Pretty Princess vs. Successful Leader: Gender Roles in Greeting Card Messages

Jiao Sun
jiaosun@usc.edu
University of Southern California
USA

Tongshuang Wu
wtshuang@cs.washington.edu
University of Washington
USA

Yue Jiang
yuejiang@mpi-inf.mpg.de
Max Planck Institute for Informatics
Germany

Ronil Awalegaonkar
ronil.awale@gmail.com
Latin School of Chicago
USA

Xi Victoria Lin
victorialin@fb.com
Meta AI
USA

Diyi Yang
diyi.yang@cc.gatech.edu
Georgia Institute of Technology
USA

ABSTRACT

People write personalized greeting cards on various occasions. While prior work has studied gender roles in greeting card messages, systematic analysis at scale and tools for raising the awareness of gender stereotyping remain under-investigated. To this end, we collect a large greeting card message corpus covering three different occasions (birthday, Valentine's Day and wedding) from three sources (exemplars from greeting message websites, real-life greetings from social media and language model generated ones). We uncover a wide range of gender stereotypes in this corpus via topic modeling, odds ratio and Word Embedding Association Test (WEAT). We further conduct a survey to understand people's perception of gender roles in messages from this corpus and if gender stereotyping is a concern. The results show that people want to be aware of gender roles in the messages, but remain unconcerned unless the perceived gender roles conflict with the recipient's true personality. In response, we developed GreetA, an interactive visualization and writing assistant tool to visualize fine-grained topics in greeting card messages drafted by the users and the associated gender perception scores, but without suggesting text changes as an intervention.

CCS CONCEPTS

• **Human-centered computing** → **Visualization toolkits**; **Collaborative and social computing design and evaluation methods**.

KEYWORDS

gender role awareness, greeting card messages, visualization system

1 INTRODUCTION

> *Now, the first guests will be arriving in a few minutes, and they are going to find you perfectly behaved, sweet, charming, innocent, attentive, delightful in every way. I particularly wish for that, Lyra, do you understand me?*
>
> —— *Mrs. Coulter, "His Dark Materials"*

Gender stereotyping creates widely accepted biases about certain characteristics of a gender group and perpetuates the notion that gender-associated behaviors are binary.

In many cultures, women are valued by physical attractiveness while men are valued by professional success [27]. Personality-wise, women are supposed to be nurturing and avoid dominance, while men are supposed to be agentic and avoid weakness [18]. Such diminishing, and sometimes negative conceptions of gender groups are one of the greatest barriers for equality and need to be tackled wherever they appear [31].

However, gender stereotypes are deeply rooted in society and arise in all types of media and social interactions [18, 24, 31]. For instance, advertisement, television and movies have all been shown to be glutted with damaging portraits of gender [31, 37, 51, 59?] before governments start to ban such content [28]. In this work, we focus on greeting cards messages, a media type that is under-investigated in this context and is exchanged among billions of people during holidays and special occasions to express affection, gratitude, sympathy, or other sentiments [34].

Greeting cards messages form a compelling source for studying gender stereotypes as they contain both descriptive components (sender's perception of the receiver) and prescriptive components (sender's expectations of the receiver). It is easy to obtain gender information from greeting card messages. People reveal the receiver's identity and gender, sometimes the sender, and the receiver's relationship to the sender in the messages. Greeting card messages significantly impact our everyday life. Since people often share greeting card messages out of the goodwill, gender stereotypes may be enforced without being noticed. We show an example where the greeting for a female recipient is about appearance while for a male recipient is about leadership for their birthdays in Table 1. Thus, it is important to help people prevent unconscious gender

Recipient	Greeting Card Message
Brother (M)	Having a brother who cares like a mother is scarce; you have exhibited great leadership qualities which makes you a successful father, husband, and sibling. Happy Birthday to you.
Niece (F)	My beautiful niece is growing a year older that makes me one of the happiest uncles in the world. Have a grand birthday sweetheart; you will always be my princess.

Table 1: Examples of gender roles in greeting card messages for birthday from OccasionMessage [36].

stereotypes. Besides, greeting card messages can be used on various social occasions (e.g., birthdays, weddings, etc.), which offers more opportunities to observe and analyze the gender stereotypes in different settings. This research seeks to answer the following research questions:

RQ1 Would there be gender stereotypes in greeting card messages? To women lean towards their beauties and household, and to men lean towards their work achievement?

RQ2 If RQ1 stands, would the gender association of topics in greeting card messages relate to recipients' age?

RQ3 Do people want to be informed of the gender role (masculinity and femininity for our study) in their greeting card messages?

RQ4 If RQ3 stands, how can we help increase the gender role awareness to people? What features do they expect?

To quantify these problems, we scraped over 18,000 greeting card messages from eight popular websites (including HallMark [32] and American Greetings [38]) to form our *template dataset*, which covers three different scenarios: Birthday, Valentine's Day and Wedding. We extract topics from greeting card messages, further select the ones that have gender associations, and do fine-grained analysis for different age groups. Besides templates, we are also interested in studying greeting messages written by people. We approximate greetings written by people using those generated by the state-of-the-art language model GPT-2 [41], which are trained over an ocean of text on the web. Our analysis shows that greeting card messages to women lean towards their appearance and household and to men lean towards work achievement, and greetings to the elderly have less association with gender compared to other age groups. We obtained similar results on greeting messages generated by language models.

To understand how people perceive the potential gender association of topics in their own messages, we designed a pilot survey and collected feedback from twenty users with a diverse ethnical background spanning from their 20s to 50s. According to the survey, most people would like to be informed of the gender role and avoid potential gender stereotypes in their greeting card messages, but do not want the machine to intervene much or modify their messages. Therefore, we develop GreetA (**Greet**ing with Gender Role **A**wareness), a visualization tool that helps users write greeting card messages with the gender role awareness.

We further verify the effectiveness of GreetA by conducting comprehensive qualitative and quantitative user studies. Our qualitative study indicates that most of participants agree that GreetA is useful for writing greetings, is easy to learn and use, and they would like to use it in the future. In addition, three contrast surveys in the quantitative study show that GreetA helps people increase the gender role awareness when they write greeting card messages. These user studies explore people's perceptions of gender association in greeting card messages and show that GreetA effectively assists people in being more aware of gender association at scale.

In summary, our work demonstrates the following contributions:

- **Dataset.** We collect a large-scale greeting template dataset with gender information, with over eighteen thousand messages and six greeting scenarios. [1] We also contribute an AI-generated (i.e., GPT-2) greeting message corpus together with a birthday greeting Tweets dataset to facilitate future research on greeting card messages.
- **Analysis.** To the best of our knowledge, we are the first to analyze gender roles in large-scale greeting card messages quantitatively using statistical NLP tools. We further analyze greeting messages among different age groups and scenarios.
- **System.** Based on the analysis of users' feedback and requirements, we design and build a supportive visualization tool, named *GreetA*, to help users write greeting card messages with gender role awareness.
- **Evaluation.** To evaluate GreetA, we conduct both a qualitative user study and a quantitative user study. We found that GreetA is easy to use and helps users increase gender awareness when writing greeting card messages.

2 RELATED WORK

Gender Related Research. In the HCI community, Burtscher and Spiel [4] provide a staring point for HCI researchers to explore questions and issues around gender. Stumpf et al. [52] give a conceptual review and provide some evidence for the impact of gender in thinking and behavior which underlines HCI research and design, and Schlesinger et al. [47] introduce a framework for engaging with complexity of users' multi-faceted identities including demographic information. Meanwhile, researchers in the social computing community have been studying the gender role in various social settings. For instance, prior work found that women contribute far less frequently than men in the Question and Answer sites and analyzed how the community cultures might be impacting men and women differently [10, 54]. Hannák et al. [14] showed that the perceived gender is significantly correlated with worker evaluations. In addition to the inequalities caused by the external factors, Foong et al. [13] find that the female workers set a self-determined hourly wage lower than males' in an online labor marketplace. Males also express higher negativity and lower desire for social support when they face mental illness [22]. Similarly, McGregor et al. [21] find that male candidates may see more and female candidates see less strategic benefits in personalizing campaign politics on social media. More relatable to our research, Reifman et al. [42] find gender and age differences when expressing vulnerable emotions such as love in anniversary greetings delivered on Twitter.

---

[1] We discarded the "congratulations", "baby shower" and "sympathy" scenarios for analysis because of their small size. In result, we kept and used "Birthday", "Valentine's Day" and "Wedding" scenarios for analysis throughout the paper.

Gender-related research has also been an emerging topic for Artificial Intelligence (AI). In the computer vision field, Scheuerman et al. [46] found that facial analysis technologies performed consistently worse on transgender individuals and were unable to classify non-binary genders. In the natural language processing field, researchers measured the gender gap in various downstream tasks (e.g., authorship and citations [23]), raising ethical considerations to use gender as a variable in NLP. In response, researchers have been addressing the gender awareness and making efforts to alleviate the gender bias in models (e.g.,machine translation field [45, 50]). Besides, researchers have discovered gender bias in pretrained models (word embeddings) [60], data [16], and algorithms themselves [1]. Cryan et al. [7] compared the lexicon and supervised methods for detecting gender stereotypes. Compared to prior work that tries to mitigate gender bias [57], our work *provides gender role awareness* and *prevent potential gender stereotypes.*

Gender Role in Greetings Card Messages. Psychology studies showed that greeting card messages are an important source of the self-concept establishment [24], which has a long-term impact on individuals [20]. West [58] illustrated that greeting card communication reflects the highly gendered division in the U.S. culture. Few prior research has been focused on analyzing greeting card messages due to limited data. Murphy [24] analyzed 180 Hallmark children's cards in a real store and found out that girls and boys are perceived as being different in terms of interests, activity levels, and characteristics. They further identified children's greeting cards as an arena of clear gender-associated messages. To facilitate further exploration, we collect a large dataset from popular websites of greeting message suggestion websites along with AI-generated greeting messages. We analyze the dataset quantitatively to discover and understand the gender association of topics in greeting card messages. Although the topic of greeting cards and gender has been visited in a series of gender research literature [17], to our best knowledge, we are the first to analyze this issue quantitatively using statistical NLP tools.

