Gender and Politeness Perception: A Novel Approach for Exploring Annotations Disagreement

Ahmad Aljanaideh
Bentley University
Waltham, MA, USA
aaljanaideh@bentley.edu

Abstract

Politeness is an important social phenomenon which influences the flow of conversations. Several studies proposed models to discover and analyze linguistic cues associated with (im)polite language. However, no prior work computationally studied how politeness perception interacts with other social dimensions such as gender. We propose a model for automatic discovery of linguistic patterns which correlate with disagreement in politeness annotations, specifically focusing on gender differences. The model discovers fine-grained context patterns of words which correlate with disagreement in politeness annotations between men and women annotators. We apply the proposed model on emails annotated for politeness. Results show women rate emails which contain formal cues (e.g. *To whom it may concern*) more polite than men annotators rate them, while men rate emails exhibiting informal language cues (e.g. *haven't seen my new swing*) more polite than women annotators rate them. Our findings highlight the importance of studying politeness through multiple demographic perspectives.

1 Introduction

Politeness is a well-studied phenomenon in the field of computational linguistics due to its role in shaping communication in both offline and online conversations (Clark, 1979; Clark and Schunk, 1980; Burke and Kraut, 2008; Zhang et al., 2018). Its interpretation can vary across social dimensions including race and gender (Brown and Levinson, 1987), necessitating studying the interplay between politeness and such dimensions.

Previous studies on politeness focused on analyzing the linguistic cues which correlate with (im)polite language (i.a. Danescu-Niculescu-Mizil et al., 2013; Aubakirova and Bansal, 2016; Aljanaideh et al., 2020). For example, Danescu-Niculescu-Mizil et al. (2013) found that direct questions (e.g. *What do you mean by that?*) correlate with impoliteness. However, these work did not study how perception of politeness might vary across different demographic dimensions such as gender. This leads to limited sociolinguistic insight. Table 1 shows an example email from the POPQUORN dataset (Pei and Jurgens, 2023), where emails were annotated for politeness by 5 annotators on a scale [1, 5] (1 = very impolite, 5 = very polite). Two women annotators annotated the email with scores of 1 and 3, while 3 men annotators annotated the email with scores of 3, 4 and 5. This leads to the following research questions: how can we discover nuanced linguistic patterns which correlate with high disagreement between different demographic groups?

Email	Gender	Score
Can you anticipate what's next in the WWF? I can and it's the Rock whoopin ass on the WCW	Woman	1
	Man	3
	Woman	3
	Man	4
	Man	5

Table 1: An example email from the POPQUORN dataset. The email was annotated by 5 annotators for politeness. The gender of each annotator was provided.

In this work, we introduce a model for discovering politeness cues which exhibit high disagreement in annotations from annotators of different demographic group, focusing on gender. This helps highlight the role of social identity in politeness interpretation and in obtaining sociolinguistic insights. We focus on gender due to its importance in social contexts. The model automatically discovers patterns which correlate with high disagreement in

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3996–4005

annotations between men and women via clustering of pre-trained BERT (Devlin et al., 2018) embeddings. We measure disagreement in annotations using the Earth Mover's Distance (EMD) (Rubner et al., 1998) between distributions of annotations from men and women annotators.

We applied the proposed model on the POPQUORN dataset which contains emails annotated for politeness where the gender of each annotator was provided. Results indicate that women annotators tend to rate emails containing formal language (e.g. *to whom it may concern*) more polite than men annotators rate them while men annotators rate emails containing informal language (e.g. *haven't seen my new swing*) more polite than women annotators. We performed quantitative validation of this hypothesis by first obtaining formality scores for emails by prompting GPT-4 (Achiam et al., 2023), and then comparing the correlation between those scores and politeness annotations of men and women annotators separately. Results show politeness annotations from women annotators correlate with formality scores significantly higher politeness annotations from men annotators. We also evaluated the predictive power of features discovered using our model by training two predictive models: 1) one classifier to predict whether an email is likely to show high disagreement between men and women annotators in politeness annotations, and 2) a model which predicts distribution over politeness annotations. Results show discovered features improve over baseline politeness features from the literature (Danescu-Niculescu-Mizil et al., 2013).

