Women in the Workplace: Analyzing Gender Biases in Corporate Email Communications

Anonymous ACL Submission

Abstract

Gender disparities in the workplace hinder women in career advancement and equity. Communication within companies reflect gender norms and discrimination that affect organizational structures. Gender biases are exhibited in different forms, including unequal treatment, associations of gender with certain concepts, and stereotyping language. We approach different angles of considering linguistic gender biases to provide an extensive analysis on the role of gender in workplace emails (1) Determine how receivers' genders affect language use in the emails through computational text analysis with LIWC and model explainability investigations. (2) Examine gender disparities through representation biases in word embeddings to find asymmetric associations of gender with profession words. (3) Identify biased emails in the workplace and create an NLP tool to identify and flag phrases in emails that express gender biases. We study corporate interactions through the Enron Corpus, a uniquely available database of 500K real workplace emails of Enron employees. Our results find significant presence of biases in all three paths, reveal gender inequalities through a case study of a corporation, and show the effectiveness of natural language processing methods to avoid such occurrences in further workplaces.

1 Introduction

Language both reflects society and influences the perspectives and structures within it. In this way, language holds power for advancing social justice, yet is also capable of bearing harmful biases. Notability, with the pervading gaps in the representation of women across many areas, gender biases present the potential dangers of such disparities. The pervasiveness of gender biases across language has been a prevalent focus in NLP research. Research concerning gender equality has expanded across sectors to reveal the inequalities in news and media (Dacon and Liu, 2021), politics (Stańczak et al., 2021), literature (Babaeianjelodar et al., 2020), and legal practices (Gillis, 2021). These biases have not changed in more than 20 years (Cépeda et al., 2021), reflecting the persistent lack of diversity and the relevance of its discussion and research.

In particular, email communications play an especially important role in the transmission of information and the administrivia within organizations. Language used in email communication is reflective of values in everyday work (Habil and Rafik-Galea, 2005). In such practices prevalent to organizational conduct, linguistic biases have a reach that constrain the careers of women (Holmes, 2000). Gender biases in the workplace have been studied in attempts to identify and address the inequality of women in corporate environments; they have been found to hinder women in job applications, recommendations, and progression to managerial positions (Strol, 2020). However, despite the scale in the use of email and its relevance in organization discourse, email is under-studied in workplace language and gender research (Mullany, 2011).

Gender biases are exhibited in different forms, including unequal treatment, associations of gender with certain concepts, and stereotyping language. In this work we study gender biases in corporate email communications through different angles to provide a comprehensive analysis and propose a tool to promote inclusivity in organizations. Our contributions are three-fold:

- We examine bias first through the role of receivers' genders on the language use of emails. This is done by extracting linguistic features from emails and analyzing whether they hold any predictive power for the genders of their senders and receivers.

- We identify asymmetric associations of gender with occupations in email contents to find biases in how gender is discussed in the workplace through word embedding models.

- We develop an NLP model to tag inappropriate gender biases in emails and apply it to workplace emails to determine its relevance in a real work setting and propose it for corporate use.

2 Research Questions and Analysis

2.1 Gender Bias in Language Directed to Men Versus Women

Our first question examines gender bias through differences in language use toward men and women. We ask whether men and women are treated differently at work over email by determining whether language use is predictive of email receivers' genders.

In our study, we work with the Enron Corpus (Klimt and Yang, 2004) as a case study of corporate communications. The Enron Corpus is a uniquely available dataset, consisting of nearly half a million emails from Enron employees, as it is of the only open source corpus of real emails and enables the naturalistic study of real workplace discourse. Such data is difficult to access due to privacy restrictions. The Enron Corpus has been valuable to research of organizational communications, providing insights into workplace structures and expanding natural language processing tools (Peterson et al., 2011; Trieu et al., 2017).