Quantify Gender Bias. Previous research has explored how to measure gender bias. Researchers [19, 61] have been widely using Word Embedding Association Test (WEAT) scores [5] to quantify gender bias. The WEAT score links bias in word embeddings to human bias. It compares two sets of target words (e.g., art and science words) and a pair of opposing attribute words (e.g., female and male names). Besides, the WEAT score measures the association strength between the target words group and the attribute word group using vector similarities in the word embedding. In this paper, we use the WEAT score to qualify gender bias in greeting messages.

Text Generation with Artificial Intelligence (AI). With the astounding growth of AI, people have been using it for dialogue systems [49], summarization [25], story generation [44], etc. GPT-2, openAI's publicly available language model, and BERT [8] are the state-of-the-art AI models for text generation. Basta et al. [1] find that GPT-2 performs better than BERT. In our work, we use GPT-2 to generate greeting card messages given some prompts as input, reflecting real greetings it has trained on. We found that GPT-2 generated greeting card messages also have similar gender associations to the template dataset.

3 ANALYSIS

We aim to explore the following two research questions:

RQ1 *Would greeting card messages to females lean towards their beauties and household, and to males lean towards their work achievement?* Prior work [24] has shown that gender signals exist in greeting card messages and make the rejection of gender stereotypes difficult. We are interested in understanding whether such distinctions hold across various messages from multiple scenarios.

RQ2 *If RQ1 stands, would the gender association of topics in greeting card messages relate to recipients' age?* Murphy [24] focused on messages whose recipients are children; However, messages sent to adults are equally important, as the accumulation of gender signals makes the stereotype difficult to overcome and has long-lasting effects in various aspects including career, mental health, etc [20]. We are interested in understanding whether the distinctions are stably maintained across age groups, or whether it diminishes or strengthens.

To answer the two questions, we collect and analyze messages from three data sources. First, we collect a large-scale greeting card message corpus under six scenarios using templates from greeting message suggestion webpages (we refer to it as *Template* Dataset). As these examples demonstrating "ideal" sample messages, we believe they reflect people's expectations on greeting message writing. Second, we also generate artificial greeting messages from large natural language models (referred to as *GPT-2* Dataset). These models are usually trained on large text corpus crawled from the Internet. Therefore, their generation– if high quality– can reflect real messages users write and post. Third, we manually collect 500 tweets from real users on Twitter to analyze how users write greeting card messages naturally in the real world. Below, we first introduce our data collection strategy and then illustrate the analysis process.

3.1 Data

Template Dataset. We crawled eight websites [29, 30, 32, 33, 35, 36, 38, 39]. Our collected greeting card messages are all publicly available. We only use collected data for personal research purposes, which is compliant with all platforms' terms of service. Among them, Hallmark and AmericanGreetings are the two largest greeting card producers globally, and the greeting messages released on their web pages cover the widely-held social values regarding various topics [24]. We use other websites as supplements that enlarge the variety of our datasets. In total, we collected 18,559 messages of 6 common scenarios across all websites, indicating the popularity of these scenarios, among which we choose 3 for our analysis. We further split these messages based on their gender association. We assume that the association is implied by the recipient's gender mentioned in the message: If a message mentions an indicator in the general female group in Table 2 (or variations of `mother` or `grandmother`), we categorize it as a *female-associated message*. We mark the messages with unknown recipients as *neutral* and will be using them to study how people write greeting card messages to recipients without gender-specific indicators (e.g., "sincere birthday messages" in HallMark and "for Coworkers" in Americangreetings). The resulting distribution is in Table 3. Note that we do not consider or differentiate greeting card messages

Group	Gender Indicators
General `female`	daughter, hers, lady, grandma, grandmother, female, aunt, wife, sis, niece, mother, she, girl, her, granny, granddaughter, girlfriend, woman, mom, sister
General `male`	dude, godfather, grandson, stepbrother, boy, sir, he, uncle, man, male, soninlaw, boyfriend, brother, grandpa, him, nephew, son, papa, exboyfriend, granddad, husband, stepson, dad, fatherinlaw, daddy, stepdad, father, grandfather, bro, his
`Mother` variations	mother, mom, mama, mommy, mum, mumsy, mamacita, ma, mam, mammy
`Father` variations	father, dad, dada, daddy, baba, papa, pappa, papasita, pa, pap, pop
`Grandmother` variations	grandmother, grandma, grandmom, grandmama, grama, granny, gran, nanny, nan, mammaw, meemaw, grammy
`Grandfather` variations	grandfather, grandpa, gramp, gramps, grampa, grandpap, granda, grampy, granddad, grandad, granddaddy, grandpappy, pop, pap, pappy, pawpaw

Table 2: We used a list of recipient indicators (i.e., "attributes" [5]) to 1) prompt the message generation in GPT-2 dataset collection (Section 3.1), and 2) distinguish female- and male- associated messages in the template dataset. The General `female` and `male` indicators are from Caliskan et al. [5] and addresses from the template dataset, used to generate general wishes. Other variations are terms of endearment from their corresponding Wikipedia pages and used to generate wishes for different age groups.

Scenario	Template Dataset	GPT-2 Dataset
Birthday	13,338 (4138/4208/4992)	40,200 (19,600/20,600/0)
Valentine's Day	2,496 (610/654/1232)	28,200 (13,600/14,600/0)
Wedding	1,360 (121/122/1117)	28,200 (13,600/14,600/0)

Table 3: Statistics showing the number of greeting card messages we collected in the Template and the GPT-2 datasets. The numbers in parentheses indicate the number of messages for female, male, and neutral messages respectively. In the GPT-2 dataset, for each scenario, we generated 200 messages for each recipient described in Table 2.

based on senders' gender, meaning that our collected dataset may include greeting card messages among the same gender.

GPT-2. The rapid development of AI has enabled text generation models to generate fluent content like humans. Therefore, we are also interested in studying whether greeting card messages written by AI will have similar patterns to human-written greeting card messages. In this work, we use a state-of-the-art language model GPT-2 as a black box tool to generate greeting card messages. Following similar approaches from prior work (e.g., Vig et al. [55] studying gender bias), we feed the GPT-2 model with different prompts (i.e., partial keywords or phrases like "Happy birthday mom!"), and collect the continuation generated by the model. The intuition is that, if the model is trained on gender-distinctive messages, prompts

with clear gender indications will trigger it to generate messages that have reflect such distinctions. We configured the model to use top-p [15] ($p = 0.1$) sampling, and constrained the length of the generated sentence to be 200 characters. Such setting maximizes the chance of getting non-repetitive and natural greeting messages based on our experiment. We designed the prompts in the following ways. *For general wishes,* we use two kinds of prompts:

- "`[scenario prefix] [female/male indicator]!`", with the `scenario prefix` being appropriate wish sentence for one of the six scenarios from Table 3 (e.g., "Happy birthday" for `birthday`, "Congratulations on getting married" for `wedding`), and the female/male attributes and variations from Table 2;
- "`[scenario prefix] [female/male name]!`", to cope with the fact that people often refer to their loved ones by name. We adopted common female/male names from Ribeiro et al. [43].

We further generate messages for babies, parents *and* grandparents to check if the gender association of topics varies for different age groups.

- For *babies*, we use "`[scenario prefix] my little baby [girl/boy] [female/male name]!`" as prompts to generate birthday messages.
- For *parents* and *grandparents*, we use "`[scenario prefix] [corresponding terms of endearment]!`" as prompts. We show terms of endearments for parents and grandparents age groups in Table 2.

For each scenario and gender indicator, we generated two hundred messages. As shown in Table 3, GPT-2 generates sufficient messages so that we can generate enough data for all age groups. Note that we do not generate greeting messages under *Valentine's Day* and *wedding* scenarios for baby age group, which leads to the statistical difference in Table 3.

Twitter Data. Template dataset represents what people regard as ideal messages, and *GPT-2* generated greeting card messages represent how model would write after seeing abundant online text. But either case might not represent well how people actually write greeting card messages in the real life, which is usually private and not accessible. Although people have been using social media to greet others, it is hard to acquire recipients' gender automatically because of the difficulty of extracting the recipient from the Tweet content and getting gender identity information from another linked account (i.e., recipient account). To address this issue, we design three criteria to collect Tweets of birthday greetings where we can identify recipients' gender: 1) the content of Tweets have to be birthday greetings to another recipient; 2) senders need to explicitly mention the recipient (by using @) in their posts, and 3) recipients need to self-identify their gender by explicit means (e.g., mention gender pronouns) in their public profiles. In our collection process, we search for keyword *"birthday"* and put time range from *2019/01/01* to *2021/03/01.* To ensure the high data quality, three of coauthors go to Twitter and manually collected 500 tweets about birthday greetings that satisfy both our criteria, so that we can acquire recipients' gender identities. Among 500 tweets we collected, 263 are for women and 236 are men.