Our findings motivate the need to study politeness from multi-perspective lenses. Understanding how different social groups perceive politeness in text can help in enriching sociolinguistic theories and interpersonal communication dynamics. It can also help in developing customized writing assistants and chatbots which consider the recipients preferences.

2 Related Work

This paper connects to various work in Computational Linguistics, Sociolinguistics and Natural Language Processing (NLP). In the Computational Politeness literature, several studies introduced models to detect politeness in text and discover politeness cues. Danescu-Niculescu-Mizil et al. (2013) developed a model which uses politeness features inspired by theories on politeness (Brown and Levinson, 1978). Examples include hedging (e.g. the use of *might, could* etc.) and the presence of *please* (e.g. *Could you please help me?*). Other work focused on developing neural networks to predict politeness in text (i.a. Aubakirova and Bansal, 2016; Niu and Bansal, 2018). Aljanaideh et al. (2020) developed a model which discovers fine-grained context patterns of words from text annotated for politeness using clustering of contextualized BERT embeddings (Devlin et al., 2018). They found that texts containing the same word (e.g. *please*) can be interpreted differently depending on the context (surrounding words). For example, requests containing *please* after a greeting (e.g. *Hello, could you please help me?*) were deemed polite while requests containing *please* with direct words (e.g. *Could you please stop?*) were deemed impolite. However, these work combine multiple annotations of a text item into a single category (polite or impolite). Our work focuses on discovering patterns which correlate with disagreement between annotators of different demographics.

In the sociolinguistics literature, Eliasoph (1987) indicated that women's politeness is often interpreted as powerlessness, despite it serving as a function of cooperation and connection. Mills (2003) and Chalupnik et al. (2017) challenged gendered stereotypes in politeness use from the speaker's perspective. They indicated that politeness shaped by multiple attributes including race, class and context, and that not all women use politeness in the same way. In our work, the goal is not on politeness use, but rather on politeness *interpretation* and how it differs between men and women.

Several studies in NLP focused on developing approaches for using demographic information for personalization, improving performance, and mitigating bias in NLP. Welch et al. (2020) developed a method for integrating demographic attributes into word embeddings. They found the resulting embeddings help improve performance of language modeling. (Mishra et al., 2020) showed that Named Entity Recognition (NER) models perform better at identifying names of specific demographic groups. Sun et al. (2023) showed that Large Language Models (LLMs) align more specific demographic groups than others in subjective NLP tasks such as politeness rating and offensives rating. Qian et al. (2022) showed that training on perturbed data (e.g. replacing *he* with *she*) leads to fairer language models.

Other studies in NLP focused on analyzing and predicting disagreement in annotations of datasets. Jiang and Marneffe (2022) used a fine-tuned RoBERTa for predicting if an item is likely to exhibit high disagreement in Natural Language Inference annotations. Jiang and Marneffe (2022) showed that including demographic information can help predict disagreement in annotations in various tasks including offensive language detection. Our goal focuses on discovering interpretable patterns correlating with high disagreement among different demographic groups of annotators for politeness, focusing on gender.

3 Dataset

We use the POPQUORN dataset (the Potato-Prolific dataset for Question Answering, Offensiveness, Text Rewriting and Politeness Rating with Demographic Nuance) (Pei and Jurgens, 2023), which is publicly available for NLP research. This dataset contains different sub-datasets for different tasks including politeness among others. The politeness sub-dataset consists of 3,178 emails originally provided by Shetty and Adibi (2004). Each email is annotated by 5-7 human annotators for politeness with an integer score in the range [1, 5]. The race, age, gender, occupation and education-level of each annotator is also provided[1]. In this work, we focus on gender due to the fact that it's an important social dimension and the availability of sufficient annotations from men and women. 254 annotators identified as women, 236 as men and 13 as non-binary[2]. We split the dataset into training, development and testing using a 70/10/20 split.

