Positive AI with Social Commonsense Models

Maarten Sap

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2021

Reading Committee:
Yejin Choi, Co-Chair
Noah A. Smith, Co-Chair
Christopher Althoff
Sapna Cheryan

Program Authorized to Offer Degree:
Computer Science and Engineering

University of Washington

Abstract

Positive AI with Social Commonsense Models

Maarten Sap

Co-chairs of the Supervisory Committee:

Associate Professor Yejin Choi
Computer Science and Engineering

Professor Noah A. Smith
Computer Science and Engineering

To effectively understand language and safely communicate with humans, machines must not only grasp the surface meanings of texts, but also their underlying social meaning. This requires understanding interpersonal social commonsense, such as knowing to thank someone for giving you a present, as well as accounting for harmful social biases and stereotypes. While understanding these implied social dynamics is easy for most humans, it remains an elusive goal for AI and NLP systems. Importantly, systems that fail to account for these social and power dynamics risk producing redundant, rude, or even harmful outputs.

In this dissertation, we take several steps towards making NLP systems more human-centric, socially aware, and equity driven, motivated by the increased prowess and prevalence of AI and NLP technology. In the first part, we investigate methods for enabling NLP systems to reason about and revise the commonsense implications of text. We introduce ATOMIC, the first large-scale social commonsense knowledge graph for machines to reason about the causes and effects of everyday situations, and POWERTRANSFORMER, a system to revise the social implications of text using connotation frames of power and agency.

In the second part, we tackle the problem of detecting and representing social biases and toxicity in language with socially aware NLP models. We examine shortcomings of existing toxic language detection tools, uncovering strong racial biases which causes text written by African American authors to be flagged as toxic more often than by white authors. Then, we introduce Social Bias Frames, a new structured linguistic representation for distilling the harmful or biased implications of text in free-text explanations. We conclude by discussing the contributions of this dissertation as well as future directions towards improving the social awareness and equity of NLP systems.

Acknowledgements

Completing this PhD was one of the hardest things I have ever done, and there is no way I ever would have been able to do it if it weren't for all the people who helped and supported me, distracted me with fun things, and who listened to my anxious tirades.

First, I would like to thank my advisors Yejin Choi and Noah Smith, for their mentorship and support, and for taking a chance on me hiring me as their PhD student. I want to thank Yejin specifically for her unwavering belief in me and for challenging me to grow, and Noah for always being a voice of calm and reason in response to my research fears and anxieties. Special thanks to my committee members Tim Althoff for the helpful and insightful conversations and Sapna Cheryan for inspiring me to study social biases in language. I also want to thank other mentors that I've had throughout my career, from Andrew Schwartz and Lyle Ungar who taught me what NLP research was and encouraged me to pursue a PhD, to Eric Horvitz, Mari Ostendorf, Roy Schwartz, Katharina Reinecke, Dan Jurasfky, and James Pennebaker for the mentorship during our collaborations.

Obviously, none of my PhD research would have been possible without my friends and collaborators. I want to extend a special thanks to Hannah Rashkin, whom I learned a tremendous amount from during our several collaborations, to Lucy Lin for all the support and paper draft editing, and to Elizabeth Clark for helping me stay sane during research endeavors. Additionally, I want to thank Saadia Gabriel, Chandra Bhagavatula, Ronan LeBras, Antoine Bosselut, Emily Allaway, Hao Fang, Hao Cheng, Ari Holtzman, Suchin Gururangan, Alisa Liu, Ximing Lu, Albert Xu, Eshaan Pathak, Eric Wallace, Dan Klein, Xuhui Zhou, Xinyao Michelle Ma, Sam Gehman, Max Forbes, Jena D. Hwang, Vered Shwartz, Lianhui Qin, Tal August, Derek Chen, Dallas Card, Chaitanya Malaviya, Asli Celikyilmaz, Nicholas Lourie, Brendan Roof, Kevin Knight, Marcella Cindy Prasetio, Ioannis Konstas, Li Zilles. Also thanks to all the undergraduate and masters students I got to mentor, and all the crowdworkers that worked on my datasets.

Within the CSE community, I cannot stress enough how important my espresso room and brunch crew—Lucy Lin, Emily Furst, Amrita Mazumdar, Kira Goldner, Elise Dorough— and my basically-therapist friends—Swabha Swayamdipta and Jesse Dodge— have been throughout these six years.

Extending beyond CSE, I would like to give special thanks to Cole Allick, James Burkman, and Andrew Bossick for being there with me through the ups and downs of this PhD.

Thanks to my parents for the enthusiasm, interest and support for my research endeavors, and to my sister, Liesbeth, for always being there for me despite being on the other side of the globe.

Finally, to Nic Morden, I could not have finished my PhD and my job search without your support.

Land Acknowledgement The work that went into this PhD and thesis was accomplished at the University of Washington in Seattle, Washington, which sits on the traditional land of the Coast Salish people, including the Duwamish People past and present. For more on land acknowledgments, please read âpihtawikosisân (2016).

DEDICATION

To my sister, Liesbeth,
I don’t know what I’d do without you

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Throughout our lives, we use social cues and world knowledge to interpret and navigate the situations we encounter or read about (Apperly, 2010). For example, if a store owner "turns on the security cameras," we can easily make inferences about their intent and reactions–that they "are worried" and "want to protect their property". However, if the owner turns on the cameras "because someone with a headscarf just walked in," we can also understand that this statement evokes the harmful implication or stereotype that "people with headscarves are seen as threatening." We make these types of inferences about situations by constructing mental models (Graesser et al., 1981) informed by our commonsense knowledge about the world (Kintsch, 1988), our knowledge of social groups and inequality (McGarty, 2018), and our personal experiences (Conway and Pleydell-Pearce, 2000).

Such understanding of social dynamics and social commonsense remains an elusive goal for artificial intelligence (AI) and natural language processing (NLP) systems (Gunning, 2018). But as these systems become increasingly prevalent in society (e.g., GMail's writing assistant; Chen et al., 2019), their effectiveness hinges on them being able to understand and reason about the social dynamics and social commonsense that govern our world (Pereira et al., 2016). For example, an AI system can assist humans better if it can infer that "if an elderly person falls," it should "call for help" (Pollack, 2005). Many other types of AI assistants, such as therapeutic counseling systems and assistive technologies for people with cognitive disabilities, require social commonsense reasoning abilities in order to operate more effectively (Kearns et al., 2020; Lewis, 2020).

On a darker note, as prevalence of AI technology increases, so does the risk for negative or harmful consequences, which calls for methods to ensure that our systems serve members of society as equitably as possible (Crawford et al., 2016; Horvitz, 2017; Floridi et al., 2018). Failure to anticipate the potentially harmful social implications of a system's output can indeed lead to catastrophic results, such as a racist neo-nazi Twitter chatbot (Vincent, 2016), systems that generate toxicity or radicalizing propaganda (Gehman et al., 2020; McGuffie and Newhouse, 2020), or hate speech detection systems that disproportionately censor minority voices (Dixon et al., 2018; Sap et al., 2019a; Oliva et al., 2021). At the core of many of these issues is that AI systems are trained on data that contains toxicity, social biases, and other undesirable content, yet such content is often not explicitly accounted for in training or at deployment time (Benjamin, 2019). This calls for algorithms to control or steer AI systems so as to avoid undesirable content.

In this dissertation, we explore ways to make NLP systems more human-centric, socially aware, and equity driven, motivated by the increased prowess and prevalence of AI and NLP technology. Specifically, we are motivated by the recent improvements in several NLP tasks yielded by large language models (LM), such as ELMo (Peters et al., 2018), OpenAI-GPT (Radford et al., 2018), GPT-2 (Radford et al., 2019), and BERT (Devlin et al., 2019), pretrained on large amounts of web-scraped text using a self-supervised language modeling objective. While these pretrained

LMs have opened the door for exciting text-to-text applications, their abilities still suffer from several issues (Bender et al., 2021). For example, these systems tend to rely on spurious lexical correlations between inputs and outputs instead of learning a task (Schwartz et al., 2017; Gururangan et al., 2018), often struggling to disentangle people mentioned in text (Sap et al., 2019c; Sakaguchi et al., 2020). Additionally, reporting biases in their pretraining data limit their ability to learn commonsense knowledge (Gordon and Van Durme, 2013), as such knowledge is often omitted from text (Grice, 1975). Finally, their pretraining data often contains toxicity, social biases, and other undesirable content (Fast et al., 2016; Gehman et al., 2020), yet such content is often not explicitly accounted for in training or at deployment time (Benjamin, 2019). Thus, in this dissertation, we investigate ways to account for these shortcomings to enable NLP systems to reason about the (potentially undesirable) social implications of text.

In Part I, we investigate formalisms and algorithms for NLP systems to reason about and revise the commonsense implications of text. Specifically, in Chapter 2, we introduce ATOMIC, an atlas of machine commonsense containing 880k *if-then* knowledge tuples (e.g., *if* "X drinks coffee", *then* "X likely wanted to wake up"). In contrast to existing approaches, ATOMIC focuses on a new type of inferential commonsense reasoning about the causes and effects of everyday situations, which we represent in short natural language phrases. By training on ATOMIC tuples, we investigate neural models' ability to generate inferences given a previously unseen situation. We find a significant gap in reasoning ability between models and humans, which has since been narrowed by Bosselut et al. (2019).

Then, in Chapter 3, we introduce POWERTRANSFORMER, new model for unsupervised controllable revision that alters the portrayal of characters in story sentences. We debias portrayals through the lens of connotation frames of power and agency (Sap et al., 2017), a formalism that encodes pragmatic knowledge of implied power and agency dynamics with respect to predicates (e.g., "X plays football" portrays X as high agency, active, and decisive). POWERTRANSFORMER was trained using a self-supervised denoising objective and an auxiliary paraphrasing objective, overcoming the lack of parallel training data for controllable revision tasks. Through ablation studies and human evaluation, we showed that our model benefits from both objectives independently and outperforms existing text revision baselines. Importantly, we used POWERTRANSFORMER to revise a set of modern movie scripts and successfully mitigate the bias we previously uncovered, raising the power and agency that female characters are portrayed with.

In Part II, we tackle the problem of detecting and representing social biases and toxicity in language with socially aware NLP models. In Chapter 5, we examine the fairness of several toxic language detection datasets and models with respect to race and dialect. We uncover strong racial bias, finding that toxicity detection models disproportionately flag sentences in African American English (AAE) and by African American authors as more often as toxic compared to by white authors. We propose dialect and race priming as ways to reduce the racial bias in annotation, showing through a user study that when annotators are made explicitly aware of an AAE sentence's dialect they are significantly less likely to label the sentence as offensive.

Finally, in Chapter 6, we introduce SOCIAL BIAS FRAMES, a new pragmatic formalism to capture the (potentially biased) social and power dynamics implied in language. Given a statement like "the all-Muslim movie was a box office bomb," SOCIAL BIAS FRAMES combines categorical inferences about biased intention and offensiveness, as well as free-text explanations of who is targeted by the statement ("Muslims") and what is implied (that "Muslims are terrorists"). We create a crowdsourcing framework to gather 150k SOCIAL BIAS FRAMES annotations from online posts to enable large-scale modelling. Then, we explore neural approaches to spelling out biased implications in terms of SOCIAL BIAS FRAMES. We find that models are more effective at classifying high-level categories of offensiveness than at generating the implied biased meaning behind statements.

We conclude this thesis with a summary of the contributions and some future directions towards NLP systems that are more human-centric, socially aware, and equity driven.

Part I

Algorithms for NLP with Social Commonsense

Chapter 2

ATOMIC: an Atlas of Machine Commonsense

This chapter discusses work originally published in Sap et al. (2019b).

In the pursuit of the ambitious goal of social commonsense reasoning with machines, having access to commonsense knowledge to reason with is a crucial component. In this chapter, we present ATOMIC,[1] an atlas of everyday commonsense reasoning, organized through 877k textual descriptions of inferential knowledge. Compared to existing resources that center around taxonomic knowledge, ATOMIC focuses on inferential knowledge organized as typed *if-then* relations with variables (e.g., "*if* X repels Y's attack, *then* Y will likely want to attack again", Figure 2.1). By generatively training on the rich inferential knowledge described in ATOMIC we show that neural models can acquire simple commonsense capabilities and reason about previously unseen events. Experimental results demonstrate that multitask models that incorporate the hierarchical structure of *if-then* relation types lead to more accurate inference compared to models trained in isolation, as measured by both automatic and human evaluation.

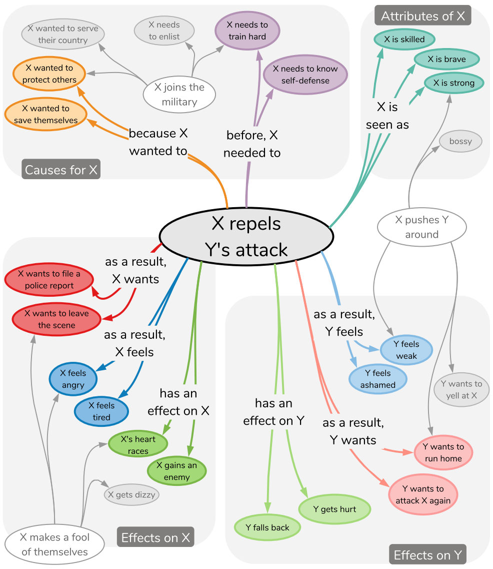


Figure 2.1: A tiny subset of ATOMIC, an atlas of machine commonsense for everyday events, causes, and effects.

2.1 Introduction

Given a snapshot observation of an event, people can easily anticipate and reason about unobserved causes and effects in relation to the observed event: what might have happened just before, what might happen next as a result, and how different events are chained through causes and effects. For instance, if we observe an event "X repels Y's attack" (Figure 2.1), we can immediately infer various plausible facts surrounding that event. In terms of the *plausible motivations* behind the event, X probably wants to protect herself. As for the *plausible pre-conditions* prior to the event, X may have been trained in self-defense to successfully fend off Y's attack. We can also infer the *plausible characteristics* of X; she might be strong, skilled, and brave. As a *result* of the event, X

[1] An **AT**las **O**f **M**ach**I**ne **C**ommonsense, available to download or browse at https://homes.cs.washington.edu/~msap/atomic/.

probably feels angry and might want to file a police report. Y, on the other hand, might feel scared of getting caught and want to run away.

The examples above illustrate how day-to-day commonsense reasoning can be operationalized through a densely connected collection of inferential knowledge. It is through this knowledge that we can watch a two-hour movie and understand a story that spans over several months, as we can reason about a great number of events, causes, and effects, while observing only on a small fraction of them. It also enables us to develop Theories of Mind about others (Moore, 2013). However, this ability, while common and trivial for humans, is lacking in today's AI systems. This is in part because the vast majority of AI systems are trained for task-specific datasets and objectives, which lead to models that are effective at finding task-specific correlations but lack simple and explainable commonsense reasoning (Davis and Marcus, 2015; Lake et al., 2017; Marcus, 2018).

In this chapter, we introduce ATOMIC, an atlas of machine commonsense, as a step toward addressing the rich spectrum of inferential knowledge that is crucial for automated commonsense reasoning. In contrast with previous efforts (Lenat, 1995; Speer and Havasi, 2012) that predominantly contain taxonomic or encyclopedic knowledge (Davis and Marcus, 2015), ATOMIC focuses on inferential *if-then* knowledge. The goal of our study is to create a knowledge repository that meets three requirements: scale, coverage, and quality. Therefore, we focus on crowdsourcing experiments instead of extracting commonsense from corpora, because the latter is subject to the significant reporting bias in language that can challenge both the coverage and quality of the extracted knowledge (Gordon and Van Durme, 2013).

We propose a new taxonomy of *if-then* reasoning types as shown in Figure 2.2. One way to categorize the types is based on the content being predicted: (1) *If-Event-Then-Mental-State*, (2) *If-Event-Then-Event*, and (3) *If-Event-Then-Persona*. Another way to categorize is based on their causal relations: (1) "causes", (2) "effects", and (3) "stative". Using this taxonomy, we gather over 877K instances of inferential knowledge.

We then investigate neural network models that can acquire simple commonsense capabilities and reason about previously unseen events by embedding the rich inferential knowledge described in ATOMIC. Experimental results demonstrate that neural networks can abstract away commonsense inferential knowledge from ATOMIC such that given a previously unseen event, they can anticipate the likely causes and effects in rich natural language descriptions. In addition, we find that multitask models that can incorporate the hierarchical structure of if-then relation types lead to more accurate inference compared to models trained in isolation.

2.2 *If-Then* Relation Types

To enable better reasoning about events, we improve upon existing resources of commonsense knowledge by adding nine new causal and inferential dimensions. Shown in Figure 2.2, we define dimensions as denoting a particular type of *If-Then* knowledge, answers to questions about an event, collected through crowdsourcing. Contrary to most previous work, ATOMIC also characterizes knowledge of events and their *implied* participants (e.g., "Alex calls for help" implies someone will answer the call), in addition to explicitly mentioned participants (e.g., "Alex calls Taylor for help").

Illustrated in Table 2.1, our nine dimensions span three types of *If-Then* relations, outlined below.

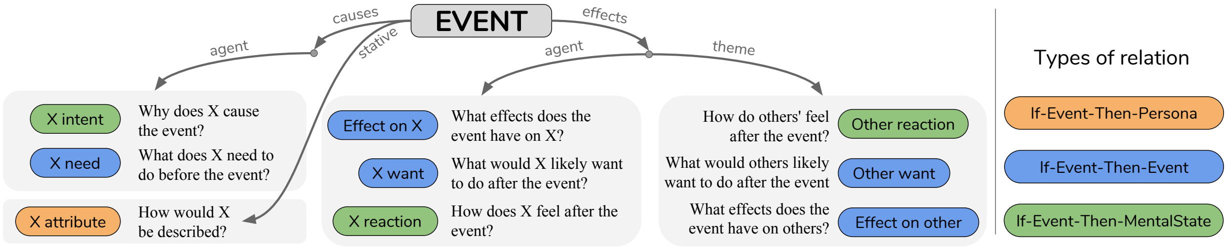

EVENT
causes
effects
stative
agent
agent
theme
X intent
Why does X cause the event?
X need
What does X need to do before the event?
X attribute
How would X be described?
Effect on X
What effects does the event have on X?
X want
What would X likely want to do after the event?
X reaction
How does X feel after the event?
How do others' feel after the event?
Other reaction
What would others likely want to do after the event
Other want
What effects does the event have on others?
Effect on other
Types of relation
If-Event-Then-Persona
If-Event-Then-Event
If-Event-Then-MentalState

Figure 2.2: The taxonomy of *if-then* reasoning types. We consider nine *if-then* relations that have overlapping hierarchical structures as visualized above. One way to categorize the types is based on the type of content being predicted: (1) **If-Event-Then-Mental-State**, (2) **If-Event-Then-Event**, and (3) **If-Event-Then-Persona**. Another way is to categorize the types based on their causal relations: (1) **"causes"**, (2) **"effects"**, and (3) **"stative"**. Some of these categories can further divide depending on whether the reasoning focuses on the "agent" (X) or the "theme" (Other) of the event.

If-Event-Then-Mental-State We define three relations relating to the mental pre- and post-conditions of an event. Given an event (e.g., "X compliments Y"), we reason about (i) likely *intents* of the event (e.g., "X wants to be nice"), (ii) likely *(emotional) reactions* of the event's subject ("X feels good"), and (iii) likely *(emotional) reactions* of others ("Y feels flattered").

If-Event-Then-Event We also define five relations relating to events that constitute probable pre- and post-conditions of a given event. Those relations describe events likely required to precede an event, as well as those likely to follow. For instance, people know that "X needs to put coffee in the filter" before "X makes Y's coffee". For post-conditions, we focus on both voluntary ("X adds cream and sugar") and involuntary ("X gets thanked by Y") possible next events. We also define voluntary and involuntary possible next events for (implied) participants.

If-Event-Then-Persona In addition to pre- and post-conditions, we also define a stative relation that describes how the subject of an event is described or perceived. For instance, when "X calls the police", X is seen as "lawful" or "responsible".

An Alternative Hierarchy The above relation types can be categorized via a different *hierarchical structure* as shown in Figure 2.2 in the appendix. In particular, they can be categorized based on their causal relations: (1) "causes", (2) "effects", and (3) "stative". Each of these categories can be further divided depending on whether the reasoning focuses on the "agent" or the "theme" of the event. We omit cases where the combination is unlikely to lead to commonsense anticipation. For example, it is usually only the "agent" who causes the event, rather than the "theme", thus we do not consider that branching. We later exploit this hierarchical structure of inferential relations for designing effective neural network architectures that can learn to reason about a given event.

2.3 ATOMIC Data

To build ATOMIC, we create a crowdsourcing framework that allows for scalable, broad collection of *If-Then* knowledge for given events.

Event	Type of relations	Inference examples	Dim.
"PersonX pays PersonY a compliment"	If-Event-Then-Mental-State	PersonX wanted to be nice PersonX will feel good PersonY will feel flattered	xIntent xReact oReact
	If-Event-Then-Event	PersonX will want to chat with PersonY PersonY will smile PersonY will compliment PersonX back	xWant oEffect oWant
	If-Event-Then-Persona	PersonX is flattering PersonX is caring	xAttr xAttr
"PersonX makes PersonY's coffee"	If-Event-Then-Mental-State	PersonX wanted to be helpful PersonY will be appreciative PersonY will be grateful	xIntent oReact oReact
	If-Event-Then-Event	PersonX needs to put the coffee in the filter PersonX gets thanked PersonX adds cream and sugar	xNeed xEffect xWant
	If-Event-Then-Persona	PersonX is helpful PersonX is deferential	xAttr xAttr
"PersonX calls the police"	If-Event-Then-Mental-State	PersonX wants to report a crime Others feel worried	xIntent oReact
	If-Event-Then-Event	PersonX needs to dial 911 PersonX wants to explain everything to the police PersonX starts to panic Others want to dispatch some officers	xNeed xWant xEffect oWant
	If-Event-Then-Persona	PersonX is lawful PersonX is responsible	xAttr xAttr

Table 2.1: Examples of **If-Event-Then-X** commonsense knowledge present in ATOMIC. For inference dimensions (dim.), "x" and "o" pertain to PersonX and others, respectively (e.g., "xAttr": attribute of PersonX, "oEffect": effect on others).

2.3.1 Compiling Base Events

As *base events* for our annotations, we extract 24K common event phrases from a variety of corpora. To ensure broad and diverse coverage, we compile common phrases from stories, books, Google Ngrams, and Wiktionary idioms (Mostafazadeh et al., 2016a; Gordon and Swanson, 2008; Goldberg and Orwant, 2013). Following Rashkin et al. (2018), we define events as verb phrases with a verb predicate and its arguments ("drinks dark roast in the morning"). If a verb and its arguments do not co-occur frequently enough,[2] we replace the arguments with a blank placeholder ("drinks ___ in the morning"). In order to learn more general representations of events, we replace tokens referring to people with a Person variable (e.g. "PersonX buys PersonY coffee"). In future work, other types of variables could be added for other entity references (e.g. "PersonX moves to CityX").

For events with multiple people explicitly involved, we run a short annotation task to help resolve coreference chains within phrases. Disambiguating the participants is important, since it can drastically change the meaning of the event (e.g., "PersonX breaks PersonX's arm" vs. "PersonX

[2] We use frequency thresholds of 5 and 100 for stories and blogs, respectively, and limit ourselves to the top 10,000 events in Google Ngrams.

breaks PersonY's arm" have very different implications). Three workers selected whether each "Person" mention in an event refers to PersonX, PersonY, or PersonZ, and we keep base events with combinations that at least two workers selected as valid (ppa=77%).

2.3.2 Crowdsourcing Framework

To ensure scalability, we implement a free-form text annotation setup which asks workers to write answers to questions about a specific event. We chose free-text over structured or categorical annotation for two reasons. First, categorical annotations with a large labeling space have a substantial learning curve, which limits the annotation speed and thereby the coverage of our knowledge graph. Second, the categorical labels are likely to limit the ability to encode the vast space of commonsense knowledge and reasoning as depicted in Figure 2.1 and Table 2.1.

	Count	#words
# triples: If-Event-Then-*	877,108	-
- Mental-State	212,598	-
- Event	521,334	-
- Persona	143,176	-
# nodes: If-Event-Then-*	309,515	2.7
- Mental-State	51,928	2.1
- Event	245,905	3.3
- Persona	11,495	1.0
Base events	24,313	4.6
# nodes appearing > 1	47,356	–

Table 2.2: Statistics of ATOMIC. Triples represent distinct <event, relation, event>. #words represents the average number of words per node.

We create four tasks on Amazon Mechanical Turk (MTurk) (sample task in Figure A.1) for gathering commonsense annotations.[3, 4] For each dimension, up to three workers are asked to provide as many as four likely annotations for an event, covering multiple possible situations (e.g., if "PersonX drinks coffee", then "PersonX needed to brew coffee" or "PersonX needed to buy coffee"; both are distinct but likely). Note that some events are not caused by PersonX, and some do not affect other people, making annotations for certain dimensions not necessary (specifically, for xIntent, xNeed, oReact, oEffect, and oWant) for all events. For those dimensions, we first ask workers whether this specific inference dimension is relevant given an event.

2.3.3 ATOMIC Statistics

Table 2.2 lists descriptive statistics of our knowledge graph. Our resulting knowledge graph contains over 300K nodes, collected using 24K base events. Nodes in the graph are short phrases (2.7 tokens on average), ranging from 1 token for stative events (attributes) to 3.3 and 4.6 tokens on average for more active events. Unlike denotational tasks where experts would only consider one label as correct, our annotations correspond to a distribution over *likely* inferences (de Marneffe et al., 2012). To measure the degree of agreement, we run a small task asking turkers to determine whether an individual annotation provided by a different turker is valid. Table 2.4 shows that annotations are deemed valid on average 86.2% of the time for a random subset of events. For quality control, we manually and semi-automatically detected and filtered out unreliable workers.

[3] The tasks were used to collect the following four sets of dimensions: (1) intent and reaction, (2) need and want, (3) effects, and (4) attributes.

[4] Our payment rate was above $12/hour, going well beyond the federal minimum rate of $8/hour.