3.2 Analysis Methodology

In this section, we illustrate our analysis pipeline in Figure 1 under the binary gender setting. It is worth mentioning that our pipeline is generic, and can be easily adapted to analyze any other gender groups by replacing female recipients and male recipients, e.g., recipients with gender indicators and recipients without specific gender indicators. Our analysis pipeline is as follows:

First, we modify and apply Empath [11] to extract topics from greeting card messages. Then, we apply the odds ratio [53] to extract gender distinct topics. Finally, we quantify and verify the association of topics and gender with the WEAT score [5]. We describe the detailed analysis steps as below.

3.2.1 Step 1: Topic modeling.

While counting word frequencies might be most intuitive for distinguishing female- and male- associated messages, word-level analysis can be too diverging and coarse to reveal higher-level themes of the messages. Instead, we choose to associate topics with messages and analyze greeting card messages from the topic level.

To model the topics, we first tried Latent Dirichlet Allocation (LDA) [2], a classical generative statistical model that extracts topics in an unsupervised manner. However, Figure 2 clearly shows that LDA tends to deliver repetitive topics. Moreover, with the topics being exact keywords extracted from the documents, they tend to be ambiguous and incomprehensible. We instead applied Empath [11] to mitigate these problems. Built on top of neural embeddings and crowdsourcing, Empath provides mappings between 60k common tokens and 200 high-level topics; Accordingly, we can extract its topic by counting the keyword occurrence in a message. For example, keywords *house, clean, tidy, family* point to the topic **domestic work** in Figure 2). We conduct a small-scale qualitative evaluation of Empath versus LDA on 30 greeting card messages randomly sampled from the template dataset. First, we run Empath and LDA separately for every greeting card message, and get the first 5 topics together with all words in each topic from the model output. Then, we hide the model name and ask two coauthors to qualitatively choose which topic model outputs better topics with corresponding words under each topic. Two coauthors unanimously choose Empath for all 30 messages, which qualitatively shows the superiority of Empath over LDA.

Since the associated keywords in Empath are taken out-of-context.[2] directly counting the occurrence of these words does not always yield reasonable topics. For example, in Table 4, while keywords *"Enlighten", "wisdom", "intellectual", "religious"* correctly reflect the *"competing"* topic, it seems unreasonable to highlight the topic *"crime"*, just because a single keyword, *"witness"* or *"killing"*, is used as metaphors. To discourage outliers, we filter the topics for each message based on the diversity and frequency of their corresponding keywords: We only keep a topic for a message, if 1) more than 5 of its unique keywords under the specific topic occurring in the message, and 2) the average occurrence frequency of each keyword is more than 3. We refer to the Empath with filtering as

Topic	Message
Competing	You have enlightened our world with your love and smile... You will continue to rise in wisdom and all the good things of life will come to you... You're filled with wisdom and special activities of yours always revealed this... You process both physical and intellectual qualities...
Crime	... We are highly delighted to witness today... ... You're renewed by the privilege given to witness another year... I already rescheduled all my appointment for today to enable me to witness your birthday party... Your cuteness is the most killing weapon... ... Your smile is my killing weapon...

Table 4: Examples of suggested top topics generated from Empath [11] and original messages, where we highlight words that Empath catches as related to the suggested topic. Most topics extracted and suggested by Empath correctly reflect gist from greeting card messages (*e.g., "competing"*), but can also be misled by the frequent occurrence of some keywords and extracts unwanted topics (*e.g., "crime"*). We propose Empath* to alleviate the impact of having frequent outlier keywords on extracting topics.

Empath*. Note that we choose the thresholds based on running multiple experiments and manually check if there are strange topics as outliers from the output, as there is no ground truth in such an unsupervised setting. Although the empirical numbers may not apply to other scenarios, we advocate researchers to consider using *number of unique keywords under one topic* and *average occurrence frequency of each keyword* as filters adapting to the new setting.

By applying Empath* to female and male greeting card messages separately, we get two topic dictionaries, with both topic names as keys and their corresponding keyword occurrence frequencies. As in Figure 1, *"celebration"* is the most frequent topic in messages for both men and women.

3.2.2 Step 2: Use the odds ratio to extract gender distinct topics.

With the topic list, we apply odds ratio (OR) [53] to understand if the topics indeed associate with recipients' genders. Bornmann et al. [3] use OR to analyze the gender differences in grant award procedures. In our case, OR quantifies the strength of gender association between the two topic lists. Intuitively, OR equal to one means that the odds of having one topic in messages to women is the same as in those to men. If OR is greater than one, the topic is more likely to occur in messages to males, and vice versa. Sorting the topics by its computed OR gives us top topics associated with messages to male or female recipients. To focus our analysis on the most distinctive topics, we filter topics that are under 30% quantile and keep the same number of topics associated with female/male recipients. Note that we decide the threshold by running multiple sets of experiments after checking the output topics manually. We also calculate the odds ratio gap between the most masculine

---

[2]The list of topic-keywords mapping from existing work Fast et al. [11] can be found here https://github.com/Ejhfast/empath-client/blob/master/empath/data/categories. tsv. Although Empath demonstrates its efficiency in extracting topics, the out-of-context mapping will sometimes lead to unsatisfying topics. We will introduce a critical evaluation of Empath as a limitation in Section 7.2.
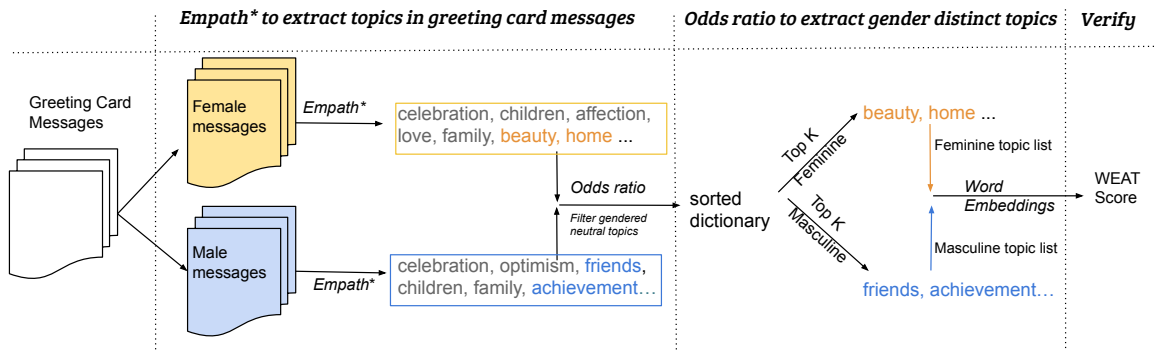
Figure 1: Analysis pipeline. We first apply a modified version of Empath [11] (Empath*) to extract topics from greeting card messages. We then apply the odds ratio [53] to extract gender distinct topics. Finally, we quantify and verify the association of topics and gender with WEAT scores [5].
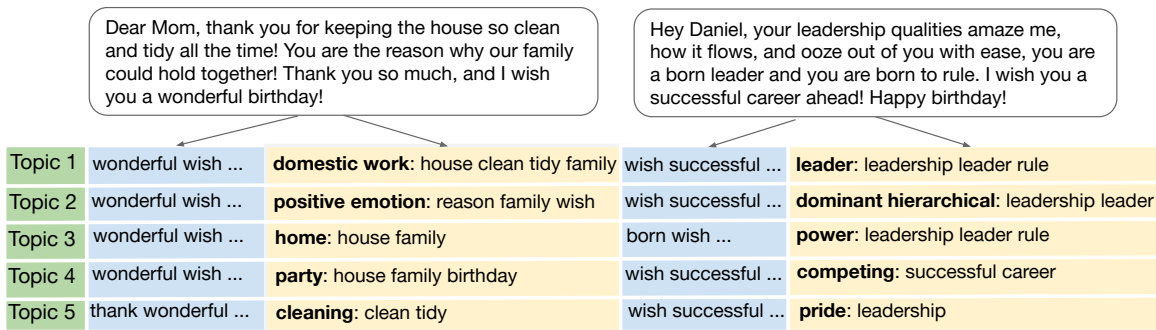

Figure 2: Topics generated by LDA [2] (in blue) severely suffer from repetition and ambiguity issues. While Empath [11] (in yellow), used in our pipeline, could accurately generate suggested topics and corresponding words falling into each topic.

and most feminine words to quantify the polarity of gender role in greeting card messages.

3.2.3 Step 3: Calculate WEAT scores to confirm the distinction.
Are the topics we pick indeed associated recipients' genders? We further validate whether the topics we select are associated with gender attributes in popular neural word embeddings like GloVe [40]. We quantify this with the Word Embedding Association Test (WEAT) [5], a popular method for measuring biases in word embeddings. Intuitively, WEAT takes a list of tokens that represent a concept (in our case, *keywords* for each *topic*) and verifies whether these tokens have a shorter distance towards female attributes or male attributes (in our case, the *indicators* from Table 2). We calculate the score on three versions of pretrained GloVe embeddings, including Google News, and Wikipedia, and Gigaword [26]. To better represent words in our corpus, we also fine-tune GloVe on our corpus with Mitten [9]. We refer to the fine-tuned GloVe embedding on our corpus as GloVe*.

3.3 Analysis Result
Table 5-7 show the top five masculine and feminine topic lists for the birthday, Valentines' day, and wedding scenario. In the tables, we use the dark-yellow dots ● to denote that the most frequent topics (frequency among top 33%) in female messages. We use light-yellow dots ● to denote less frequent topics (from top 33% to top 66%) and grey dots ● for the least frequent topics (after top 66%) in the female messages. Similarly, we use the blue color coding for male messages (● : among top 33%, ● : from top 33% to top 66% and ● : after top 66%). The *Gap* column represents the polarity of odds ratios between the most masculine and the most feminine topics discussed in Section 3.2. Besides the entire dataset (denoted as "all"), we also split the analysis for different age groups (denoted as "babies", "parents", and "grand".) We also use <u>underline</u> to mark the topics in age groups (*e.g.,* feminine in T-babies) if they are repeating the extracted topics in the full corpus (*e.g.,* feminine in T-all).