4 Model

We introduce an approach to discover features which correlate with high disagreement in politeness annotations between men and women annotators. First, we describe the metric by which disagreement is quantified. We then describe the baseline politeness features, and also introduce a model for discovering features which correlate with high disagreement. We then describe the criteria used to select features for analysis from a pool that includes both baseline features and those discovered by the proposed model.

[1]Full demographic information of the annotators can be found in the original work (Pei and Jurgens, 2023)

[2]We focus this analysis on men and women annotators since 13 points is not sufficient to draw solid conclusions.

4.1 Measuring Feature Disagreement

Here we describe the approach we adopt to quantify disagreement in politeness annotations between men and women annotators for items which contain a specific feature (e.g. items which contain the word *please*). Given a list of text items which contain a feature, we map annotations of each group (e.g. men or women) to a probability distribution. This is done by calculating the relative frequency of each score in those items. The Earth Mover's Distance (EMD) (Rubner et al., 2000) is then calculated between the two distributions. We consider this EMD value to be the Feature Disagreement value. The EMD is a distance metric for quantifying dissimilarity between two distributions by considering the amount of shifting needed to make them the same. Given two distributions $x = \{x_1, x_2, \ldots, x_n\}$ and $y = \{y_1, y_2, \ldots, y_n\}$, the EMD (denoted by E) between the two distributions is computed as:

$$E(x, y) = \sum_{i=0}^{n} |E_i|, \tag{1}$$

$$E = 0, \tag{2}$$

$$E_i = x_i + E_{i-1} - y_i. \tag{3}$$

The metric is ideal for comparing ordinal distributions, which is the case here given the ordinal nature of politeness annotations. For example, the EMD between [1, 0, 0, 0, 0] and [0, 1, 0, 0, 0] is 1, while the EMD between [1, 0, 0, 0, 0] and [0, 0, 1, 0, 0] is 2. Feature Disagreement (denoted by FD) is calculated using the following:

$$\text{FD} = \text{E}(P_{men}, P_{women})$$

where P_{men} and P_{women} represent the distributions over annotations of men and women annotators, respectively, of items which contain the feature.

We find this metric appropriate in this context because it accounts for shifts in distributions. For instance, it reasonably treats [1, 0, 0, 0, 0] and [0, 1, 0, 0, 0] as more similar than [1, 0, 0, 0, 0] and [0, 0, 0, 1, 0].

4.2 Politeness Features

We describe the features on which we perform the disagreement analysis. Those features include politeness features from the literature (Danescu-Niculescu-Mizil et al., 2013), and features automatically discovered using a model we introduce.

4.2.1 Politeness Strategies

Danescu-Niculescu-Mizil et al. (2013) developed a model for predicting politeness using unigrams in addition to 21 features inspired by theories in the politeness literature (Brown and Levinson, 1978). Those features include the use of gratitude words (e.g. *thank, appreciate*), direct questions (e.g. *What do you mean?*) and the use of *please* (e.g. *Could you please help me?*). We include unigrams and the 21 politeness strategies in the feature disagreement analysis.

4.2.2 Proposed Feature Discovery Model

We describe the model by which interpretable patterns of high Feature Disagreement values are discovered from the training data. To achieve this, we leverage clustering of contextualized pre-trained BERT (Devlin et al., 2018) embeddings. Clustering BERT embeddings of the same word proved useful for discovering fine-grained patterns of using words in (im)polite contexts (Aljanaideh et al., 2020). For example, Aljanaideh et al. (2020) found clustering embeddings of *please* helps in discovering polite (*Could you please help?*) contexts and impolite (e.g. *please do not do that again*). However, their approach maps annotations of same item into a single category, which is not suitable for our problem setup given we consider multiple annotations per item. Moreover, they use a recursive decision-tree where the model looks for the best split by calculating the pair-wise distance between all pairs of embeddings of a word. This can be inefficient especially for very frequent words. Instead, we use the k-means algorithm, Moreover, we associate each embedding with multiple politeness scores from men and women annotators. Our algorithm is applied on embeddings of every word as follows:

1. we apply k-means clustering on the embeddings of the word (e.g. *think*) using k values in the range $[2, n^{\log_{10}(2)}]$, where n is the number of embeddings. This is done to avoid an extremely high number of clusters for a word since obtaining an extremely high number of small clusters hurts interpretability.