	Model	xIntent	xNeed	xAttr	xEffect	xReact	xWant	oEffect	oReact	oWant
DEV	9ENC9DEC	**8.35**	17.68	5.18	**10.64**	**5.38**	**13.24**	6.49	5.17	12.08
	NearestNeighbor	6.14	11.36	3.57	5.81	4.37	7.73	**8.02**	**6.38**	8.94
	EVENT2(IN)VOLUNTARY	7.51	**17.80**	5.18	10.51	4.78	12.76	7.04	4.84	**12.48**
	EVENT2PERSONX/Y	7.31	17.08	**5.26**	9.78	4.83	12.14	6.38	4.84	11.45
	EVENT2PRE/POST	7.58	17.17	–	10.50	4.73	11.78	6.71	4.87	11.52
TEST	9ENC9DEC	**8.68**	18.15	**5.18**	**10.34**	**5.43**	**14.50**	6.61	5.08	12.73
	NearestNeighbor	6.64	11.35	3.37	5.52	4.59	8.17	**7.58**	**5.88**	9.18
	EVENT2(IN)VOLUNTARY	7.94	**18.22**	5.02	9.78	4.78	13.67	7.16	4.71	**13.23**
	EVENT2PERSONX/Y	7.67	17.33	5.09	9.45	4.82	13.19	6.59	4.68	11.70
	EVENT2PRE/POST	7.96	17.42	–	9.79	4.75	12.85	6.90	4.76	11.97

Table 2.3: Average BLEU score (reported as percentages) for the top 10 generations for each inference dimension: comparison of multitask models to single-task model. Note that BLEU scores are known to be brittle to generations worded differently from the references (Liu et al., 2016b). We embolden the best performing model for each dimension.

2.4 Methods

Our goal is to investigate whether models can learn to perform *If-Then* commonsense inference given a previously unseen event. To this extent, we frame the problem as a conditional sequence generation problem: given an event phrase $\mathbf{e}$ and an inference dimension c, the model generates the target $\mathbf{t} = f_\theta(\mathbf{e}, c)$. Specifically, we explore various multitask encoder-decoder setups.

Encoder We represent the event phrase as a sequence of n word vectors $\mathbf{e} = \{e_0, e_1, \ldots, e_{n-1}\} \in \mathbb{R}^{n \times i_{enc}}$ where each word is an i_{enc}-dimensional vector. The event sequence is compressed into a hidden representation $\mathbf{h}$ through an encoding function $f_{enc} : \mathbb{R}^{i \times h_{enc}} \to \mathbb{R}^h$.

In this work, we use 300-dimensional static GloVe pre-trained embeddings (Pennington et al., 2014a) as our base word vectors. We augment these embeddings with 1024-dimensional ELMo pre-trained embeddings (Peters et al., 2018). ELMo provides deep contextualized representation of words using character-based representations, which allows robust representations of previously unseen events. The encoding function is a bidirectional GRU (Cho et al., 2014) of hidden size h_{enc}.

Decoder Each decoder is a unidirectional GRU of hidden size h_{dec}, with a hidden state initialized to $\mathbf{h}_{dec}^{(0)} = \mathbf{h}$. The target is represented by a sequence of vectors $\mathbf{t} = \{t_0, t_1, \ldots\}$, where each $t_i \in \mathbb{R}^h_{dec}$ is based on a learned embedding. The decoder then maximizes $p(t_{i+1} \mid \mathbf{h}_{dec}^{(i)}, t_0, \ldots, t_i) = \text{softmax}(W_o \times \text{GRU}(\mathbf{h}_{dec}^{(i)}, t_i) + b_o)$.

Single vs. Multitask Learning We experiment with various ways to combine the commonsense dimensions with multitask modeling. We design models that exploit the hierarchical structure of the commonsense dimensions (depicted in Figure 2.2), sharing encoders for dimensions that are related. Specifically, we explore the following models:

- EVENT2(IN)VOLUNTARY: We explore grouping dimensions together depending on whether they denote voluntary (e.g., xIntent, oWant) or involuntary (e.g., xReact, oEffect) events.

Model	xNeed	xIntent	xAttr	xEffect	xReact	xWant	oEffect	oReact	oWant	avg.
9ENC9DEC	48.74	51.70	52.20	47.52	63.57	51.56	22.92	32.92	35.50	45.32
EVENT2(IN)VOLUNTARY	49.82	**61.32**	52.58	46.76	71.22	52.44	**26.46**	**36.04**	34.70	**47.93**
EVENT2PERSONX/Y	**54.04**	53.93	**52.98**	**48.86**	66.42	**54.04**	24.72	33.80	35.08	46.41
EVENT2PRE/POST	47.94	57.77	52.20	46.78	**72.22**	47.94	26.26	34.48	35.78	46.76
gold ATOMIC annotations	81.98	91.37	78.44	83.92	95.18	90.90	84.62	86.13	83.12	86.18

Table 2.4: Precision at 10 (%) of generated inferences as selected by human judges for four models, averaged and broken down by dimension. We embolden the best performing model for each dimension. EVENT2(IN)VOLUNTARY outperforms all other models significantly ($p < 0.05$). For comparison, we show precision of gold ATOMIC annotations. Note that there is a varying number of gold annotations per event/dimension, while all models were constrained to make 10 predictions.

This model has one encoder for four "voluntary" decoders, as well as another encoder for five "involuntary" decoders.

- EVENT2PERSONX/Y: We dissociate dimensions relating to the event's agent (PersonX) from those relating to the event's theme (others, or PersonY). This model has one encoder for six "agent" decoders as well as another encoder for three "theme" decoders.
- EVENT2PRE/POST: We split our dimensions based on whether they are related to causes (xNeed, xIntent) or effects (e.g., xWant, oEffect, xReact). In this model, there are two encoders and eight decoders.[5]

As a single task baseline, we train nine separate encoder-decoders, one for each dimension (9ENC9DEC).

Training Details To test our models, we split seed events into training, validation, and test sets (80%/10%/10%), ensuring that events that share the same first two content words are in the same set. As is common in generation tasks, we minimize the cross entropy of the distribution over predicted targets compared to the gold distribution in our data.[6] During multitask training, we average the cross entropy of each task. Since multiple crowdworkers annotated each event, we define our training instances to be the combination of one worker's annotations. During experiments, we use the 300-dimensional GloVe embeddings, yielding an encoder input size of i_{enc} = 1324 once concatenated with the 1,024-dimensional ELMo embeddings. In the encoder, ELMo's character-level modeling allows for an unlimited vocabulary. We set the encoder and decoder hidden sizes to $h_{enc} = 100$ and $h_{dec} = 100$.

2.5 Results

We evaluate models on their ability to reason about previously unseen events. Given an unseen event, models generate natural language expressions for each of the nine dimension of *if-then* inferences. We report performance using automatic scores and a human evaluation of the generated inferences.

[5] We omit xAttr in this model, as it is trivially covered in the single task baseline.

[6] All our experiments were run using AllenNLP (Gardner et al., 2017).

2.5.1 Automatic Scores

We automatically evaluate the sequence generation for each model and each inference dimension using BLEU scores. Specifically, we compute the average BLEU score ($n = 2$, Smoothing1; Chen and Cherry, 2014) between each sequence in the top 10 predictions and the corresponding set of MTurk annotations. As an event may not involve all nine inference dimensions (e.g., "PersonX sees PersonX's house" has no implications for anybody other than "PersonX"), annotators may decide to leave an inference dimension *empty*. When computing BLEU scores, we omit instances with one-third or more *empty* annotations.

Table 2.3 presents the results on both DEV and TEST datasets. The experiments show that models that exploit the hierarchical structure of the commonsense relations perform better than the model that uses separate parameters (9ENC9DEC). Importantly, BLEU is a crude measure of performance as it is based on the exact match of n-grams and fails to capture semantically relevant generations that are worded differently (Liu et al., 2016b). As shown in Figure 2.3, the generated samples depict varying word and phrase choices, thus we also perform human evaluation to complement automatic evaluations.

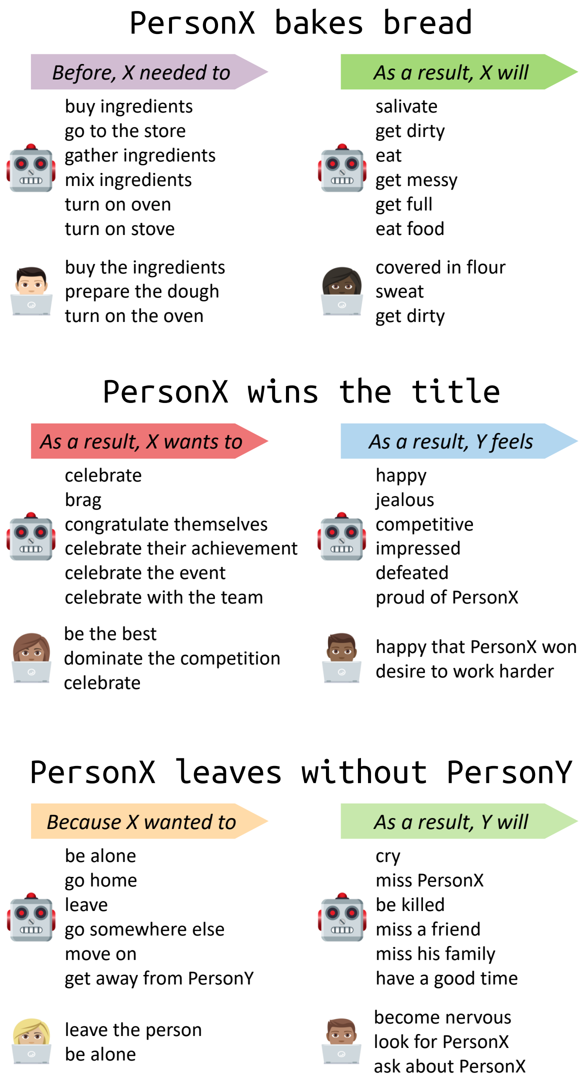


Figure 2.3: Examples of machine (🤖) generated inferences for three events from the development set, ordered from most likely (top) to least likely (bottom) according to the EVENT2(IN)VOLUNTARY model. Human (🧑‍💻) generated inferences are also shown for comparison.

2.5.2 Human Evaluation

Since automatic evaluation of generated language is an open research question (Liu et al., 2016b), we also assess our models' performance through human evaluation. We randomly select 100 events from the test set and use beam search to generate the 10 most likely inferences per dimension. We present five crowdworkers with the 10 generated inferences, and ask them to select all inferences they think are valid. Table 2.4 shows each model's precision at 10, computed as the average number of correct generations per dimension. Following the same crowdsourcing setup, we also assess the quality of the gold ATOMIC annotations for the same set of test events. Human evaluation (last line of Table 2.4) indicates that 86.2% of the descriptions are valid, showcasing the quality of commonsense knowledge contained in ATOMIC.

Human evaluation supports our conclusion from automatic evaluation – that models that leverage the *if-then* hierarchy perform better than models that don't. Specifically, explicitly modeling whether inference dimensions describe voluntary actions (e.g., what X wants to do next) or involuntary effects (e.g., X or Y's reactions) yields more sensible generations, as evidenced by the performance of EVENT2(IN)VOLUNTARY.

2.5.3 Qualitative Results

We present sample commonsense predictions in Figure 2.3. Given an event "PersonX bakes bread", our model can correctly infer that X probably *needs* to "go to the store" or "mix ingredients" or "turn on the oven". Our model also correctly predicts that the likely effect of this event would be that X will "get dirty" or "eat food".

2.5.4 Comparison with ConceptNet

ConceptNet (Speer et al., 2017) represents commonsense knowledge as a graph of *concepts* connected by *relations*. Concepts consist of words or phrases, while relations come from a fixed set of edge types.

While ConceptNet captures general commonsense knowledge—much of which is taxonomic in nature[7]—ATOMIC focuses on sequences of events and the social commonsense relating to them. This focus means that while events and dimensions in ATOMIC loosely correspond to concepts and relations from ConceptNet, individual dimensions, such as *intents*, can't be mapped cleanly onto any combination of ConceptNet's relations. The correspondence is neither one-to-one nor one-to-many. Still, in order to empirically investigate the differences between ConceptNet and ATOMIC, we used the following best-effort mappings between the dimensions and relations:

- **Wants**: MOTIVATEDBYGOAL, HASSUBEVENT, HASFIRSTSUBEVENT, CAUSESDESIRE
- **Effects**: CAUSES, HASSUBEVENT, HASFIRSTSUBEVENT, HASLASTSUBEVENT
- **Needs**: MOTIVATEDBYGOAL, ENTAILS, HASPREREQUISITE
- **Intents**: MOTIVATEDBYGOAL, CAUSESDESIRE, HASSUBEVENT, HASFIRSTSUBEVENT
- **Reactions**: CAUSES, HASLASTSUBEVENT, HASSUBEVENT
- **Attributes**: HASPROPERTY

We then computed the overlap of `<event1, dimension, event2>` triples in ATOMIC with the `<concept1, relation, concept2>` triples in ConceptNet. We found the overlap to only be as high as 7% for *wants*, 6% for *effects*, 6% for *needs*, 5% for *intents*, 2% for reactions, and 0% for attributes. Moreover, only 25% of the events in ATOMIC are found in ConceptNet. Thus, ATOMIC offers a substantial amount of new inferential knowledge that has not been captured by existing resources.

[7] While ConceptNet includes various inferential relations (e.g., "entails", "causes", "motivated by"), their instances amount to only about 1% of ConceptNet.

2.6 Related Work

2.6.1 Descriptive Knowledge from Crowdsourcing

Knowledge acquisition and representation have been extensively studied in prior research (Espinosa and Lieberman, 2005; Speer and Havasi, 2012; Lenat, 1995). However, most prior efforts focused on taxonomic or encyclopedic knowledge (Davis and Marcus, 2015), which, in terms of epistemology, corresponds to *knowledge of "what"*. Relatively less progress has been made on *knowledge of "how"* and *"why"*. For example, OpenCyc 4.0 is a large commonsense knowledge base consisting of 239,000 concepts and 2,039,000 facts in LISP-style logic (Lenat, 1995), known to be mostly taxonomic (Davis and Marcus, 2015). In fact, only 0.42% of ATOMIC events appear in OpenCyc, which we found contains 99.8% relations that are either taxonomic (isA), string formatting relations, or various definitional relations. A typical example is shown below:

```
(genls (LeftObjectOfPairFn
 SuperiorLobeOfLung) LeftObject)
(isa (WordNetSynsetReifiedFn
 460174) WordNetSynset)
(genls (AssociatesDegreeInFn
 EngineeringField) AssociatesDegree)
```

Importantly, these LISP-based representations of OpenCyc are non-trivial to integrate into modern neural network based models, as it is not straightforward to compute their embedding representations. In contrast, the natural language representations in ATOMIC can be readily used to obtain their neural embeddings, which can also be mixed with pretrained embeddings of words or language models.

Similarly, ConceptNet (Speer et al., 2017) represents commonsense knowledge as a graph that connects words and phrases (*concepts*) with labeled edges (*relations*). While ConceptNet provides relatively more inferential relations (e.g., "entails", "causes", "motivated by"), they still amount to only about 1% of all triples in the graph. In contrast, ATOMIC is centered around events represented with natural language descriptions. While events and dimensions in ATOMIC loosely correspond to concepts and relations in ConceptNet, the two represent very different information and ultimately have relatively small overlap as discussed in the Results section.

Recent work by Gordon and Hobbs (2017) compiles a list of nearly 1,400 commonsense axioms in formal logic, which connect abstract concepts to each other. For example, they define an event as being made up of subevents, expressed by:

```
(forall (e)
 (iff (event e)
  (or (exists (e1 e2)
   (and (nequal e1 e2)(change' e e1 e2)))
    (exists (e1)
     (subevent e1 e)))))
```

These axioms are abstract in that they are not grounded with respect to specific objects, events, or actions. In contrast, our work presents 880K triples of commonsense knowledge expressed in natural language and fully grounded with concrete events, actions, mental states.

The recent work of Rashkin et al. (2018) introduced a commonsense inference task about events and mental states: given an event described in natural language, the task is to generate the reaction and intent of actors involved in the event. ATOMIC is inspired by this work, but substantially scales up (i) the crowdsourcing procedure to nine dimensions per event, and (ii) the size of the knowledge graph—from 77K events in Event2Mind to 300K events in ATOMIC. Moreover, while the primary focus of (Rashkin et al., 2018) was inferential knowledge, its scope was limited to mental states.

2.6.2 Acquired Knowledge from Extraction and Induction

More generally, the goal of moving beyond static commonsense knowledge to enable automated commonsense reasoning has inspired much research. Several projects have sought to extract commonsense inferential rules from naturally occurring resources such as large corpora (Schubert, 2002), movie scripts (Tandon et al., 2017), and web how-tos (Chu et al., 2017). Such systems must inevitably deal with reporting bias (Gordon and Van Durme, 2013), or the fact that the frequency and selection of phenomena represented in natural language systematically differ from what occurs in the real world. Other approaches have sought to induce commonsense rules from large knowledge bases (Galárraga et al., 2013; Yang et al., 2015). While these approaches have also had success, the choice of schema and information represented in current knowledge bases limits the scope of propositions such systems can learn.

2.6.3 Scripts and Narrative Reasoning

Other work has focused more specifically on representing and reasoning about sequences of events, similarly to ATOMIC. Early work on event sequences studied *scripts*, a kind of structured representation for prototypical sequences of events (Schank and Abelson, 1977). More recently, *narrative event chains* have been proposed as a similar formalism for prototypical sequences of events that may be learned from raw text (Chambers and Jurafsky, 2008). This work additionally proposed the *Narrative Cloze Test* as a benchmark for story understanding. In contrast to narrative event chains, the *ROC Stories Corpus* crowdsources event sequences represented as natural language stories rather than using a specific formalism (Mostafazadeh et al., 2016a). Additionally, the *Story Cloze Test* adapts these stories into a new benchmark by requiring systems to choose between the true and a false ending to the story. Our work interpolates between these two approaches by representing events in natural language while structuring the relationships between events into the edges of a graph. The *Choice of Plausible Alternatives* (COPA) task offers a similar benchmark for commonsense understanding of events and their relationships (Roemmele et al., 2011). In COPA, a system is presented a premise and two alternatives that might have a causal relationship with the premise. While COPA, like ATOMIC, represents events as free-form text with structured relationships, it covers only a limited number of relations (cause and effect) and is smaller in scale (contains only 1,000 instances).

2.7 Summary

In this chapter, we presented ATOMIC, an atlas of everyday commonsense inferential knowledge about events described in natural language and associated with typed *if-then* relations. ATOMIC consists of over 300k events associated with 877k inferential relations, making it the largest knowledge graph of its kind. Our crowdsourcing framework allowed for the gathering of annotations in the form of free-form textual responses to simple questions which enables large-scale high quality collection of commonsense about events. We also presented neural network models that can learn to reason about previously unseen events to generate their likely causes and effects in natural language.

Shortly after the release of ATOMIC, large pretrained language models (Radford et al., 2018; Devlin et al., 2019, e.g., OpenAI-GPT, BERT) emerged, showing promising ability to capture certain types of knowledge about the world (Petroni et al., 2019; Zhou et al., 2020). This raised the question: can pretrained LMs produce ATOMIC-style commonsense inferences better than randomly initialized encoder-decoder models considered in this chapter. To answer this question, we created COMET (Bosselut et al., 2019), by finetuning OpenAI-GPT, a Transformer-based (Vaswani et al., 2017a) language model pretrained on a large corpus of books (Zhu et al., 2015), on linearized ATOMIC knowledge tuples.[8] Using the same automatic and human evaluation setups as in §2.5, we find that not only does COMET produce better inferences than RNN-based models from §2.4, but also compared to a non-pretrained, i.e., randomly initialized, version. These results suggest that the knowledge in ATOMIC is somewhat learned by these pretrained models, but that perhaps finetuning on the commonsense inference task is needed to harness it.

Overall, our findings in ATOMIC showed that machines can help make inferences about previously unseen situations written in language, especially when using pretrained language models. This has opened the door for many ATOMIC-based applications. For example, Kearns et al. (2020) used ATOMIC and COMET to enhance a counselor response platform by providing commonsense inferences related to client utterances, as a way to encourage empathetic responses from counselors. Additionally, several follow up works have explored the use of ATOMIC and COMET for improving automated story (Ammanabrolu et al., 2021), sarcasm (Chakrabarty et al., 2020a), and simile (Chakrabarty et al., 2020b) generation. On the knowledge side, ATOMIC-style taxonomies could be extended to cover more object-related knowledge (Hwang et al., 2021) as well as commonsense implications of negated events (Jiang et al., 2021).

[8] Adding special tokens for each of the 9 relations. For further details, please see Bosselut et al. (2019).

Chapter 3

PowerTransformer: Controllable Revision for Biased Language Correction

This chapter discusses work originally published in Ma et al. (2020).

With ATOMIC, neural models are able to *generate* the social implications around situations described in text. In this chapter, we investigate whether we can *revise* text to change its social implications, specifically focusing on the implications captured by *connotation frames of power and agency* (Sap et al., 2017), This pragmatic formalism can uncover a specific type of gender biases in movies, namely, that male characters are portrayed with more power and agency than female characters (see Sap et al., 2017, for more details).

Thus, we introduce a new text revision task of controllable debiasing, to help debias the portrayal of characters through the lens of connotation frames of power and agency. To this end, we create PowerTransformer, a transformer-based encoder-decoder trained on a joint reconstruction and paraphrasing objective. Our approach demonstrates promising results to revise sentences with targeted power and agency, and outperforms ablations and baselines on both automatic and human evaluations. Finally, as a case study, we show the feasibility for controllable debiasing at debiasing the portrayal of characters in movie scripts.

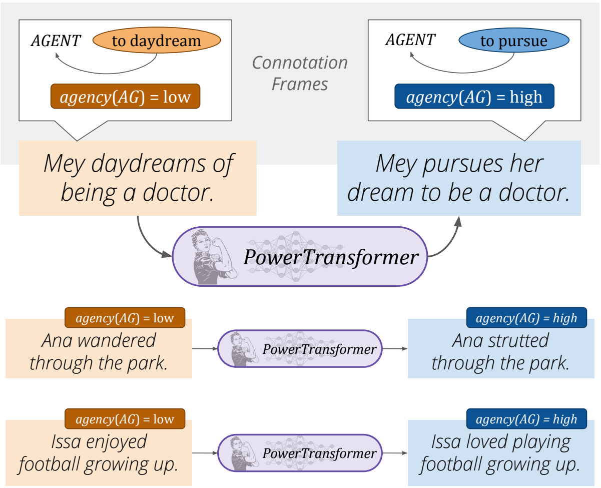

AGENT
to daydream
agency(AG) = low
Connotation Frames
AGENT
to pursue
agency(AG) = high
Mey daydreams of being a doctor.
Mey pursues her dream to be a doctor.
PowerTransformer
agency(AG) = low
Ana wandered through the park.
PowerTransformer
agency(AG) = high
Ana strutted through the park.
agency(AG) = low
Issa enjoyed football growing up.
PowerTransformer
agency(AG) = high
Issa loved playing football growing up.

Figure 3.1: Examples of using connotation frames (Sap et al., 2017) for controllable revisions to portray characters with more agency and power. In the second example, "Ana strutted" implies that she is more active and decisive, compared to "Ana wandered" which portrays her as aimless and passive.

3.1 Introduction

Narratives and news texts often reflect societal biases and stereotypes, such as the traditional gender role that women are passive and submissive (Lakoff, 1973; Fiske, 1993; Fast et al., 2016). The task of *controllable text revision*, i.e., rephrasing text to a targeted style or framing, can help correct for these biases by altering and equalizing the way people are described. For example, automatically rewriting *"Mey daydreamed about being a doctor"* as *"Mey pursued her dream to be a doctor"* portrays Mey with more authority and decisiveness (Figure 3.1). Such controllable revision methods could be used to help reshape how gender roles are portrayed in media (e.g., through machine-in-the-loop writing systems; Clark et al., 2018).

To edit such biases out of text, a controllable rewriting model faces three key challenges. First, a model should be able to make edits beyond surface-level paraphrasing, as simple paraphrasing will often not adequately debias the underlying events described. For example, Mey's portrayal in Figure 3.1 carries both overt bias (the choice of action) and subtle bias (the framing of the action), both of which require rewriting to be adequately debiased. Second, a model's debiasing revisions should be purposeful and precise and should not make unnecessary changes to the underlying meaning of the original text. Lastly, since parallel data does not exist, models must learn to revise and debias text without supervised data, thereby preventing straightforward machine translation-style modelling.

We formulate *controllable debiasing* as a new controllable text revision task that aims to correct the implicit and possibly unwanted bias against or towards a specific character portrayed in text (§3.2). As shown in Figure 3.1 (top), we study the portrayal biases through the lens of connotation frames of *power and agency* (Sap et al., 2017), which provide pragmatic knowledge about implied power and agency levels projected onto characters by a predicate.

We create POWERTRANSFORMER, an encoder-decoder model that rewrites sentences with a desired portrayal using agency connotation frames (§3.3). We combine a reconstruction and paraphrase objective into our model to overcome the lack of parallel supervised data, building off of the denoising autoencoder setup from Li et al. (2018a). To steer the revisions, we endow the model with connotation frame knowledge both at training time using control tokens, and at generation time using agency-based vocab boosting.

Our findings show that POWERTRANSFORMER is effective at rewriting sentences with desired agency connotations while only making minimal changes to their meaning, as measured through both human and automatic evaluations (§3.4). We also show that POWERTRANSFORMER significantly outperforms existing stylistic rewriting methods (Prabhumoye et al., 2018; Dathathri et al., 2020) on those metrics. Additionally, through ablations studies, we establish the usefulness of each component of the model, finding benefits from both the joint objective (47% gain in accuracy) and the agency scaling (12% gain in accuracy).

Finally, in §3.5, we apply controllable debiasing to a corpus of modern English movies (Gorinski and Lapata, 2015) as a step towards removing gender bias in character portrayal established by prior work (Sap et al., 2017). Using POWERTRANSFORMER, we revise the movie scripts and significantly increase the agency levels of female characters, thereby reducing the gender bias. Our findings show promise for using modern NLP tools to help mitigate societal biases in text. We release our preprocessed data and code at `http://maartensap.com/controllable-debiasing`.