We filter the dataset in order to identify linguistic differences based on the directed receiver gender. Due to the constraints of the study, we limit to binary gender classes, which, while unrepresentative of the true diversity, allows us to focus the scope of the analysis. Furthermore, as a result of the confines of our dataset, without labeled gender classes, gender, for the purpose of the investigation, was determined based on the assumed genders of employees' first names according to the Gender Guesser package[1]. To obtain our subset, we extract the assumed sender and receiver gender and narrow down the dataset to emails which (1) have a receiver with a name with a classifiable name and (2) do not contain a header for forwarded or replied information in their bodies, which do not represent the language of the sender. This resulted in 150,490 emails to study. We applied LIWC, a dictionary based text analysis software, to conduct computational linguistic analysis and obtain 118 features of study that describe the linguistic and structural characteristics of these email bodies.

We determine whether the linguistic and structural characteristics of the email can predict the assumed gender of its reciever in binary classification. The model developed is based on the H2O library's ensemble algorithm[2]. A second model was also developed to predict a class that included both the sender's and receiver's assumed gender (e.g. male_female or male_male). The model was interpreted to identify variables that reflected biased language. SHAP, or SHapley Additive exPlanations, interpreted predictions to describe the most important features in our model's decisions as well as their impact (Lundberg and Lee, 2017). SHapley values enabled an understanding of the characteristics of emails to males and females.

The model performances demonstrate the relationship between language use in emails and gender and establish linguistic gender biases in workplace communication. With the LIWC features, the model attains a test F1 score of 0.86 in classifying the receiver's gender. Email language use holds high predictive power of the receiver gender, indicating gender bias in how workers are addressed in organizational discourse. Even further, when taking into account the sender and predicting the gender of both, the model attains a high F1 score of 0.79. The language of emails to men and women are predictably different, indicating bias in the interactions of men and women in the workplace.

[1] https://pypi.org/project/gender-guesser/

[2] https://docs.h2o.ai/h2o/latest-stable/h2o-py/docs/

Feature Name	Feature Definition/Example	Feature Importance	Associated Gender
BigWords	Percent words >= 7 letters	0.1502	M
achieve	Achievement (ex: better, best)	0.0670	M
reward	Reward (ex: opportun*, win)	0.0326	M
Clout	Language of leadership, status	0.0256	M
Tone	Degree of positive tone	0.0070	M
work	Work (ex: work, working)	0.0899	F
number	Numbers (ex: one, two)	0.0640	F
prosocial	Prosocial behavior (ex: care)	0.0513	F
WPS	Average words per sentence	0.0282	F
i	1st person singular (ex: I, my)	0.0078	F

Table 1: Feature importances of select features. Features are sorted by associated receiver gender and feature importance (FI) on the model.

The distinguishing patterns that make up the biases exhibited in workplace emails, as demonstrated through feature impact on the model, can be observed in Table 1. Various characteristics in emails associated with emails directed to men and women are shown. Emails to men contained more achievement–oriented language, captured by the *achieve* and *reward* features. Men were more likely, in this way, to be praised and have achievements acknowledged, than women. Bias in how the accomplishments of men and women are recognized make an impactful appearance in the workplace. Language shifts indicate the differences in the perceptions, roles, and power dynamics of men and women.

2.2 Gender Disparities Through Representation Bias

The second question explored gender bias through the asymmetric associations of gender with profession words as captured by work embeddings. We study the gender disparities in the content of emails to determine the imbalance in how men and women are addressed in email, identifying genderedness in ungendered profession words. We compile a single large corpus consisting of the email bodies from the Enron Corpus, and train a Word2Vec model on the discourses to generate embeddings based on the data.