2. For each k, Feature Disagreement is calculated within each cluster using the annotations of the items the word appeared in, and those Feature Disagreement values are averaged across clusters to obtain a value which reflects the extent to which the clustering results in features which correlate with high disagreement. This is given by:

$$\text{FD}(k) = \frac{1}{|C_k|} \sum_{c \in C_k} \text{E}(P_{\text{men}}(c), P_{\text{women}}(c)),$$

   where C_k is the set of clusters for k, and $P_{\text{men}}(c)$ and $P_{\text{women}}(c)$ represent the distributions over annotations within cluster c of men and women annotators, respectively.

3. The value of k corresponding to the largest increase in Feature Disagreement in comparison with that of k-1 is selected as the optimal number of clusters. This is done instead of selecting the k value which leads to the largest Feature Disagreement since the Feature Disagreement value continues to increase as k increases, and thus would result in clusters, each of which is of size 1 which does not provide any interpretability. The process mimics the elbow method, except that the value calculated for each k is the Feature Disagreement value instead of the standard within-cluster-sum-of-square (WCSS).

   The above process is applied for each word in the training set of emails. The result is a number of clusters for every word. Each cluster represents a feature and is associated with a Feature Disagreement value. A high Feature Disagreement value indicates the items which contain the feature exhibit high disagreement in politeness perception between men and women annotators, while a low Feature Disagreement value indicates the items which contain the feature exhibit low disagreement.

4.3 Selecting Features for Analysis

We select features which exhibit high disagreement among men and women annotators, and perform analysis over those features. We sort all features described in the previous sections by their Feature Disagreement values, and perform analysis on features which exhibit high such values.

5 Results and Discussion

In this section we present and analyze our findings. We first describe preliminary analysis on

how men and women rate politeness on average. We then present features associated with high disagreement in politeness annotations. Driven by our findings, we perform analysis on correlation between politeness and formality. We finally validate the discovered features quantitatively by evaluating their predictive power in two tasks: predicting if an email is likely to exhibit high disagreement between men and women annotators, and predicting a full distribution over annotations.

5.1 Preliminary Comparison

We first report the overall average politeness score by men and women annotators. For men annotators, the average politeness score is 3.33. For women annotators, the average politeness score is 3.30. Figure 1 shows the overall annotation distributions for men and women annotators. The EMD between the two distributions is 0.09. As the figure shows, the distributions are similar, with women annotators leaning more for scores of 1 and 5, and men annotators leaning more towards scores of 3 and 4. Next we perform analysis on fine-grained features which exhibit significant differences among men and women annotators.

5.2 Feature Disagreement Analysis

Figure 2 shows examples of features which exhibited high Feature Disagreement (represented by the EMD value on the top-left of each plot) between men and women annotators. Words without an underscore refer to a unigram feature (i.e. all emails which contain the word are considered), whereas words with an underscore and a number refer to a cluster of emails which contained the word in a *specific context* (obtained using the model described in Section 4).

Among the top features, we noticed a mix of unigrams and certain word clusters. Emails containing the word *Whom* were rated more polite by women than men annotators. Those emails exhibit a formal tone (e.g. *to whom it may concern*). Emails containing the word *seen* (e.g. *haven't seen my new swing*, *These guys are acting like they have never seen something like that before*) exhibited an informal tone and were rated more impolite by women annotators than men annotators. Emails containing the word *nice* exhibited an interesting pattern with annotations from men annotators exhibiting a bell-shaped pattern, while annotations from women annotators exhibit a U-shaped distribution. This signals that averaging annotations may hide interesting patterns, and examining distributions over annotations provides more insight. Those emails were mostly informal (e.g. *I'm just trying to be nice*, *god, i'm trying to be nice*).

