Causally Modeling the Linguistic and Social Factors that Predict Email Response

Yinuo Xu[*1], **Hong Chen**[*1], **Sushrita Rakshit**[*1], **Aparna Ananthasubramaniam**[*1], **Omkar Yadav**[*1], **Mingqian Zheng**[*2], **Michael Jiang**[*1], **Lechen Zhang**[*1], **Bowen Yi**[*1], **Kenan Alkiek**[*1], **Abraham Israeli**[*1], **Bangzhao Shu**[*3], **Hua Shen**[*4], **Jiaxin Pei**[*5], **Haotian Zhang**[*1], **Miriam Schirmer**[*6], and **David Jurgens**[1]
[1]University of Michigan, [2]Carnegie Mellon University, [3]Northeastern University
[4]University of Washington, [5]Stanford University, [6]Northwestern University

Abstract

Email is a vital conduit for human communication across businesses, organizations, and broader societal contexts. In this study, we aim to model the intents, expectations, and responsiveness in email exchanges. To this end, we release SIZZLER, a new dataset containing 1800 emails annotated with nuanced types of intents and expectations. We benchmark models ranging from feature-based logistic regression to zero-shot prompting of large language models. Leveraging the predictive model for intent, expectations, and 14 other features, we analyze 11.3M emails from GMANE to study how linguistic and social factors influence the conversational dynamics in email exchanges. Through our causal analysis, we find that the email response rates are influenced by social status, argumentation, and in certain limited contexts, the strength of social connection.

1 Introduction

From greeting friends to discussing work, emails have become a vital component of human communications in modern society. A key intent behind many emails is to get a reply from the recipient. When making requests of other people, multiple theories have suggested specific factors that predict whether a request will be successful, e.g., its politeness (Brown and Levinson, 1987). Prior work in NLP has focused on testing the effects of specific factors like power and face-saving as predictors (e.g., Danescu-Niculescu-Mizil et al., 2013; Prabhakaran et al., 2014). Yet, these studies often lack the big picture of the environment in which a request is made to understand how multiple factors interact to influence successful requests. Email represents an ideal setting for studying such requests due to social structure (revealed through communication) and the flexible content length. As a result, here, we propose a new study of requests by using longitudinal email data with causal inference to holistically test what predicts a successful request, i.e., a received reply.

Figure 1: The overview of analyzing the social dynamics of conversation-ending emails. By introducing a novel dataset and associated classifiers, we develop a set of social and linguistic factors to conduct causal analysis to predict conversation-ending emails.

Email communications have been studied for many years (Lang, 1995; Baron, 2002; Dürscheid et al., 2013), with recent research focused on the structure, dynamics, and intent of email exchanges (Wang et al., 2019; Shu et al., 2020a; Robertson et al., 2021b; Shah et al., 2023). However, despite existing research studying email communication (Di Castro et al., 2016; Kooti et al., 2015; Zhang et al., 2020), it is still unclear how the social features like the sender's position in the network, writing style, and tone shape the structure and dynamics of email exchanges. In this work, we address the following research question: *In what ways do social and linguistic factors influence the likelihood of an email receiving a response?*

[*] denotes equal contribution

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 11842–11866

To model the dynamics of email conversation, this study analyzes over 11M emails from the GMANE corpus (Bevendorff et al., 2020), a dataset of email chains on a public forum. We introduce a new dataset, SIZZLER, that captures the intent and reply-expectation of emails, and create a comprehensive list of social and linguistic factors for each email, such as the sender's position in the GMANE network and the argument quality of their text. Figure 1 illustrates our modeling framework.

