarcastic post could be interpreted either way depending on whether the sarcasm was meant positively or negatively.

We also investigated how GPT-3 handles such ambiguous cases. We find that GPT-3 tends to classify them as neutral, even for explicitly empowering posts such the above example. Instances in which the posts carried a sarcastic tone were commonly interpreted by GPT-3 as neutral as well, indicating that simultaneously empowering and ambiguous language is poorly understood by the model. The fact that ambiguity is still challenging for large models motivates the need for further work in this area, and TalkUp provides diverse examples of ambiguous language that can be used to to work towards this end.

	% Empower		% Disempower	
	Post	Response	Post	Response
r/AskReddit	12.0	14.1	6.8	5.6
r/relationship	38.7	27.2	12.7	11.4
r/Fitness	30.0	28.3	7.2	5.6
r/teenager	24.2	24.8	6.3	5.7
CasualConversation	25.6	29.2	2.8	2.3
Overall	15.2	16.5	6.9	5.8

Table 7.3: The percentage of empowering and disempowering posts and responses in each subreddit.

7.5 Example Application: Unearthing Empowerment Patterns on Reddit

As a case study, we demonstrate how TalkUp and the trained empowerment classifier can be used to uncover interesting patterns in how people use empowering language. Specifically, we apply the trained classifier in §7.4.5 to generate empowerment labels of *all* Reddit posts and responses in RtGender, to learn about how both posters and responders communicate.[5] We analyze empowering and disempowering posts in different subreddits and by different gender of the poster and responder.

By Subreddit Table 7.3 shows the percentage of empowering and disempowering posts and responses in thefive subreddits of TalkUp. The results indicate that the subreddits have significantly different degrees of empowerment, and that and certain subreddits (e.g. relationship, Fitness) are significantly more empowering than others (e.g. AskReddit). Our model can be used to monitor the overall empowerment level of communities and identify unusual patterns, such as a significant rise in disempowerment. Furthermore, we find that there are more empowering responses than posts in total. On the contrary, there are more disempowering posts than responses across all subreddits. This may be because responses are often directed towards specific posts or users, and as a result, the writer may be more conscious of their tone and try to be more empowering compared to posts.

By poster and responder gender Table 7.4 shows the percentage of empowering and disempowering context by the gender of posters and responders. Overall, women seem to post and interact with more empowering content. Unsurprisingly, the results show that of all the posts predicted to be empowering, women wrote a considerably higher percentage of them than men. Interestingly, however, women are also responsible for a slightly higher percentage of *disempowering* posts than men. Another surprising finding is that posts written by men that were

[5]Given that responses are only available for the posts and not for the responses, and that some samples in the data do not provide the gender of the responder, we used a model that only incorporates subreddit information as additional context to the text itself.

		Post		Response	
Poster	Responder	%E	%D	%E	%D
Man	Man	13.4	6.5	13.8	5.9
	Woman	16.2	7.1	18.1	6.0
Woman	Man	16.5	6.9	16.7	6.3
	Woman	20.2	7.3	20.4	6.4

Table 7.4: The percentage of empowering (%E) and disempowering (%D) posts and responses in RtGender classified by the model trained with TalkUp, broken down by the gender of both the poster and responder.

commented on by women tend to be more empowering or more disempowering than those commented on by men, suggesting that women not only post more empowerment-charged language, but they also *engage* with more empowerment-charged posts. This may be tied to factors like the topics or types of posts that women tend to engage with and could be used to answer sociological questions about gender and social media.

7.6 Related Work

To our knowledge, Mayfield et al. [144] is the only prior work exploring empowerment in NLP, but the contributions of our works are quite different. Mayfield et al. [144] primarily focus on an algorithm for predicting rare classes and use empowerment as an example. In contrast, we focus on understanding empowering language itself, before developing automated detection tools. We explore the reasons behind empowerment, considering multiple dimensions of social context such as gender, topic, and poster-commenter alignment. Mayfield et al. [144] use non-public data from a specific cancer support group, while TalkUp spans diverse topics and user bases, making our scope broader and more generalizable.

As empowering language is not well understood in NLP, our work has also drawn insights from research on related concepts:

Power. Danescu-Niculescu-Mizil et al. [44] develop a framework for analyzing power differences in social interactions based on how much one conversant echoes the linguistic style of the other. Prabhakaran and Rambow [171, 172] predict power levels of participants in written dialogue from the Enron email corpus, and several other of their works explore power dynamics in other contexts, such as gender [174] and political debates [170].

Our work studies *empowerment* rather than power. Power is certainly a closely related concept, but empowerment is a distinct linguistic phenomenon – it concerns not just static power levels, but interactions that *increase or decrease* a person's power, and it is also a broader concept that encompasses things like self-fulfillment and self-esteem. While power has primarily been analyzed at the word level, such as by examining connotations of particular verbs [166, 191], our work attempts to look at higher-level pragmatics – implications that may not be captured by word choice alone, but suggested between the lines.

Condescension. The closest concept to empowerment that has been more thoroughly studied

in NLP is *condescension*. Prior works have defined condescension as language that is not overtly negative, but that assumes a status difference between the speaker and listener that the listener disagrees with [102]. Intuitively, condescension can be interpreted as roughly the *opposite* of empowerment: it implicitly suggests that the listener has lower status or worth.