People write more about appearances for women while more about careers for men, as shown in the first row of Table 5). We clearly see *"feminine"*, *"attractive"*, and *"beauty"* appearing in all the feminine topic list. This suggests that people tend to send wishes about appearance regardless of the scenario. Interestingly, GPT-2 dataset also displays a tendency of generating *"domestic work"* related messages for women. In contrast, the topics for men are relatively more diverse. Still, we observe *"leader"* tend to occur frequently across scenarios (T-all and G-all in Table 5, G-all in Table 6 and 7). The color of the dots also suggests that these topics frequently occur in the messages. For example, there are 246

Group	Feminine topics	Masculine topics	Gap
T-all	● feminine, ● appearance, ● attractive, ● aggression, ● beauty	● masculine, ● leader, ● work, ● philosophy, ● social media	3.50
T-babies	● feminine, ● childish, ● affection, ● shape/size, ● friends	● achievement, ● home, ● cold, ● ancient, ● violence	1.46
T-parents	● feminine, ● appearance, ● domestic work, ● beauty, ● confusion	● masculine, ● ancient, ● heroic, ● business, ● fun	4.19
T-grand	● love, ● affection, ● home, ● family, ● ancient	● healing, ● celebration, ● optimism, ● friends, ● rural	**0.94**
G-all	● feminine, ● attractive, ● appearance, ● childish, ● white collar job	● masculine, ● leader, ● play, ● wedding, ● driving	3.13
G-babies	● feminine, ● childish, ● royalty, ● ocean, ● magic	● play, ● pet, ● technology, ● animal, ● hipster	9.05
G-parents	● domestic work, ● phone, ● dispute, ● home, ● family	● leader, ● tourism, ● weather, ● restaurant, ● art	3.59
G-grand	● occupation, ● domestic work ● white collar job, ● feminine, ● hygiene	● pet, ● masculine, ● warmth, ● eating, ● fun	**1.34**
Tweets	● beauty, ● affection, ● body, ● fear, ● friends	● zest, ● joy, ● anticipation, ● pride, ● contentment	1.47

Table 5: Feminine and Masculine Topics for the Birthday Scenario. "T" refers to the template dataset and "G" refers to the GPT-2 generated dataset. We use yellow dots ● to denote the most frequent topics (among top 33%), light-yellow dots ● to denote less frequent topics (from top 33% to top 66%) and grey dots ● for the least frequent topics (after top 66%) in the female messages. Similarly, we use the blue color coding for male messages (●: among top 33%, ●: from top 33% to top 66% and ●: after top 66%). We also calculate the odds ratio gap between the most masculine and most feminine words shown in *Gap* column. The higher *Gap* scores are, the more polarized greeting card messages are for women and men on discussed topics.

Group	Feminine topics	Masculine topics	Gap
T-all	● feminine, ● attractive, ● beauty, ● affection, ● friends	● nervousness, ● cold, ● pain, ● body, ● youth	1.53
G-all	● wedding, ● family, ● children, ● pride, ● worship	● leader, ● play, ● eating, ● cooking, ● restaurant	2.98
G-parents	● domestic work, ● furniture, ● family, ● home, ● hipster	● fire, ● art, ● emotional, ● writing, ● real estate	1.32
G-grand	● occupation, ● domestic work, ● dance, ● messaging, ● furniture	● masculine, ● tourism, ● sports, ● art, ● zest	**1.20**

Table 6: Feminine and Masculine Topics for the Valentine's Day Scenario.

messages for women while only 112 for men about the *"appearance"*, and there are 47 messages about the leadership for women while 134 for men. Similarly, we can also see that people talk more about appearance (e.g., beauty and body) to women in tweets.

In addition, we analyze topics in greeting card messages to recipients without specific gender indicators in the birthday scenario with our analysis pipeline, by replacing greeting card messages to female recipients and male recipients in the pipeline with the ones to recipients with and without specific gender indicators. We utilize messages in the template dataset with no gender indicators to conduct the study, denoted as "neutral" in Table 3. On the other

hand, we sample the same number of messages for recipients with specific gender indicators (i.e., messages to female and male recipients combined). The top 5 distinct topics for recipients with specific gender indicators (i.e., binary people) over ones without specific gender indicators are *"home"*, *"royalty"*, *"family"*, *"domestic work"*, *"masculine"*, and the top 5 topics for recipients without specific gender indicators compared to binary people are *"work"* (e.g., we pray for numerous years of meaning and *accomplishments* for you.), "business" (e.g., If you are half as productive as you are in the *office*, your life is going to be awesome.), "meeting" (e.g., happy birthday to one of the greatest people I have ever had the pleasure

Group	Feminine topics	Masculine topics	Gap
T-all	● sadness, ● feminine, ● attractive, ● beauty, ● children	● wedding, ● childish, ● celebration, ● positive emotion, cheerfulness	0.85
G-all	● feminine, ● fashion, ● childish, ● appearance, ● fabric	● leader, ● play, ● wedding, ● valuable, ● computer	6.41
G-parents	● domestic work, ● home, ● sports, ● music, ● pride	● leader, ● hipster, ● sexual, ● fashion, ● swearing terms	3.49
G-grand	● occupation, ● domestic work, ● fabric, ● clothing, ● medical emergence	● journalism, ● cold, ● music, ● sports, ● musical	**1.17**

Table 7: Feminine and Masculine Topics for the wedding scenario.

of *meeting*.), "law" (e.g., your *discipline* stands amongst the most needed professionals across all niche), "health" (e.g., we wish you good *health*.), with the *Gap* between most distinctive topics for two groups is 2.95. The topics difference indicates that when people write greeting card messages to binary people, they tend to write more stereotypical topics compared to when they write to recipients without specific gender indicators. From the top 5 distinct topics we extract for recipients without specific gender indicators, we can see that most topics are work-orientated and more similar to what we extracted for men, compared to appearance and domestic work-related topics for women.

Greetings generated by GPT-2 amplifies gender stereotypes. The gap for `G-all` is either higher than or comparable to that for `T-all`. More specifically, while the differences between the two scores in *Birthday* is trivial, those in *Valentine's day* and *Wedding* double or even becomes six times larger.

The gender association differs across different age groups. The gender distinction is the smallest when people write to the elderly. While all the non-grandparents age groups have repeating topics from the complete corpus (i.e., "-all" rows), the topic overlaps greatly diminishes in the grandparent age group. For example, *"healing"*, as the most masculine topic, appears only 3 times for male recipients more than it is for female recipients in the elderly group. It means that the gender-associated topics for grandparents are not as strong as for other age groups. We further checked if there are potential semantic noises in our text prompt. For instance, our text prompt "`[scenario prefix] my little baby [girl/boy] [female/male name]!`" might not refer to literal babies and can be used for one's adult children or one's boyfriend or girlfriend. We random sampled 50 GPT-2 generated greeting card messages for baby boy and baby girl separately. Then, we manually evaluate 1) if the generated messages are referring to literal babies or not and 2) if the generated messages are fluent and natural. Among 100 generated messages, there are 13 invalid ones (6 for baby girls and 7 for baby boys). The high validity rate (i.e., 87%) shows few semantic noises in our text prompt and generated greeting card messages are of high quality.

While we were not able to split out or generalize baby-related messages in the other two scenarios, the `grand` group has the lowest gap score among all age groups, which also holds for the `Valentine's Day` scenario (Table 6) and the `Wedding` scenario

	GloVe	GloVe Google News	GloVe*
Birthday	0.932	0.958	0.624
Valentine's Day	1.104	0.903	0.598
Wedding	1.105	0.969	0.768

Table 8: The WEAT scores of feminine/masculine topic lists for different events in the template dataset based on pretrained GloVe [40] embedding on different data sources ("GloVe": original Glove word embedding; "GloVe Google News": pretrained GloVe embedding trained on Google News dataset; "GloVe*": GloVe embedding fine-tuned on template dataset).

(Table 7). It again shows that the gender distinction is smallest when people write to the elderly compared to the other groups.

The WEAT score verifies that the topics we found have a clear gender association in word embeddings. The WEAT score is in the range of −2 to 2. A high positive score indicates that detected feminine topics are more associated with female attributes in the current embedding space. A high negative score means that detected feminine topics are more associated with male attributes. As shown in Table 8, we find that WEAT scores across all scenarios and embeddings are all positive, which indicates and verifies the positive association between gender attributes and selected topics.

4 GREETA SYSTEM

4.1 Pilot Survey

Our analysis shows that greeting card messages to women lean towards appearance and household while towards career and leadership to men. AI also exhibits the same characteristics when generating greeting messages. However, whether people are aware of the gender role when greeting and care about having the gender role in their greeting card messages is under-investigated. Here, we want to answer the research questions RQ3 raised in Section 1: (*Do people want to be informed of the gender role in their greeting card messages?*) and RQ4 (*If RQ3 stands, how can we help users write greeting card messages? What features do they expect in such a tool?*).