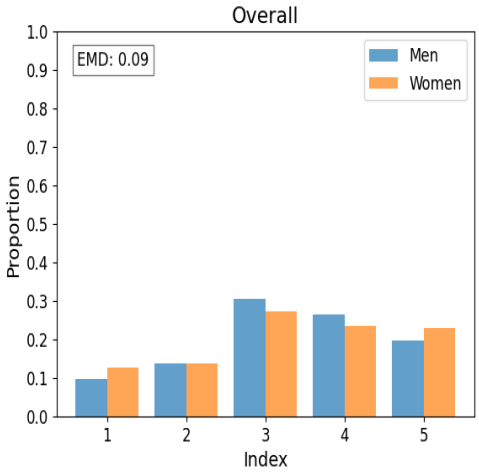

Figure 1: Overall politeness annotation distributions for men and women annotators.

Love_3 correspond to emails containing the word *love* used specifically in a romantic tone (e.g. *Love you, ..*). Those emails exhibited were rated more polite by men annotators than women annotators. Emails containing the word *could* in a social context (e.g. *Perhaps you could find something to do at home?*) were rated more polite by women annotators than men annotators. Emails containing the word *attention* exhibited a formal tone (e.g. *I appreciate you bringing this to my attention*) and were rated more polite by women annotators than men annotators.

Emails containing *'d* (the short version of *would*) when occurring in a somewhat informal and direct context (e.g. *You'd better recognize...*) were rated more impolite by women annotators than men annotators. Emails containing the word *has* in a formal context (e.g. *scheduling has been resolved*) were rated more polite by women annotators than men annotators. *thank_5* corresponds to a cluster of emails which contain an informal use of the word *thank*. Examples of such emails include *Thank goodness!* and *THANK YOU!!*. Those emails were rated more polite by men annotators than women annotators. Overall, we noticed than men annotators rate informal language more polite than women while women rate formal language more polite than men.

Interestingly, none of Danescu-Niculescu-Mizil et al. (2013)'s politeness strategies were among

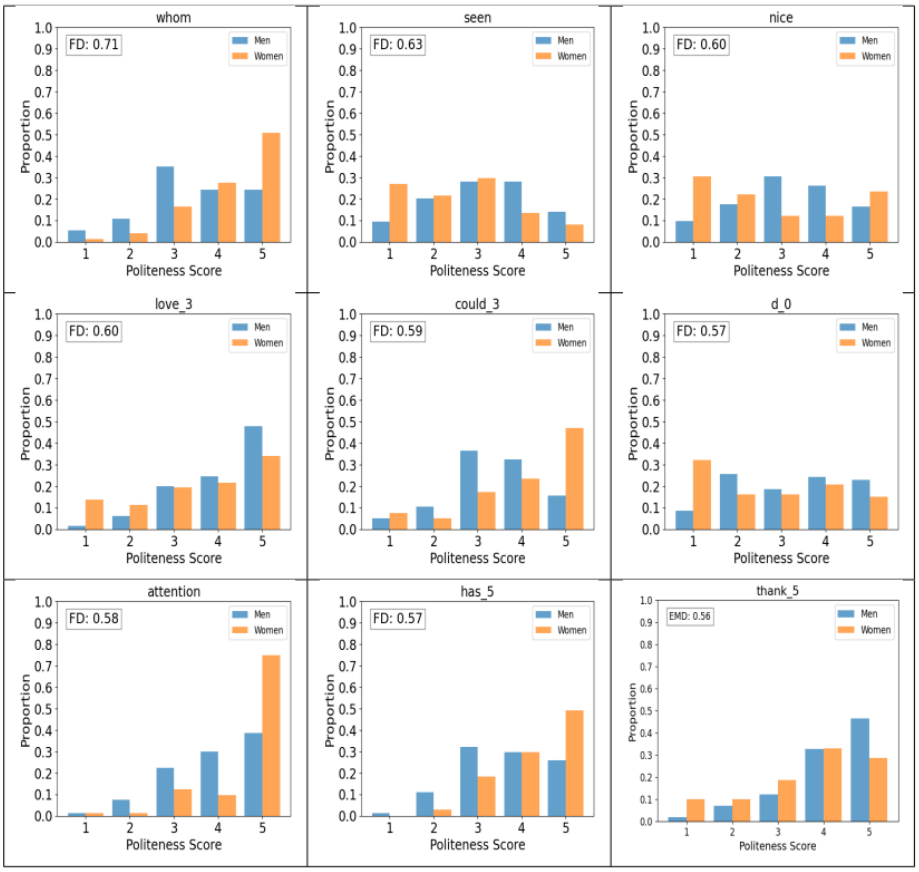

Figure 2: Distributions over annotations for items which contain a specific feature.