Our work particularly builds upon Wang and Potts [218]: they develop TalkDown, a dataset of Reddit posts labeled as "condescending" or "not condescending." Specifically, they identify condescending *posts* by looking for *replies* that indicate the original post is condescending. Our approach is parallel to this work: we likewise surface Reddit posts whose *responses* indicate that the *original post* is empowering (thus aligning with our definition of empowerment in §7.2 as an effect on the listener). TalkUp complements TalkDown by focusing on the positive aspect of such language: instead of only identifying text as condescending or not condescending, we distinguish between disempowering, empower, and neutral posts.

7.7 Conclusions and Future Work

We explore the problem of empowerment detection, grounding it in relevant social psychology and linguistics literature. To facilitate studies of empowerment, we create TalkUp, a high-quality dataset of Reddit posts labeled for empowerment and other contextual information. Our preliminary analyses demonstrate that empowerment is not captured by existing NLP methods and models, but that it can be detected with our dataset. Furthermore, we demonstrate the importance of social context in understanding empowering language with different genders, poster-commenter alignments, and topics of discussion. In studying empowerment, we work towards bigger open challenges in pragmatics, implicature, and social context in NLP.

While our primary contribution lies in empowerment classification and detection, TalkUp offers potential beyond these tasks. We believe that a classifier trained with our data could be used not only for detection but also to *generate* more empowering language. Similar to the approach in Sharma et al. [195], such a classifier could assign rewards to tailor a generation model to produce more empowering outputs. Moreover, an empowerment classifier could be employed for controllable text generation with constrained decoding, as demonstrated by Yang and Klein [225], Liu et al. [138], and Kumar et al. [124]. Additionally, a model capable of controlling for empowerment could suggest edits to human-written text, making it more empowering—a feature with potential applications in real-world dialogue settings such as education and psychotherapy.

TalkUp currently focuses on simple two-turn interactions involving three social variables (gender, alignment, and topic), but its framework can be extended to more complex social interactions. For example, power dynamics can be influenced by various other social roles, including occupation (e.g., manager vs. employee), race (e.g., White vs. Person of Color), and age (e.g., older vs. younger person). Different combinations of these identities may result in further intersectional dynamics, as discussed by Crenshaw [42], Collins and Bilge [39], and Lalor et al. [128]. Additionally, since most real-world conversations involve long back-and-forth exchanges, we encourage future work to explore empowerment in multi-turn dialogues, expanding the scope and applicability of empowerment research in NLP.

Ethical Considerations

In constructing our study, we took precautions to ensure the task design, data collection and handling are done ethically and according to current recommended practices and guidelines [13, 82, 151, 210]. Specifically, we ensured fair compensation by calculating the pay based on minimum wage in CA (higher than then the average pay worldwide, including most U.S. states). To avoid exposing the annotators to potentially offensive or otherwise harmful content from social media, we manually checked every data sample. Beyond scientific goal of our work to understand sociolinguistic characteristics of empowering language and open new directions to NLP research on deeper pragmatic phenomena, the practical goal is to advance NLP technologies with positive impact through understanding and incorporating empowerment in practical applications including education, therapy, medicine, and more.

Chapter 8

Ethical Considerations

While this thesis contributes to the advancement of socially aware NLP, it is essential to acknowledge the limitations and potential risks associated with this research. By understanding these challenges and proactively addressing them, we can work towards developing NLP technologies that are both beneficial and ethical.

Limitations While this thesis aims to address the broader implications of social context in NLP, several limitations constrain the scope of our research. First, our focus on a generalizable model has necessitated a selection of representative cultures and communities, which may not fully capture the nuances and complexities of all human societies. In particular, our analysis was primarily centered around languages and communities with abundant digital resources, such as those represented on platforms like Reddit. Additionally, our exploration of personal context was limited by the available data, often focusing on demographic information like gender and age. Finally, as researchers situated within a specific academic and cultural context, our perspectives inevitably carry biases that may not fully represent the diverse experiences of those impacted by these technologies.

Simplifications and Categorizations To facilitate computational analysis, we have employed simplifying assumptions and categorical divisions. For example, our use of gender and cultural group categories, while useful for operationalizing our research, may oversimplify the complexities of individual and group identities. It is important to acknowledge that these categorizations can obscure nuances and potentially reinforce harmful stereotypes.

Potential Misuse and Risks The methodologies and frameworks developed in this thesis, while intended for beneficial purposes, carry inherent risks of misuse. For instance, techniques designed to measure public sentiment and cultural similarities could be exploited to manipulate public opinion or reinforce harmful stereotypes. Similarly, the ability to detect community norms might be used to target specific groups. Moreover, the development of socially aware language models could lead to the creation of tools for generating personalized misinformation or perpetuating biases. To mitigate these risks, it is imperative to develop robust safeguards, promote transparency, and foster ongoing critical evaluation of these technologies.

Furthermore, the customization and tailoring required to make models more responsive to social contexts may increase their complexity, leading to higher energy and carbon costs. As we advance towards more socially aware and adaptive models, it is imperative to concurrently invest in research focused on making these models more energy-efficient and sustainable.