To investigate people's attitudes towards gender roles in greeting card messages, we designed and distributed a pilot survey. We recruited twenty participants (gender: 6 female, 12 male and two prefer not to self-identify; age: span from 10-20 to 40-50. ethical

background: 11 are Asian, 2 are Latino or Hispanic and 6 are White or Caucasian) by posting volunteer recruitment information on co-authors' social platforms. All recruited participants did not know any content of the project, including any prior findings in our project or our research questions. To set up the context and get a chance to digest the actual messages written by participants — we first asked them to choose a loved one (or to others) and write a short birthday greeting message. Afterward, we collected their opinions on showing the gender role in greeting card messages. In specific, we asked two questions: (1) whether they would be concerned if their message(s) show a strong gender role, and (2) what features they would expect if they were using a greeting writing tool that informs them about masculine or feminine languages in their written message. We come to some **C**onclusions derived from our pilot survey:

C1: Users would like to be informed about the gender role. Some users (14 out of 20) are curious about the gender role in their messages, , *e.g.* ., "*I would be curious to know if what I am saying in a greeting does have a strong gender role. I am not sure how the tool would work beyond traditional binary stereotypes, though. I would be concerned about how objective you can make a tool that measures gender orientation in language*" (*P*16). In addition, users prefer to know the gender role to avoid offending the receivers of their messages, *e.g.*, "*I would like to be informed to ensure I'm not offending anyone in any sense. Moreover, it will help me understand when I can avoid these remarks*" (*P*3).

C2: However, most users will not be concerned about the gender role. Some users think greetings are personal and they know their addresses' pronouns, *e.g.*, "*I would say my style of address to someone depends on my individual experiences with that person, rather than any potential stereotypes.*" (*P*1). They also stated that the inclusion of the gender role is usually intended, *e.g.*, "*I would be concerned. My sister does not necessarily fit stereotypes, and my birthday messages are pretty similar across gender. Adding gender-specific language would likely be different from my usual style*" (*P*5). Statistically, 90% users (18 out of 20) expressed that they would not be concerned about gender roles in their greeting card messages.

C3: Users would like to get some suggestions, but are against the system changing things for them. Users do not want to be intervened with modifying their greeting messages. For example, *P*4 responded to the second question, saying that "*I'd be careful about adding a feature that suggests an alternative message because sometimes the person for whom the greeting is meant does not follow traditional gender roles.*" *P*3 and *P*16 further emphasized that they do not want to be criticized by the machine. However, they are not against the idea of getting suggestions in two aspects: (1) To diversify their message. *P*8 mentioned "*I'd love to get some new ideas on what I can include in a greeting message.*"; and (2) to know what contributes to their manifested gender role, o that they can make informed decisions on what to keep or change their messages.

4.2 Design Requirements

Based on Section 4.1, we find that users would like to have an assistant tool to help users write greeting card messages with gender role awareness. Such a tool should satisfy the following requirements:

*R*1: **Bring the gender role awareness to greeting card messages.** More specifically, we need to provide an overall gender perception of a continuous scale (based on users' requirements of going beyond binary gender type) and highlight writings that contribute to the point.

*R*2: **Avoid explicit word attribution and neutralization.** People do not want to be judged, and the appearance of gender roles in greeting card messages is expected.

*R*3: **Associate gender orientation with topics in messages for both analysis and suggestions.** Users want to understand the connection between gender perception and content. They prefer the system to suggest changes but make the modifications themselves.

After digesting users' requirements, we design an interactive analysis tool, GreetA, to increase users' awareness without intervening in their writings. *The target of GreetA is to bring gender role awareness to users when writing greeting card messages.* As *P*17 mentioned during the pilot survey, "You can always just ignore the tool if you want. It is just a warning anyways". We want to emphasize that there is nothing wrong with having gender roles in greeting card messages. Our design is to increase gender role awareness and help users better understand gender roles in their messages to prevent unconscious gender stereotypes. Users always have full control over their messages and can decide whether to modify greeting card messages on their wishes.

4.3 GreetA

Figure 3 shows an overview of GreetA. Users can choose whom they are writing to and write greeting messages in panel *a*. We use the blue and yellow coding in GreetA to be colorblind-friendly. The color-coding is consistent in the system. Blue, yellow and grey represent more masculine, more feminine and neutral respectively.

Message input panel a and overall gender perception score panel b. Users can write greeting messages in panel *a*. After a user clicks on the analyze text button, GreetA would automatically analyze the user's message and show the analysis result. We provide an overall gender perception score (*R*1), and the lengths of blue and yellow circle fragments represent how much the message leans towards being masculine or feminine. The score inside fragments indicates how much the message leans towards being feminine.

Topic analysis panel c. We provide top topics that the input message is associated with from the Empath output and sort them based on their occurrence frequencies (*R*3) in panel *c*. Furthermore, we also list all words falling into each topic to the right of the topic name. In the weight column, the weight bar's length indicates the occurrence frequency of the corresponding topics. When the user clicks on the checkbox in front of each topic, we will highlight the corresponding words related to that topic in the original message with the same color-coding.

Topic exploration panel d. In panel *d*, We randomly sampled some topics from the collected template datasets. Users potentially be interested in adding the topics to the input message. Each displayed topic contains words that fall under the topic based on Empath (e.g., for examples under *competing* topic in Figure 3, there are keywords *achieved, outstanding, winning* in 3 sentences that all fall
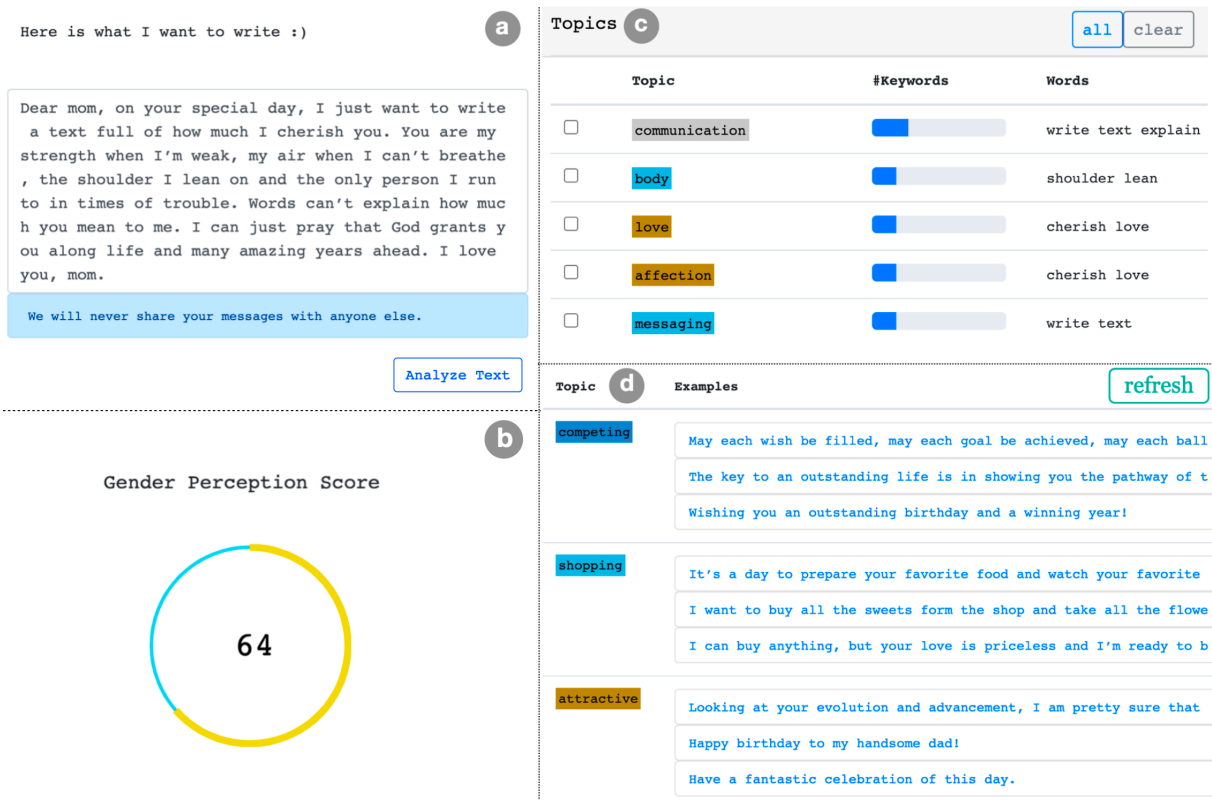
Figure 3: GreetA overview. GreetA contains four panels: message input panel *a*, overall gender perception score panel *b*, topic analysis panel *c*, and topic exploration panel *d*.

under *competing* topic based on Empath). All examples we show are retrieved from the collected template datset. If there are more than 3 examples falling under the same topic, we randomly select 3 of them to display in panel *d*. Users can refresh to get another set of topics and corresponding examples if they find the current examples unsatisfactory and want to explore more, and expand the previews to see the complete example content. We highlight suggested topics with the same color-coding to show the gender role it relates to based on our analysis result.

5 QUALITATIVE USER STUDY

To understand whether GreetA helps people write greeting messages with the gender role awareness, we explore two aspects via a qualitative user study: how much GreetA can help people become more aware of their messages' gender roles (*Q*1), and whether people want to edit their messages after perceiving the awareness (*Q*2). To answer these two questions, we conduct a qualitative user study and show our results in this section. We have taken standard practice consideration with anonymous data collection and annotations. This research study has been approved by the Institutional Review Board (IRB) at the researchers' institution.

Participants and Apparatus. We recruited seven student volunteers from the authors' university (age 21-27, $\mu = 24$, $\theta^2 = 3.5$) with a diverse major background, including acting, chemistry, computer science, education, etc., denoted as *P*1-*P*7. All of them have never heard of the project or seen GreetA before the user study. We make GreetA temporarily publicly available during the study period. Participants used their laptops to conduct the study. It was an online one-on-one study by connecting with participants via video calls. During the study, we encouraged participants to think aloud, share the screen and feel free to stop sharing whenever they type anything personal or do not feel comfortable to share. Please note that we did not save participants' greeting card messages during the study. We also asked for participants' consent if we wanted to quote part of their messages in the paper. All surveys distributed during the study are available online, and every participant could only submit once.