3.2 Controllable Debiasing

Controllable debiasing is a novel formalization of stylistic rewriting that aims to debias the portrayal of characters through controllable revision. To achieve the desired character portrayal, a system must be able to change the underlying meaning of events, unlike certain formalizations (e.g., politeness transfer; Rao and Tetreault, 2018) where full meaning preservation is required. Without this, systems run the risk of merely paraphrasing the biases in text. However, revisions

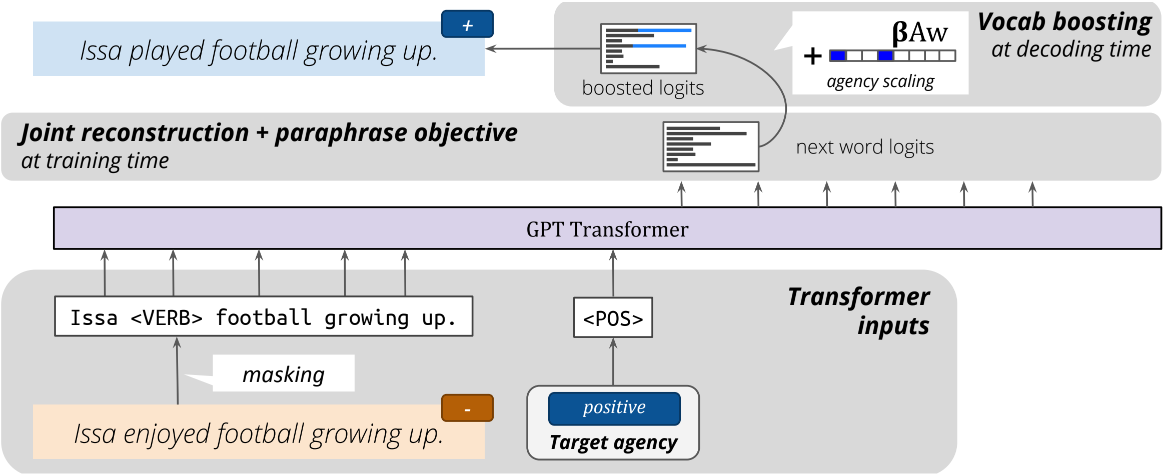


Figure 3.2: Overview of the full POWERTRANSFORMER model. An input sentence is masked for verb tokens indicative of agency. Masked inputs and target agency are used as GPT inputs. We use a joint objective using both paraphrase data and masked input sentences for training. At decoding time, we employ a vocab boosting technique to steer generations towards the target agency.

must be precise and avoid unnecessary meaning changes, which can often occur in stylistic rewriting (e.g., reversing the sentiment of a review drastically changes its underlying meaning).

For our new rewriting task of changing portrayal bias, we focus on connotation frames that measure the *power* and *agency* ascribed to characters through the actions they take. Connotation frames (Rashkin et al., 2016; Sap et al., 2017) distill implicit relations between a verb, its agent, and its theme. In this work, we use the positive, neutral, and negative agency dimensions, where agency is defined as the capacity to intentionally make changes or act upon one's environment (Dennett, 1989). For example, illustrated in Figure 3.1, "X pursued Y" implies that X has positive agency.[1] Using machine-in-the-loop writing systems (e.g., Ghazvininejad et al., 2016, 2017; Clark et al., 2018, Textio[2]), models trained on this task could help authors write news, stories, or movies that portray characters in less biased ways, and thereby help mitigate the negative effects of stereotypical portrayals in media (Behm-Morawitz and Mastro, 2008; Field et al., 2019a).

3.3 POWERTRANSFORMER

We present a new approach for controllable debiasing called POWERTRANSFORMER, which addresses two key challenges: the paucity of parallel supervised data for training and the difficulty of incorporating fine-grained control for steering the agency of the output. Our approach (Figure 3.2) jointly learns to reconstruct partially masked story sentences while also learning to paraphrase from an external corpus of paraphrases (§3.3.2). At generation time, we also include a boosting method for fine-grained steering towards the desired agency level as described in §3.3.3.

[1] Future work could explore using the power dimension instead of agency, or alternative operationalizations of biases, e.g., Social Bias Frames (Sap et al., 2020) or *regard* towards minorities as introduced by Sheng et al. (2019).

[2] https://textio.com/

3.3.1 Model Overview

POWERTRANSFORMER is an encoder-decoder style model with an OpenAI-GPT transformer model (Radford et al., 2018) as the base. The input sentence $\mathbf{x}$ is converted to a sequence of byte pair encodings (BPE) $\{x_1, ..., x_n\}$, and given to the encoder after being scrubbed of its agency markers as described below. To steer the model, we also give the encoder the target agency t, which we represent as one of three special tokens {<Pos>,<Equal>,<Neg>}.[3]

3.3.2 Joint Objective

We train our model on both a reconstruction and a paraphrasing task, for which inputs are masked and paraphrased versions of the output, respectively.

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{para}} \tag{3.1}$$

Masking and Reconstructing Inspired by the delete-retrieve-generate model from Li et al. (2018a), this objective teaches the model to recover masked out agency-associated verbs in sentences. We first assign an agency level to an input sentence by counting verbs in the agency lexicon from Sap et al. (2017).[4] Then, we mask out all verbs indicative of the agency level, replacing them with a special <VERB> token. In this setup, the target output is the original sentence $\mathbf{x} = \{x_1, ..., x_n\}$, with the masked sentence $\hat{\mathbf{x}}$ and the target agency level t as inputs. During training, we minimize the cross entropy of the target output sentence given the inputs:

$$\mathcal{L}_{\text{recon}} = -\frac{1}{n}\sum_{i=1}^{n} \log p(x_i | x_{<i}, \hat{\mathbf{x}}, t) \tag{3.2}$$

Paraphrasing To go beyond reconstructing sentences, we add a paraphrasing objective using an out-of-domain paraphrase corpus (§3.4.1). We extract agency levels for each sentence and its paraphrase and mask out the agency verbs in the input, using the same methods as described above. Here, the inputs are the masked sentence $\hat{\mathbf{x}}$ and the target agency t, while the target output $\mathbf{y} = \{y_1, ..., y_m\}$ is the paraphrase. As with reconstruction, we minimize the cross entropy of the target output given the inputs:

$$\mathcal{L}_{\text{para}} = -\frac{1}{m}\sum_{i=1}^{m} \log p(y_i | y_{<i}, \hat{\mathbf{x}}, t) \tag{3.3}$$

3.3.3 Controlled Decoding with Vocab Boosting

We employ a vocab-boosting technique during generation to encourage models towards generating with the desired agency, inspired by Ghosh et al. (2017). At each decoding timestep i, we re-scale the unnormalized token probabilities (logits $l_i \in \mathbb{R}^V$, where V is the vocabulary size) to

[3] In earlier experiments, we also provided the original agency as an input to the model during training and decoding, but found that it made little difference in performance.

[4] For sentences that have multiple verbs, we assign the agency level that the most verbs in the sentence have (e.g., a sentence with two positive agency verbs and one negative agency verb will be assigned positive agency).

boost the likelihood of predicting words with the target agency. The next token probabilities are then computed using the "boosted" logits:

$$P(y_i|y_{<i}, x, t) \propto \text{softmax}(l_i + \beta \cdot Aw) \tag{3.4}$$

where A is a $\mathbb{R}^{V\times 3}$ matrix that represents a 3-dimensional {positive, equal, and negative} agency embedding for each token in the vocabulary, w is a $\mathbb{R}^3$ one-hot vector denoting the target agency for the output, and β is a scalar hyperparameter representing the boosting strength. We create A manually using the verbs in the agency lexicon (Sap et al., 2017).[5] Used only at decoding time, this method effectively increases the likelihood of using a word with the target agency level.

3.4 Controllable Debiasing Experiments

In this section, we describe three experiments for investigating POWERTRANSFORMER performance. First, we evaluate performance of our full model and ablated baselines, using automatic metrics to quantify the effectiveness of each modelling component (§3.4.4). Next, we compare our full model to baselines from related work (§3.4.5). Lastly, given the limitations of automated metrics for evaluating generations (Liu et al., 2016a; Mir et al., 2019), we obtain human judgments of model performance through crowdsourcing (§3.4.6). We additionally include examples of generations in Table 3.4.

3.4.1 Datasets

In our experiments, we use a dataset of short stories for the reconstruction task and a parallel corpus of paraphrases for both paraphrase and reconstruction tasks. We show data statistics in Table 3.1, with additional preprocessing details in Appendix B.1.

	Type	# Instances	Pos	Neutral	Neg
ROC	train	10721	3834	4151	2736
	dev	1803	633	710	460
	test	899	325	350	224
Para.	train	45000	16410	14153	14437
	dev	10000	3645	3328	3127

Table 3.1: Statistics for our main story sentences dataset (ROC) and for the external paraphrase corpus (Para.).

ROC story corpus The main focus of our study is controllable revision of story sentences; therefore, we select sentences from the ROC story corpus (ROC; Mostafazadeh et al., 2016b). After extracting agency levels for all sentences from the training stories, we sample roughly equal amounts of all three agency levels, and randomly split sentences into training, development, and test sets.[6]

Paraphrase corpus As additional training data, we use the corpus of automatically aligned paraphrases of TV subtitles (Creutz, 2018, Para.). As with the ROC story corpus, we extract agency levels for each sentence and its paraphrase, then sample roughly equal amounts of pairs with all

[5] Since our model operates on BPE tokens, we manually set the first BPE token of every tense of every verb to the desired agency. We also experimented with learning A from data, but found no improvement over manually setting it.

[6] We use a 80:13:7 train, development, test ratio.

Ablations using the Development Set

	Main Metrics		Additional Metrics		
	Agency	**Meaning**	**Fluency**	**Repetition**	**Diversity**
PowerTransformer variants	Acc (↑)	BertScore (↑)	PPL (↓)	w/ Rep (↓)	Unique (↑)
(*ParaOnly+noBoost*)	.30	.95	**58.76**	.002	.54
(*ParaOnly+Boost*)	.42	.90	76.25	**.001**	.59
(*Joint+noBoost*)	.77	**.96**	70.61	.007	.87
(*Joint+noBoost*)+*SupplyVerb*	.77	**.96**	94.54	.004	.92
Full = (*Joint+Boost*)	**.89**	**.96**	76.78	.015	**.99**

Table 3.2: Ablation study results on the development set. We present separate metrics for evaluating the change in agency, the meaning preservation, fluency, repetitiveness and diversity of the output (bolding the best performance). (↑) indicates that higher is better and (↓) indicates that lower is better.

different sentence-paraphrase agency combinations (further details in §B.1.2). We randomly split the data into 45k train and 10k dev. instances (Table 3.1).[7]

3.4.2 Metrics

In addition to human evaluations, we also use a variety of automated evaluation metrics to characterize different aspects of performance. We measure the accuracy of the change in agency by comparing the target agency level with that of the output (extracted using the connotation frames lexicon). As a measure of meaning preservation, we use BERT-score F1 metrics (Zhang et al., 2020) to compare the semantic similarity of the input sentence with the machine output.

As additional metrics, we measure the fluency, the repetitiveness, and diversity of the output. Following previous work (Dai et al., 2019), we measure fluency as perplexity (*PPL*) of the output sentence using a pre-trained GPT model that has not been fine-tuned for this task. As an additional metric of potential text degeneration, we compute the fraction of output sentences that have a bigram that is repeated two or more times (*w/ rep*). Finally, we compute the fraction of generations that are unique with respect to the rest of the output, to ensure diverse, input-specific generations (*unique*).

3.4.3 Experimental Setup

We randomize ROC story and paraphrase data, and use OpenAI GPT LM as our pretrained model. For decoding, we use top-p=0.4 nucleus sampling (Holtzman et al., 2020), and a boosting strength of β=5 (hyperparameters and details in §B.2.1).

3.4.4 Investigating Effectiveness of Approach

We first establish our model's effectiveness at controllable debiasing on our dev. set, and investigate the importance of various components in our approach through ablation analyses. For qualitative analyses, we also show example revisions in Table 3.4 (and Table B.2 in the appendix).

[7] Since this is just additional training data, we do not test our models on this corpus, but do use the dev. set for selecting some hyperparameters.

Test Set Comparisons (pos-to-neg and neg-to-pos set)

	Main Metrics		Additional Metrics		
	Agency Acc (↑)	**Meaning** BertScore (↑)	**Fluency** PPL (↓)	**Repetition** w/ rep (↓)	**Diversity** unique (↑)
PPLM (Dathathri et al., 2020)	.13	.95	106.12	.053	**1.00**
BST (Prabhumoye et al., 2018)	**.88**	.83	**91.22**	.053	0.79
PowerTransformer	.86	**.96**	95.19	**.015**	**1.00**

Table 3.3: Performance of different re-writing methods on the neg-to-pos and pos-to-neg subsets of the test set (bolding the best performance). We evaluate the change in agency and the meaning preservation. As secondary metrics, we include fluency, repetitiveness, and diversity of output.

Ablated Baselines

We first investigate the importance of the reconstruction objective, by comparing our joint objective model (*Joint*) with a model trained with just the paraphrasing objective (without masking, *ParaOnly*). Then, to quantify the effect of boosting, we compare models with (*Boost*) and without (*noBoost*) agency-specific vocab boosting. Note that *ParaOnly+noBoost* is equivalent to a GPT-based encoder-decoder model, similar to seq2seq frameworks commonly used in paraphrasing tasks (Cao et al., 2017; Li et al., 2018b; Prakash et al., 2016).

As a final comparison, we implement a model variant that more closely mirrors the delete-retrieve-generate paradigm (Li et al., 2018a) by adding a "retrieve" step in which we concatenate transformer input with a verb retrieved from the verb agency lexicon that is most similar to the masked out verb (*SupplyVerb*).[8]

Results

In Table 3.2, our results show that the full model (*Joint+Boost*) yields text revisions with the most accurate target agency and the most meaning preservation. In general, we find that both the joint objective and vocab boosting (*Boost*) substantially increase the target agency accuracy, as also illustrated in examples (d) and (e) in Table 3.4. However, unsurprisingly, vocab boosting also slightly lowers fluency, yielding higher perplexities than models' non-boosted counterparts. Our results also show that using the joint objective with boosting increases the diversity of output, but causes marginally more repetition of bigrams.

Counterintuitively, our ablations show that supplying a verb to the model as an explicit retrieval step (*SupplyVerb*) does not improve the agency or meaning metrics and actually hurts the fluency of the output (as measured by higher perplexities). Upon qualitative investigation (Table B.2 in the appendix), the retrieved verb is often related to a different word sense of the masked verb, breaking the grammaticality of the sentence.

[8] We retrieve a verb from the Sap et al. (2017) lexicon that has the target agency and is most similar to the masked out verb, where similarity is defined as cosine distance between word embeddings using GloVe 300-d embeddings (Pennington et al., 2014b).

3.4.5 Comparison with External Approaches

To further validate our approach, we compare against two baselines from related style transfer and stylistic generation tasks. As these models were designed for binary style transfer, we only report our baseline and model results on the positive and negative agency portions of our data.

Baselines

BST We compare to the backtranslation style transfer model from Prabhumoye et al. (2018). This model first translates input sentences to a pivot language (preserving the meaning but losing language-specific style), then relies on style-specific decoder-translators for generating the output sentence. We include set-up details in §B.2.3.

PPLM Recent work in controllable generation has introduced PPLM, a new plug-and-play technique with promising results for decoding stylistic text (Dathathri et al., 2020). This method operates on an underlying neural language model at decoding time. It uses backpropagation from a stylistic discriminator to update the past and present hidden representations to be more consistent with the targeted style or domain. We adapt the approach to controllable revision by replacing the base language model with an autoencoder trained on a reconstruction objective, described in detail in §B.2.2.

Results

We present results in Table 3.3. Our experiments show that POWERTRANSFORMER performs better than the baselines overall. Specifically, while the BST revisions obtain slightly higher accuracy on the output agency levels, these revisions have the both the lowest diversity and meaning preservation, suggesting the model ignores the input (Table 3.4). PPLM shows opposite trends, yielding the lowest accuracy with high meaning preservation and high diversity of generations. Illustrated in Table 3.4, this model often makes less purposeful and less concise alterations.

3.4.6 Evaluating with Human Judgements

To validate our automatic evaluations, we collect human judgments of the controllable revisions from several baselines and POWERTRANSFORMER (*Joint+Boost*).

Human Evaluation Task

We design a head-to-head[9] crowdsourcing task on Amazon Mechanical Turk where we ask raters to compare two outputs from different models given the same input sentence and target agency (see Figure B.1 in the appendix). We first ask them to judge whether either output is gibberish, then, in two questions, choose which revision has better targeted agency and which better preserves the meaning of the original sentence. For consistency, each pair is rated by three judges. To ensure the quality of our evaluations, we selected workers who could reliably distinguish high from low agency sentences in a qualification task (see Figure B.2 in the appendix).

[9] We use head-to-head evaluations as those have been shown to be more reliable than scale-rating evaluations (Kiritchenko and Mohammad, 2017).

dir.		Input	Model	Revised Sentence (out)	Agency(out)
agency (+ → −)	(a)	After the party I headed home.	PPLM	after the party my classmate, Kayla and the Tgirls of the Universe.	=
			BST	please 's , i have a word of this .	−
			PowerT $_{Joint+NoBoost}$	after the party i stayed home.	−
			PowerT $_{Joint+Boost}$	after the party i stayed home.	−
	(b)	A friend asked me to watch her two year old child for a minute.	PPLM	a friend asked me to watch her two year old child for a minute.	+
			BST	l didn 't have a word of this , you 're .	−
			PowerT $_{Joint+NoBoost}$	a friend needed me to watch her two year old child for a minute.	−
			PowerT $_{Joint+Boost}$	a friend needed me to watch her two year old child for a minute.	−
	(c)	After filling in the data it looked quite sharp.	PPLM	before filling the last question it it it it looked quite sharp. before filling the last question it it	+
			BST	when the 't you want a word ?	−
			PowerT $_{Joint+NoBoost}$	after analyzing in the data it looked quite sharp.	=
			PowerT $_{Joint+Boost}$	after seeing in the data it seemed quite sharp.	−
agency (− → +)	(d)	Allie was failing science class.	PPLM	allie was failing science grade.	−
			BST	do you want me ?	+
			PowerT $_{Joint+NoBoost}$	allie was failing science class.	−
			PowerT $_{Joint+Boost}$	allie was taking science class.	+
	(e)	Darla wanted a soft drink.	PPLM	darla wants a hard hard drink.	−
			BST	don 't take me a man .	+
			PowerT $_{Joint+NoBoost}$	darla ordered a soft drink.	+
			PowerT $_{Joint+Boost}$	darla ordered a soft drink.	+
	(f)	Clint paused on the trail.	PPLM	clint was on the trail.	=
			BST	don 't you want me ,	−
			PowerT $_{Joint+NoBoost}$	clint hiked on the trail.	=
			PowerT $_{Joint+Boost}$	clint walked on the trail heading down.	+

Table 3.4: Example sentences from our dev. set, along with their revisions from various models and the achieved agency levels (Agency(out)). Examples (a)-(c) should be rewritten from high to low agency, and (d)-(f) from low to high agency. Confirming our quantitative results in Tables 3.2 and 3.3, PowerTransformer (*Joint+Boost*) is the most effective at making purposeful and precise changes to the input sentences to alter their agency while minimally changing their meaning. Revisions from more models are listed in Table B.2 (in the appendix).

For this evaluation, we generate three revisions–one for each target agency level–for a random subset of 100 test examples. We compare the output of our full PowerTransformer model with two external baselines (PPLM and BST). For further comparison, we also include the most competitive ablated baseline from Table 3.2 (i.e., *Joint+noBoost*).

Results

In Figure 3.3, we show the percentages of times in which POWERTRANSFORMER was preferred over the three baseline models.[10] Percentages >50% indicate a preference towards POWERTRANSFORMER.

Overall, the sentence revisions by POWERTRANSFORMER are preferred over all of the baselines in obtaining the desired agency level. For meaning preservation, our model is always selected over BST, mirroring BertScores in Table 3.3. The difference is less stark when comparing to PPLM which sometimes makes no changes or irrelevant changes to the input sentence, and reversed when comparing to the ablated *noBoost*.

Additionally, BST revisions were marked as gibberish substantially more than those by other models (63% vs. 3-7%). While this seemingly contradicts BST's low perplexity scores, this is in line with previous work showing automatic fluency metrics can favor degenerate, bland, or repetitive language (Holtzman et al., 2020).

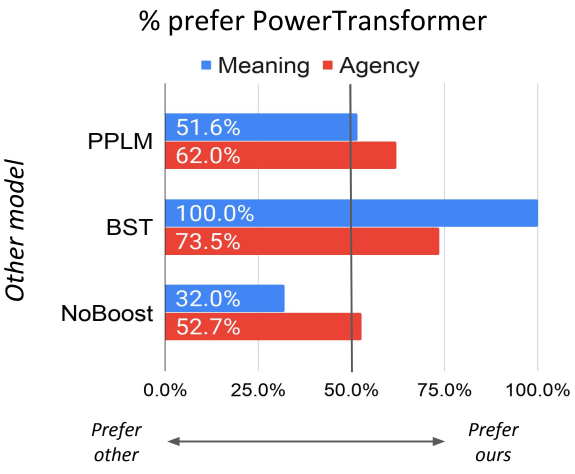


Figure 3.3: Human judgements of target agency and meaning preservation in POWERTRANSFORMER vs. three other model variants. Selection rates >50% indicate preference towards our model.

3.5 Gender Bias in Movies

As a proof-of-concept of controllable debiasing, we investigate whether gender biases in portrayals of movie characters can be mitigated using POWERTRANSFORMER.

3.5.1 Movie Scripts Corpus

We draw our data from the 767 modern English movie scripts by Gorinski and Lapata (2015), focusing on the narrations which describe characters and their actions (as opposed to the character's dialogue utterances). Described in further detail in §B.3 in the appendix, we automatically extract characters and assign them a binary[11] gender (man, woman) using a list of highly gendered names (e.g., “Sarah”, “William”) and a list of gendered words (e.g., “waiter,” “waitress”). Following previous work (Ramakrishna et al., 2017; Sap et al., 2017), we assign narration sentences to characters if their name appears in them.

Our corpus contains 16,763 characters from 767 different English movies. Of those characters,

[10] Judgments in our evaluation task had an average pairwise agreement of 75% (Krippendorf's α=.52).

[11] Note that gender is a social construct that goes beyond the man-woman binary (Lorber et al., 1991), however more inclusive analyses (e.g., with non-binary genders) are not possible given the limited information about the individuals mentioned in our data.

68% are inferred to be men and only 32% to be women,[12] consistent with known gender skews in movie characters (Google, 2017). This bias in representation is also present at the narrative level. Specifically, female characters are only mentioned in $n_{narr,f}$ =27 narrations on average, compared to $n_{narr,m}$ =34 narrations for male characters (Cohen's $|d| = 0.13$, $p < 0.001$). Similarly, compared to their male counterparts, female characters are described in significantly fewer words ($n_{words,f} = 329$, $n_{words,m} = 435$, $|d| = 0.14$, $p < 0.001$) and with fewer verbs ($n_{verbs,f} = 41$, $n_{verbs,m} = 54$, $|d| = 0.13$, $p < 0.001$).

3.5.2 Debiasing Portrayal in Movies

Given the known bias that female characters are portrayed with less agency (Sap et al., 2017), our goal is to re-balance their agency levels to be more on par with those of male characters. Therefore, we revise only the sentences describing female characters to have higher agency, using POWERTRANSFORMER. Then we extract connotation frames of agency for revised script sentences, and aggregate per character. Shown in Figure 3.4, revisions successfully increase the instances of positive agency of female characters, and decrease their negative agency or passiveness.

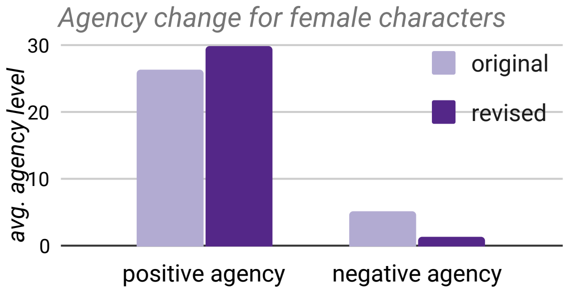


Figure 3.4: Average agency levels (i.e., number of agency verbs) for female characters in original and revised scripts. POWERTRANSFORMER can revise the portrayals of female characters in movies to give them higher positive agency and lower negative agency.

We further examine the change in gender association of positive and negative agency, to verify the effectiveness of controllable debiasing. We first count all the positive and negative agency verbs used to describe characters (in original or rewritten sentences). Following Sap et al. (2017), we then fit a logistic regression model to quantify the association between character's gender with their agency levels, controlling for their number of words, verbs, and narrations. For better interpretation of the β coefficients, we z-score all the continuous variables.

We confirm that indeed, controllable debiasing using POWERTRANSFORMER can reverse the bias in portrayal in movies. In original scripts, male characters were portrayed with significantly higher positive agency ($\beta_{\text{pos}} = 1.2$, $p < 0.001$) and lower negative agency ($\beta_{\text{neg}} = -0.3$, $p < 0.001$) than female characters. However, our model successfully reverses this gender bias, portraying women with significantly more positive agency ($\beta'_{\text{pos}} = -62.6$, $p < 0.001$) and significantly less negative agency ($\beta'_{\text{neg}} = 8.7$, $p < 0.001$).

Our findings on movie scripts show the promise of using controllable debiasing to successfully mitigate gender biases in portrayal of characters, which could be extended to other domains (e.g., news or fiction, Field and Tsvetkov, 2019; Fast et al., 2016). Additionally, future work could consider alternative views of portrayal biases (e.g., "regard" or bias directed at different demographic groups; Sheng et al., 2019; Sap et al., 2020), or use more holistic views of gender roles (e.g., "masculine default" cultures; Cheryan and Markus, 2020).

[12] There were 2597 characters for which the gender could not be inferred.

3.6 Related Work

controllable debiasing is a new formalization of the unsupervised stylistic rewriting task, contrasting with supervised approaches which benefit from parallel corpora (e.g., Xu et al., 2012, 2015; Rao and Tetreault, 2018; Pryzant et al., 2020). In unsupervised settings, a majority of work has dealt with the dearth of parallel data by using encoder-decoder setups paired with discriminators to disentangle style from content and steer generations (e.g., Shen et al., 2017; Zhang et al., 2018a; Fu et al., 2018; Yang et al., 2018; Niu and Bansal, 2018; Romanov et al., 2019; Dai et al., 2019; John et al., 2019) or backtranslation setups (Prabhumoye et al., 2018; Lample et al., 2018). In contrast, Li et al. (2018a) introduce a modular approach (later adapted to transformer models by Sudhakar et al., 2019) that relies on drop-in replacement of attribute markers followed by language correction. POWERTRANSFORMER improves on this approach with an additional out-of-domain paraphrasing objective.

While a majority of related existing stylistic rewriting work defines style as sentiment (e.g., on reviews), a notable exception is Nogueira dos Santos et al. (2018), who use stylistic rewriting to make text less hateful or offensive. Similar in spirit, controllable debiasing is a novel formalization that aims to address and revise social biases expressed in text, but using the nuanced implications distilled in connotation frames of power and agency instead of binary offensiveness.

Our work also draws inspiration from controllable generation methods (e.g., Koncel-Kedziorski et al., 2016; Hu et al., 2017; Ficler and Goldberg, 2017). While those methods steer the generation output to contain desired attributes, controllable revision is constrained to revise an input sentence in addition to generating with desired attributes.