We follow the methods of Bolukbasi et al. (2019) to measure direct gender bias. We first determine the gender direction, g, in the embeddings, based on a definitional set of 10 gendered word pairs (e.g. she-he, woman-man). The center of each gendered pair is calculated with an average of the vectors. From each word in the pairs, we find the difference to the center. To the matrix, we apply Principal Component Analysis to reduce the dimensions, and use the top principal component to draw the essential information indicating genderedness. We calculate the unit vector principal component as the gender direction *g*. We determine the bias in *N*, a list of 312 neutral profession words. The following formula is used to computer direct gender bias from Bolukbasi et al.:

$$DirectBias_c = \frac{1}{|N|} \sum_{w \in N} |\cos(w, g)|^c$$

We compute the cosine similarity of each profession word to the gender direction to identify the extent that it is gendered and biased in its use. The strictness of the bias is represented by c, which we set as c = 1, as in Babaeianjelodar et al. (2020).

From the procedure, we determine a direct gender bias of 0.08 in the emails. According to Bolukbasi et al., this value confirms occupation words to have significant components along the gender direction. The substantial associations of ungendered professions with genders in the language of workplace communications presents further evidence of bias in an important aspect of corporate structures. This suggests greater implications on how roles are distributed in the workplace as well as how assumptions based on stereotypes are prevalent discussing individuals in email. Gender biases and inequality are prevalent in the content of workplace emails.

2.3 Gender Bias Through Detecting Sexist Phrases

Our third question aims to identify sexist phrases in workplace emails, creating a classification model to analyze the distribution of such statements in organizational discourse and propose its use as a tool for flagging problematic language during email composition in the workplace.

We work with the ISEP dataset (Grosz and Conde-Cespedes, 2020), which contains examples of statements of workplace sexism manually filtered from Twitter, work-related quotes, and faculty/student submissions. In the initial work that presented the dataset, Grosz and Conde-Cespedes developed a BiLSTM model with attention using GloVe embeddings. We develop a model based on newer state-of-the-art architectures to perform the task of classifying the statements for sexism.

Our models are based on various pretrained language model architectures with attention which have been established as one of the best available language models in various NLP tasks, like BERT (Devlin, et al., 2019). We fine-tune the models on the ISEP dataset to build a tool for predicting sexist comments common in the workplace.

Once we determine the effectiveness of NLP tools for flagging sexist statements, we examine the prevalence of such comments in real workplace emails, applying the model on over 100K sentences from randomly sampled Enron emails to classify whether they are sexist.

Model (+Attn)	*F1-score*	*ROC_AUC*
BERT	0.91	0.97
DeBERTa	0.92	0.96
DistilBERT	0.92	0.97
RoBERTa	**0.94**	**0.97**

Table 2: Model performances on predicting sexist statements with the ISEP dataset.

The performance of our models, summarized in Table 2., shows the best model, RoBERTa+Attn, to be highly effective at detecting biased statements.

Identified Phrase
This one has no volume but be careful. Why women can't be mechanics...
all women are noisy fucks.
This must have been created by one of your fellow engineers. You guys just have a bad case of penis envy.
I do not know anything about Kristen. I prefer at least one aggressive person on the desk, Monte and Ashley are kind of shy.

Table 3: Biased comments identified with the RoBERTa+Attn model in the Enron dataset.

On the Enron dataset, the model identified about 10% of sentences to be sexist, revealing that such comments were quite common in real organizational communications. Examples of sexist emails in the workplace can be observed in Table 3. With the pervasiveness of such comments and the potential of NLP tools to advance equity, integration of progressive technologies is much needed. We propose the application of our model in organizations as a tool to flag inappropriate phrases during email composition to promote equality and respect. Use of this tool would thus be able to reduce a substantial number of sexist comments in the workplace, as demonstrated by its application Enron.

3 Conclusion

In this paper, we have examined the presence of gender biases in workplace email communications on multiple dimensions. Our analyses show that language use in the workplace differs to men and to women. Linguistic features of emails were predictive of the receiver gender, and identified characteristics in language addressing men and women. Furthermore, we found gender disparities in email contents, finding an imbalance genderedness of professions. Finally, we develop a model that effectively identifies sexist workplace statements that reveal a frequent presence of biased language in the workplace. Our extensive analysis reveals gender biases on multiple levels confirm the inequality faced by women in workplaces that affect women's careers.