5.1 Procedure

Preparation. Before the study, we asked every user to write birthday greetings consisting of 6-10 sentences to their loved one and recommended them to include specific events.

Prior Survey. We asked participants to fill out the prior survey besides basic demographic information of the user and the pronouns of their loved ones.

Using GreetA. We first played a three-minute introduction video of GreetA to show the users how to interact with. We asked them to play with GreetA for fifteen to thirty minutes. During this process, we encouraged the users to think aloud and voice out (e.g., which features are useful, which features are unnecessary, what obstacles they are encountering, etc).

Evaluation Survey. After playing with GreetA, we asked them to fill out the evaluation survey. The survey contains two parts: 1) To evaluate *Q*1, we first applied the usability metric for user experience (UMUX) [12], a four-item Likert scale, to assess participants' satisfaction with the perceived usability of GreetA. We then added additional Likert scale items to further measure the level of helpfulness participants perceived with our system. The additional items mainly focus on measuring *whether GreetA is useful for writing greeting card messages, easy to learn, easy to use, helps increase the gender role awareness, and whether they are willing to use it in the future.* 2) For *Q*2, we use open questions including *Why they changed/kept (a certain part of) the message? What features make you want to change (not change) your message? What do you like about our tool? What do you dislike about our tool? Will you consider using it in the future and why?*

5.2 Result and Analysis

We show survey responses to usability questions in Figure 4(a) as a diverging stacked bar chart, where we take "neutral" as a baseline so that positive responses are stacked to the right while the negative responses to the left. We describe key insights as follows:

GreetA is easy to learn and use. All users agreed that GreetA is easy to learn and use.

GreetA helps increase the gender role awareness (*Q*1). As shown in Figure 4(a), six out of seven users agreed that GreetA could help increase the gender role awareness. Some users think GreetA could help decrease gender bias in their messages, *e.g.*, "*If I need to write the greetings to someone I am not very familiar with, like collaborators or colleagues or leaders, I would get the suggestions from the tool to make my message not gender-biased.*" (P1). Other users stated that GreetA could avoid offense to message recipients, *e.g.*, "*As I always watch lots of heroic movies, I am worried that I may write some gender-stereotyped phrasings. I need some tools like this to help me before sending my greeting email to someone else so they will not feel offended or comfortable*" (P6).

Users may not be willing to change their greeting card messages. Some users think that if the message recipient is a close friend, they are not concerned about gender bias, *e.g.*, "*since we know each other well and the recipient will not get offended if I use some feminine or masculine words in my wishes.*" Many users also stated that if the messages contain common experience or hobbies that are gender-biased, it does not matter. *P*6 shared her message to her close male friend with us when she noticed that cooking was associated with domestic work, which is very feminine in our analysis result. She mentioned that she would like to keep it as it is one of the most precious memories between her and her friend,

and her friend would not think it was inappropriate. Similarly, P4 showed us part of his message to his female friend "*No wonder why my game was far more advanced than yours on the court. At least you now have an excuse to be blocked and dunked on.*", which GreetA extracted as a very masculine topic "game", but *P*4 would not want to change it as they indeed played basketball a lot together.

People lean to use GreetA in the future (*Q*2). About half of the users either strongly agreed or agreed that they would use GreetA in the future. *P*2 and *P*4 both indicated that they would GreetA to modify their messages since GreetA helped them identify some words related to negative emotions (e.g., sadness). After consideration, they find the wordings inappropriate and decided to change them to be more suitable for celebrations. On the other hand, some users also impressed their disagreement, *e.g.*, "*If the message is for someone close like friends or family members, I will write on my own.*" (P1).

We noticed that *P*5 chose "disagree" with both *GreetA could help increase the gender role awareness* and *willing to use.* Reasons *P*5 gave were that "*I think maybe because the message I wrote was quite neutral, the tool did not point anything out.*" As we aim to bring the gender role awareness to writing instead of providing editing suggestions, we expect that users like *P*5 who have high gender-role awareness when writing would find GreetA less insightful.

Furthermore, we asked users' opinions on which components they think are most useful. One user can select multiple components as the most useful features; then, we calculate the percentage of votes on one specific feature over total votes for all features in Figure 4(b). The result shows that users think *topic analysis* is the most useful function, followed by *topic suggestion* and *overall gender role score.* For example, *P*2 and *P*4 both found their wordings reveal negative emotions (e.g., sadness) in their greeting card messages and decided to modify their messages. *P*1 and *P*3 also pointed out that they took some topics in the topic exploration panel *d* and put them into their messages.

Conclusion of the user study. Although people may not want to change their messages if they know the recipient well, most users found that GreetA helps them increase awareness of the gender role when writing greeting messages, and they are willing to use the system to improve their messages to avoid offending recipients.

6 QUANTITATIVE USER STUDY

Although over 85% of users in qualitative study (Section 5) self-report that GreetA helps increase the gender role awareness, we conduct another quantitative study to examine it by setting up three contrast surveys on Amazon Mechanical Turk (MTurk). Ideally, more and more workers can identify the gender role in the greeting card messages and agree with the ground truth labels with different levels of aids from GreetA. The ground-truth labels are created and agreed upon among 3 co-authors. We will explain the details of experiments set up in the following.

- **Survey 1: Unaided.** Given just two greeting messages without the help from GreetA, we ask MTurk annotators to select the one that is more likely to be sent to female recipients and explain the reason why they think it is more feminine.

(a) Usability
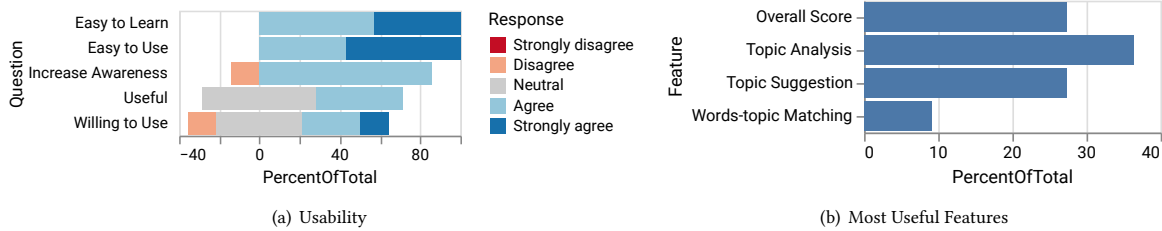

(b) Most Useful Features

Figure 4: The qualitative user study results: including a) diverging stacked bar chart of users' responses on evaluating whether GreetA is useful, easy to learn and use, and increases the gender role awareness. We also asked users if they are willing to use GreetA in the future; b) we collected users' opinions about the most useful features in GreetA. One user can select multiple components as the most useful features; then, we calculate the percentage of votes on one specific feature over total votes for all features.

- **Survey 2: Topics.** In addition to the two greeting messages, we provide the topics and corresponding words for each message extracted by GreetA to MTurk workers.
- **Survey 3: Topics + Overall Score.** In addition to the two greeting messages and the extracted topics, we also provide a Gender Perception Score by GreetA showing how each message leans towards masculine or feminine. The score is higher if the message leans towards the feminine.

6.1 Message Selection Criteria

Each message has an assigned value in the template dataset showing this gender association where greeting messages with values > 50 are more feminine and < 50 are more masculine. So we divided the corpus into 3 classes with values a) > 51 [feminine], b) < 49 [masculine], c) between 49 and 51 [no gender preference]. We then picked samples based on topics. People tend to talk more about business and leadership to male recipients while talking more about appearance and family-related topics when writing to female recipients. Thus, we picked appearance and family-related topics in the feminine class and business and leadership-related topics in the masculine class. Furthermore, we asked 3 co-authors to pair-wisely pick more feminine messages and check whether the results match the result GreetA provides, which we use as the ground truth for later. The gender perception scores from GreetA are all consistent with ground-truth labels created by co-authors. In the end, we picked 20 pairs of greeting messages, including 3 types: female v.s. male, female v.s. neutral and neutral v.s. male.

6.2 Procedure

We recruited 25 high-quality (approval rate: >=98%, HITs: >=1000) MTurk workers located in the United States for each survey. Each MTurk worker in each survey (Unaided, Topic, or Topic + Overall Score) is assigned to read 20 pairs of greeting messages with/without additional information from GreetA and select the more feminine ones and explain the reasons. Each pair took an average of half a minute to finish complete with compensation of $0.15 (20 pairs will take 10 minutes with a compensation of $3), above the US Federal minimum wage. An annotator can participate only once in one survey and will be filtered out as a qualification for the other two surveys. In our instruction, we clarified that our surveys are used to evaluate people's perception of gender distinction in the greeting

messages and describe the information we provide in the surveys along with a sample example.

6.3 Result

GreetA helps to increase gender role awareness in general. We compared results from the three surveys. As shown in Figure 5, the average accuracies across all comparisons for three surveys are 70.1%, 80.0%, and 87.7%, respectively. The result indicates that the correctness of identifying gender perception got improved after introducing GreetA. The *topic+score* condition yielded the highest accuracy, which may not be too surprising, given the summary score essentially reveals the gender association.