Data and Privacy The use of social context inherently involves the handling of sensitive personal information. While the datasets used in this research were publicly available, it is essential to recognize that the individuals who generated this content did not explicitly consent to its use in our specific study. To protect user privacy, we have taken steps to anonymize data and avoid identifying individuals. However, as the field of NLP continues to evolve, it is crucial to develop more sophisticated techniques for ensuring data privacy and ethical use.

Looking forward, the thesis advocates for ongoing research into the interpretability and controllability of NLP models. Such research is essential to developing models that are not only socially aware but also respectful of user privacy, ensuring they are neither overly intrusive nor harmful. By advancing in these areas, future research can contribute to the development of ethical, socially aware models that balance innovation with the protection of individual rights.

Chapter 9

Conclusions

9.1 Summary of Contributions

This thesis presents the efforts to demonstrate the vital connection between NLP and sociocultural knowledge by presenting a series of models, datasets, and analyses. We demonstrate how incorporating cultural and community context can improve NLP models and how NLP models can be used to study social phenomena and validate social science theories quantitatively. These endeavors hold great potential in shedding light on various aspects of NLP and social science and particularly in how they can complement and enrich one another. Specific contributions of this thesis include:

- Developed methods to measure public sentiment towards different cultures and connect its impact on language change.
- Demonstrated how knowledge of cultural similarities between languages can guide cross-lingual transfer to improve NLP model performance.
- Introduced a framework to measure implicit norms and values within communities to better understand community-specific language norms.
- Proposed a new task, community-sensitive norm violation detection, and improved community moderation models by incorporating community context
- Explored the use of individuals' social attributes to identify and model social biases in corpora, focusing on personal context.
- Proposed methods to incorporate personal contexts into text classification, such as empowering language detection, significantly improving model performance.

9.2 Discussion and Future Work

This thesis has laid the groundwork for developing socioculturally aware NLP technologies that can engage with complex social contexts. While our research has made significant progress, there are several promising avenues for future exploration.

Interpretable and Controllable Personalization A promising avenue for future research is the development of personalized NLP models that are both interpretable and controllable. By understanding and incorporating individual context, these models can create highly tailored user experiences. However, this personalization comes with risks, as models may access sensitive information that, if misused,could lead to discomfort or mistrust.

To mitigate these risks, models must be transparent about the attributes or contexts used for personalization and provide users with control over the information they share. This approach balances the benefits of personalization with privacy concerns. A critical aspect of achieving this goal is the way context is provided to models. Moving forward, it will be important to develop models that can accept context as raw input (such as through retrieval-augmented generation) rather than relying on distributional representations of context, which are often difficult to interpret or control. This shift would allow users greater transparency and agency in how their personal data is used by the model.

Adaptable and Interactive Socially-Aware AI Our research has primarily assumed that social context is readily available. However, in real-world situations, this information is often incomplete or uncertain. Future work should focus on creating models that can adaptively adjust to varying social contexts, acknowledging the inherent uncertainty and fluidity of human interactions. Additionally, there is significant potential for developing interactive systems that actively seek out relevant social context. Instead of passively relying on pre-existing information, these models could engage in dynamic conversations with users, asking clarifying questions or requesting additional details. This would extend our current static models, making them more adaptable and interactive, reflecting the complexity of human communication.

Pluralistic and Dynamic Values Another important consideration for future research is the recognition of pluralistic and dynamic values within social contexts. In the work presented in this thesis, there is an implicit assumption that cultural values and norms are relatively static and distinct between groups. However, in reality, individuals often possess multicultural backgrounds, and their values can be fluid, evolving over time and across different contexts. This recognition is particularly relevant in the context of community norms and values, which are central to many of the models presented in this thesis. While the methodologies presented here can be applied to different time periods and can measure shifts in norms, the dynamic nature of these values was not the primary focus. Future research could explore how to better integrate these pluralistic and evolving perspectives about social context into NLP models, allowing them to more accurately reflect the complexity of human values and social identities.

Social Context and Multimodality The work in this thesis has focused exclusively on textual input and output, yet much of social context is conveyed through other modalities—such as intonation, facial expressions, and visual cues. These non-verbal signals are integral to understanding social interactions and are essential for creating more socially aware language agents. A promising direction for future research is the development of multimodal models that can process and integrate inputs from multiple sources, such as vision and speech. This would significantly enhance the ability of NLP models to understand and respond to social context, particularly in

applications involving robots or other tangible machines that must interact with humans in a three-dimensional world. However, the incorporation of multimodal signals also raises concerns about the potential for increased social biases. As models begin to process more complex and diverse inputs, there is an increased risk of reinforcing stereotypes and implicit biases. Future research must carefully consider these risks and develop strategies to mitigate them, ensuring that the expansion into multimodality enhances rather than undermines the social awareness of AI systems.

Bibliography

[1] 2003—2023. The russian national corpus. https://ruscorpora.ru. Accessed: 2023-10-08.

[2] Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard H. Hovy, Kai-Wei Chang, and Nanyun Peng. 2018. Near or far, wide range zero-shot cross-lingual dependency parsing. *CoRR*, abs/1811.00570.