the top features in terms of disagreement between men and women annotators. Feature Disagreement values of those features ranged from 0.09 (the presence of a positive word) to 0.24 (direct questions) which are significantly lower than the Feature Disagreement values shown in the table. This signals those features do *not* capture the cues that drive the difference in politeness perception between men and women annotators in this domain. However, it also signals they are perceived relatively consistently across men and women annotators in this domain.

5.3 Politeness and Formality Correlation by Gender

We quantitatively validate findings in the previous section which suggest that women annotators rate formal language more polite than men annotators. Given that the dataset does not contain formality annotations, we use the GPT-4 (Achiam et al., 2023) model to obtain a formality score for each email. Table 3 shows the prompt we used. After that, the politeness scores for each email are averaged separately for men and women annotators. We then compare the strength of the correlation between politeness scores from men and formality scores, and politeness scores from women and formality scores for the training set of emails. Table 2 shows those correlation values. Women politeness annotations correlate with formality scores higher than men annotations do. We confirmed the statistical significance in the difference between the two correlations using the Fisher's r-to-z transformation ($p<0.05$). This result further confirms that in this domain, women annotators rate formal language more polite than men annotators.

Gender	Corr. with formality
Men	0.40
Women	0.47

Table 2: Correlation between politeness scores from annotators of each gender and formality scores.

5.4 Predictive Analysis of Features

To evaluate the predictive power of discovered features, we train two models using the discovered features for two tasks: 1) predicting if an email

On a scale [1, 5], rate the formality of text where 1 is very informal and 5 is very formal. Do not include any other output, just the formality score. Here is the text: [email]

Table 3: The prompt we used to obtain the formality score for each email using GPT-4.

exhibits high disagreement in annotations between men and women, and 2) predicting the full distribution over annotations. First, we describe the method by which word clusters are used to extract features. We then describe the two models. Finally, we describe the results.

5.4.1 Feature Extraction via Word Clusters

We follow Aljanaideh et al. (2020)'s approach of generating features based on word clusters. Given a training email, its features are the cluster ids of its constituent words. For an unseen (development or test) email, each word embedding in the email is assigned to the closest cluster among the training clusters for that word. We use the euclidean distance as a proximity measure. This results in a contextualized unigram-like features (e.g. *Please_1, let_2...*) for each email where 1 and 2 correspond to the cluster id the embedding was assigned to for the word).

5.4.2 Predictive Model - Classification

We quantitatively validate our model by developing an SVM classifier which uses the word clusters as features to predict if an email exhibits high disagreement or not among men and women annotators. Specifically, we label each item with high disagreement if its EMD value falls in the top quartile, and with low disagreement if its EMD value falls in the bottom quartile. The remaining items are not used. Table 4 shows the classification results when using word clusters as features in addition to other baseline features. Word clusters when combined with Danescu-Niculescu-Mizil et al. (2013)'s features achieve the highest accuracy [3]. These results indicate that word clusters could help predict a more nuanced form of labels than classification categorical labels.

5.5 Discussion

Our insights imply that gender plays a role in how politeness is interpreted, which has important implications in professional contexts. Understanding differences in politeness perception between different groups can help in tailoring emails to different audiences to ensure politeness aligns with the recipient's expectations. However, our results correspond to emails in used work settings. We reserve exploring other domains for future work.

Features	Accuracy
DNM	56.9
unigrams	58.3
DNM + unigrams	60.9
word clusters (this work)	62.0
DNM + word clusters	**62.3**
unigrams + word clusters	61.3
DNM + unigrams + word clusters	61.1

Table 4: Classification accuracy of the SVM model used for predicting if an email exhibits high disagreement or not using different feature sets.