GreetA offers useful information for workers to identify gender roles in greeting card messages. To better understand whether workers' rationales differ across conditions, we coded their self-reported explanations for the ranking. Three of the coauthors independently read a subset of the responses to identify emergent codes and created a codebook (Table 9) using a discussion period. Using this codebook, for each survey, we coded a sample of 180 random worker responses (out of 500): 160 were unique, and 20 overlapped between annotators, allowing us to compute inter-annotator agreement and manually annotate what workers rely on to make decisions based on the codebook. We achieved reasonably high agreements (Cohen's kappa score [6] $\kappa = 0.76$). We observed that, as expected, annotators in *Topic* and *Topic+score* relied more on the additional information provided by GreetA. For example, those in the *Topics* condition relied more heavily on the topics (with almost 50% annotators), compared to the *Unaided* group (with 15%), and *Topics+score* displays an increment in `Overall` (which includes the of reliance on the score.) Curiously, the reliance on topics dropped in *Topics+score*; Instead, annotators seem to rely more heavily on keywords in this case. We suspect that annotators would search for keywords that supply the ranking given by the score.

People have a clearer mental model on female associated topics. We further separate the evaluation for different types of questions. Figure 5 shows that although comparisons that involve female-oriented messages yielded reasonable accuracies, the accuracy of male vs. neutral pairs was considerably harder, regardless of the condition. This observation is also aligned with our results in Table 5-7, i.e., masculine topics show more diversity than feminine topics. It might result from 1) people may have a more fixed

Codes	Explanations and Examples	Unaided	Topics	Topics+score
Words	Pick on certain words and phrasings *"everyday hero is more masculine."*	133 (73.9%)	130 (72.2%)	146 (81.1%)
Topics	Based on conveyed topics *"it speaks to the person's nurturing qualities"*	26 (14.4%)	85 (47.2%)	48 (26.7%)
Overall	Perceived the overall gender role *"This sounds like a message to a female sibling"*	29 (16.1%)	22 (12.2%)	52 (28.9%)
Experience	Based on their experience or gender stereotype *"less likely for men to be thanked for cleaning"*	35 (19.4%)	25 (13.9%)	20 (11.1%)

Table 9: The codebook of how annotators choose the more feminine message, with the number of self-reports.
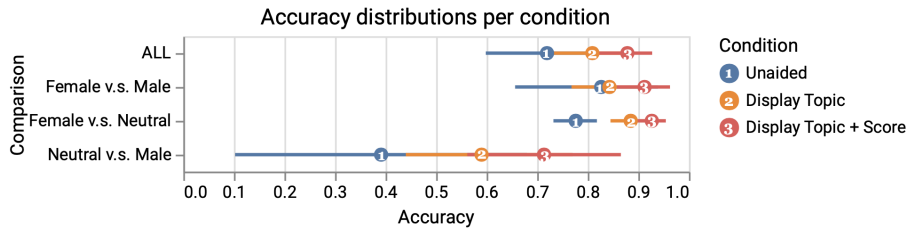

Figure 5: The accuracy distribution of question pairs in 3 surveys, where dots are the average accuracy, and lines crossing the dots represent each condition's confidence interval. The average accuracy across all comparisons for three surveys is consistently improved for all combinations. It indicates the effectiveness of GreetA on improving gender role awareness.

impression on how to greet women, and 2) it is inherently harder to choose a more "feminine" message when there are only "neutral" and "masculine" messages.

Culture factors and experience influence how people perceive gender role [48, 56], which we also observe in our study for greeting card messages. When inspecting explanations, we also noticed interesting contrasts that are likely due to different cultural backgrounds and experiences. For example, for the message "I propose a toast to you, darling friend, and our incredible friendship. You are aging like a fine wine, and it is an honor to be a part of your life." While one annotator perceived "aging like fine wine" as offensive and unwanted for women (and think it should be sent to males), another put "darling and fine wine are more often used by women' girl friends" (and think it should be sent to women). Future research is needed to consider cultural aspects for analysis.

It is important to avoid gender stereotypes in greeting card messages. Explanations that annotators wrote during our quantitative study further prove the importance of preventing gender stereotypes in greeting card messages. About 15% annotators expressed that they identified gender association based on their own experience or gender stereotypes. For example, "*Domestic work and cleaning typically refer to a female's role at home. Women are generally the ones who are expected to do the household chores and take care of the family.*". Besides, about 5% of participants also explicitly criticized the message, , e.g., "*Most women are infantilized due to being in a male-dominated society. Women are not taken seriously as leaders.*"; "*Many women are expected to have so many responsibilities and make sure they look good. Especially women who have children to raise. They are often overworked and do not get much recognition for all of the things they do.*"; "*Most women have to be multitaskers and*

multitalented in ways that are not expected of men. An example is the expectation for a woman to be a good mom, a good wife, never need to take a break, and still find time to keep herself looking attractive. A female recipient should be seen as someone to admire when she can do all things and make it look effortless." This collected feedback further reinforces the importance of our work, and how GreetA could be used to advocate a more fair and diverse environment.

7 DISCUSSION

7.1 Implication

Our work provides a quantitative understanding of the gender roles in greeting card messages by analyzing both the template dataset of greeting messages and the AI-generated (GPT-2) message dataset. Greetings to women lean towards their appearances and household, and messages to men lean towards their career achievement. We design and develop GreetA to help visualize such gender related information, and and provide relevant example greetings as references for users. Both our qualitative and quantitative user studies demonstrated the effectiveness and usability of GreetA. We hope that GreetA could bring more insights into how to avoid potential gender stereotypes when users write greeting template messages. We envision that GreetA can have big impacts on multiple aspects to different stakeholders. Individuals, companies or institutes need to send greetings and wishes to their friends, families, employers or partnership on various scenarios. We hope that GreetA could also help them better polish their messages and avoid potential unconscious stereotypes to advocate and encourage a more fair and diverse environment.

From a generalization perspective, GreetA can be useful in many scenarios. The topic analysis and overall gender perception analysis can be extended to analyze and understand the gender perception in the general text. For example, our system can be adapted to benefit a broad set of gender bias analysis tasks from Wikipedia pages, blogs, Twitter, etc. Researchers who work on the natural language generation tasks can also use GreetA to check if their results may include potential gender stereotypes. It will contribute to preventing unwanted bias before we put AI into real-world production. Moreover, our analysis pipeline and GreetA are general and can be applied in other applications beyond checking greeting card messages, such as game design, helping and raising awareness about imposter syndrome and checking algorithmic bias.

Note that GreetA might be used for malicious purposes. We have the topic exploration feature in GreetA to assist users in coming up with potential topics to talk about in their greeting card messages. The source is the template dataset that we collected from the greeting card websites. As our analysis shows, the greeting card messages from the collected corpus may contain potential stereotypes. Thus, GreetA might serve as a tool that highlights stereotyped messages if users consistently refresh topics and pick messages with potential gender stereotypes on purpose.

Our analysis shows that GPT-2 generated messages amplify the bias in human-written messages. People should be more careful when considering AI in real-world applications. AI researchers should also put more effort into mitigating the bias from different levels (*e.g.*, corpus level for our case) and avoiding biased models.

7.2 Limitation and Future work

This work is subject to several limitations. First, we applied Empath [11] to analyze topics; although it covers various topics, its measures mainly rely on individual lexicons and do not consider the contextual information. For example, Empath would extract topic *weather* for message *"You are a cool person"* based on linguistic information. However, here *"cool"* indicates the personality rather than the weather. Besides, Empath is bad at dealing with figurative speeches. For example, Empath would extract topic *body* for message *"You are the shoulder I lean on"*, but here shoulder is a figurative saying of emotional support. To quantitatively assess the performance of Empath, we randomly select 50 greeting card messages, put selected messages to GreetA, and we evaluate how many among top 5 topics Empath suggested are natural and correct to humans. As a result, 228 out of 250 (i.e., 91.2%) word-topic suggestions are correct. It again shows the efficiency of using Empath in our task, but also shows the room for improvement. Future work can utilize human annotation or other topic extraction techniques to identify fine-grained topics and better deal with figurative speech.

Secondly, we mainly collect and use the template dataset that people perceive as "ideal" messages to analyze how people write greeting card messages. They might be different from what people write in the real world, as they might be more formal and contain the less personal experience. However, the real greeting card messages are hard to get as they are inherently private, and people may not want to share with others. Although we collect 500 real greeting card messages from Twitter, how large these messages can represent is unsure. Future work that collects and utilizes a larger-scale of greeting card messages in social platforms (e.g., Twitter and Reddit) could be complementary to our work. It is also important to understand and analyze greeting card messages to non-binary people, missing from our work because of the scarce source available online. Instead, our work utilizes collected greeting card messages with no specific gender indicators on the birthday scenario and understands the topic difference between birthday greeting messages for binary people and recipients without specific gender indicators. Future research can build upon our work, incorporate more experiments on analyzing greeting card messages to non-binary people across various scenarios and provide assistance when greeting non-binary people. We hope that our work can contribute to building a more inclusive and diverse environment.

8 CONCLUSION

In this work, we collected a large-scale greeting card messages corpus from eight popular greeting suggestion websites and generated an artificial dataset using the natural language model GPT-2. Via a set of thorough analyses, we found that the topics in greeting card messages are indeed gender oriented and greetings to women lean towards their appearance, and men lean towards their career achievement. Our pilot survey showed that people want to be aware of gender role in their greeting messages. In response, we designed and developed GreetA based on a deep understanding of users' requirements. GreetA visualizes the gender orientation in greeting messages and their topical aspects, as well as recommends potential topics for writing better greetings. The qualitative and quantitative studies showed that GreetA effectively brings the gender role awareness to users. Our work is the first to quantitatively analyze the gender role using statistical NLP tools in greeting messages. GreetA is also the first interactive visualization system that helps bring gender role awareness to people when writing greeting messages.