[3] Joanna Almeida, Renee M. Johnson, Heather L. Corliss, Beth E. Molnar, and Deborah Azrael. 2009. Emotional distress among lgbt youth: The influence of perceived discrimination based on sexual orientation. *Journal of Youth and Adolescence*, 38(7):1001–1014.

[4] Hind Almerekhi, Supervised by Bernard J Jansen, and co-supervised by Haewoon Kwak. 2020. Investigating toxicity across multiple reddit communities, users, and moderators. In *Companion Proceedings of the Web Conference 2020*, pages 294–298.

[5] Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

[6] Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. In *Proc. ACM Hum.-Comput. Interact.*, volume 3, page 88. ACM.

[7] René Appel and Pieter Muysken. 2005. *Language contact and bilingualism*. Amsterdam University Press.

[8] Kimberly F. Balsam, Yamile Molina, Blair Beadnell, Jane Simoni, and Karina Walters. 2011. Measuring multiple minority stress: The lgbt people of color microaggressions scale. *Cultur Divers Ethnic Minor Psychol.*, 17(2):163–174.

[9] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

[10] Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proc. of WMT*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

[11] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

[12] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

[13] Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

[14] Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. 2024. Stela: a community-centred approach to norm elicitation for ai alignment. *Scientific Reports*, 14(1):6616.

[15] Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew R. Gormley. 2023. Unlimiformer: Long-range transformers with unlimited length input. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[16] Cristina Bicchieri, Ryan Muldoon, and Alessandro Sontuoso. 2023. Social Norms. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2023 edition. Metaphysics Research Lab, Stanford University.

[17] Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

[18] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2021. Improving language models by retrieving from trillions of tokens. *CoRR*, abs/2112.04426.

[19] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*.

[20] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

[21] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

[22] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In

Proc. of EMNLP-IJCNLP, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

[23] Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[24] Andrew D Brown. 2022. Identities in and around organizations: Towards an identity work perspective. *Human relations*, 75(7):1205–1237.

[25] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

[26] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

[27] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

[28] Ewa S Callahan and Susan C Herring. 2011. Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10):1899–1915.

[29] Rafael Monroy Casas and JM Hernández Campoy. 1995. A sociolinguistic approach to the study of idioms: Some anthropolinguistic sketches. *Cuadernos de Filología inglesa*, 4.

[30] Logan S. Casey, Jesse Chandler, Adam Seth Levine, Andrew Proctor, and Dara Z. Strolovitch. 2017. Intertemporal differences among mturk workers: Time-based sample variations and implications for online data collection. *SAGE Open*, 7(2):2158244017712774.

[31] Judi Chamberlin. 1997. A working definition of empowerment. *Psychiatric Rehabilitation Journal*, 20(4):43–46.

[32] Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. Norms matter: Contrasting social support around behavior change in online weight loss communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

[33] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. In *Proc. of CSCW*, pages 1–22, New York. ACM.

[34] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

[35] Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4743–4754, Hong Kong, China. Association for Computational Linguistics.

[36] Yo Joong Choe, Kyubyong Park, and Dongwoo Kim. 2020. word2word: A collection of bilingual lexicons for 3,564 language pairs. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*.

[37] Munmun Choudhury, Shagun Jhaver, Benjamin Sugar, and Ingmar Weber. 2016. Social media participation in an activist movement for racial equality. *Proceedings of the ... International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media*, 2016:92–101.

[38] Heather M Clarke and Kara A Arnold. 2018. The influence of sexual orientation on the perceived fit of male applicants for both male- and female-typed jobs. *Frontiers in Psychology*, 9:656.

[39] Patricia Hill Collins and Sirma Bilge. 2020. *Intersectionality*. John Wiley & Sons.

[40] Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

[41] Walter Coutu. 1951. Role-playing vs. role-taking: An appeal for clarification. *American sociological review*, 16(2):180–187.

[42] Kimberle Crenshaw. 1990. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.*, 43:1241.

[43] Béatrice Daille. 1994. *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. Ph.D. thesis, Ph. D. thesis, Université Paris 7.

[44] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2011. Echoes of power: Language effects and power differences in social interaction.

[45] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and

Christopher Potts. 2013. No country for old members: user lifecycle and linguistic change in online communities. *Proceedings of the 22nd international conference on World Wide Web*.

[46] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: user lifecycle and linguistic change in online communities. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 307–318. International World Wide Web Conferences Steering Committee / ACM.

[47] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proc. of FAT*, pages 120–128. ACM.

[48] Erik W de Kwaadsteniet, Toko Kiyonari, Welmer E Molenmaker, and Eric van Dijk. 2019. Do people prefer leaders who enforce norms? reputational effects of reward and punishment decisions in noisy social dilemmas. *Journal of Experimental Social Psychology*, 84:103800.

[49] Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Long papers*.

[50] Marco Del Tredici and Raquel Fernández. 2018. The road to success: Assessing the fate of linguistic innovations in online communities. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1591–1603, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[51] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[52] Larry Dignan. 2024. Reddit's data licensing play: Do you want your llm trained on reddit data?

[53] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18.