5.5.1 Predictive Model - Distribution Prediction

We further validate the proposed model quantitatively by evaluating the predictive power of features discovered by the clustering model. Specifically, we train a model to predict distributions over politeness annotations using the discovered features. This is done to analyze whether these features help predict a more nuanced form of labels than classical categorical labels. We use a Logistic Regression model trained to minimize EMD loss between true and predicted distributions instead of the cross-entropy loss. The EMD is a more suitable loss function than the cross-entropy since the EMD takes the shifting in distributions into account. For example, if the true distribution is [1, 0, 0, 0, 0], then it is plausible to penalize a predicted distribution of [0, 0, 0, 0, 1] more than a predicted distribution of [0, 1, 0, 0, 0]. On the other hand, the cross entropy would assign the same penalty to those two predictions. Moreover, the EMD is used as an evaluation metric, therefore, we use it as an optimization criteria.

We evaluate different combinations of feature sets and compare them using the average EMD on the test set of emails. Table 5 shows the average EMD values on the test set for different feature combinations of unigrams, Danescu-Niculescu-Mizil et al. (2013)'s politeness features (denoted in the table by DNM), and word clusters (our work). We show the percentage decrease in average EMD of our model variations in comparison

[3] We tried using fine-tuning BERT but it result in poor performance potentially due to the fact the training set is relatively small

Features	EMD	Delta vs DNM+uni (%↓)
DNM	0.63	-5.00%
unigrams	0.61	-1.67%
DNM + unigrams	0.60	0.00%
word clusters (this work)	0.57†	5.00%
DNM + word clusters	**0.56**†	6.67%
unigrams + word clusters	0.58	3.33%
DNM + unigrams + word clusters	0.56†	6.67%

Table 5: EMD values on the test set for different feature types (lower is better). Delta column shows percentage decrease relative to the **DNM + unigrams** baseline. † indicates statistically significant improvement over the baseline (Wilcoxon signed-rank test, $p < 0.01$). DNM refers to Danescu-Niculescu-Mizil et al. (2013)'s politeness features.

with the baseline (Danescu-Niculescu-Mizil et al., 2013). Using unigrams and DNM's politeness features alone results in a relatively poor performance. However, combining them helps improve performance. Using word clusters helps obtain a 3.33% improvement (p<0.05, Wilcoxon signed-rank test) in comparison with Danescu-Niculescu-Mizil et al. (2013)'s features combined with unigrams. Combining Danescu-Niculescu-Mizil et al. (2013)'s features with word clusters results in the best performance overall. Adding unigrams to Danescu-Niculescu-Mizil et al. (2013)'s features and word clusters did not result in an improvement. This could be due to the fact that unigrams do not take context of words into account, and thus adding them to features which encode context (i.e. word clusters) hurts performance.[4]

6 Conclusions

In this work, we introduced a model for discovering features which correlate with high disagreement among annotators of different demographics, focusing on gender. The model uses clustering of BERT embeddings and the Earth Mover's distance to discover features which correlate with high disagreement in politeness annotations among men and women. We found that men annotators rate emails containing informal cues more polite than women annotators, while women annotators rate emails containing formal cues more polite than men. The findings were quantitatively validated by performing correlation analysis between politeness annotations and formality scores obtained from GPT-4. We performed further quantitative analysis which showed the features carry a predictive power in the task of predicting if an email exhibits disagreement among men and women annotators, and in the task of predicting the full distribution over annotations for the email.

Politeness is important phenomenon which can shape conversations. Our findings highlight the importance of studying politeness from different perspectives. In the future, we plan to expand this analysis to more demographic dimensions such as race and age. We also plan to use LLMs in order to discover more nuanced linguistic cues which correlate with disagreement in annotations.