3.7 Summary

In this chapter, we tackled the challenge of changing the social implications of text using controllable text revision. Specifically, we focused on the a new text revision task of controllable debiasing, to help debias the portrayal of characters through the lens of connotation frames of power and agency. To this end, we created POWERTRANSFORMER, a transformer-based encoder-decoder trained on a joint reconstruction and paraphrasing objective. Our approach demonstrated promising results to revise sentences with targeted power and agency, and outperformed ablations and baselines on both automatic and human evaluations. Finally, as a case study, we showed the feasibility for controllable debiasing at mitigating the gender biases in character portrayals in movie scripts.

Our findings highlight the potential of neural NLP models as a tool for editing text to obtain a desired social meaning. Specifically, within a human-AI collaborative writing setup, such text editing systems could provide alternative phrasings that can expand the creativity of the resulting text (Ghazvininejad et al., 2016, 2017; Clark et al., 2018). Additionally, our promising controllable debiasing results using connotation frames of power and agency opens the door for other socially aware text revision systems that can correct a wider range of social biases (e.g., microaggressions, toxic language).

Part II

Understanding and Detecting Social Biases in Language

Warning, this part discusses content that is sensitive or offensive in nature

Chapter 4

A Primer on Social Biases and Toxicity in Language

As we know, the way we interpret the meaning of an utterance or text heavily relies on our background knowledge about the world and its social dynamics and inequalities (Kintsch, 1988; McGarty, 2018). But language can also help reinforce these dynamics and this inequality (Lakoff, 1973; Fiske, 1993), e.g., by invoking stereotypes or calling for discrimination against minority groups (Figure 4.1). By being trained on large amounts of text, NLP systems will indubitably learn to produce statements with biased or toxic meanings (Sheng et al., 2019; Gehman et al., 2020), which significantly hinders their fairness and safety (Bender et al., 2021).

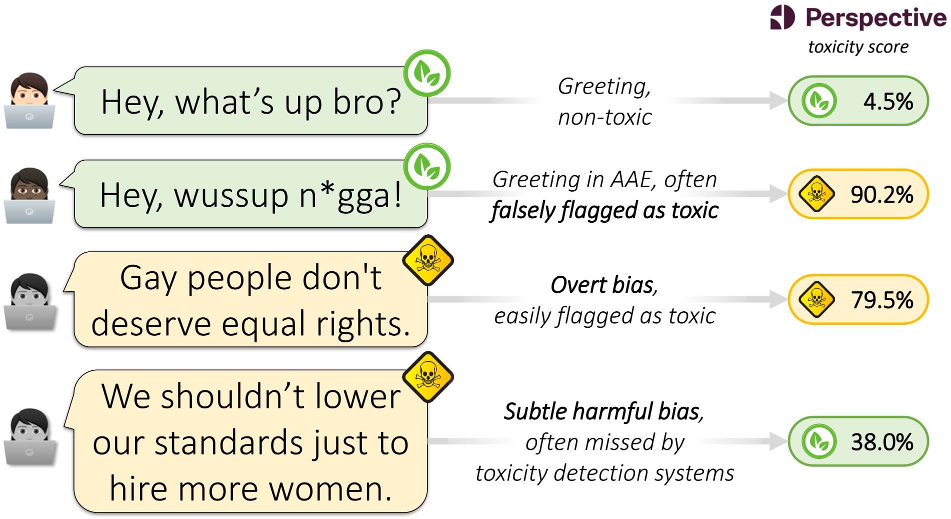

Perspective
toxicity score
Hey, what's up bro?
Greeting, non-toxic
4.5%
Hey, wussup n*gga!
Greeting in AAE, often falsely flagged as toxic
90.2%
Gay people don't deserve equal rights.
Overt bias, easily flagged as toxic
79.5%
We shouldn't lower our standards just to hire more women.
Subtle harmful bias, often missed by toxicity detection systems
38.0%

Figure 4.1: Examples of utterances that showcase how complex the toxicity detection task is, along with toxicity scores from a popular toxicity detection tool (PerspectiveAPI). The first two greetings are virtually equivalent in meaning, yet the African American English (AAE) greeting can often be falsely flagged as toxic by humans or machines. The bottom two utterances both have harmful implications, but the subtle bias is often missed by toxicity detection systems which only look for overt biases.

However, there are many challenges to achieving a machine that can reason about biased or harmful implications of text, as illustrated in Figure 4.1. One of the main challenges, as discussed in Chapter 5, is that determining what is offensive, biased, or harmful, depends on the social context of the utterance. For example, in Figure 4.1, the harmless greeting "Hey wussup n*gga" in African American English (AAE) can be misinterpreted as more offensive by non-Black listeners compared to its "general" English counterpart (Green, 2002; Rosa, 2019). Ignoring the context of speech can in turn cause NLP systems to perpetuate these biases, which often leads to minority speech being flagged as toxic more often (Sap et al., 2019a; Oliva et al., 2021).

Another big challenge is that distilling the biased or harmful meanings of text has been simplified down to a binary task of flagging toxicity. Such a task framing misses the subjectivity of the problem at hand, often cannot handle nuanced or subtle biases, and lacks interpretability (Ross et al., 2017; Dinan et al., 2019; Breitfeller et al., 2019). Instead, in Chapter 6, we propose SOCIAL BIAS FRAMES, a new conceptual formalism that aims to capture the biased and harmful implications in language, whether overt or subtle. For example, our framework aims to explain that the statement "We shouldn't lower our standards just to hire more women" carries the subtle but biased implication that "women candidates are less qualified," which perpetuates gender inequality. Distilling these harmful implications constitutes a more nuanced view of toxicity and social bias, and incorporates explanations of bias for humans and machines to use.

There are many other challenges to capturing these harmful implications in text, such as the impact of social context on the meaning of utterances (e.g., speaker and listener identity; Hovy and Yang, 2021), as we discuss in Chapter 7.

Chapter 5

Uncovering Racial Bias in Hate Speech Detection

This chapter discusses work originally published in Sap et al. (2019a).

5.1 Introduction

Toxic language (e.g., hate speech, abusive speech, or other offensive speech) primarily targets members of minority groups and can catalyze real-life violence towards them (O'Keeffe et al., 2011; Cleland, 2014; Mozur, 2018). Social media platforms are under increasing pressure to respond (Trindade, 2018), but automated removal of such content risks further suppressing already-marginalized voices (Yasin, 2018; Dixon et al., 2018). Thus, great care is needed when developing automatic toxic language identification tools.

The task is especially challenging because what is considered toxic inherently depends on social context (e.g., speaker's identity or dialect). Indeed, terms previously used to disparage communities (e.g., "n*gga", "queer") have been reclaimed by those communities while remaining offensive when used by outsiders (Rahman, 2012). Figure 5.1 illustrates how phrases in the African American English dialect (AAE) are labelled by a publicly available toxicity detection tool as much more toxic than general American English equivalents, despite their being understood as non-toxic by AAE speakers (Spears, 1998, see §5.2).

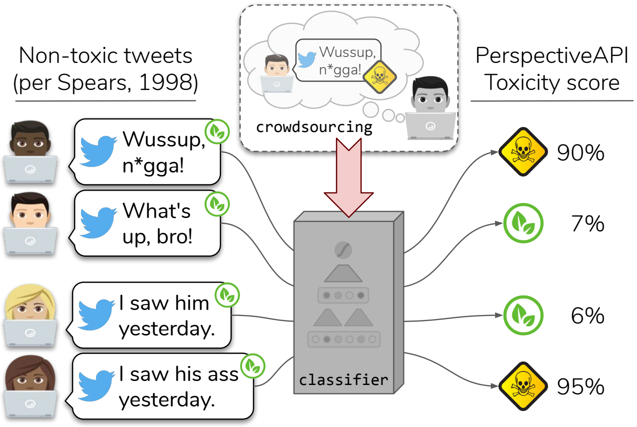


Figure 5.1: Phrases in African American English (AAE), their non-AAE equivalents (from Spears, 1998), and toxicity scores from PerspectiveAPI.com. Perspective is a tool from Jigsaw/Alphabet that uses a convolutional neural network to detect toxic language, trained on crowdsourced data where annotators were asked to label the toxicity of text without metadata.

In this chapter, we first empirically characterize the racial bias present in several widely used Twitter corpora annotated for toxic content, and quantify the propagation of this bias through models trained on them (§5.3). We establish strong associations between AAE markers (e.g., "n*ggas", "ass") and toxicity annotations, and show that models acquire and replicate this bias: in other corpora, tweets inferred to be in AAE and tweets from self-identifying African American users are more likely to be classified as offensive.

Second, through an annotation study, we introduce a way of mitigating annotator bias through *dialect* and *race priming*. Specifically, by designing tasks that explicitly highlight the inferred dialect of a tweet or likely racial background of its author, we show that annotators are significantly less likely to label an AAE tweet as offensive than when not shown this information (§5.4).

Our findings show that existing approaches to toxic language detection have racial biases, and that text alone does not determine offensiveness. Therefore, we encourage paying greater attention to the confounding effects of dialect and a speaker's social identity (e.g., race) so as to avoid unintended negative impacts.

5.2 Race and Dialect on Social Media

Since previous research has exposed the potential for other identity-based biases in offensive language detection (e.g., gender bias; Park et al., 2018), here we investigate *racial* bias against speech by African Americans, focusing on Twitter as it is a particularly important space for Black activism (Williams and Domoszlai, 2013; Freelon et al., 2016; Anderson et al., 2018). Race is a complex, multi-faceted social construct (Sen and Wasow, 2016) that has correlations with geography, status, dialect, and more. As Twitter accounts typically do not have self-reported race information, researchers rely on various correlates of race as proxies. We use the African American English *dialect* (AAE) as a proxy for race. AAE is a widely used dialect of English that is common among, but not unique to, those who identify as African American,[1] and is often used in written form on social media to signal a cultural identity (Green, 2002; Edwards, 2004; Florini, 2014).

Dialect estimation In this work, we infer dialect using a lexical detector of words associated with AAE or white-aligned English. We use the topic model from Blodgett et al. (2016a), which was trained on 60M geolocated tweets and relies on US census race/ethnicity data as topics. The model yields probabilities of a tweet being AAE (p_{AAE}) or White-aligned English (p_{white}).[2]

5.3 Biases in Toxic Language Datasets

To understand the racial and dialectic bias in toxic language detection, we focus our analyses on two corpora of tweets (Davidson et al., 2017a; Founta et al., 2018a) that are widely used in hate speech detection (Park et al., 2018; van Aken et al., 2018; Kapoor et al., 2018; Alorainy et al., 2018; Lee et al., 2018; Waseem et al., 2018).[3] Different protocols were used to collect the tweets in these corpora, but both were annotated by Figure-Eight[4] crowdworkers for various types of toxic language, shown in Table 5.1.

DWMW17 (Davidson et al., 2017a) includes annotations of 25K tweets as *hate speech*, *offensive* (but not hate speech), or *none*. The authors collected data from Twitter, starting with 1,000 terms

[1] Of course, many African Americans might not use AAE in every context, or at all. For further discussion of AAE, please refer to Blodgett et al. (2016a).

[2] The model yields AAE, Hispanic, Asian/Other and White-aligned dialect probabilities, but for the purpose of our study we only focus on AAE and White-aligned dialects.

[3] Our findings also hold for the widely used data from Waseem and Hovy (2016a). However, because of severe limitations of that dataset (see Schmidt and Wiegand, 2017; Klubička and Fernandez, 2018), we relegate those analyses to supplementary (§C.2).

[4] www.figure-eight.com

from HateBase (an online database of hate speech terms) as seeds, and crowdsourced at least three annotations per tweet.

FDCL18 (Founta et al., 2018a) collects 100K tweets annotated with four labels: *hateful*, *abusive*, *spam* or *none*. Authors used a bootstrapping approach to sampling tweets, which were then labelled by five workers on the FigureEight crowdsourcing platform.

5.3.1 Data Bias

	category	count	AAE corr.
DWMW17	hate speech	1,430	-0.057
	offensive	19,190	0.420
	none	4,163	-0.414
	total	**24,783**	
FDCL18	hateful	4,965	0.141
	abusive	27,150	0.355
	spam	14,030	-0.102
	none	53,851	-0.307
	total	**99,996**	

Table 5.1: Number of tweets in each category, and correlation with AAE (Pearson r, $p \ll 0.001$). We assign tweets to categories based on the label for FDCL18, and majority class for DWMW17. Correlations are colored for interpretability.

To quantify the racial bias that can arise during the annotation process, we investigate the correlation between toxicity annotations and dialect probabilities given by Blodgett et al. (2016a).

Table 5.1 shows the Pearson r correlation between p_{AAE} and each toxicity category. For both datasets, we uncover strong associations between inferred AAE dialect and various hate speech categories, specifically the "offensive" label from DWMW17 ($r = 0.42$) and the "abusive" label from FDCL18 ($r = 0.35$), providing evidence that dialect-based bias is present in these corpora. As additional analyses, we examine the interaction between unigrams indicative of dialect and hate speech categories.

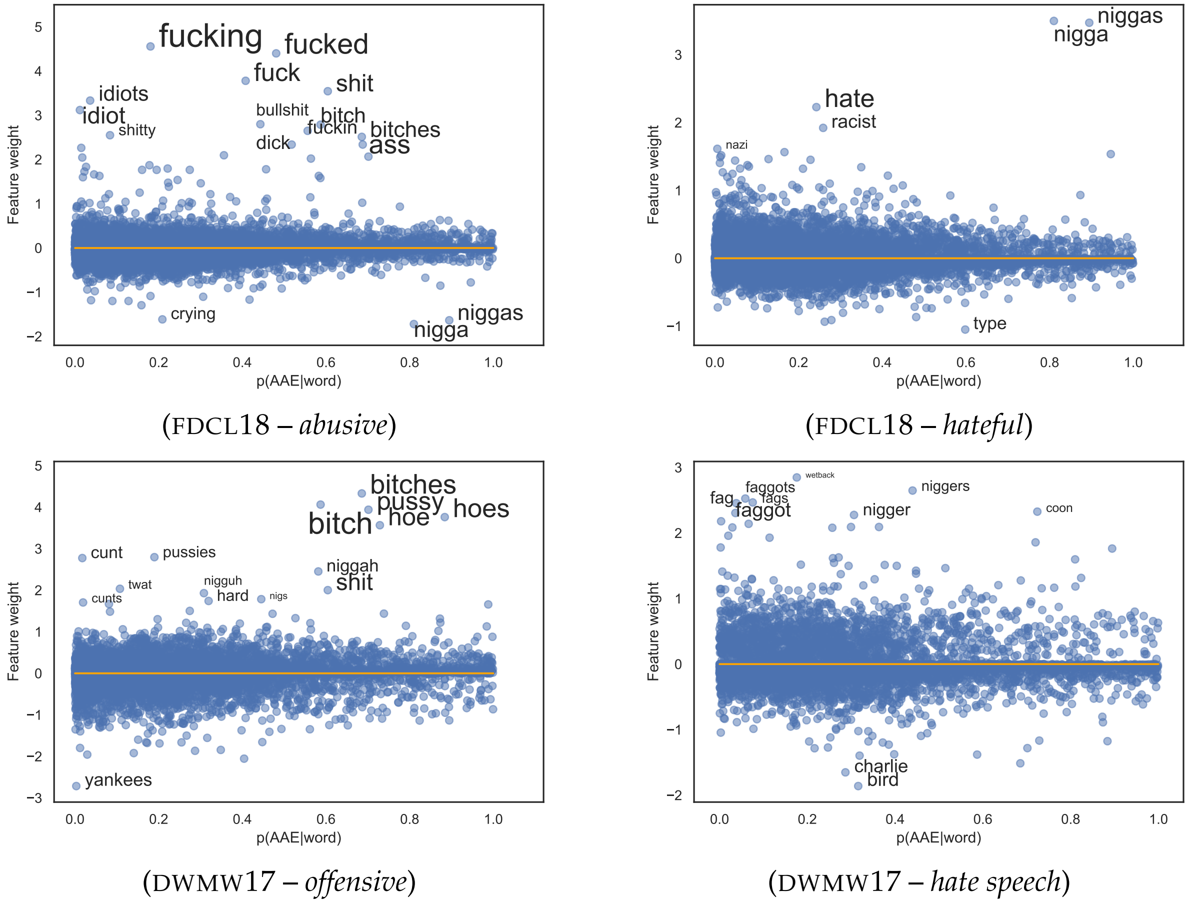
To better understand the correlations between inferred dialect and the annotated hate speech categories (abusive, offensive, etc.) we use simple linear models to look for influential terms. Specifically, we train l_2-regularized multiclass logistic regression classifiers operating on unigram features for each of DWMW17 and FDCL18 (tuning the regularization strength on validation data). We then use the Blodgett et al. (2016a) model to infer p_{AAE} for each individual vocabulary term in isolation. While this does not completely explain the correlations observed in section §5.3.1, it does allow us to identify individual words that are both strongly associated with AAE, and highly predictive of particular categories.

Figure 5.2 shows the feature weights and p_{AAE} for each word in the models for FDCL18 (top) and DWMW17 (bottom), with the most highly weighted terms identified on the plots. The size of words indicates how common they are (proportional to the log of the number of times they appear in the corpus).

These results reveal important limitations of these datasets, and illustrate the potential for discriminatory impact of any simple models trained on this data. First, and most obviously, the most highly weighted unigrams for predicting "hateful" in FDCL18 are "n*gga" and "n*ggas", which are strongly associated with AAE (and their offensiveness depends on speaker and context; Spears, 1998). Because these terms are both frequent and highly weighted, any simple model trained on this data would indiscriminately label large numbers of tweets containing either of these terms as "hateful".

By contrast, the terms that are highly predictive of "hate speech" in DWMW17 (i.e., slurs) partly reflect the HateBase lexicon used in constructing this dataset, and the resulting emphasis is differ-

Warning, due to the nature of this research, these figures contain offensive or upsetting terms (e.g. racist, sexist, or homophobic slurs).



(FDCL18 – *abusive*)

(FDCL18 – *hateful*)

(DWMW17 – *offensive*)

(DWMW17 – *hate speech*)

Figure 5.2: Feature weights learned by l_2-regularized multiclass logistic regression models with unigram features, plotted against p_{AAE} for each term, based on Blodgett et al. (2016a). Top: weights for predicting *abusive* (left) and *hateful* (right) from a model trained on FDCL18. Bottom: weights for predicting *offensive* (left) and *hate speech* (right) from a model trained on DWMW17. Labels are shown for the most heavily-weighted terms, with label size proportional to the log count of the term in validation data. *Note*: "c*nt", "n*gger," "f*ggot," and their variations are considered sexist, racist, and homophobic slurs, respectively, and are predictive of hate speech DWMW17.

ent. (We also see artefacts of the dataset construction in the negative weights placed on "charlie", "bird", and "yankees" — terms which occur in HateBase, but have harmless primary meanings.)

To verify that no single term is responsible for the correlations reported in Table 5.1, we consider each word in the vocabulary in turn, and compute correlations excluding tweets containing that term. The results of this analysis (not shown) find that almost all of the correlations we observe are robust. For example, the correlation between p_{AAE} and "abusive" in FDCL18 increases the most if we drop tweets containing "fucking" (highly positively weighted, but non-AAE aligned), and decreases slightly if we drop terms like "ass" or "bitch". The one exception is the correlation between "hateful" and p_{AAE} in FDCL18: if we exclude tweets which contain "n*gga" or "n*ggas", the correlation drops to r=0.047. However, this also causes the correlation between p_{AAE} and "abusive" to increase to r=0.376.

5.3.2 Bias Propagation through Models

To further quantify the impact of racial biases in hate speech detection, we investigate how these biases are acquired by predictive models. First, we report differences in rates of false positives (FP) between AAE and White-aligned dialect groups for models trained on DWMW17 or FDCL18. Then, we apply these models to two reference Twitter corpora, described below, and compute average rates of reported toxicity, showing how these biases generalize to other data.[5]

DEMOGRAPHIC16 (Blodgett et al., 2016a) contains 56M tweets (2.8M users) with dialect estimated using a *demographic-aware topic model* that leverages census race/ethnicity data and geocoordinates of the user profile. As recommended, we assign dialect labels to tweets with dialect probabilities greater than 80%.

			% false identification		
DWMW17	Group	Acc.	None	Offensive	Hate
	AAE	94.3	1.1	**46.3**	0.8
	White	87.5	**7.9**	9.0	**3.8**
	Overall	91.4	2.9	17.9	2.3

			% false identification		
FDCL18	Group	Acc.	None	Abusive	Hateful
	AAE	81.4	4.2	**26.0**	**1.7**
	White	82.7	**30.5**	4.5	0.8
	Overall	81.4	20.9	6.6	0.8

Table 5.2: Classification accuracy and per-class rates of false positives (FP) on test data for models trained on DWMW17 and FDCL18, where the group with highest rate of FP is bolded.

USERLEVELRACE18 (Preoţiuc-Pietro and Ungar, 2018) is a corpus of 5.4M tweets, collected from 4,132 survey participants (3,184 White, 374 AA) who reported their race/ethnicity and Twitter user handle. For this dataset, we compare differences in toxicity predictions by *self-reported race*, instead of inferring message-level dialect.[6]

For each of the two toxic language corpora, we train a classifier to predict the toxicity label of a tweet. Using a basic neural attention architecture (Wang et al., 2016; Yang et al., 2016), we train a classifier initialized with GloVe vectors (Pennington et al., 2014a) to minimize the cross-entropy of the annotated class conditional on text, $\boldsymbol{x}$:

$$p(\text{class} \mid \boldsymbol{x}) \propto \exp(\boldsymbol{W}_o \boldsymbol{h} + \boldsymbol{b}_o), \tag{5.1}$$

with $\boldsymbol{h} = f(\boldsymbol{x})$, where f is a BiLSTM with attention, followed by a projection layer to encode the tweets into an H-dimensional vector.[7] We refer the reader to §C.1 for experimental details.

Results Table 5.2 shows that while both models achieve high accuracy, the false positive rates (FPR) differ across groups for several toxicity labels. The DWMW17 classifier predicts almost 50%

[5] We assume *a priori* that the average tweet is not inherently more toxic in a particular dialect. Assessing the veracity of this assumption requires a deep understanding of socio-cultural norms of profane and toxic speech.

[6] Note that lexical dialect inferences of AAE (p_{AAE}) significantly correlate with both the AAE group from DEMOGRAPHIC16 (Pearson $r = 0.61$, $p \ll 0.001$) and self-reported AA race from USERLEVELRACE18 (Pearson $r = 0.21$, $p \ll 0.001$).

[7] In preliminary experiments, our findings held regardless of our choice of classifier.

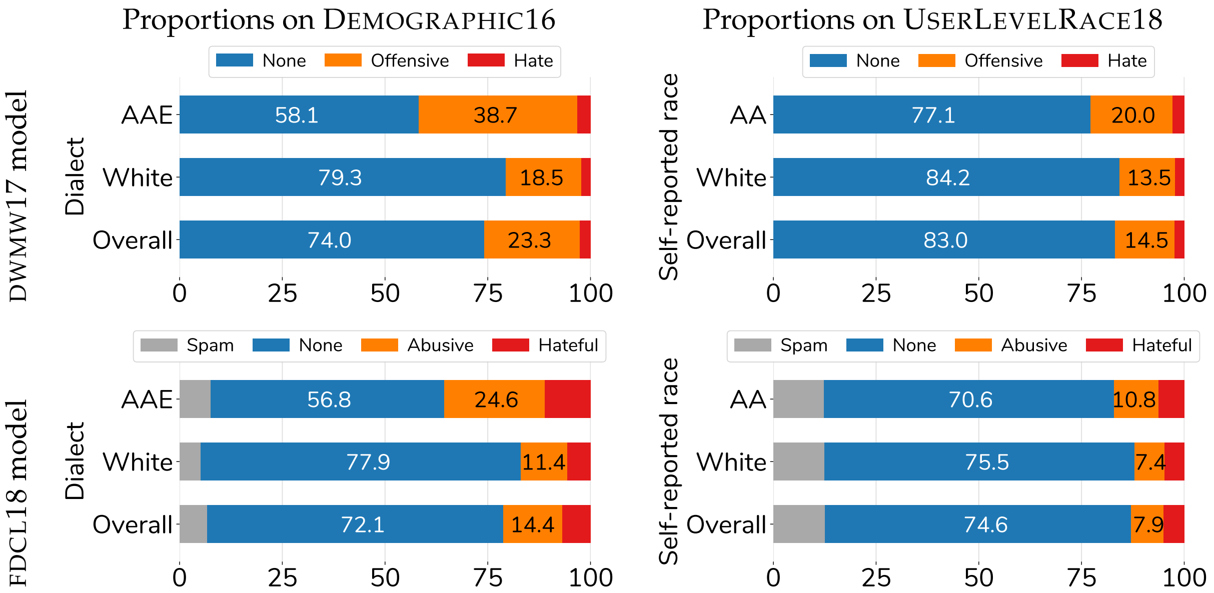


Figure 5.3: Average probability mass of toxicity classes in Demographic16 and UserLevelRace18, respectively, as given by classifiers trained on DWMW17 (top) and FDCL18 (bottom).

of non-offensive AAE tweets as being offensive, and FDCL18 classifier shows higher FPR for the "Abusive" and "Hateful" categories for AAE tweets. Additionally, both classifiers show strong tendencies to label White tweets as "none". These discrepancies in FPR across groups violate the *equality of opportunity* criterion, indicating discriminatory impact (Hardt et al., 2016).

We further quantify this potential discrimination in our two reference Twitter corpora. Figure 5.3 shows that the proportions of tweets classified as toxic also differ by group in these corpora. Specifically, in Demographic16, AAE tweets are more than twice as likely to be labelled as "offensive" or "abusive" (by classifiers trained on DWMW17 and FDCL18, respectively). We show similar effects on UserLevelRace18, where tweets by African American authors are 1.5 times more likely to be labelled "offensive". Our findings corroborate the existence of racial bias in the toxic language datasets and confirm that models propagate this bias when trained on them.

5.3.3 Toxicity and AAE in the PerspectiveAPI

We further showcase that other hate speech detection models acquire and propagate racial biases, by examining the biases in the PerspectiveAPI. This commercially deployed toxicity detection system was trained on a proprietary corpus of comments from Wikipedia, *New York Times*, and other news sites and boasts an strong performance of 0.97 AUC.[8] We obtain toxicity scores for all tweets in DWMW17 and FDCL18, as well as for 100K random tweets from Demographic16 and UserLevelRace18.[9]

In Table 5.3, we show correlations (Pearson r) between dialects/groups in our datasets and the Perspective toxicity scores. As with our own classification models, we find significant racial bias

[8] https://github.com/conversationai/perspectiveapi

[9] The API (http://perspectiveapi.com) was accessed in December 2018

in the PerspectiveAPI, with tweets in AAE or tweets by African Americans having a higher correlation with toxicity compared to white-aligned tweets, confirming observations by Chung (2019).