[54] Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

[55] MeiXing Dong, David Jurgens, Carmen Banea, and Rada Mihalcea. 2019. Perceptions of social roles across cultures. In *International Conference on Social Informatics*, pages 157–172. Springer.

[56] Bryan Dosono and Bryan C. Semaan. 2019. Moderation practices as emotional labor in

sustaining online communities: The case of AAPI identity work on reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 142. ACM.

[57] William Downes. 1998. *Language and society*. Cambridge University Press.

[58] Timothy Dozat and Christopher D. Manning. 2016. Deep biaffine attention for neural dependency parsing. *CoRR*, abs/1611.01734.

[59] Penelope Eckert. 1989. *Jocks and burnouts: Social categories and identity in the high school*. Teachers College Press.

[60] Penelope Eckert and Sally Mcconnell-Ginet. 1999. New generalizations and explanations in language and gender research. *Language in Society*, 28:185 – 201.

[61] Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and Gender*, 2 edition. Cambridge University Press.

[62] Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and gender*. Cambridge University Press.

[63] Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2018. How to make causal inferences using texts. *Working Paper*.

[64] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[65] Ethan Fast and Eric Horvitz. 2016. Identifying dogmatism in social media: Signals and models. In *Proc. of EMNLP*, pages 690–699, Austin, Texas. Association for Computational Linguistics.

[66] Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14014–14031, Toronto, Canada. Association for Computational Linguistics.

[67] Christiane Fellbaum. 2012. Wordnet. *The encyclopedia of applied linguistics*.

[68] Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. 2019. Contextual affective analysis: A case study of people portrayals in online #metoo stories. In *ICWSM*.

[69] Anjalie Field, Chan Young Park, Kevin Z. Lin, and Yulia Tsvetkov. 2022. Controlled analyses of social biases in Wikipedia bios. In *Proc. The ACM Web Conference '22*.

[70] Anjalie Field and Yulia Tsvetkov. 2019. Entity-centric contextual affective analysis. In *Proc. of ACL*, pages 2550–2560, Florence, Italy. Association for Computational Linguistics.

[71] Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised discovery of implicit gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 596–608, Online. Association for Computational Linguistics.

[72] Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised discovery of implicit gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 596–608.

[73] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

[74] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit rules! characterizing an ecosystem of governance. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

[75] Casey Fiesler and Nicholas Proferes. 2018. "participant" perceptions of Twitter research ethics. *Social Media + Society*, 4(1):2056305118763366.

[76] Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert Mullins. 2024. Inverse constitutional ai: Compressing preferences into principles.

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

[78] Andrew R. Flores. 2019. Social acceptance of lgbt people in 174 countries. https://williamsinstitute.law.ucla.edu/publications/global-acceptance-index-lgbt/. Accessed: 2021-04-17.

[79] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

[80] Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response rankingtraining with large-scale human feedback data. In *EMNLP*.

[81] Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying cross-cultural differences in word usage. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 674–683, Osaka, Japan. The COLING 2016 Organizing Committee.

[82] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

[83] Carroll J Glynn and Michael E Huge. 2007. Opinions as norms: Applying a return potential model to the study of communication behaviors. *Communication Research*, 34(5):548–568.

[84] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

[85] Venkata S Govindarajan, Kyle Mahowald, David I Beaver, and Junyi Jessy Li. 2023. Counterfactual probing for the influence of affect and specificity on intergroup bias. *arXiv preprint arXiv:2305.16409*.

[86] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90.

[87] Emre Güvendir. 2015. Why are males inclined to use strong swear words more than females? an evolutionary explanation based on male intergroup aggressiveness. *Language Sciences*, 50:133–139.

[88] Edward Twitchell Hall. 1989. *Beyond culture*. Anchor.

[89] Angelique Harris, Juan Battle, J.R. Pastrana, Antonio (., and Jessie Daniels. 2013. The sociopolitical involvement of black, latino, and asian/pacific islander gay and bisexual men. *Journal of Men's Studies*, 21(3):236–254.

[90] Martin Haspelmath. 2008. Loanword typology: Steps toward a systematic cross-linguistic study of lexical borrowability. *Empirical Approaches to Language Typology*, 35:43.

[91] Martin Haspelmath and Uri Tadmor, editors. 2009. *Loanwords in the World's Languages: A Comparative Handbook*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

[92] Einar Haugen. 1950. The analysis of linguistic borrowing. *Language*, pages 210–231.

[93] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

[94] Libby Hemphill and Jahna Otterbacher. 2012. Learning the lingo? gender, prestige and linguistic adaptation in review communities. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, page 305–314, New York, NY, USA. Association for Computing Machinery.

[95] David B Henry, Jennifer Cartland, Holly Ruchross, and Kathleen Monahan. 2004. A return potential measure of setting norms for aggression. *American Journal of Community Psychology*, 33(3-4):131–149.

[96] John Herrman. 2021. Everything's a joke until it's not.

[97] Jack Hessel and Lillian Lee. 2019. Something's brewing! early prediction of controversy-causing posts from discussion features. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1648–1659, Minneapolis, Minnesota. Association for Computational Linguistics.

[98] Jack Hessel, Chenhao Tan, and Lillian Lee. 2016. Science, askscience, and badscience: On the coexistence of highly related communities. In *International Conference on Web and Social Media*.