7 Limitations

There are several limitations to our approach. First, politeness interpretation can vary based on other social dimensions including race, age and national norms. We reserve studying those other dimensions for future work. Third, the emails on which the model was applied come from a specific work-setting domain. Therefore, the conclusions we reached may not be observed in other domains or contexts. Finally, we used GPT-4 to obtain formality scores for emails in order to perform politeness correlation analysis. However, LLMs outputs can be inconsistent and can also exhibit biases. One potential risk in this study is reinforcing gender stereotypes. Another risk is that the model might be used to fuel controversy or division since it discovers linguistic cues which correlate with disagreement between different groups.

[4]We did not run Aljanaideh et al. (2020)'s model on this data since their setup assumes all annotations are combined into a single label, and thus doesn't suit our setup.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman,

Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Ahmad Aljanaideh, Eric Fosler-Lussier, and Marie-Catherine de Marneffe. 2020. Contextualized Embeddings for Enriching Linguistic Analyses on Politeness. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2181–2190.

Malika Aubakirova and Mohit Bansal. 2016. Interpreting Neural Networks to Improve Politeness Comprehension. *In Empirical Methods in Natural Language Processing*.

Penelope Brown and Stephen C Levinson. 1978. Universals in Language Usage: Politeness Phenomena. In *Questions and politeness: Strategies in social interaction*, pages 56–311. Cambridge University Press.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some Universals in Language Usage*, volume 4. Cambridge university press.

Moira Burke and Robert Kraut. 2008. Mind your ps and qs: the impact of politeness and rudeness in online communities. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 281–284.

Malgorzata Chalupnik, Christine Christie, and Louise Mullany. 2017. (Im) Politeness and Gender. *The Palgrave handbook of linguistic (im) politeness*, pages 517–537.

Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Z Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. *arXiv preprint arXiv:2005.04246*.

Herbert H Clark. 1979. Responding to Indirect Speech Acts. *Cognitive psychology*, 11(4):430–477.

Herbert H Clark and Dale H Schunk. 1980. Polite responses to polite requests. *Cognition*, 8(2):111–143.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of The 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Nina Eliasoph. 1987. Politeness, power, and women's language: Rethinking study in language and gender. *Berkeley Journal of Sociology*, 32:79–103.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Sara Mills. 2003. Rethinking politeness, impoliteness and gender identity. In *Gender identity and discourse analysis*, pages 69–89. John Benjamins Publishing Company.

Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing demographic bias in named entity recognition. *arXiv preprint arXiv:2008.03415*.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Jiaxin Pei and David Jurgens. 2023. When Do Annotator Demographics Matter? Measuring the Influence of Annotator Demographics with the POPQUORN Dataset. *The 17th Linguistic Annotation Workshop*.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 1998. A metric for distributions with applications to image databases. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 59–66. IEEE.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121.

Jitesh Shetty and Jafar Adibi. 2004. The enron email dataset database schema and brief statistical report. *Information sciences institute technical report, University of Southern California*, 4(1):120–128.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2023. Aligning with whom? Large Language Models have Gender and Racial Biases in Subjective NLP Tasks. *arXiv preprint arXiv:2311.09730*.

Charles Welch, Jonathan K Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. *Empirical Methods in Natural Language Processing*.

Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

A Appendix

A.1 Experimental Details

For clustering and classification, we used the scikit-learn implementations of k-means and SVMs. For the SVM classifier, the hyper-parameters we used include regularization strengths of **0.01**, 0.1, 1, 10, 100, and **linear** and RBF kernels (best parameters found are in bold). For the EMD-loss Logistic Regression, we used the TensorFlow library, using the Adam optimizer and a batch size of 32 (best out of 16, 32 and 64). For extracting baseline politeness features, we used the Convokit library (Chang et al., 2020). Results were obtained with a single run.

A.2 Artifacts and Licensing

The code and derived artifacts used in this work were created for research and evaluation purposes and are not publicly released. All third-party datasets and tools are used in accordance with their original licenses and terms of use.

A.3 Ethical Considerations

We use publicly available datasets released by prior work. These datasets were collected and anonymized by the original authors, and do not contain personally identifiable information.

A.4 Use of AI Assistant

An AI Assistant was used to correct the language and discover potential bugs in code.