5.4 Effect of Dialect

To study the effect of dialect information on ratings of offensiveness, we run a small controlled experiment on Amazon Mechanical Turk where we prime annotators to consider the dialect and race of Twitter users. We ask workers to determine whether a tweet (a) is offensive *to them*, and (b) could be seen as offensive *to anyone*. In the *dialect priming* condition, we explicitly include the tweet's dialect as measured by Blodgett et al. (2016a), as well as extra instructions priming workers to think of tweet dialect as a proxy for the author's race. In the *race priming* condition, we encourage workers to consider the likely racial background of a tweet's author, based on its inferred dialect (e.g., an AAE tweet is likely authored by an African American Twitter user; see §C.3 for the task instructions). For all tasks, we ask annotators to optionally report gender, age, race, and political leaning.[10]

dataset	dial./group	corr.
DWMW17	White	−0.320
	AAE	0.310
FDCL18	White	−0.340
	AAE	0.453
DEMOGRAPHIC16	White	−0.096
	AAE	0.056
USERLEVELRACE18	White	−0.046
	AA	0.042

Table 5.3: Correlations (Pearson r) between dialects/groups in the datasets and the PerspectiveAPI TOXICITY scores. All correlations are significant ($p \ll 0.001$, Holm-corrected for multiple comparisons.)

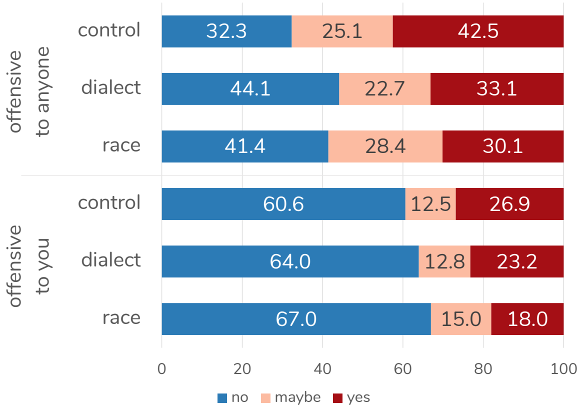


Figure 5.4: Proportion (in %) of offensiveness annotations of AAE tweets in control, dialect, and race priming conditions. Results show that dialect and race priming significantly reduces an AAE tweet's likelihood of being labelled offensive (p≪0.001).

With a distinct set of workers for each condition, we gather five annotations apiece for a sample of 1,351 tweets stratified by dialect, toxicity category, and dataset (DWMW17 and FDCL18).[11] Despite the inherent subjectivity of these questions, workers frequently agreed about a tweet being offensive to anyone (76% pairwise agreement, $\kappa = 0.48$) or to themselves (74% p.a., $\kappa = 0.30$).

Results Figure 5.4 shows that priming workers to think about dialect and race makes them significantly less likely to label an AAE tweet as (potentially) offensive to anyone. Additionally, race priming makes workers less likely to find AAE tweets offensive to them.

To confirm these effects, we compare the means of the control condition and treatment conditions,[12] and test significance with a t test. When rating offensiveness to anyone, the mean

[10] This study was approved by the Institutional Review Board (IRB) at the University of Washington.

[11] Annotations in the control setting agreed moderately with toxicity labels in DWMW17 and FDCL18 (Pearson r = 0.592 and r = 0.331, respectively; $p \ll 0.001$).

[12] We convert the offensiveness labels to real numbers (0: "no", 0.5: "maybe", 1: "yes").

for control condition ($M_c = 0.55$) differs from dialect ($M_d = 0.44$) and race ($M_r = 0.44$) conditions significantly ($p \ll 0.001$). For ratings of offensiveness to workers, only the difference in means for control ($M_c = 0.33$) and race (M_d =0.25) conditions is significant ($p \ll 0.001$).

Additionally, we find that overall, annotators are substantially more likely to rate a tweet as being offensive to *someone*, than to rate it as offensive to *themselves*, suggesting that people recognize the subjectivity of offensive language.

Our experiment provide insight into racial bias in annotations and shows the potential for reducing it, but several limitations apply, including the skewed demographics of our worker pool (75% self-reported White). Additionally, research suggests that motivations to not seem prejudiced could buffer stereotype use, which could in turn influence annotator responses (Plant and Devine, 1998; Moskowitz and Li, 2011).

5.5 Related Work

A robust body of work has emerged trying to address the problem of hate speech and abusive language on social media (Schmidt and Wiegand, 2017). Many datasets have been created, but most are either small-scale pilots (∼100 instances; Kwok and Wang, 2013; Burnap and Williams, 2015; Zhang et al., 2018b), or focus on other domains (e.g., Wikipedia edits; Wulczyn et al., 2017). In addition to DWMW17 and FDCL18, published Twitter corpora include Golbeck et al. (2017), which uses a somewhat restrictive definition of abuse, and Ribeiro et al. (2018), which is focused on network features, rather than text.

Past work on bias in hate speech datasets has exclusively focused on finding and removing bias against explicit identity mentions (e.g., woman, atheist, queer; Park and Fung, 2017; Dixon et al., 2018). In contrast, our work shows how insensitivity to dialect can lead to discrimination against minorities, even without explicit identity mentions.

5.6 Summary

In this chapter, we analyzed racial bias in widely-used corpora of annotated toxic language, establishing correlations between annotations of offensiveness and the African American English (AAE) dialect. We showed that models trained on these corpora propagate these biases, as AAE tweets are twice as likely to be labelled offensive compared to others. Finally, we introduced *dialect* and *race priming*, two ways to reduce annotator bias by highlighting the dialect of a tweet in the data annotation, and showed that it significantly decreases the likelihood of AAE tweets being labelled as offensive.

The findings in this chapter uncovered a previously unknown shortcoming of the existing paradigm used to detect toxic and hateful language, namely, racial bias in hate speech classification. This racial bias is alarming as it constitutes representational harms against African Americans (Barocas et al., 2017; Blodgett et al., 2020), and perpetuating the myth that AAE is a more toxic or less proper variety of English is a form of linguistic discrimination that upholds racial hierarchies (Rosa and Flores, 2017; Rosa, 2019).

Altogether, our results suggest that extra attention be paid to the varieties or dialects of English that text is written in, and that the offensiveness of the meaning of an utterance is much more complex and nuanced than a simple classification task can capture.

Chapter 6

SOCIAL BIAS FRAMES: Reasoning about Social and Power Implications of Language

This chapter discusses work originally published in Sap et al. (2020).

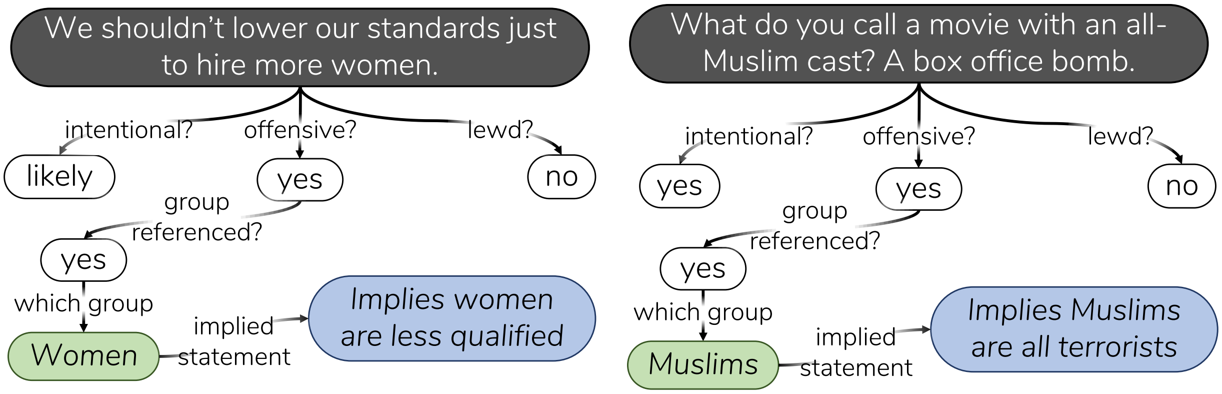

We shouldn't lower our standards just to hire more women.
intentional?
likely
offensive?
yes
lewd?
no
group referenced?
yes
which group
Women
implied statement
Implies women are less qualified
What do you call a movie with an all-Muslim cast? A box office bomb.
intentional?
yes
offensive?
yes
lewd?
no
group referenced?
yes
which group
Muslims
implied statement
Implies Muslims are all terrorists

Figure 6.1: SOCIAL BIAS FRAMES aim to represent the various pragmatic meanings related to social bias implications, by combining categorical and free-text annotations, e.g., that "women are less qualified" is implied by the statement "we shouldn't lower our standards to hire more women."

Biased and toxic meanings in language can arise in very subtle ways, and attempting to simply classify whether an utterance is toxic or not can lead to false flagging of non-offensive minority speech, as discussed in the previous chapter.

In this chapter, we tackle the much deeper problem of distilling the biased or toxic implications of text, moving beyond simple binary classification, and broadening our scope to both overt toxicity and subtle biases and microaggressions. For example, given a statement that "we shouldn't lower our standards to hire more women," we aim to capture the implication of this utterance — that "women (candidates) are less qualified."

We introduce SOCIAL BIAS FRAMES, a new formalism for modelling the implied or evoked social biases and stereotypes in language (illustrated in Figure 6.1). In addition, we introduce the Social Bias Inference Corpus to support large-scale modelling and evaluation with 150k structured annotations of social media posts, covering over 34k implications about a thousand demographic groups. We establish baseline performance of state-of-the-art neural models at predicting SOCIAL BIAS FRAMES from previously unseen text. Our results show that models can somewhat predict high-level categories of offensiveness, but they struggle to effectively generate more detailed explanations in terms of SOCIAL BIAS FRAMES.

6.1 Introduction

Language has enormous power to project social biases and reinforce stereotypes on people (Fiske, 1993). The way such biases are projected is rarely in what is stated explicitly, but in all the implied layers of meanings that frame and influence people's judgments about others. For example, on hearing a statement that an all-Muslim movie was a "box office bomb", most people can instantly recognize the implied demonizing stereotype that "Muslims are terrorists" (Figure 6.1). Understanding these biases with accurate underlying explanations is necessary for AI systems to adequately interact in the social world (Pereira et al., 2016), and failure to do so can result in the deployment of harmful technologies (e.g., conversational AI systems turning sexist and racist; Vincent, 2016).

Most previous approaches to understanding the implied harm in statements have cast this task as a simple *toxicity* classification (e.g., Waseem and Hovy, 2016b; Founta et al., 2018b; Davidson et al., 2017b). However, simple classifications run the risk of discriminating against minority groups, due to high variation and identity-based biases in annotations (e.g., which cause models to learn associations between dialect and toxicity; Sap et al., 2019a; Davidson et al., 2019). In addition, detailed explanations are much more informative for people to understand and reason about *why* a statement is potentially harmful against other people (Gregor and Benbasat, 1999; Ribeiro et al., 2016).

Thus, we propose SOCIAL BIAS FRAMES, a novel conceptual formalism that aims to model pragmatic frames in which people project social biases and stereotypes on others. Compared to semantic frames (Fillmore and Baker, 2001), the meanings projected by pragmatic frames are richer, and thus cannot be easily formalized using only categorical labels. Therefore, as illustrated in Figure 6.1, our formalism combines hierarchical categories of biased implications such as *intent* and *offensiveness* with implicatures described in free-form text such as *groups referenced* and *implied statements*. In addition, we introduce SBIC,[1] a new corpus collected using a novel crowdsourcing framework. SBIC supports large-scale learning and evaluation with over 150k structured annotations of social media posts, spanning over 34k implications about a thousand demographic groups.

We then establish baseline approaches that learn to recover SOCIAL BIAS FRAMES from unstructured text. We find that while state-of-the-art neural models are effective at making high-level categorization of whether a given statement projects unwanted social bias (80% F_1), they are not effective at spelling out more detailed explanations by accurately decoding SOCIAL BIAS FRAMES. Our study motivates future research that combines structured pragmatic inference with commonsense reasoning on social implications.

Important implications of this study. We recognize that studying SOCIAL BIAS FRAMES necessarily requires us to confront online content that may be offensive or disturbing (see §6.6 for further discussion on the ethical implications of this study). However, deliberate avoidance does not eliminate such problems. Therefore, the important premise we take in this study is that assessing social media content through the lens of SOCIAL BIAS FRAMES is important for automatic flagging or AI-augmented writing interfaces, where potentially harmful online content can be analyzed with detailed explanations for users or moderators to consider and verify. In addition, the collective analysis over large corpora can also be insightful for educating people on reducing

[1] SBIC: **S**ocial **B**ias **I**nference **C**orpus, available at http://tinyurl.com/social-bias-frames.

post	off?	int?	lewd?	group?	targeted group	implied statement	ingroup?
Kendrick just dropped the coldest video of all fucking time	no	no	no				
What a fucking stupid bitch shut the fuck up already	yes	yes	no				
need to fuck sum bad , I got some pressure built up :P	no	no	yes				
I take a bitch home give her long dick	yes	yes	yes	no			
You annoyed me every day for months you're a fucking moron	yes	yes	no	no			
My problem with Korean artists: I dont know how to pronounce your name I can't hanggul to save my life	yes	no	no	yes	Korean folks	have weird names	no
I'll continue this thread till non-lifting fags stop lifting like idiots at the gym... Which is never.	yes	yes	no	yes	gay men	are weak and not manly	no
I thought drugs were the only things black people could shoot up Boy was I wrong	yes	yes	no	yes	Black folks	do drugs kill people commit shootings	no

Table 6.1: Examples of inference tuples in SBIC. The types of inferences captured by Social Bias Frames cover (potentially subtle) offensive implications about various demographic groups.

unconscious biases in their language.

6.2 Social Bias Frames Definition

To better enable models to account for socially biased implications of language,[2] we design a new pragmatic formalism that distinguishes several related but distinct inferences, shown in Figure 6.1. Given a natural language utterance, henceforth, *post*, we collect both categorical as well as free text inferences (described below), inspired by recent efforts in free-text annotations of commonsense knowledge (e.g., Speer et al., 2017; Rashkin et al., 2018; Sap et al., 2019b) and argumentation (Habernal and Gurevych, 2016; Becker et al., 2017). The free-text explanations are crucial to our formalism, as they can both increase trust in predictions made by the machine (Kulesza et al., 2012; Bussone et al., 2015; Nguyen et al., 2018) and encourage a poster's empathy towards a targeted group, thereby combating biases (Cohen-Almagor, 2014).

We base our initial frame design on social science literature of pragmatics (Lakoff, 1973; de Marneffe et al., 2012) and impoliteness (Kasper, 1990; Gabriel, 1998; Dynel, 2015; Vonasch and Baumeister, 2017). We then refine the frame structure (including number of possible answers to questions) based on the annotator (dis)agreement in multiple pilot studies. We describe each of the included variables below.

Offensiveness is our main categorical annotation, and denotes the overall rudeness, disrespect, or toxicity of a post. We consider whether a post could be considered "offensive to anyone", as

[2] In this work, we employ the U.S. sociocultural lens when discussing bias and power dynamics among demographic groups.

previous work has shown this to have higher recall (Sap et al., 2019a). This is a categorical variable with three possible answers (*yes*, *maybe*, *no*).

Intent to offend captures whether the perceived motivation of the author was to offend, which is key to understanding how it is received (Kasper, 1990; Dynel, 2015), yet distinct from offensiveness (Gabriel, 1998; Daly, 2018). This is a categorical variable with four possible answers (*yes*, *probably*, *probably not*, *no*).

Lewd or sexual references are a key subcategory of what constitutes potentially offensive material in many cultures, especially in the United States (Strub, 2008). This is a categorical variable with three possible answers (*yes*, *maybe*, *no*).

Group implications are distinguished from individual-only attacks or insults that do not invoke power dynamics between groups (e.g., "F*ck you" vs. "F*ck you, f*ggot"). This is a categorical variable with two possible answers: individual-only (*no*), group targeted (*yes*).

Targeted group describes the social or demographic group that is referenced or targeted by the post. Here we collect *free-text answers*, but provide a seed list of demographic or social groups to encourage consistency.

Implied statement represents the power dynamic or stereotype that is referenced in the post. We collect *free-text answers* in the form of simple Hearst-like patterns (e.g., "*women are* ADJ", "*gay men VBP*"; Hearst, 1992).

In-group language aims to capture whether the author of a post may be a member of the same social/demographic group that is targeted, as speaker identity changes how a statement is perceived (O'Dea et al., 2015). Specifically, in-group language (words or phrases that (re)establish belonging to a social group; Eble, 1996) can change the perceived offensiveness of a statement, such as reclaimed slurs (Croom, 2011; Galinsky et al., 2013) or self-deprecating language (Greengross and Miller, 2008). Note that we do not attempt to categorize the identity of the speaker. This variable takes three possible values (*yes*, *maybe*, *no*).

6.3 Collecting Social Bias Annotations

To create SBIC, we design a crowdsourcing framework to distill the biased implications of posts at a large scale.

6.3.1 Data Selection

We draw from various sources of potentially biased online content, shown in Table 6.2, to select posts to annotate. Since online toxicity can be relatively scarce (Founta et al., 2018b),[3]

[3] Founta et al. (2018b) find that the prevalence of toxic content online is <4%.

we start by annotating English Reddit posts, specifically three intentionally offensive subReddits and a corpus of potential microaggressions from Breitfeller et al. (2019). By nature, the three offensive subreddits are very likely to have harmful implications, as posts are often made with intents to deride adversity or social inequality (Bicknell, 2007). Microaggressions, on the other hand, are likely to contain subtle biased implications—a natural fit for SOCIAL BIAS FRAMES.

type	source	# posts
Reddit	r/darkJokes	10,095
	r/meanJokes	3,483
	r/offensiveJokes	356
	Microaggressions	2,011
	subtotal	*15,945*
Twitter	Founta et al. (2018b)	11,864
	Davidson et al. (2017b)	3,008
	Waseem and Hovy (2016b)	1,816
	subtotal	*16,688*
Hate Sites	Gab	3,715
	Stormfront	4,016
	Banned Reddits	4,308
	subtotal	*12,039*
SBIC	**total # posts**	44,671

Table 6.2: Breakdown of origins of posts in SBIC.

In addition, we include posts from three existing English Twitter datasets annotated for toxic or abusive language, filtering out @-replies, retweets, and links. We mainly annotate tweets released by Founta et al. (2018b), who use a bootstrapping approach to sample potentially offensive tweets. We also include tweets from Waseem and Hovy (2016b) and Davidson et al. (2017b), who collect datasets of tweets containing racist or sexist hashtags and slurs, respectively.

Finally, we include posts from known English hate communities: Stormfront (de Gibert et al., 2018) and Gab,[4] which are both documented white-supremacist and neo-nazi communities (Bowman-Grieve, 2009; Hess, 2016), and two English subreddits that were banned for inciting violence against women (r/Incels and r/MensRights; Fingas, 2017; Center, 2012).

6.3.2 Annotation Task Design

We design a hierarchical annotation framework to collect biased implications of a given post (shown in Figure D.1 in the appendix) on Amazon Mechanical Turk (MTurk). For each post, workers indicate whether the post is offensive, whether the intent was to offend, and whether it contains lewd or sexual content. Only if annotators indicate potential offensiveness do they answer the group implication question. If the post targets or references a group or demographic, workers select or write which one(s); per selected group, they then write two to four stereotypes. Finally, workers are asked whether they think the speaker is part of one of the minority groups referenced by the post.

We collect three annotations per post, and restrict our worker pool to the U.S. and Canada. We ask workers to optionally provide coarse-grained demographic information.[5]

Annotator demographics In our final annotations, our worker pool was relatively gender-balanced and age-balanced (55% women, 42% men, <1% non-binary; 36±10 years old), but racially skewed (82% White, 4% Asian, 4% Hispanic, 4% Black).

[4] https://files.pushshift.io/gab/GABPOSTS_CORPUS.xz

[5] This study was approved by our institutional review board.

Annotator agreement Overall, the annotations in SBIC showed 82.4% pairwise agreement and Krippendorf's α=0.45 on average, which is substantially higher than previous work in toxic language detection (e.g., α=0.22 in Ross et al., 2017). Broken down by each categorical question, workers agreed on a post being offensive at a rate of 76% (Krippendorf's α=0.51), its intent being to offend at 75% (α=0.46), and it having group implications at 74% (α=0.48). For categorizing posts as lewd, workers agreed substantially (94%, α=0.62). However, flagging potential in-group speech had lower agreement, likely because this is a very nuanced annotation, and because highly skewed categories (only 5% "yes"; see Table 6.3) lead to low αs (here, α=0.17 with agreement 94%).[6] Finally, workers agreed on the exact same targeted group 80.2% of the time (α=0.50).

total # tuples		147,139
# unique	posts	44,671
	groups	1,414
	implications	32,028
	post-group	48,923
	post-group-implication	87,942
	group-implication	34,333
skews	offensive	44.8%
	intent	43.4%
	lewd	7.9%
	group targeted	50.9%
	in-group	4.6%

Table 6.3: Statistics of the SBIC dataset. Skews indicate the number of times a worker annotated a post as offensive, etc.

6.3.3 SBIC Description

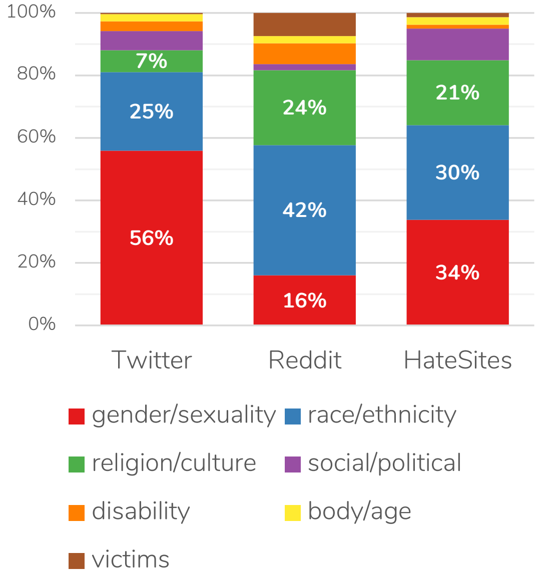


Figure 6.2: Breakdown of targeted group categories by domains. We show percentages within domains for the top three most represented identities, namely gender/sexuality (e.g., women, LGBTQ), race/ethnicity (e.g., Black, Latinx, and Asian), and culture/origin (e.g., Muslim, Jewish).

After data collection, SBIC contains 150k structured inference tuples, covering 34k free text group-implication pairs (see Table 6.3). We show example inference tuples in Table 6.1.

Additionally, we show a breakdown of the types of targeted groups in Figure 6.2. While SBIC covers a variety of types of biases, gender-based, race-based, and culture-based biases are the most represented, which parallels the types of discrimination happening in the real world (RWJF, 2017).

We find that our dataset is predominantly written in White-aligned English (78% of posts), as measured by a lexical dialect detector by Blodgett et al. (2016b), with <10% of posts having indicators of African-American English. We caution researchers to consider the potential for dialect- or identity-based biases in labelling (Davidson et al., 2019; Sap et al., 2019a) before deploying technology based on SBIC (see Section 6.6).

6.4 Social Bias Inference

Given a post, we establish baseline performance of models at inferring Social Bias Frames. An ideal model should be able to both *generate* the implied

[6] Given our data selection process, we expect the rate of in-group posts to be very low (see the following section).

power dynamics in textual form, as well as *classify* the post's offensiveness and other categorical variables. Satisfying these conditions, we use the OpenAI-GPT transformer networks (Vaswani et al., 2017b; Radford et al., 2018, 2019) as a basis for our experiments, given their recent successes at classification, commonsense generation, and conditional generation (Bosselut et al., 2019; Keskar et al., 2019).

Training We cast our frame prediction task as a hybrid classification and language generation task, where we linearize the variables following the frame hierarchy.[7] At training time, our model takes as input a sequence of N tokens:

$$\mathbf{x} = \{[\texttt{STR}], w_1, w_2, ..., w_n, [\texttt{SEP}], w_{[\texttt{lewd}]}, w_{[\texttt{off}]}, w_{[\texttt{int}]}, w_{[\texttt{grp}]}, [\texttt{SEP}], w_{[\texttt{G}]_1}, w_{[\texttt{G}]_2}, ..., [\texttt{SEP}], w_{[\texttt{S}]_1}, w_{[\texttt{S}]_2}, ..., [\texttt{SEP}], w_{[\texttt{ing}]}, [\texttt{END}]\} \quad (6.1)$$

where $[\texttt{STR}]$ is our start token, $w_{1:n}$ is the sequence of tokens in a post, $w_{[\texttt{G}]_i}$ the tokens representing the group, and $w_{[\texttt{S}]_i}$ the implied statement. We add two task-specific vocabulary items for each of our five classification tasks ($w_{[\texttt{lewd}]}$, $w_{[\texttt{off}]}$, $w_{[\texttt{int}]}$, $w_{[\texttt{grp}]}$, $w_{[\texttt{ing}]}$), each representing the negative and positive values of the class (e.g., for offensiveness, [offY] and [offN]).[8]

The model relies on a stack of transformer blocks of multi-headed attention and fully connected layers to encode the input tokens (for a detailed modelling description, see Radford et al., 2018, 2019). Since GPT is a forward-only language model, the attention is only computed over preceding tokens. At the last layer, the model projects the embedding into a vocabulary-sized vector, which is turned into a probability distribution over the vocabulary using a softmax layer.

We minimize the cross-entropy of the contextual probability of the correct token in our full linearized frame objective (of length N):

$$\mathcal{L} = -\frac{1}{N} \sum_i \log p_{\texttt{GPT}}(w_i \mid w_{0:i-1})$$

During training, no loss is incurred for lower-level variables with no values, i.e., variables that cannot take values due to earlier variable values (e.g., there is no targeted group for posts marked as non-offensive).

In our experiments we use pretrained versions of OpenAI's GPT and GPT2 (Radford et al., 2018, 2019) for our model variants, named SBF-GPT$_1$ and SBF-GPT$_2$, respectively. While their architectures are similar (stack of Transformers), GPT was trained on a large corpus of fiction books, whereas GPT2 was trained on 40Gbs of English web text.