[99] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan

Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2024. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

[100] Geert H Hofstede, Gert Jan Hofstede, and Michael Minkov. 2005. *Cultures and organizations: Software of the mind*, volume 2. Mcgraw-hill New York.

[101] Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

[102] Thomas Huckin. 2002. Critical discourse analysis and the discourse of condescension. *Discourse studies in composition*, pages 155–176.

[103] Jay Jackson. 1966. A conceptual and measurement model for norms and roles. *The Pacific Sociological Review*, 9(1):35–47.

[104] Jay Jackson. 1975. Normative power and conflict potential. *Sociological Methods & Research*, 4(2):237–263.

[105] Joshua Conrad Jackson, Joseph Watts, Teague R. Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray, and Kristen A. Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

[106] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

[107] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–35.

[108] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):1–33.

[109] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

[110] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the looking glass: Study of transparency in reddit's moderation practices. *Proceedings of the ACM on Human-Computer Interaction*, 4(GROUP):1–35.

[111] David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Flo-

rence, Italy. Association for Computational Linguistics.

[112] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

[113] Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. 2017. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 262–272, Valencia, Spain. Association for Computational Linguistics.

[114] Anna Kasunic and Geoff Kaufman. 2018. " at least the pizzas you make are hot": Norms, values, and abrasive humor on the subreddit r/roastme. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

[115] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc.

[116] Katherine A. Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proc. of ACL*, pages 5332–5344, Online. ACL.

[117] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

[118] Daniel Kershaw, Matthew Rowe, and Patrick Stacey. 2016. Towards modelling language innovation acceptance in online social networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, page 553–562, New York, NY, USA. Association for Computing Machinery.

[119] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

[120] Charles Kiene, Jialun Aaron Jiang, and Benjamin Mako Hill. 2019. Technological frames and user innovation: exploring technological change in community moderation teams. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23.

[121] Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. Surviving an "eternal september": How an online community managed a surge of newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San*

Jose, CA, USA, May 7-12, 2016, pages 1152–1156. ACM.

[122] Fajri Koto, Tilman Beck, Zeerak Talat, Iryna Gurevych, and Timothy Baldwin. 2024. Zero-shot sentiment analysis in low-resource languages using a multilingual sentiment lexicon. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 298–320, St. Julian's, Malta. Association for Computational Linguistics.

[123] Alex Krouglov. 2002. War and peace: Ukrainian and Russian in Ukraine. *Journal of Language and Politics*, 1(2):221–239.

[124] Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled text generation as continuous optimization with multiple constraints. In *Advances in Neural Information Processing Systems*, volume 34, pages 14542–14554. Curran Associates, Inc.

[125] Zoltán Kövecses. 2003. Language, figurative thought, and cross-cultural comparison. *Metaphor and Symbol*, 18(4):311–320.

[126] Zoltán Kövecses. 2010. *Metaphor: A practical introduction*. Oxford University Press.

[127] Sanford Labovitz and Robert Hagedorn. 1973. Measuring social norms. *Pacific Sociological Review*, 16(3):283–303.

[128] John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States. Association for Computational Linguistics.

[129] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik W. Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Gov. Inf. Q.*, 31:317–326.

[130] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

[131] Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. *arXiv preprint arXiv:2203.07450*.

[132] Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.

[133] Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China. Association for Computational Linguistics.

[134] Bill Yuchen Lin, Frank F. Xu, Kenny Zhu, and Seung-won Hwang. 2018. Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 709–719, Melbourne, Australia. Association for Computational Linguistics.

[135] Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

[136] Laura Linnan, Anthony D LaMontagne, Anne Stoddard, Karen M Emmons, and Glorian Sorensen. 2005. Norms and their relationship to behavior in worksite settings: an application of the jackson return potential model. *American journal of health behavior*, 29(3):258–268.

[137] Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

[138] Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

[139] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[140] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[141] Saskia Locke, Anthony Bashall, Sarah Al-Adely, John Moore, Anthony Wilson, and Gareth B. Kitchen. 2021. Natural language processing in medicine: A review. *Trends in Anaesthesia and Critical Care*, 38:4–9.

[142] Li Lucy and David Bamman. 2021. Characterizing English variation across social media communities with BERT. *Transactions of the Association for Computational Linguistics*, 9:538–556.

[143] Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

[144] Elijah Mayfield, David Adamson, and Carolyn Penstein Rosé. 2013. Recognizing rare social phenomena in conversation: Empowerment detection in support group chatrooms. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 104–113, Sofia, Bulgaria. Association for Computational Linguistics.

[145] Charles G McClintock. 1978. Social values: Their definition, measurement and development. *Journal of Research & Development in Education*.

[146] Rachel I McDonald and Christian S Crandall. 2015. Social norms and social influence. *Current Opinion in Behavioral Sciences*, 3:147–151.

[147] April McMahon. 1994. *Understanding language change*. Cambridge University Press.

[148] Ellen Hawley McWhirter. 1991. Empowerment in counseling. *Journal of Counseling & Development*, 69(3):222–227.