The prominent role of gender in workplace organization carries implicit gender bias and jeopardizes equality. Further implications of these findings in the Enron dataset expand to potential discrimnation persisting in the present day's companies. Representation of women in the C-suite and high corporate positions is scarce, and understanding the everyday gender biases that influence women provides insight into how the views of surrounding individuals may dictate such gaps. This establishes a need for ways to address prejudice and promote diversity in corporations.

We propose a tool for use in organizations to flag inappropriate phrases while composing emails to promote inclusive language. In future works, we look for different and more inclusive approaches to studying gender without assuming a binary definition. We hope our work brings awareness of the importance of working towards building inclusive workplaces and the potential of NLP tools to further study the area.

References

Marzieh Babaeianjelodar, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag. 2020. Quantifying Gender Bias in Different Corpora. In *Companion Proceedings of the Web Conference 2020*, pages 752–759, New York, NY, USA, April. Association for Computing Machinery.

Noa Baker Gillis. 2021. Sexism in the Judiciary: The Importance of Bias Definition in NLP and In Our Courts. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 45–54, Online, August. Association for Computational Linguistics.

Paola Cépeda, Hadas Kotek, Katharina Pabst, and Kristen Syrett. 2021. Gender bias in linguistics textbooks: Has anything changed since Macaulay & Brice 1997? *Language*, 97(4):678–702.

Jamell Dacon and Haochen Liu. 2021. Does Gender Matter in the News? Detecting and Examining Gender Bias in News Articles. In *Companion Proceedings of the Web Conference 2021*, pages 385–392, New York, NY, USA, April. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. May. arXiv:1810.04805 [cs].

Dylan Grosz and Patricia Conde-Cespedes. 2020. Automatic Detection of Sexist Statements Commonly Used at the Workplace. In Wei Lu and Kenny Q. Zhu, editors, *Trends and Applications in Knowledge Discovery and Data Mining*, pages 104–115, Cham. Springer International Publishing.

Hadina Habil and Shameem Rafik-Galea. 2005. Communicating at the Workplace: Insights into Malaysian Electronic Business Discourse. , January.

Janet Holmes. 2000. *Gendered Speech in Social Context: Perspectives from Gown and Town*. Victoria University Press. Google-Books-ID: kIbaW2KOoqwC.

Bryan Klimt and Yiming Yang. 2004. Introducing the Enron Corpus. In *CEAS 2004 - First Conference on Email and Anti-Spam, July 30-31, 2004, Mountain View, California, USA*.

Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. November. arXiv:1705.07874 [cs, stat].

Louise Jane Mullany. 2011. Gender, language and leadership in the workplace. , December.

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email Formality in the Workplace: A Case Study on the Enron Corpus. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 86–95, Portland, Oregon, June. Association for Computational Linguistics.

Leslie Peterson, Lucy Grogan-Ripp, Gwendolyn Smith, Caroline Walz, Cole White, Martha J. Fay, and Kristine Knutson. 2019. Gender and Communication : Perceptions of Diffuse Status Characteristics in Workplace Email. , May.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing Embeddings for Reduced Gender Bias in Text Classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 69–75, Florence, Italy, August. Association for Computational Linguistics.

Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, and Isabelle Augenstein. 2021. Quantifying Gender Bias Towards Politicians in Cross-Lingual Language Models. April. arXiv:2104.07505 [cs, stat].

Oksana O. Strol. 2020. Gender-Biased Language of the Workplace. , January.

Lap Q. Trieu, Trung-Nguyen Tran, Mai-Khiem Tran, and Minh-Triet Tran. 2017. Document Sensitivity Classification for Data Leakage Prevention with Twitter-Based Document Embedding and Query Expansion. In *2017 13th International Conference on Computational Intelligence and Security (CIS)*, pages 537–542. December.