ACKNOWLEDGEMENT

REFERENCES

[1] C. Basta, M. R. Costa-jussà, and Noe Casas. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. *ArXiv* abs/1904.08783 (2019).
[2] D. Blei, A. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022.
[3] L. Bornmann, R. Mutz, and H. Daniel. 2007. Gender differences in grant peer review: A meta-analysis. *J. Informetrics* 1 (2007), 226–238.
[4] Sabrina Burtscher and Katta Spiel. 2020. "But where would I even start?": developing (gender) sensitivity in HCI research and practice. *Proceedings of the Conference on Mensch und Computer* (2020).
[5] A. Caliskan, J. Bryson, and A. Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356 (2017), 183 – 186.
[6] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
[7] J. Cryan, Shiliang Tang, Xinyi Zhang, M. Metzger, Haitao Zheng, and B. Y. Zhao. 2020. Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).
[8] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[9] Nicholas Dingwall and Christopher Potts. 2018. Mittens: An Extension of GloVe for Learning Domain-Specialized Representations. *ArXiv* abs/1803.09901 (2018).

[10] Patrick Marcel Joseph Dubois, Mahya Maftouni, Parmit K. Chilana, Joanna Mc-Grenere, and Andrea Bunt. 2020. Gender Differences in Graphic Design Q&As: How Community and Site Characteristics Contribute to Gender Gaps in Answering Questions. *Proceedings of the ACM on Human-Computer Interaction* 4 (2020).

[11] Ethan Fast, Binbin Chen, and M. Bernstein. 2016. Empath: Understanding Topic Signals in Large-Scale Text. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016).

[12] Kraig Finstad. 2010. The Usability Metric for User Experience. *Interacting with Computers* 22 (09 2010), 323–327. https://doi.org/10.1016/j.intcom.2010.04.004

[13] Eureka Foong, Nicholas Vincent, Brent Hecht, and Elizabeth M Gerber. 2018. Women (still) ask for less: Gender differences in hourly rate in an online labor marketplace. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–21.

[14] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1914–1933. https://doi.org/10.1145/2998181.2998327

[15] Ari Holtzman, Jan Buys, M. Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. *ArXiv* abs/1904.09751 (2020).

[16] Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *ACL*.

[17] Genevieve Amaris Keith. 2009. Hailing gender: The rhetorical action of greeting cards.

[18] Anne M. Koenig. 2018. Comparing Prescriptive and Descriptive Gender Stereotypes About Children, Adults, and the Elderly. *Frontiers in Psychology* 9 (2018), 1086. https://doi.org/10.3389/fpsyg.2018.01086

[19] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. *ArXiv* abs/1903.10561 (2019).

[20] Cade McCall and Nilanjana Dasgupta. 2007. The malleability of men's gender self-concept. *Self and Identity* 6 (2007), 173 – 188.

[21] Shannon C. McGregor, Regina G. Lawrence, and Arielle Cardona. 2017. Personalization, gender, and social media: gubernatorial candidates' social media strategies. *Information, Communication & Society* 20 (2017), 264 – 283.

[22] S. McKenzie, S. Collings, G. Jenkin, and J. River. 2018. Masculinity, Social Connectedness, and Mental Health: Men's Diverse Patterns of Practice. *American Journal of Men's Health* 12 (2018), 1247 – 1261.

[23] Saif M. Mohammad. 2020. Gender Gap in Natural Language Processing Research: Disparities in Authorship and Citations. In *ACL*.

[24] Bren Ortega Murphy. 1994. Greeting cards & gender messages. *Women and language* 17, 1 (1994), 25–30.

[25] Ramesh Nallapati, Bowen Zhou, C. D. Santos, Çaglar Gülçehre, and B. Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *CoNLL*.

[26] Courtney Napoles, Matthew R. Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *AKBC-WEKEX@NAACL-HLT*.

[27] Online. 2020. Americans see different expectations for men and women. https://www.pewsocialtrends.org/2017/12/05/americans-see-different-expectations-for-men-and-women/. (Accessed on 10/15/2020).

[28] Online. 2020. First ads banned for contravening UK gender stereotyping rules. https://www.theguardian.com/media/2019/aug/14/first-ads-banned-for-contravening-gender-stereotyping-rules. (Accessed on 10/15/2020).

[29] Online. 2020. Free greeting card messages. https://www.greeting-card-messages.com/. (Accessed on 09/03/2020).

[30] Online. 2020. Free Wishes, Messages, Quotes, Greeting Cards | 143 Greetings. https://www.143greetings.com/. (Accessed on 09/03/2020).

[31] Online. 2020. Gender Stereotyping in Advertising. https://twfhk.org/blog/gender-stereotyping-advertising. (Accessed on 10/15/2020).

[32] Online. 2020. Greeting Card Industry Facts and Figures. https://www.liveabout.com/greeting-card-industry-facts-and-figures-2905385. (Accessed on 09/03/2020).

[33] Online. 2020. Greeting Card Messages. https://www.bestcardmessages.com/. (Accessed on 09/03/2020).

[34] Online. 2020. Greeting Card Messages and Ideas | Hallmark Ideas & Inspiration. https://ideas.hallmark.com/tag/greeting-card-messages/. (Accessed on 09/03/2020).

[35] Online. 2020. How-to-Birthday Wishes and Messages by Davia. https://www.holidaycardsapp.com/wishes/category/how-to/. (Accessed on 09/03/2020).

[36] Online. 2020. Occasions Messages - Wishes, Messages, Greetings and Cards. https://www.occasionsmessages.com/. (Accessed on 09/03/2020).

[37] Online. 2020. WATCHING GENDER: How Stereotypes in Movies and on TV Impact Kids' Development — Betsy Bozdech reports. https://awfj.org/blog/2017/06/26/watching-gender-how-stereotypes-in-

[38] Online. 2020. What To Write In A Card | American Greetings. https://www.americangreetings.com/inspiration/what-to-write. (Accessed on 09/03/2020).

[39] Online. 2020. Wishes and Messages - WishesMsg. https://www.wishesmsg.com/. (Accessed on 09/03/2020).

[40] Jeffrey Pennington, R. Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*.

[41] A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

[42] Alan Reifman, Mykaela Ursua-Benitez, Sylvia Niehuis, Emma Willis-Grossmann, and McKinley Thacker. 2020. #Happyanniversary: Gender and age differences in spouses' and partners' Twitter greetings. *Interpersona: an international journal on personal relationships* 14 (2020), 54–68.

[43] Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP models with CheckList. In *ACL*.

[44] Melissa Roemmele. 2016. Writing Stories with Help from Recurrent Neural Networks. In *AAAI*.

[45] Danielle Saunders and Bill Byrne. 2020. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7724–7736. https://doi.org/10.18653/v1/2020.acl-main.690

[46] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 144 (Nov. 2019), 33 pages. https://doi.org/10.1145/3359246

[47] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional HCI: Engaging Identity through Gender, Race, and Class. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017).

[48] Jilwan Soltanpanah, E. J. Parks-Stamm, S. Martiny, and Floyd W. Rudmin. 2018. A Cross-Cultural Examination of the Relationship between Egalitarian Gender Role Attitudes and Life Satisfaction. *Sex Roles* 79 (2018), 50–58.

[49] Alessandro Sordoni, Michel Galley, M. Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and W. Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. *ArXiv* abs/1506.06714 (2015).

[50] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1679–1684. https://doi.org/10.18653/v1/P19-1164

[51] Josefine Steinhagen, Martin Eisend, and Silke Knoll. 2010. *Gender Stereotyping in Advertising on Public and Private TV Channels in Germany*. Gabler, Wiesbaden, 285–295. https://doi.org/10.1007/978-3-8349-6006-1_19

[52] Simone Stumpf, Anicia N. Peters, Shaowen Bardzell, Margaret M. Burnett, Daniela K. Busse, Jessica R. Cauchard, and Elizabeth F. Churchill. 2020. Gender-Inclusive HCI Research and Design: A Conceptual Review. *Found. Trends Hum. Comput. Interact.* 13 (2020), 1–69.

[53] M. Szumilas. 2010. Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent* 19 3 (2010), 227–9.

[54] Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. 2013. Gender, Representation and Online Participation: A Quantitative Study. *Interacting with Computers* 26, 5 (09 2013), 488–511. https://doi.org/10.1093/iwc/iwt047 arXiv:https://academic.oup.com/iwc/article-pdf/26/5/488/9644215/iwt047.pdf

[55] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265* (2020).

[56] F. J. Vijver. 2007. Cultural and Gender Differences in Gender-Role Beliefs, Sharing Household Task and Child-Care Responsibilities, and Well-Being Among Immigrants and Majority Members in The Netherlands. *Sex Roles* 57 (2007), 813–824.

[57] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and V. Ordonez. 2019. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 5309–5318.

[58] Emily M. West. 2009. Doing Gender Difference Through Greeting Cards. *Feminist Media Studies* 9 (2009), 285 – 299.

[59] Eike Wille, Hanna Gaspard, Ulrich Trautwein, Kerstin Oschatz, Katharina Scheiter, and Benjamin Nagengast. 2018. Gender Stereotypes in a Children's Television Program: Effects on Girls' and Boys' Stereotype Endorsement, Math Performance, Motivational Dispositions, and Attitudes. *Frontiers in Psychology* 9 (2018), 2435. https://doi.org/10.3389/fpsyg.2018.02435

[60] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, V. Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *NAACL-HLT*.

[61] P. Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining Gender Bias in Languages with Grammatical Gender. In *EMNLP/IJCNLP*.