[149] Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Front. Artif. Intell.*

[150] Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

[151] Alan Mislove and Christo Wilson. 2018. A practitioner's guide to ethical web data collection. In Brooke Foucault Welles and Sandra González-Bailón, editors, *The Oxford Handbook of Networked Communication*. Oxford University Press.

[152] Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proc. of ACL*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

[153] György Molnár and Zoltán Szüts. 2018. The role of chatbots in formal education. In *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 000197–000202.

[154] Laetitia B Mulder. 2008. The difference between punishments and rewards in fostering moral concerns in social decision making. *Journal of Experimental Social Psychology*, 44(6):1436–1443.

[155] Korac-Kakabadse Nada, Kouzmin Alexander, Korac-Kakabadse Andrew, and Savery Lawson. 2001. Low-and high-context communication patterns: towards mapping cross-cultural encounters. *Cross Cultural Management: An International Journal*, 8(2):3–24.

[156] Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

[157] Yair Neuman and Yochai Cohen. 2023. Ai for identifying social norm violation. *Scientific Reports*, 13(1):8103.

[158] Dat Quoc Nguyen, Dai Quoc Nguyen, Dang Duc Pham, and Son Bao Pham. 2014. RDR-POSTagger: A ripple down rules-based part-of-speech tagger. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 17–20, Gothenburg, Sweden. Association for Computational Linguistics.

[159] Joakim Nivre, Mitchell Abrams, Željko Agić, and et al. 2018. Universal dependencies 2.2. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[160] Jessica M Nolan. 2015. Using jackson's return potential model to explore the normativeness of recycling. *Environment and Behavior*, 47(8):835–855.

[161] Stephen P. Osborne. 1994. The language of empowerment. *International Journal of Public Sector Management*, 7(3):56–62.

[162] Ulrike Oster. 2019. Cross-cultural semantic and pragmatlc profiling of emotion words. regulation and expression of anger in Spanish and German. *Current Approaches to Metaphor Analysis in Discourse*, 39:35.

[163] Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. 2021. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2855–2870, Online. Association for Computational Linguistics.

[164] Chan Young Park, Julia Mendelsohn, Karthik Radhakrishnan, Kinjal Jain, Tushar Kanakagiri, David Jurgens, and Yulia Tsvetkov. 2021. Detecting community sensitive norm violations in online conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3386–3397, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[165] Chan Young Park, Xinru Yan, Anjalie Field, and Yulia Tsvetkov. 2021. Multilingual contextual affective analysis of LGBT people portrayals in wikipedia. In *Proc. ICWSM'21*.

[166] Chan Young Park, Xinru Yan, Anjalie Field, and Yulia Tsvetkov. 2021. Multilingual contextual affective analysis of lgbt people portrayals in wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 479–490.

[167] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

[168] Shana Poplack. 2017. *Borrowing: Loanwords in the Speech Community and in the Grammar*. Oxford University Press.

[169] Shana Poplack, David Sankoff, and Christopher Miller. 1988. The social correlates and linguistic processes of lexical borrowing and assimilation.

[170] Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. 2014. Power of confidence: How poll scores impact topic dynamics in political debates. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 77–82, Baltimore, Maryland. Association for Computational Linguistics.

[171] Vinodkumar Prabhakaran and Owen Rambow. 2014. Predicting power relations between participants in written dialog from a single thread. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages

339–344, Baltimore, Maryland. Association for Computational Linguistics.

[172] Vinodkumar Prabhakaran and Owen Rambow. 2014. Predicting power relations between participants in written dialog from a single thread. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 339–344, Baltimore, Maryland. Association for Computational Linguistics.

[173] Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012. Predicting overt display of power in written dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–522, Montréal, Canada. Association for Computational Linguistics.

[174] Vinodkumar Prabhakaran, Emily E. Reid, and Owen Rambow. 2014. Gender and power: How gender and gender environment affect manifestations of power. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1965–1976, Doha, Qatar. Association for Computational Linguistics.

[175] Martin Pütz, Justyna A Robinson, and Monika Reif. 2014. *Cognitive sociolinguistics: Social and cultural variation in cognition and language use*, volume 59. John Benjamins Publishing Company.

[176] Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

[177] Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *International Conference on Web and Social Media*.

[178] Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 557–568.

[179] Kavel Rao, Liwei Jiang, Valentina Pyatkin, Yuling Gu, Niket Tandon, Nouha Dziri, Faeze Brahman, and Yejin Choi. 2023. What makes it ok to set a fire? iterative self-distillation of contexts and rationales for disambiguating defeasible social and moral situations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12140–12159, Singapore. Association for Computational Linguistics.

[180] Hannah Rashkin, Eric Bell, Yejin Choi, and Svitlana Volkova. 2017. Multilingual connotation frames: A case study on social media for targeted sentiment analysis and forecast. In *Proc. of ACL*, pages 459–464, Vancouver, Canada. Association for Computational Linguistics.

[181] Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proc. of ACL*, pages 311–321, Berlin, Germany. Association for Computational Linguistics.

[182] Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

[183] Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science (Forthcoming)*, 64:887–903.

[184] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[185] Paul R. Rosenbaum and Donald B. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

[186] Han saem Kim. 2016. Construction of yonsei 20th century corpus. *Language Facts and Perspectives*, (37):229.

[187] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Info. processing & management*, 24(5):513–523.