Inference We frame our inference task as a conditional language generation task. Conditioned on the post, we generate tokens one-by-one either by greedily selecting the most probable one, or by sampling from the next word distribution, and appending the selected token to the output. We stop when the $[\texttt{END}]$ token is generated, at which point our entire frame is predicted. For greedy decoding, we only generate our frames once, but for sampling, we repeat the generation

[7] We linearize following the order in which variables were annotated (see Figure D.1). Future work could explore alternate orderings.

[8] We binarize our categorical annotations, assigning 1 to "yes," "probably," and "maybe,", and 0 to all other values.

	model	offensive			intent			lewd			group			in-group		
		42.2% pos. (dev.)			*44.8% pos (dev.)*			*3.0% pos (dev.)*			*66.6% pos (dev.)*			*5.1% pos (dev.)*		
		F_1	pr.	rec.	F_1	pr.	rec.	F_1	pr.	rec.	F_1	pr.	rec.	F_1	pr.	rec.
dev.	SBF-GPT$_1$-gdy	75.2	88.3	65.5	74.4	89.8	63.6	75.2	78.2	72.5	62.3	74.6	53.4	–	–	–
	SBF-GPT$_2$-gdy	77.2	88.3	68.6	**76.3**	89.5	66.5	77.6	81.2	74.3	**66.9**	67.9	65.8	**24.0**	85.7	14.0
	SBF-GPT$_2$-smp	**80.5**	84.3	76.9	75.3	89.9	64.7	**78.6**	80.6	76.6	66.0	67.6	64.5	–	–	–
test	SBF-GPT$_2$-gdy	78.8	89.8	70.2	78.6	90.8	69.2	80.7	84.5	77.3	69.9	70.5	69.4	–	–	–

Table 6.4: Experimental results (%) of various models on the classification tasks (gdy: argmax, smp: sampling). Some models did not predict the positive class for "in-group language," their performance is denoted by "–". We bold the F_1 scores of the best performing model(s) on the development set. For easier interpretation, we also report the percentage of instances in the positive class in the development set.

procedure to yield ten candidate frame predictions and choose the highest scoring one under our model.

In contrast to training time, where all inputs are consistent with our frames' structure, at test time, our model can sometimes predict combinations of variables that are inconsistent with the constraints of the frame (e.g., predicting a post to be inoffensive, but still predict it to be offensive to a group). To mitigate this issue, we also experiment with a constrained decoding algorithm (denoted "constr") that considers various global assignments of variables. Specifically, after greedy decoding, we recompute the probabilities of each of the categorical variables, and search for the most probable assignment given the generated text candidate and variable probabilities.[9] This can allow variables to be assigned an alternative value that is more globally optimal.[10]

6.4.1 Evaluation

We evaluate performance of our models in the following ways. For classification, we report precision, recall, and F_1 scores of the positive class. Following previous generative inference work (Sap et al., 2019b), we use automated metrics to evaluate model generations. We use BLEU-2 and RougeL (F_1) scores to capture word overlap between the generated inference and the references, which captures quality of generation (Galley et al., 2015; Hashimoto et al., 2019). We additionally compute word mover's distance (WMD; Kusner et al., 2015), which uses distributed word representations to measure similarity between the generated and target text.[11]

6.4.2 Training Details

As each post can contain multiple annotations, we define a training instance as containing one post-group-statement triple (along with the five categorical annotations). We then split our dataset into train/dev./test (75:12.5:12.5), ensuring that no post is present in multiple splits. For evaluation (dev., test), we combine the categorical variables by averaging their binarized values and re-binarizing using a .5 threshold, and compare the generated inferences (hypotheses) to all targeted groups and implied statements (references).

All experiments are carried out using HuggingFace's Transformers library.[12] We tune hyper-

[9] We only use the possible assignments in the same forward pass; we do not use assignments from different samples.

[10] In practice, as seen in Tables 6.4, 6.5, and D.1, this only slightly improves predictions.

[11] We use GloVe trained on CommonCrawl, as part of the SpaCy `en_core_web_md` package.

[12] `https://github.com/huggingface/transformers`

		group targeted			implied statement		
		BLEU	Rouge-L	WMD	BLEU	Rouge-L	WMD
dev.	SBF-GPT$_1$-gdy	69.9	60.3	1.01	**49.9**	40.2	2.97
	SBF-GPT$_1$-gdy-constr	69.2	64.7	1.05	49.0	42.8	3.02
	SBF-GPT$_2$-gdy	74.2	64.6	0.90	49.8	41.4	**2.96**
	SBF-GPT$_2$-gdy-constr	73.4	**68.2**	0.89	49.6	**43.5**	**2.96**
	SBF-GPT$_2$-smp	**83.2**	33.7	**0.62**	44.3	17.8	3.31
	SBF-GPT$_2$-smp-constr	83.0	33.7	0.63	44.1	17.9	3.31
test	SBF-GPT$_2$-gdy	77.0	71.3	0.76	52.2	46.5	2.81
	SBF-GPT$_2$-gdy-constr	77.9	68.7	0.74	52.6	44.9	2.79

Table 6.5: Automatic evaluation of various models on the generation tasks. We bold the scores of the best performing model(s) on the development set. Higher is better for BLEU and ROUGE scores, and lower is better for WMD.

parameters on the dev. set, and report performance for the best performing setting (according to average F_1). We train or finetune our models using a batch size of 4, a learning rate of 5×10^{-6} for GPT and 10^{-5} for GPT2 (both with linear warm up), and consider training for $e \in \{1, 2, 5\}$ epochs.

6.5 Results

Listed in Tables 6.4 and 6.5, our modelling results indicate that making inferences about social biases in language is challenging for these models.

Classification Shown in Table 6.4, models perform well on higher-level variables such as offensiveness and lewdness, despite the latter being heavily skewed. We hypothesize that correctly predicting lewdness might require more lexical matching (e.g., detecting words with sexual connotations).

Whether a group is targeted is slightly less easy for models to predict, and whether the language is in-group is even more challenging, with most of the models defaulting to never predicting it. This highly skewed category poses a challenge for all models, likely due to subtlety of the task and the lack of positive instances.

SBF-GPT$_2$-gdy is the only model that predicts positive values for in-group language, for which it benefits from constrained decoding with a 1.9% improvement in F_1 score (we show results with all constrained decoding variants in Table D.1 in the appendix).

Generation When evaluating our models on the generation tasks (i.e., targeted group and implied statement), we find that no one model outperforms others across all metrics (Table 6.5).

Overall, models do well at generating the targeted groups, likely because of the more limited generation space (there are only 1.4k possible groups in SBIC). Conversely, for implied statement generation (where output space is much larger), model performance is slightly worse.

Similar to the classification tasks, SBF-GPT$_2$-gdy shows a slight increase in RougeL score when using constrained decoding, but we see a slight drop in BLEU scores.

Error analysis Since small differences in automated evaluation metrics for text generation sometimes only weakly correlate with human judgments (Liu et al., 2016c), we manually perform an

	post	predicted group	predicted implication	reference groups	reference implications
(a)	All-female casts bother me but I have literally never noticed the hundreds of movies with 0 female characters I have already watched	—	—	women	women aren't good actors; women can't be in leading roles
(b)	As expected, when the terrorist group Hamas won the election in Gaza it was the usual one man, one vote, one time, Islamist election. #Islam	muslim folks	muslims are terrorists	muslim folks; arabic folks	bad people; islam promotes controlling governments; muslims are authoritarians; not fair
(c)	"Black guy in class: *attempts to throw a paper ball into the trash* *misses* Mrs. Mundy: "You're a disgrace to your race, Marcus"	black folks	black people are trash	black folks	are good at basketball; black men are defined by athletic skill.; black men should be good at basketball.; do not take things seriously
(d)	How can a Libtard speak truth to power when the only thing they vomit from their mouths is lies?	liberals	liberals are stupid	liberals	liberals are incompetent; liberals are stupid; liberals only speak lies
(e)	I hate fat bitches	overweight / fat folks	fat people are ugly	overweight / fat folks	are not pleasant people; fat folks are all the same; fat folks are less than others; not likable

Table 6.6: Examples of SBF-GPT$_2$-gdy-constr model predictions. The model struggles to pick up on subtle biases (a), and tends to generate generic stereotypes rather than implications that are entailed by the post (b, c).

error analysis on a manually selected set of generated development-set examples from the SBF-GPT$_2$-gdy-constr model (Table 6.6). Overall, the model seems to struggle with generating textual implications that are relevant to the post, instead generating very generic stereotypes about the demographic groups (e.g., in examples b and c). The model generates the correct stereotypes when there is high lexical overlap with the post (e.g., examples d and e). This is in line with previous research showing that large language models rely on correlational patterns in data (Sap et al., 2019c; Sakaguchi et al., 2020).

6.6 Ethical Considerations

Risks in deployment Automatic detection of offensiveness or reasoning about harmful implications of language should be done with care. When deploying such algorithms, ethical aspects should be considered including which performance metric should be optimized (Corbett-Davies et al., 2017), as well as the fairness of the model on speech by different demographic groups or in different varieties of English (Mitchell et al., 2019). Additionally, deployment of such technology should discuss potential nefarious side effects, such as censorship (Ullmann and Tomalin, 2019) and dialect-based racial bias (Sap et al., 2019a; Davidson et al., 2019). Finally, offensiveness could be paired with promotions of positive online interactions, such as emphasis of community standards (Does et al., 2011) or counter-speech (Chung et al., 2019; Qian et al., 2019).

Risks in annotation Recent work has highlighted various negative side effects caused by annotating potentially abusive or harmful content (e.g., acute stress; Roberts, 2016). We mitigated these by limiting the number of posts that one worker could annotate in one day, paying workers above minimum wage ($7–12), and providing crisis management resources to our annotators.[13] Additionally, we acknowledge the implications of using data available on public forums for research (Zimmer, 2018) and urge researchers and practitioners to respect the privacy of the authors of posts in SBIC (Ayers et al., 2018).

6.7 Related Work

Bias and toxicity detection Detection of hateful, abusive, or other toxic language has received increased attention recently (Schmidt and Wiegand, 2017), and most dataset creation work has cast this detection problem as binary classification (Waseem and Hovy, 2016b; Davidson et al., 2017b; Founta et al., 2018b). Moving beyond a single binary label, Wulczyn et al. (2017) and the PerspectiveAPI use a set of binary variables to annotate Wikipedia comments for several toxicity-related categories (e.g., identity attack, profanity). Similarly, Zampieri et al. (2019) hierarchically annotate a dataset of tweets with offensiveness and whether a group or individual is targeted. Most related to our work, Ousidhoum et al. (2019) create a multilingual dataset of 13k tweets annotated for five different emotion- and toxicity-related aspects, including a 16-class variable representing social groups targeted. In comparison, Social Bias Frames not only captures binary toxicity and hierarchical information about whether a group is targeted, but also *free-text* implications about 1.4k different targeted groups and the implied harm behind statements.

Similar in spirit to this work, recent work has tackled more subtle bias in language, such as microaggressions (Breitfeller et al., 2019) and condescension (Wang and Potts, 2019). These types of biases are in line with the biases covered by Social Bias Frames, but more narrowly scoped.

Inference about social dynamics Various work has tackled the task of making inferences about power and social dynamics. Particularly, previous work has analyzed power dynamics about specific entities, either in conversation settings (Prabhakaran et al., 2014; Danescu-Niculescu-Mizil et al., 2012) or in narrative text (Sap et al., 2017; Field et al., 2019b; Antoniak et al., 2019). Additionally, recent work in commonsense inference has focused on mental states of participants of a situation (e.g., Rashkin et al., 2018; Sap et al., 2019b). In contrast to reasoning about particular individuals, our work focuses on biased implications of social and demographic groups as a whole.

6.8 Summary

In this chapter, we introduced Social Bias Frames, a new structured commonsense formalism that distills knowledge about the biased implications of language, to help machines reason about and account for social biases in language. Our frames combine categorical knowledge about the

[13] We direct workers to the Crisis Text Line (https://www.crisistextline.org/).

offensiveness, intent, and targets of statements, as well as free-text inferences about which groups are targeted and biased implications or stereotypes. We collected a new dataset of 150k annotations on social media posts using a new crowdsourcing framework and established baseline performance of models built on top of large pretrained language models. We showed that while classifying the offensiveness of statements is easier, current models struggle to generate relevant social bias inferences, especially when implications have low lexical overlap with posts. This indicates that more sophisticated models are required for the types of people-centric reasoning in SOCIAL BIAS FRAMES.

This study showcases the promise of tackling social biases and toxicity in language as a problem of understanding the implied meaning using structured explanations, rather than using binary classification. Such an approach that generates explanations can be useful for online content moderation pipelines (Roberts, 2019), especially as it can help educate the authors of biased or toxic posts on why their posts were flagged (Myers West, 2018; Jhaver et al., 2019). It can also be more useful for characterizing the types of groups that are targeted in text corpora, by generating fine-grained biased implications for utterances.

Chapter 7

Conclusion

In this dissertation, we investigated methods for making NLP systems that can reason about and rewrite with social dynamics in language, as well as how to make holistic and equitable systems to understand harmful implications and social biases in language. However, our investigations shed light onto several shortcomings and thus exciting future directions to explore in order for NLP systems to truly reason about social commonsense and social biases in language.

7.1 Contributions

In Part I, we explore two aspects of social commonsense reasoning in language: generating the social implications about situations and rewriting situations with different social implications.

With ATOMIC (Chapter 2), we significantly bridged the gap towards endowing machines with *inferential* knowledge, in the form of a large-scale knowledge graph of 877k tuples represented in natural language. With this new resource, we showed that machines can learn to generate the implications of previously unseen situations, and can do so better using knowledge learned during pretraining. However, machines struggled to generate inferences for explicitly mentioned or implied third party participants of situations, compared to the main participant of the event. These findings motivate the need for exploring better person-centric modelling, as well as expanding the knowledge that machines have access to to allow for a broad range of applications.

Then, we tackled a different challenge: the controllable debiasing of sentence through the lens of connotations of text (Chapter 3). Due to the lack of parallel data for this task, we created an unsupervised approach: POWERTRANSFORMER, a transformer-based encoder-decoder trained on a joint reconstruction and paraphrasing objective. We showed that our model outperformed ablated versions as well as baselines from previous work both on automatic and human evaluations. Additionally, as a case study, we showed the feasibility for controllable debiasing at debiasing the portrayal of characters in movie scripts. Our findings highlight the potential of neural models as a tool for controllable editing with commonsense, specifically for mitigating social biases in text.

In Part II, we focused on a different set of social implications–namely, stereotypes, social biases, and toxicity in language– and explored NLP methods to holistically and equitably detect these harmful implications in language.

In Chapter 5, we analyzed widely-used corpora of annotated toxic language, and uncovered strong racial bias in the form of correlations between annotations of offensiveness and African American English. We showed that models trained on such corpora propagate these biases, with AAE tweets being twice as likely to be labelled offensive compared to others. Additionally, we introduced dialect and race priming, two ways to reduce annotator bias by highlighting the dialect of a tweet in the data annotation, and showed that it significantly decreases the likelihood of AAE tweets being labelled as offensive. Our results suggest that, when labelling toxicity in language, extra attention should be paid to the social factors at play (e.g., dialect, racial inequality).

Finally, to enable more holistic, explainable, and nuanced understanding of toxicity and social biases in language, we introduced SOCIAL BIAS FRAMES, a new structured commonsense formalism that distills knowledge about the biased implications of language (Chapter 6). We collected a dataset of 150k annotations on social media posts using a new crowdsourcing framework and establish baseline performance of models built on top of large pretrained language models. We showed that while classifying the offensiveness of statements is easier, models struggled to generate relevant social bias inferences, especially when implications have low lexical overlap with posts. Our findings suggest that more sophisticated models that can perform people-centric reasoning are likely required for SOCIAL BIAS FRAMES inferences.

7.2 Future Directions

While this dissertation took several steps towards this goal, there are several directions still to explore towards making NLP systems more human-centric, socially aware, and equity driven.

Human-centric NLP models of commonsense In this dissertation, we tackled complex reasoning tasks about people explicitly mentioned or implicitly referred to in text, with respect to social commonsense and social biases. However, one recurring shortcoming of our models is their inability to properly reason about different people. For example, in Chapter 2, both encoder-decoder and pretrained models struggled to produce relevant inferences with respect to mentioned or implied other participants of situations (e.g., PersonY), a shortcoming corroborated by other work (Sap et al., 2019c; Sakaguchi et al., 2020). Thus, future work should explore methods that can better distinguish between participants while performing commonsense inferences, either through explicit inductive biases (Henaff et al., 2017; Févry et al., 2020), distillation or knowledge probing (Tenney et al., 2019), or interactional or dialogue settings (Shen et al., 2019). Additionally, future work could draw from cognitive and neuroscience for inspiration on tackling this challenge (Quiroga et al., 2005; Tamir et al., 2016; Thornton et al., 2019).

Expanding knowledge of social dynamics With AI systems becoming ubiquitous in everyday situations, their knowledge of the social world needs to keep expanding and updating. While ATOMIC provided a starting point for understanding social dynamics related to events, there are other aspects related to our social world that models should be aware of. For example, AI systems should have an understanding of what is socially acceptable to do, as explored in preliminary work on distilling social norms and morality related to situations (Forbes et al., 2020). Additionally, AI systems should understand that certain statements or actions might be more appropriate when done by certain people compared to others (e.g., certain greetings are more offensive when said by a white person than a Black person; Figure 4.1; Croom, 2011; Galinsky et al., 2013).

A key challenge is to find ways of collecting the right kind of knowledge, and to be able to identify the right knowledge to use in which situation. Additionally, systems should ideally be able to personalize to a user's cultural background (e.g., social group, country, etc), as well as to their identity and experience (e.g., gender, profession, familiarity with technology) and in general to their values (Lee et al., 2020). Methodologically, a promising direction for this could rely on few-shot approaches which have shown great promise recently with pretrained language models (Brown et al., 2020; Gao et al., 2021).

Positive societal applications of NLP Finally, this results discussed in this dissertation showcased the possibility for positive societal impact with NLP models, which future work could explore further in several directions. For example, social commonsense models open the door to develop assistive technologies for people with cognitive disabilities (Lewis, 2020), who often struggle to reason about others' mental states (Korkmaz, 2011). Additionally, the promising results with POWERTRANSFORMER opens the door for other text debiasing technology, e.g., systems that rephrase text to avoid biased implications using SOCIAL BIAS FRAMES (building on recent controllable text generation techniques such as Liu et al., 2021).

However, developing these (or any) NLP systems equitably will require care. Future work should consider value-sensitive (Friedman et al., 2008) or participatory design (Sanders, 2002; DiSalvo et al., 2012; Denton et al., 2020) approaches. Additionally, NLP systems should continuously be evaluated on equity and fairness standards to make sure that they are achieving the positive societal impact that they aim to have.

Bibliography

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. *CoRR*, abs/1809.07572.

Wafa Alorainy, Pete Burnap, Han Liu, and Matthew Williams. 2018. Cyber hate classification: 'othering' language and paragraph embedding. *CoRR*, abs/1801.07495.

Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O Riedl. 2021. Automated storytelling via causal, commonsense plot ordering. In *AAAI*.

Monica Anderson, Skye Toor, Lee Rainie, and Aaron Smith. 2018. Activism in the social media ages. http://www.pewinternet.org/2018/07/11/activism-in-the-social-media-age/. Accessed: 2019-03-01.

Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. In *CSCW*.

Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

John W Ayers, Theodore L Caputi, Camille Nebeker, and Mark Dredze. 2018. Don't quote me: reverse identification of research participants in social media studies. *NPJ digital medicine*, 1(1):1–2.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *ACL*.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *SIGCIS*.

Maria Becker, Michael Staniek, Vivi Nastase, and Anette Frank. 2017. Enriching argumentative texts with implicit knowledge. In *NLDB*.

Elizabeth Behm-Morawitz and Dana E Mastro. 2008. Mean girls? the influence of gender portrayals in teen movies on emerging adults' gender-based attitudes and beliefs. *Journalism & Mass Communication Quarterly*, 85(1):131–146.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons.

Jeanette Bicknell. 2007. What is offensive about offensive jokes? *Philosophy Today*, 51(4):458–465.

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

Su Lin Blodgett, Solon Barocas, Hal Daumé, III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proc. of ACL*.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016a. Demographic dialectal variation in social media: A case study of African-American english. In *EMNLP*.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016b. Demographic dialectal variation in social media: a case study of African-American English. In *EMNLP*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *ACL*.

Lorraine Bowman-Grieve. 2009. Exploring "Stormfront": a virtual community of the radical right. *Studies in conflict & terrorism*, 32(11):989–1007.

Luke M 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 *EMNLP*.

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

Pete Burnap and Matthew L. Williams. 2015. Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7:223–242.

Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169. IEEE.

Ziqiang Cao, Chuwei Luo, Wenjie Li, and Sujian Li. 2017. Joint copying and restricted generation for paraphrase. In *AAAI*.

Southern Poverty Law Center. 2012. Misogyny: the sites. *Intelligence Report*, 145.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020a. R^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *ACL*.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020b. Generating similes ~~effortlessly~~ *like a pro*: A style transfer approach for simile generation. In *EMNLP*.

Nathanael Chambers and Daniel Jurafsky. 2008. Unsupervised learning of narrative event chains. In *ACL*.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367.

Mia Xu Chen, Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M Dai, Zhifeng Chen, et al. 2019. Gmail smart compose: Real-time assisted writing. In *KDD*.

Sapna Cheryan and Hazel Rose Markus. 2020. Masculine defaults: Identifying and mitigating hidden cultural biases. *Psychological Review*.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *SSST@EMNLP*.

Cuong Xuan Chu, Niket Tandon, and Gerhard Weikum. 2017. Distilling task knowledge from how-to communities. In *WWW*.

Anna Chung. 2019. How automated tools discriminate against black language. https://onezero.medium.com/how-automated-tools-discriminate-against-black-language-2ac8eab8d6db. Accessed: 2019-03-02.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *ACL*.

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *IUI*.

Jamie Cleland. 2014. Racism, football fans, and online message boards: How social media has added a new dimension to racist discourse in English football. *J. Sport Soc. Issues*, 38(5):415–431.

Raphael Cohen-Almagor. 2014. Countering hate on the internet. *Annual review of law and ethics*, 22:431–443.

M A Conway and C W Pleydell-Pearce. 2000. The construction of autobiographical memories in the self-memory system. *Psychological Review*, 107(2):261–288.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *KDD*.

Kate Crawford, Meredith Whittaker, Madeleine Clare Elish, Solon Barocas, Aaron Plasek, and Kadija Ferryman. 2016. The ai now report. *The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. In *LREC*. Corpus available at http://urn.fi/urn:nbn:fi:lb-201804191.

Adam M Croom. 2011. Slurs. *Language Sciences*, 33(3):343–358.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *ACL*.

Helen L Daly. 2018. On insults. *Journal of the American Philosophical Association*, 4(4):510–524.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: language effects and power differences in social interaction. In *WWW*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *ICLR*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Abusive Language Workshop*.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017a. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Thomas Davidson, Dana Warmsley, Michael W Macy, and Ingmar Weber. 2017b. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58:92–103.

Daniel Clement Dennett. 1989. *The intentional stance*. MIT press.

Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the people back in: Contesting benchmark machine learning datasets. In *ICML Workshop on Participatory Approaches to Machine Learning*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *EMNLP*.

Carl DiSalvo, Andrew Clement, and Volkmar Pipek. 2012. Communities: Participatory design for, with and by communities.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of Conference on AI, Ethics, and Society*.

Serena Does, Belle Derks, and Naomi Ellemers. 2011. Thou shalt not discriminate: how emphasizing moral ideals rather than obligations increases whites' support for social equality. *Journal of Experimental Social Psychology*, 47(3):562–571.

Marta Dynel. 2015. The landscape of impoliteness research. *Journal of Politeness Research*, 11(2):383.

Connie C Eble. 1996. *Slang & sociability: in-group language among college students*. Univ of North Carolina Press.

Walter F. Edwards. 2004. African American Vernacular English: phonology. In *A Handbook of Varieties of English: Morphology and Syntax*.

José H. Espinosa and Henry Lieberman. 2005. Eventnet: Inferring temporal relations between commonsense events. In *MICAI*.

Ethan Fast, Tina Vachovsky, and Michael S Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *ICWSM*.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *EMNLP*.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *EMNLP Workshop on Stylistic Variation*.

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

Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. 2019b. Contextual affective analysis: a case study of people portrayals in online #MeToo stories. In *ICWSM*.

Anjalie Field and Yulia Tsvetkov. 2019. Entity-centric contextual affective analysis. In *ACL*.

Charles J Fillmore and Collin F Baker. 2001. Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*.

Jon Fingas. 2017. Reddit bans misogynist community as part of anti-violence crackdown. https://www.engadget.com/2017/11/08/reddit-bans-misogynist-community-in-anti-violence-crackdown/. Accessed: 2019-12-06.

Susan T Fiske. 1993. Controlling other people. the impact of power on stereotyping. *American psychologist*, 48(6):621–628.

Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. 2018. Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707.

Sarah Florini. 2014. Tweets, tweeps, and signifyin': Communication and cultural performance on "Black Twitter". *Television & New Media*, 15(3):223–237.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *EMNLP*.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018a. Large scale crowdsourcing and characterization of twitter abusive behavior. In *ICWSM*.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018b. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *ICWSM*.

Deen Freelon, Charlton D. McIlwain, and Meredith D. Clark. 2016. Beyond the hashtags. http://cmsimpact.org/wp-content/uploads/2016/03/beyond_the_hashtags_2016.pdf. Accessed: 2019-03-01.

Batya Friedman, Peter H Kahn, and Alan Borning. 2008. Value sensitive design and information systems. *The handbook of information and computer ethics*, pages 69–101.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *AAAI*.

Yiannis Gabriel. 1998. An introduction to the social psychology of insults in organizations. *Human Relations*, 51(11):1329–1354.

Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian M. Suchanek. 2013. Amie: association rule mining under incomplete evidence in ontological knowledge bases. In *WWW*.

Adam D Galinsky, Cynthia S Wang, Jennifer A Whitson, Eric M Anicich, Kurt Hugenberg, and Galen V Bodenhausen. 2013. The reappropriation of stigmatizing labels: the reciprocal relationship between power and self-labeling. *Psychol. Sci.*, 24(10):2020–2029.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and William B. Dolan. 2015. deltaBLEU: a discriminative metric for generation tasks with intrinsically diverse targets. In *ACL*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *ACL*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform. ArXiv:1803.07640.

Sam Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP*.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *EMNLP*.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *ACL Demonstrations*.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *ACL*.

Ona de Gibert, Naiara Pérez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Abusive Language Workshop at EMNLP*.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A large labeled corpus for online harassment research. In *WebSci*, pages 229–233. ACM.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *SEM2013*.