[188] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

[189] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

[190] Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proc. of EMNLP*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

[191] Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

[192] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proc. of SocialNLP*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

[193] Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

[194] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforce-

ment learning approach. *CoRR*, abs/2101.07714.

[195] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. *CoRR*, abs/2101.07714.

[196] Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G. Lucas, Adam S. Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

[197] Qinlan Shen and Carolyn P Rosé. 2022. A tale of two subreddits: Measuring the impacts of quarantines on political engagement on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 932–943.

[198] Bernard J. Siegel. 1977. Encyclopedia of anthropology. David E. Hunter and Phillip Whitten, eds. New York. *American Anthropologist*, 79(2):452–454.

[199] Amit Singhal, Chris Buckley, and Manclar Mitra. 1996. Pivoted document length normalization. In *Proc. of SIGIR*, volume 51, pages 176–184. ACM New York, NY, USA.

[200] Laura Smith, Salvatore Giorgi, Rishi Solanki, Johannes Eichstaedt, H Andrew Schwartz, Muhammad Abdul-Mageed, Anneke Buffone, and Lyle Ungar. 2016. Does 'well-being'translate on twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2042–2047.

[201] Stacy Smith, Marc Choueiti, Katherine Pieper, Traci Gillig, Carmen Lee, and Dylan DeLuca. 2015. Inequality in 700 popular films: Examining portrayals of gender, race, & LGBT status from 2007 to 2014. *Institute for Diversity and Empowerment at Annenberg*.

[202] Aaron J Snoswell, Lucinda Nelson, Hao Xue, Flora D Salim, Nicolas Suzor, and Jean Burgess. 2023. Measuring misogyny in natural language generation: Preliminary results from a case study on two reddit communities. *arXiv preprint arXiv:2312.03330*.

[203] Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. 2023. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *AAAI Conference on Artificial Intelligence*.

[204] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

[205] Sarah Grey Thomason and Terrence Kaufman. 2001. *Language contact*. Edinburgh University Press Edinburgh.

[206] B. Thompson, S. Roberts, and G. Lupyan. 2018. Quantifying semantic similarity across languages. *Proceedings of the 40th Annual Conference of the Cognitive Science Society (CogSci 2018)*.

[207] András Tilcsik, Michel Anteby, and Carly R. Knight. 2015. Concealable stigma and occupational segregation: Toward a theory of gay and lesbian occupations. *Administrative*

Science Quarterly, 60(3):446–481.

[208] Claudio Vaz Torres. 1999. *Leadership style norms among americans and brazilians: assessing differences using jackson's return potential model*. California School of Professional Psychology-San Diego.

[209] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[210] Leanne Townsend and Claire Wallace. 2016. Social media research: A guide to ethics. *University of Aberdeen*, 1:16.

[211] Yulia Tsvetkov and Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In *Coling 2010: Posters*, pages 1256–1264, Beijing, China. Coling 2010 Organizing Committee.

[212] UNICEF. 2021. Defining social norms and related concepts.

[213] Ishan Sanjeev Upadhyay, KV Aditya Srivatsa, and Radhika Mamidi. 2022. Towards toxic positivity detection. In *Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media*, pages 75–82, Seattle, Washington. Association for Computational Linguistics.

[214] Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

[215] Jelena Vulanović. 2014. Cultural markedness and strategies for translating idiomatic expressions in the epic poem "The Mountain Wreath" into English. *Mediterranean Journal of Social Sciences*, 5(13):210.

[216] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proc. of ICWSM*.

[217] Xinyi Wang and Graham Neubig. 2019. Target conditioned sampling: Optimizing data selection for multilingual neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5823–5828, Florence, Italy. Association for Computational Linguistics.

[218] Zijian Wang and Christopher Potts. 2019. Talkdown: A corpus for condescension detection in context. *CoRR*, abs/1909.11272.

[219] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

[220] Carol Waseleski. 2017. Gender and the Use of Exclamation Points in Computer-Mediated Communication: An Analysis of Exclamations Posted to Two Electronic Discussion Lists. *Journal of Computer-Mediated Communication*, 11(4):1012–1024.

[221] Uriel Weinreich. 1979. *Languages in contact: Findings and problems*. Walter de Gruyter.

[222] Etienne Wenger-Trayner and Beverly Wenger-Trayner. 2015. Introduction to communities of practice: A brief overview of the concept and its uses.

[223] Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

[224] Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

[225] Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

[226] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[227] Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, FangYuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. In *Advances in Neural Information Processing Systems*, volume 36, pages 58478–58507. Curran Associates, Inc.

[228] Jason Shuo Zhang, Brian C. Keegan, Qin Lv, and Chenhao Tan. 2021. Understanding the diverging user trajectories in highly-related online communities during the covid-19 pandemic.

[229] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

[230] Wen Zhang, Yunhan Liu, Yixuan Dong, Wanna He, Shiming Yao, Ziqian Xu, and Yan Mu. 2023. How we learn social norms: a three-stage model for social norm learning. *Frontiers in Psychology*, 14:1153809.

[231] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

[232] Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing positive perspectives with text reframing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3682–3700, Dublin, Ireland.

Association for Computational Linguistics.

[233] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.