Google. 2017. Using technology to address gender bias in film. https://www.google.com/about/main/gender-equality-films/index.html.

Andrew S Gordon and Jerry R Hobbs. 2017. *A Formal Theory of Commonsense Psychology: How People Think People Think*. Cambridge University Press.

Andrew S Gordon and Reid Swanson. 2008. StoryUpgrade: finding stories in internet weblogs. In *ICWSM*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 25–30, New York, NY, USA. ACM.

Philip Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *NAACL*.

Arthur C Graesser, Scott P Robertson, and Patricia A Anderson. 1981. Incorporating inferences in narrative representations: A study of how and why. *Cognitive Psychology*, 13(1):1–26.

Lisa Green. 2002. *African American English: A Linguistic Introduction*, 8.3.2002 edition edition. Cambridge University Press.

Gil Greengross and Geoffrey F Miller. 2008. Dissing oneself versus dissing rivals: effects of status, personality, and sex on the Short-Term and Long-Term attractiveness of Self-Deprecating and Other-Deprecating humor. *Evolutionary Psychology*, 6(3).

Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, pages 497–530.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

David Gunning. 2018. Machine common sense concept paper. ArXiv Preprint.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL-HLT (2)*.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *EMNLP*, pages 1214–1223.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *NeurIPS*.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *NAACL-HLT*.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *ACL*, pages 539–545.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2017. Tracking the world state with recurrent entity networks. In *ICLR*.

Amanda Hess. 2016. The far right has a new digital safe space. `https://www.nytimes.com/2016/11/30/arts/the-far-right-has-a-new-digital-safe-space.html`. Accessed: 2019-12-06.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.

Eric Horvitz. 2017. Ai, people, and society.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *NAACL*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *ICML*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.

Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021. "i'm not mad": Commonsense implications of negation and contradiction. In *NAACL*.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *ACL*.

Raghav Kapoor, Yaman Kumar, Kshitij Rajput, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2018. Mind your language: Abuse and offense detection for code-switched languages. *CoRR*, abs/1809.08652.

Gabriele Kasper. 1990. Linguistic politeness: current research issues. *Journal of Pragmatics*, 14(2):193–218.

William R Kearns, Neha Kaura, Myra Divina, Cuong Vo, Dong Si, Teresa Ward, and Weichao Yuwen. 2020. A wizard-of-oz interface and persona-based methodology for collecting health counseling dialog. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: a conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Walter Kintsch. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2):163.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *ACL*.

Filip Klubička and Raquel Fernandez. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. In *LREC*.

Rik Koncel-Kedziorski, Ioannis Konstas, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2016. A theme-rewriting approach for generating algebra word problems. In *EMNLP*.

Baris Korkmaz. 2011. Theory of mind and neurodevelopmental disorders of childhood. *Pediatr Res*, 69(5 Pt 2):101R–8R.

Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1–10. ACM.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *ICML*, pages 957–966.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *AAAI*.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *The Behavioral and brain sciences*, 40:e253.

Robin Lakoff. 1973. Language and woman's place. *Language in society*, 2(1):45–79.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-Attribute text rewriting. In *ICLR*.

Min Kyung Lee, Nina Grgić-Hlača, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. 2020. Human-Centered approaches to fair and responsible AI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pages 1–8, New York, NY, USA. Association for Computing Machinery.

Younghun Lee, Seunghyun Yoon, and Kyomin Jung. 2018. Comparative studies of detecting abusive language on twitter. *CoRR*, abs/1808.10245.

Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Clayton Lewis. 2020. Implications of developments in machine learning for people with cognitive disabilities. *ACM SIGACCESS Accessibility and Computing*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018a. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *NAACL*.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018b. Paraphrase generation with deep reinforcement learning. In *EMNLP*.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. Dexperts: Decoding-time controlled text generationwith experts and anti-experts. In *ACL*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016a. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016b. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016c. How NOT to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. In *ACL*.

Judith Lorber, Susan A Farrell, et al. 1991. The social construction of gender. *Newbury Park*, 5.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. Powertransformer: Unsupervised controllable revision for biased language correction. In *EMNLP*.

Gary Marcus. 2018. Deep learning: A critical appraisal. *CoRR*, abs/1801.00631.

Marie-Catherine de Marneffe, Christopher D Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Comput. Linguist.*, 38(2):301–333.

Craig McGarty. 2018. Social categorization. In *Oxford Research Encyclopedia of Psychology*.

Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of gpt-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *NAACL*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *FAccT*.

Chris Moore. 2013. *The development of commonsense psychology*. Psychology Press.

Gordon B. Moskowitz and Peizhong Li. 2011. Egalitarian goals trigger stereotype inhibition: A proactive form of stereotype control. *J. Exp. Soc. Psychol.*, 47(1):103–116.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016a. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016b. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL*. Corpus available at `https://www.cs.rochester.edu/nlp/rocstories/`.

Paul Mozur. 2018. A genocide incited on Facebook, with posts from Myanmar's military. `https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html`. Accessed: 2018-12-6.

Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383.

An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. 2018. Believe it or not: designing a human-AI partnership for mixed-initiative fact-checking. In *The 31st Annual ACM Symposium on User Interface Software and Technology*, pages 189–199. ACM.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *TACL*.

Conor J O'Dea, Stuart S Miller, Emma B Andres, Madelyn H Ray, Derrick F Till, and Donald A Saucier. 2015. Out of bounds: Factors affecting the perceived offensiveness of racial slurs. *Language Sciences*, 52:155–164.

Gwenn Schurgin O'Keeffe, Kathleen Clarke-Pearson, and Council on Communications and Media. 2011. The impact of social media on children, adolescents, and families. *Pediatrics*, 127(4):800–804.

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture*, 25(2):700–732.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and Multi-Aspect hate speech analysis. In *EMNLP*.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on Twitter. In *Proceedings of the Workshop on Abusive Language Online*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *EMNLP*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014a. Glove: Global vectors for word representation. In *EMNLP*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014b. Glove: Global vectors for word representation. In *EMNLP*.

Gonçalo Pereira, Rui Prada, and Pedro A Santos. 2016. Integrating social power into the decision-making of cognitive agents. *Artificial Intelligence*, 241:1–44.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *EMNLP*, pages 2463–2473.

E. Ashby Plant and Patricia G. Devine. 1998. Internal and external motivation to respond without prejudice. *J. Pers. Soc. Psychol.*, 75(3):811–832.

Martha E. Pollack. 2005. Intelligent technology for an aging population: The use of ai to assist elders with cognitive impairment. *AI Magazine*.

Vinodkumar Prabhakaran, Prabhakaran Vinodkumar, and Rambow Owen. 2014. Predicting power relations between participants in written dialog from a single thread. In *ACL*.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through Back-Translation. In *ACL*. Code available at https://github.com/shrimai/Style-Transfer-Through-Back-Translation.

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual LSTM networks. In *COLING*.

Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-level race and ethnicity predictors from Twitter text. In *COLING*.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *AAAI*.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *EMNLP*.

R Quian Quiroga, L Reddy, G Kreiman, C Koch, and I Fried. 2005. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Unpublished.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Unpublished.

Jacquelyn Rahman. 2012. The N word: Its history and use in the African American community. *Journal of English Linguistics*, 40(2):137–171.

Anil Ramakrishna, Victor R Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. 2017. Linguistic analysis of differences in portrayal of movie characters. In *ACL*.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *NAACL*.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. In *ACL*.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *ACL*.

Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgílio A. F. Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on Twitter. In *ICWSM*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why should I trust you?": Explaining the predictions of any classifier. In *KDD*.

Sarah T Roberts. 2016. Commercial content moderation: digital laborers’ dirty work. In Safiya Umoja Noble and Brendesha M Tynes, editors, *The Intersectional Internet: Race, Sex, Class and Culture Online*, Media Studies Publications. Peter Lang Publishing.

Sarah T Roberts. 2019. *Behind the screen*. Yale University Press.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Alexey Romanov, Anna Rumshisky, Anna Rogers, and David Donahue. 2019. Adversarial decomposition of text representation. In *NAACL*.

Jonathan Rosa. 2019. *Looking like a language, sounding like a race*. Oxford University Press.

Jonathan Rosa and Nelson Flores. 2017. Unsettling race and language: Toward a raciolinguistic perspective. In *Language In Society*. Cambridge University Press.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: the case of the european refugee crisis. In *NLP 4 CMC Workshop*.

RWJF. 2017. Discrimination in america: experiences and views. `https://www.rwjf.org/en/library/research/2017/10/discrimination-in-america--experiences-and-views.html`. Accessed: 2019-11-5.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: an adversarial winograd schema challenge at scale. In *AAAI*.

Elizabeth Sanders. 2002. *From user-centered to participatory design approaches*, pages 1–7.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *ACL*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019a. The risk of racial bias in hate speech detection. In *ACL*.

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 *ACL*.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019b. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *EMNLP*. Connotation Frames downloaded from `http://maartensap.com/movie-bias/`.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019c. Social IQa: commonsense reasoning about social interactions. In *EMNLP*.

R.C. Schank and R.P. Abelson. 1977. *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*. The Artificial Intelligence Series. Lawrence Erlbaum Associates.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Workshop on NLP for Social Media*.

Lenhart Schubert. 2002. Can we derive general world knowledge from texts? In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 94–97, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. In *CoNLL*.

Maya Sen and Omar Wasow. 2016. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19.

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. 2019. Pragmatically informative text generation. In *NAACL*, pages 4060–4067.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from Non-Parallel text by Cross-Alignment. In *NeurIPS*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *EMNLP*.

Arthur K Spears. 1998. African-American language use: Ideology and so-called obscenity. In Salikoko S Mufwene, John R Rickford, Guy Bailey, and John Baugh, editors, *African-American English: Structure, History and Use*, pages 226–250. Routledge New York.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*.

Whitney Strub. 2008. The clearly obscene and the queerly obscene: heteronormativity and obscenity in cold war los angeles. *American Quarterly*, 60(2):373–398.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. Transforming delete, retrieve, generate approach for controlled text style transfer. In *EMNLP*.

Diana I Tamir, Mark A Thornton, Juan Manuel Contreras, and Jason P Mitchell. 2016. Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences of the United States of America*, 113(1):194–199.

Niket Tandon, Gerard de Melo, and Gerhard Weikum. 2017. Webchild 2.0 : Fine-grained commonsense knowledge distillation. In *ACL*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *ACL*.

Mark A Thornton, Miriam E Weaverdyck, and Diana I Tamir. 2019. The brain represents people as the mental states they habitually experience. *Nature communications*, 10(1):2291.

Luiz Valério P Trindade. 2018. On the frontline: The rise of hate speech and racism on social media. `https://discoversociety.org/2018/09/04/on-the-frontline-the-rise-of-hate-speech-and-racism-on-social-media/`. Accessed: 2018-12-6.

Stefanie Ullmann and Marcus Tomalin. 2019. Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *NeurIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *NeurIPS*.

James Vincent. 2016. Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist. Accessed: 2019-10-26.

Andrew J Vonasch and Roy F Baumeister. 2017. Unjustified side effects were strongly intended: taboo tradeoffs and the side-effect effect. *Journal of Experimental Social Psychology*, 68:83–92.

Yequan Wang, Minlie Huang, xiaoyan zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *EMNLP*.

Zijian Wang and Christopher Potts. 2019. TalkDown: a corpus for condescension detection in context. In *EMNLP*.

Zeerak Waseem and Dirk Hovy. 2016a. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *NAACL Student Research Workshop*.

Zeerak Waseem and Dirk Hovy. 2016b. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *NAACL Student Research Workshop*.

Zeerak Waseem, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In Jennifer Golbeck, editor, *Online Harassment*, pages 29–55. Springer International Publishing, Cham.

Apryl Williams and Doris Domoszlai. 2013. BlackTwitter: a networked cultural identity. *Harmony Institute*.

Thomas Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. Unpublished.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: personal attacks seen at scale. In *WWW*.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *TACL*.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *COLING*.

Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *NeurIPS*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL*.

Danyaal Yasin. 2018. Black and banned: Who is free speech for? https://www.indexoncensorship.org/2018/09/black-and-banned-who-is-free-speech-for/. Accessed: 2018-12-6.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *NAACL*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*.

Ye Zhang, Nan Ding, and Radu Soricut. 2018a. SHAPED: Shared-Private Encoder-Decoder for text style adaptation. In *NAACL*.

Ziqi Zhang, David Robinson, and Jonathan A. Tepper. 2018b. Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In *Proceedings of ESWC*.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *AAAI*, volume 34.

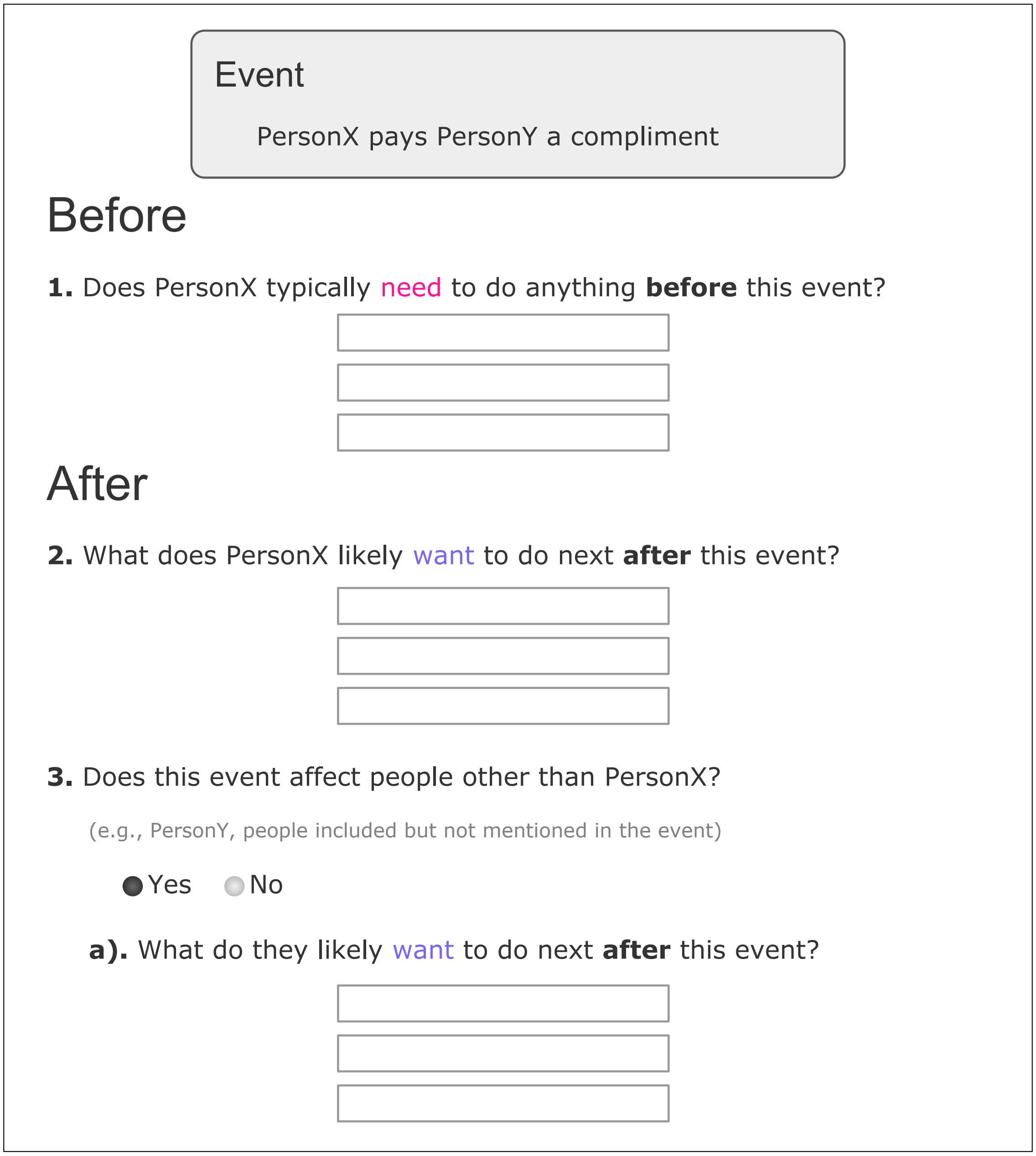
Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*.

Michael Zimmer. 2018. Addressing conceptual gaps in big data research ethics: an application of contextual integrity. *Social Media + Society*, 4(2).

âpihtawikosisân. 2016. Beyond territorial acknowledgments. Accessed 2021-07-09.

Appendix A

ATOMIC Supplementary



Event

PersonX pays PersonY a compliment

Before

1. Does PersonX typically need to do anything **before** this event?

After

2. What does PersonX likely want to do next **after** this event?

3. Does this event affect people other than PersonX?

(e.g., PersonY, people included but not mentioned in the event)

● Yes ○ No

a). What do they likely want to do next **after** this event?

Figure A.1: Template of the crowdsourcing task for gathering commonsense knowledge around events. Specific setups vary depending on the dimension annotated.

Appendix B

POWERTRANSFORMER Supplementary

B.1 Additional data description

B.1.1 ROC story corpus

This English corpus originally contains 100,000 five-sentence stories written by crowdworkers about realistic everyday scenarios. We select the data for our task by first extracting agency levels for each sentence, filtering out those with indeterminable agency. Additionally, we filter out sentences with four or more verbs, to prevent the sentence masking from deleting too many content words.

B.1.2 Paraphrase corpus

This corpus contains paraphrases of spoken dialogue extracted from movie and TV subtitles.[1] OpusParcus was created by automatically aligning the subtitles sentences using several probabilistic metrics, including likelihood under a round-trip translation paraphrasing model (Bannard and Callison-Burch, 2005) and pointwise mutual information. For our paraphrasing dataset, we apply the same filtering as with the ROC story corpus to the English portion of the OpusParcus training corpus and select the top 10% highest scoring paraphrases using the PMI scoring from the original paper. We extract agency levels for each pair of paraphrases, and select pairs to obtain roughly equal number of agency-level pairs (i.e., 1/9th positive-neutral, 1/9th positive-negative, etc.) We preprocess the text by stripping any leading periods and commas.

Hyperparameter	Value
Vocabulary Size	40486
Maximum Sequence Length	64
Training Batch Size	4
Embedding Size	768
# Attention Heads	12
# Attention Layers	12

Table B.1: POWERTRANSFORMER hyperparameters.

B.2 Experimental details

We use the Hugging Face Wolf et al. (2019) implementation of OpenAI's GPT model (117M parameters; Radford et al., 2018). our final setup uses AdamW Loshchilov and Hutter (2019) as our optimizer with a learning weight of 1e-5, batch size of 4 and maximum sequence length of 64. In preliminary results, we find that β=5 aptly steers the generation while avoiding repetition issues.

[1] From http://www.opensubtitles.org

B.2.1 POWERTRANSFORMER details

All the experiments are performed on NVIDIA TITAN card and use the model hyperparameters listed in Table B.1.

POWERT $_{ParaOnly+None}$

We train this model for 10 epochs with each epoch taking approximately an hour. The learning rate is 1e-5 with AdamW optimizer, which is tuned manually in the [1e-6, 1e-3] range for 7 times. We use $p = 0.4$ for nucleus sampling and p is tuned manually in the [0.4, 0.9] range for 5 values.

POWERT $_{ParaOnly+Static}$

The POWERT $_{ParaOnly+Static}$ loads the trained model from POWERT $_{ParaOnly+None}$ and add re-scaling to the logits. The re-scaling factor, β was tuned manually tuned in the [0, 10] range. We try 8 βs and use 5 in the final model. We use the same p as POWERT $_{ParaOnly+None}$

POWERT $_{Joint+None}$

Similar to POWERT $_{ParaOnly+None}$, we train this model for 10 epochs with each epoch taking approximately an hour. The learning rate is 1e-5 with AdamW optimizer, which is tuned manually in the [1e-6, 1e-3] range for 7 times.We use the same p as POWERT $_{ParaOnly+None}$

POWERT $_{Joint+Static}$

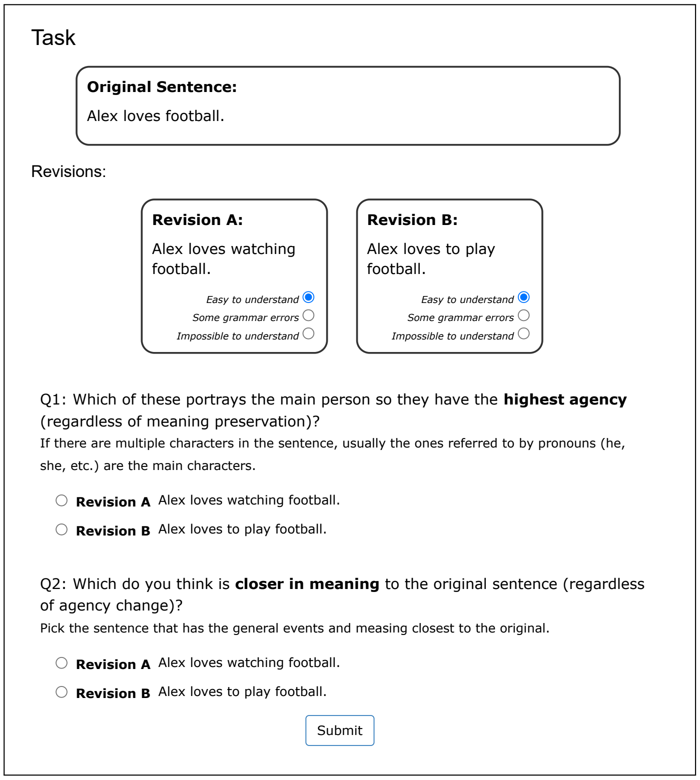
The POWERT $_{Joint+Static}$ loads the trained model from POWERT $_{Joint+None}$ and add re-scaling to the logits. The re-scaling factor, β was tuned manually tuned in the [0, 10] range. We try 8 βs and use 5 in the final model. We use the same p as POWERT $_{ParaOnly+None}$

B.2.2 PPLM details

The PPLM decoding method can be used on top of any model, but their original codebase is for use with a pre-trained language model rather than a model for paraphrasing or style transfer. We augment their techniques for this task by replacing the base model in their code with a denoising autoencoder that was trained to reconstruct the input sentence. The denoising autoencoder was implemented using the base GPT2 model (to fit with their code library and be similar size to our model). It was trained on our ROC only training data with a reconstruction objective. In order to denoise the autoencoder, we randomly "dropout" about 50% of the tokens from the context by replacing them with mask tokens. This autoencoder is trained to reconstruct input sentences, but when used with the PPLM decoding method, the input gets dynamically updated to decode a sentence that is similar in meaning but more likely to have a positive/negative agency according to a discriminator that is trained on top of the autoencoder. The PPLM decoding method also has hyperparameters that control the strength of the target label. If set too high, then the output could be degenerate. We manually set the hyperparameters to be as strong possible without producing degenerate text, using a subset of the dev. set as a guide.

B.2.3 Backtranslation details

We use the code provided by Prabhumoye et al. (2018) for running this baseline. After lowercasing all the negative and positive agency examples in our training data (ROC and OpusParcus), we translate to French using the machine translation model provided in the code base. This baseline requires training a style classifier (agency) and two decoders (one for each agency level). Since the classifier essentially re-learns the agency lexicon, we do not search for hyperparameters, and simply set a learning rate of 5, and 6 epochs. For training the decoders, we perform grid search to find the best hyperparameters. We experiment with a learning rates of {0.5, 1, 2, 5}, {2, 3, 5} epochs, a classification-loss weight of {0.5, 1, 2}, and a word-loss weight of {0.5, 1, 2}, and select the configuration with the best word-level accuracy on the dev. set. We use SGD with a batch size of 64 for all experiments, and refer the reader to the code base for other default parameters.



Task

Original Sentence:
Alex loves football.

Revisions:

Revision A:
Alex loves watching football.

Easy to understand ●
Some grammar errors ○
Impossible to understand ○

Revision B:
Alex loves to play football.

Easy to understand ●
Some grammar errors ○
Impossible to understand ○

Q1: Which of these portrays the main person so they have the **highest agency** (regardless of meaning preservation)?
If there are multiple characters in the sentence, usually the ones referred to by pronouns (he, she, etc.) are the main characters.

○ **Revision A** Alex loves watching football.
○ **Revision B** Alex loves to play football.

Q2: Which do you think is **closer in meaning** to the original sentence (regardless of agency change)?
Pick the sentence that has the general events and measing closest to the original.

○ **Revision A** Alex loves watching football.
○ **Revision B** Alex loves to play football.

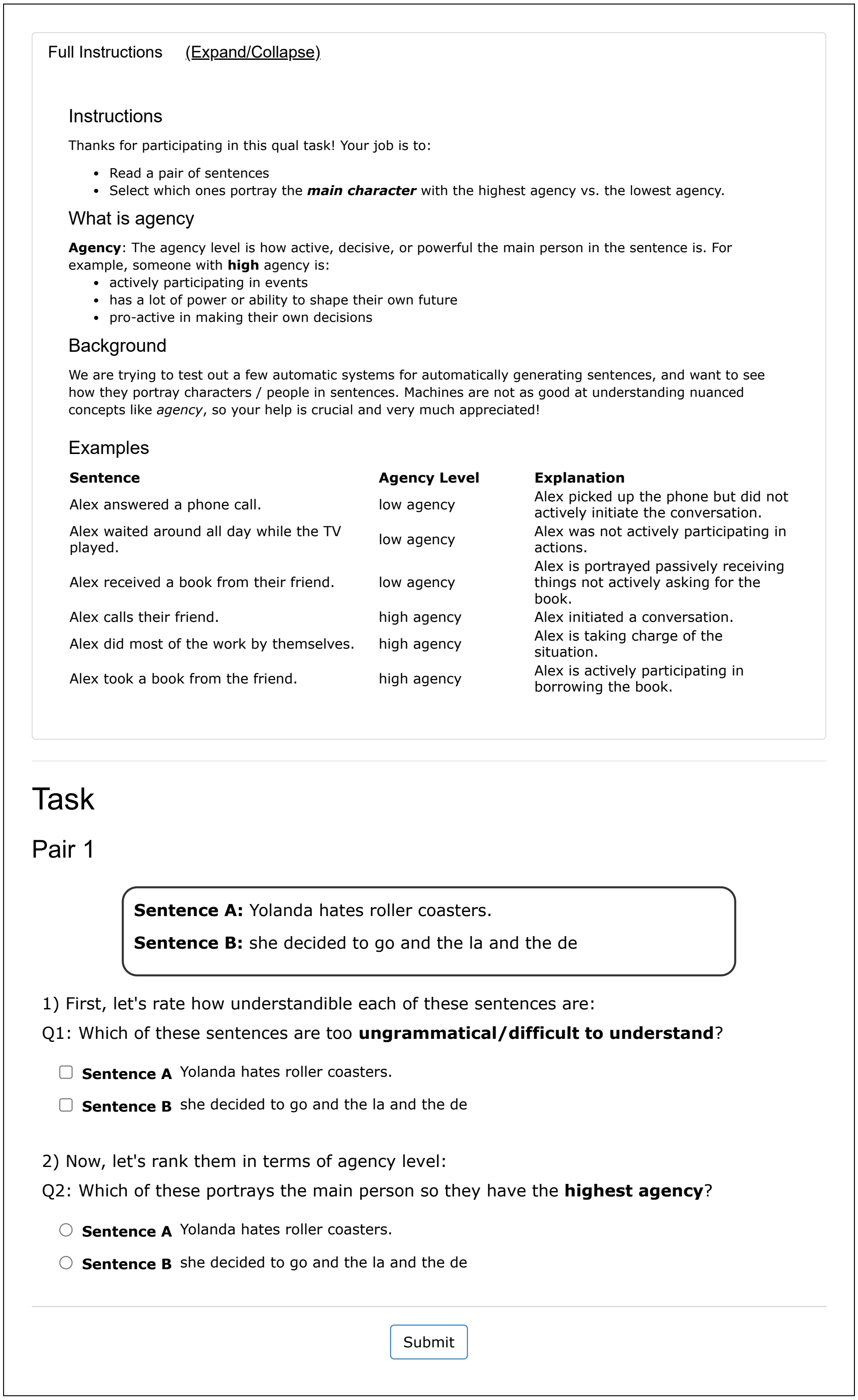
Figure B.1: Screenshot of the human evaluation annotation task.

B.3 Gender Bias in Movies

B.3.1 Extracting gender from characters

The movie scripts mention characters in all caps, making it easy to identify and extract them. We then cross reference the name (or, description for unnamed characters, e.g., “the doorman”) with a list of gendered names[2] and gendered words (e.g., “waitress,” “policeman,” “police woman”). To allow for better rewriting using our model, we split the narratives into sentences (using NLTK’s sentence tokenizer Bird et al., 2009), and assign each sentence to a character if their name appears in the sentence.

[2] http://www.cs.cmu.edu/Groups/AI/util/areas/nlp/corpora/names/0.html



Full Instructions (Expand/Collapse)

Instructions

Thanks for participating in this qual task! Your job is to:

- Read a pair of sentences
- Select which ones portray the ***main character*** with the highest agency vs. the lowest agency.

What is agency

Agency: The agency level is how active, decisive, or powerful the main person in the sentence is. For example, someone with **high** agency is:

- actively participating in events
- has a lot of power or ability to shape their own future
- pro-active in making their own decisions

Background

We are trying to test out a few automatic systems for automatically generating sentences, and want to see how they portray characters / people in sentences. Machines are not as good at understanding nuanced concepts like *agency*, so your help is crucial and very much appreciated!

Examples

Sentence	**Agency Level**	**Explanation**
Alex answered a phone call.	low agency	Alex picked up the phone but did not actively initiate the conversation.
Alex waited around all day while the TV played.	low agency	Alex was not actively participating in actions.
Alex received a book from their friend.	low agency	Alex is portrayed passively receiving things not actively asking for the book.
Alex calls their friend.	high agency	Alex initiated a conversation.
Alex did most of the work by themselves.	high agency	Alex is taking charge of the situation.
Alex took a book from the friend.	high agency	Alex is actively participating in borrowing the book.

Task

Pair 1

Sentence A: Yolanda hates roller coasters.

Sentence B: she decided to go and the la and the de

1) First, let's rate how understandible each of these sentences are:

Q1: Which of these sentences are too **ungrammatical/difficult to understand**?

□ **Sentence A** Yolanda hates roller coasters.

□ **Sentence B** she decided to go and the la and the de

2) Now, let's rank them in terms of agency level:

Q2: Which of these portrays the main person so they have the **highest agency**?

○ **Sentence A** Yolanda hates roller coasters.

○ **Sentence B** she decided to go and the la and the de

Figure B.2: Screenshot of the qualification task and its instructions. In the real task, workers rated three pairs of sentences, but only one is shown here.

ΔAgency	Input	Model	Output	Agency(out)
(a) + → −	After the party I headed home.	PPLM	After the party my classmate, Kayla and the Tgirls of the Universe.	=
		BST	please ’s , i have a word of this .	−
		POWERT $_{ParaOnly+NoBoost}$	after the party i headed home.	+
		POWERT $_{ParaOnly+Boost}$	after the party i headed home.	+
		POWERT $_{Joint+SupplyVerb}$	after the party i faced home.	−
		POWERT $_{Joint+NoBoost}$	after the party i stayed home.	−
		POWERT $_{Joint+Boost}$	after the party i stayed home.	−
(b) + → −	A friend asked me to watch her two year old child for a minute.	PPLM	A Friend asked me to watch her two year old child for a minute.	+
		BST	l didn ’t have a word of this , you ’re .	−
		POWERT $_{ParaOnly+NoBoost}$	a friend asked me to watch her two year old child for a minute.	+
		POWERT $_{ParaOnly+Boost}$	a friend asked me to watch her two year old child for a minute.	+
		POWERT $_{Joint+SupplyVerb}$	a friend told me to watch her two year old child for a minute.	+
		POWERT $_{Joint+NoBoost}$	a friend needed me to watch her two year old child for a minute.	−
		POWERT $_{Joint+Boost}$	a friend needed me to watch her two year old child for a minute.	−
(c) + → −	After filling in the data it looked quite sharp.	PPLM	Before filling the last question it it it it looked quite sharp. Before filling the last question it it	+
		BST	when the ’t you want a word ?	−
		POWERT $_{ParaOnly+NoBoost}$	after filling in the data it looked quite sharp.	+
		POWERT $_{ParaOnly+Boost}$	after filling in the data it seemed quite sharp.	+
		POWERT $_{Joint+SupplyVerb}$	after putting in the data it looked quite sharp.	=
		POWERT $_{Joint+NoBoost}$	after analyzing in the data it looked quite sharp.	=
		POWERT $_{Joint+Boost}$	after seeing in the data it seemed quite sharp.	−
(d) − → +	Allie was failing science class.	PPLM	Allie was failing science grade.	−
		BST	do you want me ?	+
		POWERT $_{ParaOnly+NoBoost}$	allie was failing science class.	−
		POWERT $_{ParaOnly+Boost}$	allie was failing science class.	−
		POWERT $_{Joint+SupplyVerb}$	allie was ignoring science class.	+
		POWERT $_{Joint+NoBoost}$	allie was failing science class.	−
		POWERT $_{Joint+Boost}$	allie was taking science class.	+
(e) − → +	Darla wanted a soft drink.	PPLM	darla wants a hard hard drink.	−
		BST	don ’t take me a man .	+
		POWERT $_{ParaOnly+NoBoost}$	darla wanted a soft drink.	−
		POWERT $_{ParaOnly+Boost}$	darla wanted a soft drink.	−
		POWERT $_{Joint+SupplyVerb}$	darla got a soft drink.	+
		POWERT $_{Joint+NoBoost}$	darla ordered a soft drink.	+
		POWERT $_{Joint+Boost}$	darla ordered a soft drink.	+
(f) − → +	Clint paused on the trail.	PPLM	clint was on the trail.	
		BST	don ’t you want me ,	−
		POWERT $_{ParaOnly+NoBoost}$	clint paused on the trail.	−
		POWERT $_{ParaOnly+Boost}$	clint stopped on the trail.	+
		POWERT $_{Joint+SupplyVerb}$	clint walked on the trail.	+
		POWERT $_{Joint+NoBoost}$	clint hiked on the trail.	=
		POWERT $_{Joint+Boost}$	clint walked on the trail heading down.	+

Table B.2: Full version of Table 3.4. Example revisions from various models for sentences from the dev. set. Columns are: the target change in agency from the original to the target agency, the input sentence, the model, generated output, and the actual agency level of the output measured by the connotation frame lexicon.

Appendix C

Racial Bias in Hate Speech Detection Supplementary

We present further evidence of racial bias in hate speech detection in this appendix.

C.1 Experimental Details for Classification

For each dataset, we randomly split the data into train/dev./test sets (73/12/15%), and perform early stopping when classification accuracy on dev. data stops increasing. For DWMW17, which has multiple annotations per instance, we use the majority class as the label, dropping instances that are tied. For both datasets, we preprocess the text using an adapted version of the script for Twitter GloVe vectors.[1] In our experiments, we set $H = 64$, and use a vocabulary size of $|V| = 19$k and $|V| = 74$k for DWMW17 and FDCL18, respectively, and initialize the embedding layer with 300-dimensional GloVe vectors trained on 840 billion tokens. We experimented with using ELMo embeddings, but found that they did not boost performance for this task. We optimize these models using Adam with a learning rate of 0.001, and a batch size of 64.

C.2 Bias in Waseem and Hovy (2016a)

category	count	AAE corr.
racism	1,976	−0.117
sexism	3,430	0.168
none	11,501	−0.064
total	**16,907**	

Table C.1: Data statistics in WH16, as well as the Pearson r correlations with the labels and inferred AAE dialect. All correlations are $p \ll 0.001$.

We replicate our analyses in §5.3 on the widely used dataset by Waseem and Hovy (2016a, henceforth, WH16), which categorizes tweets in three hate speech categories: *racist*, *sexist*, or *none*, shown in Table C.1, along with their correlations with AAE. This dataset suffers from severe sampling bias that limit the conclusions to be drawn from this data: 70% of sexist tweets were written by two users, and 99% of racist tweets were written by a single user Schmidt and Wiegand (2017); Klubička and Fernandez (2018).

In Figure C.1 (left), we show how models trained on this dataset have slightly higher false positive rates of sexism on AAE tweets, and of the "none" label for White tweets compared to AAE tweets. When predicting on our reference corpora (Figure C.1, middle and right), we see AAE tweets (or tweets by African Americans) are labelled as sexist more than White-aligned tweets or tweets by White users. Again, due to the sampling issues, these results should be interpreted cautiously.

[1] https://nlp.stanford.edu/projects/glove/preprocess-twitter.rb

WH16	% false identification		
Group | Acc.	Racism	Sexism	None
AAE | 83.8	0.9	2.8	32.5
White | 83.5	3.2	2.7	34.6
Overall | 84.1	2.7	3.0	35.9

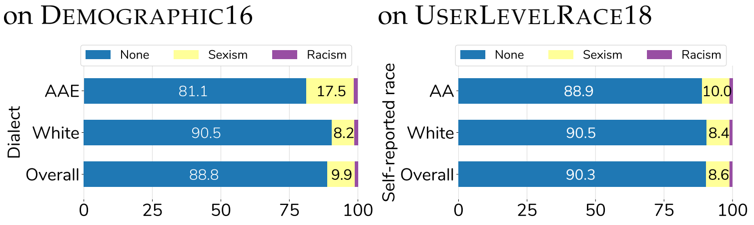
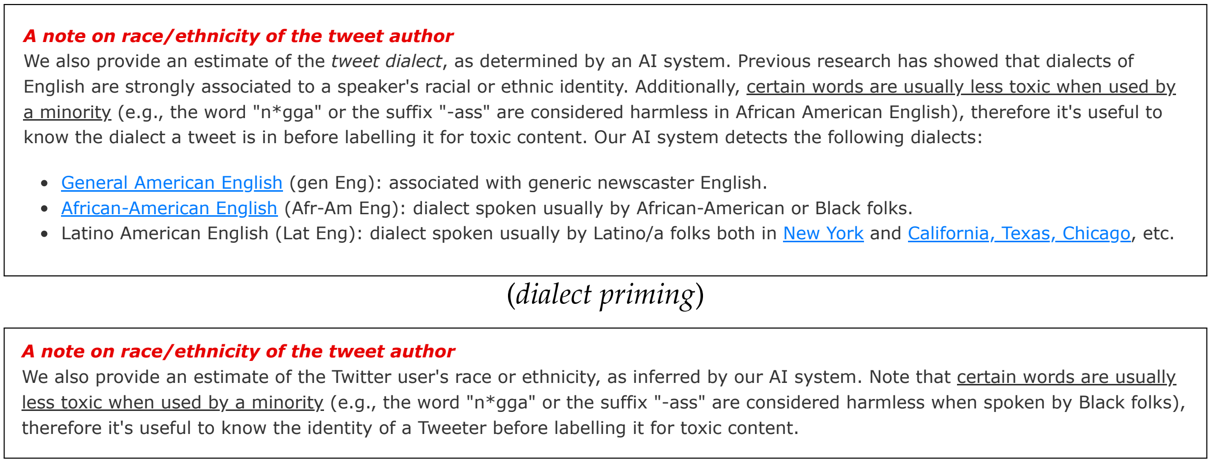


Figure C.1: *Left*: classification accuracy and per-class rates of false positives (FP) on test data for the model trained on WH16. *Middle and right*: average probability mass of toxicity classes in DEMOGRAPHIC16 and USERLEVELRACE18, respectively, as given by the WH16 classifier. As in Figure 5.3, proportions are shown for AAE, White-aligned English, and overall (all tweets) for DEMOGRAPHIC16, and for self-identified White authors, African American authors (AA), and overall for USERLEVELRACE18.

C.3 Dialect Priming Experimental Details

We collected annotations from 110 (76% White), 143 (77% White), and 81 (72% White) workers in the control, dialect, and race priming conditions, respectively. Figure C.2 shows the instruction snippet related to dialect and race shown to workers in the two treatment conditions. Additionally, Figure C.3 shows the annotation interface, with (a) and without priming (b,c).



A note on race/ethnicity of the tweet author
We also provide an estimate of the *tweet dialect*, as determined by an AI system. Previous research has showed that dialects of English are strongly associated to a speaker's racial or ethnic identity. Additionally, certain words are usually less toxic when used by a minority (e.g., the word "n*gga" or the suffix "-ass" are considered harmless in African American English), therefore it's useful to know the dialect a tweet is in before labelling it for toxic content. Our AI system detects the following dialects:

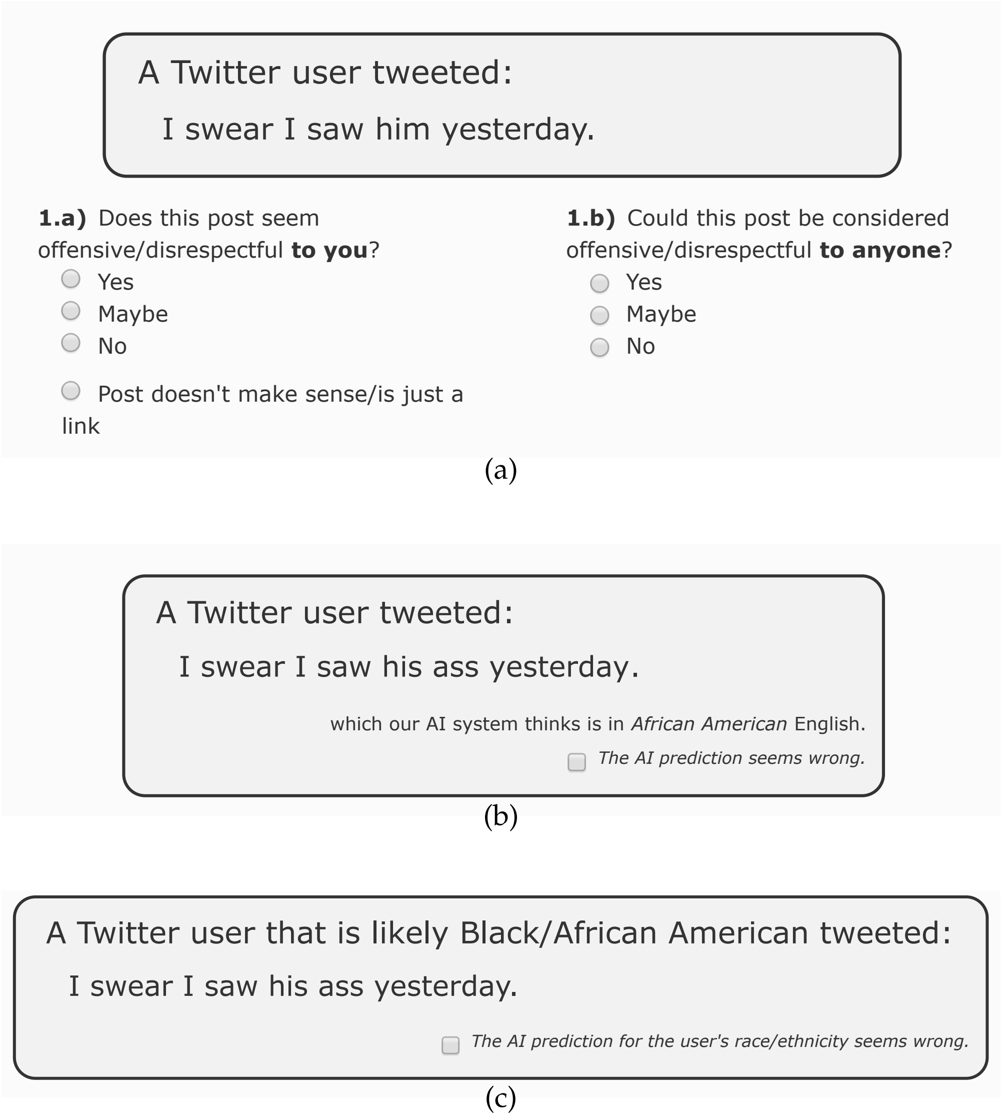
- General American English (gen Eng): associated with generic newscaster English.
- African-American English (Afr-Am Eng): dialect spoken usually by African-American or Black folks.
- Latino American English (Lat Eng): dialect spoken usually by Latino/a folks both in New York and California, Texas, Chicago, etc.

(*dialect priming*)

A note on race/ethnicity of the tweet author
We also provide an estimate of the Twitter user's race or ethnicity, as inferred by our AI system. Note that certain words are usually less toxic when used by a minority (e.g., the word "n*gga" or the suffix "-ass" are considered harmless when spoken by Black folks), therefore it's useful to know the identity of a Tweeter before labelling it for toxic content.

(*race priming*)

Figure C.2: Additional instructions shown to workers in the *dialect* and *race* priming. In the dialect condition, we provide links to the dialect Wikipedia pages.



A Twitter user tweeted:

I swear I saw him yesterday.

1.a) Does this post seem offensive/disrespectful **to you**?

○ Yes
○ Maybe
○ No

○ Post doesn't make sense/is just a link

1.b) Could this post be considered offensive/disrespectful **to anyone**?

○ Yes
○ Maybe
○ No

(a)

A Twitter user tweeted:

I swear I saw his ass yesterday.

which our AI system thinks is in *African American* English.

□ *The AI prediction seems wrong.*

(b)

A Twitter user that is likely Black/African American tweeted:

I swear I saw his ass yesterday.

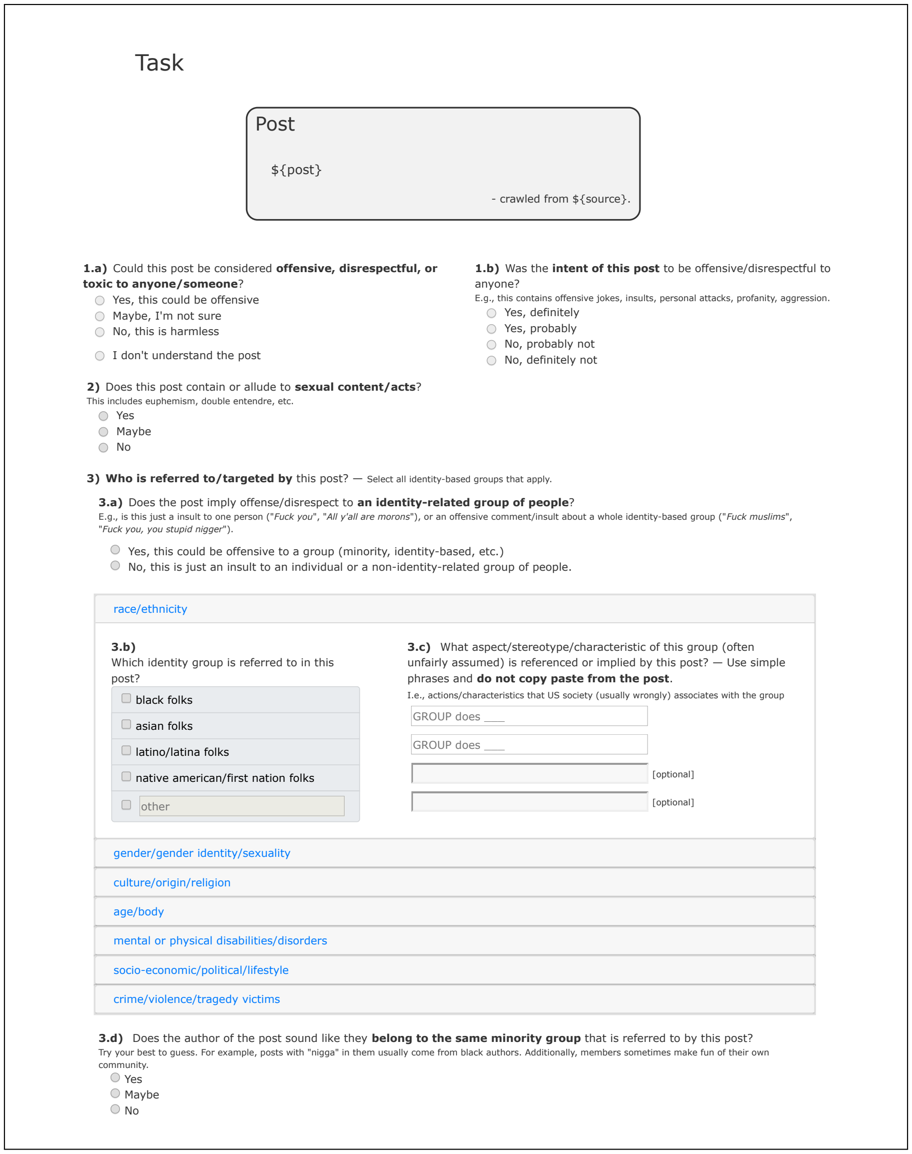
□ *The AI prediction for the user's race/ethnicity seems wrong.*

(c)

Figure C.3: Interface for the controlled experiment. (a) shows the control condition along with the offensiveness questions. (b) and (c) show the changes to the treatment interface in the *dialect* and *race priming* conditions.

Appendix D

Social Bias Frames Supplementary



Task

Post

${post}

- crawled from ${source}.

1.a) Could this post be considered **offensive, disrespectful, or toxic to anyone/someone**?

○ Yes, this could be offensive
○ Maybe, I'm not sure
○ No, this is harmless

○ I don't understand the post

1.b) Was the **intent of this post** to be offensive/disrespectful to anyone?
E.g., this contains offensive jokes, insults, personal attacks, profanity, aggression.

○ Yes, definitely
○ Yes, probably
○ No, probably not
○ No, definitely not

2) Does this post contain or allude to **sexual content/acts**?
This includes euphemism, double entendre, etc.

○ Yes
○ Maybe
○ No

3) **Who is referred to/targeted by** this post? — Select all identity-based groups that apply.

3.a) Does the post imply offense/disrespect to **an identity-related group of people**?
E.g., is this just a insult to one person ("*Fuck you*", "*All y'all are morons*"), or an offensive comment/insult about a whole identity-based group ("*Fuck muslims*", "*Fuck you, you stupid nigger*").

○ Yes, this could be offensive to a group (minority, identity-based, etc.)
○ No, this is just an insult to an individual or a non-identity-related group of people.

race/ethnicity

3.b)
Which identity group is referred to in this post?

□ black folks
□ asian folks
□ latino/latina folks
□ native american/first nation folks
□ other

3.c) What aspect/stereotype/characteristic of this group (often unfairly assumed) is referenced or implied by this post? — Use simple phrases and **do not copy paste from the post**.
I.e., actions/characteristics that US society (usually wrongly) associates with the group

GROUP does ___
GROUP does ___
[optional]
[optional]

gender/gender identity/sexuality

culture/origin/religion

age/body

mental or physical disabilities/disorders

socio-economic/political/lifestyle

crime/violence/tragedy victims

3.d) Does the author of the post sound like they **belong to the same minority group** that is referred to by this post?
Try your best to guess. For example, posts with "nigga" in them usually come from black authors. Additionally, members sometimes make fun of their own community.

○ Yes
○ Maybe
○ No

Figure D.1: Snippet of the annotation task used to collect SBIC. The collection of structured annotations for our framework is broken down into questions pertaining to offensiveness, intent of the post, targeted group and minority speaker.

		offensive			intent			lewd			group			in-group		
	model	*42.2% pos. (dev.)*			*44.8% pos. (dev.)*			*3.0% pos. (dev.)*			*66.6% pos. (dev.)*			*5.1% pos. (dev.)*		
		F_1	pr.	rec.	F_1	pr.	rec.	F_1	pr.	rec.	F_1	pr.	rec.	F_1	pr.	rec.
dev.	SBF-GPT$_1$-gdy	75.2	88.3	65.5	74.4	89.8	63.6	75.2	78.2	72.5	62.3	74.6	53.4	–	–	–
	″-constr	75.2	88.3	65.5	74.4	89.8	63.6	75.2	78.2	72.5	62.3	74.6	53.4	–	–	–
	SBF-GPT$_2$-gdy	77.2	88.3	68.6	**76.3**	89.5	66.5	77.6	81.2	74.3	**66.9**	67.9	65.8	24.0	85.7	14.0
	″-constr	77.2	88.3	68.6	**76.3**	89.5	66.5	77.6	81.2	74.3	**66.9**	67.9	65.8	**25.9**	63.6	16.3
	SBF-GPT$_2$-smp	**80.5**	84.3	76.9	75.3	89.9	64.7	**78.6**	80.6	76.6	66.0	67.6	64.5	–	–	–
	″-constr	80.4	84.3	76.8	75.3	89.9	64.7	78.5	80.6	76.5	66.0	67.6	64.5	–	–	–
test	SBF-GPT$_2$-gdy	78.8	89.8	70.2	78.6	90.8	69.2	80.7	84.5	77.3	69.9	70.5	69.4	–	–	–
	″-constr	78.8	89.8	70.2	78.6	90.8	69.2	80.7	84.5	77.3	69.9	70.5	69.4	–	–	–

Table D.1: Full experimental results (%) of various models on the classification tasks (gdy: argmax, smp: sampling; constr: constrained decoding). Some models did not predict the positive class for "in-group language," their performance is denoted by "–". We bold the F_1 scores of the best performing model(s) on the development set. For easier interpretation, we also report the percentage of instances in the positive class in the development set.