Our contribution is threefold. First, we introduce a new dataset, SIZZLER, of emails annotated for intents and sender expectations by human experts. Second, we propose a structured framework for modeling the mechanisms that potentially increase the likelihood of an email receiving a reply. We release these labels for 11M GMANE emails, along with the associated computational features (e.g., argument quality). Third, we conduct a causal analysis using propensity score matching to determine which of a comprehensive list of social network and linguistic factors influence email responses. We find that email response rates are generally more impacted by social status and argumentation than by the strength of social connection. All data and code are available on GitHub.[1]

2 The SIZZLER Dataset

We present the **S**ocial **I**nformation for analy**Z**ing and characteri**Z**ing the **L**ikelihood of **E**mail **R**eplies (SIZZLER) dataset, which is based on the Webis GMANE Email Corpus 2019 (Bevendorff et al., 2020). The GMANE corpus is one of the largest email datasets available, comprising over 153 million emails from gmane.io. To analyze the effects of social network and linguistic factors on reply behavior in emails, SIZZLER includes not only each email's reply status and the pragmatic factors of interest, but also a number of covariates that influence email replies, such as the sender's communicative intent (Cohen et al., 2004; Dabbish et al., 2005; Sappelli et al., 2016) and expectation of receiving a reply (Hanrahan et al., 2016b) (e.g., allowing us to account for confounders like: an email thanking someone for their help is less likely to receive a reply and more likely to be polite). To create this dataset, we first extract features related to email threads and networks from the GMANE corpus and then develop models to capture the sender's communicative intent expectations of receiving a reply. This section describes the data processing, case definitions, and annotation process.

[1] https://github.com/davidjurgens/per-my-previous-email

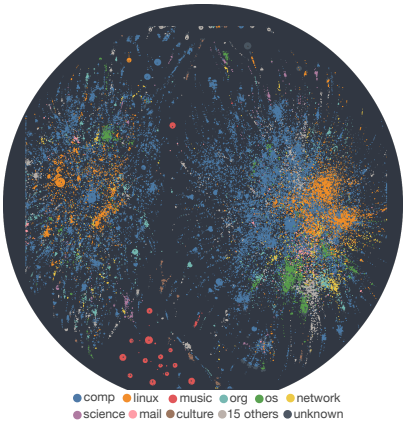


Figure 2: The email reply network for Dec 2009. The network has a clear grouping of communities across all mailing lists, shown as color.

2.1 Preprocessing and Sampling

We filter non-English content (∼15% of the emails) to focus on the dominant language in the corpus. We preprocess and analyze data from the entire twelve months of 2009[2]. Overall, our data includes 11.03M emails written by 1.78M users. The data contains 1.88M discussion threads and 7.46M unreplied emails that were either written as part of a discussion thread or singleton threads. We further extract features of the email, including the number of characters in the email's body and supporting content (e.g., code, logs, opening) as defined by Bevendorff et al. (2020); the top-level group the email was posted in; and the properties of the discussion thread the email include labeled topic, length, breadth, the depth of the post in the discussion tree; and whether the email received a reply.

2.2 GMANE Network

As a public email forum, the interactions on GMANE contain a structure representing which individuals communicate with each other. We construct a directed, weighted network from email replies; there is an edge from user i to user j if i replies to j, with the edge weight representing the log-scaled number of replies. Figure 2 shows the

[2] Even though the platform contains data through 2019, we chose to analyze data from 2009 as there was fairly stable email volume between 2007-2009 and a steep decline in email volume from 2009 to 2010 (Appendix Figure 7). As we are unaware of why these declines occur, utilizing 2009 was the best way to ensure the most recent and reliable data.

network based on 731,252 email posts from Dec 2009, with components with fewer than 100 users removed (138,346 users, 232,277 edges). The network shows two clear core-periphery structures in the GMANE network that are largely uncorrelated with the top-level group that users most often post in. From this network, we calculate each author's relative position in the network. In large social networks, an individual's social status and connection are frequently related to their position in the structure of the network, with high-status individuals closer to the core and disproportionately having ties to other high-status individuals (Ball and Newman, 2013). These structural differences lend themselves to differences in social connectedness and status and, as such, may influence whether a particular user is likely to get a response to their email.

2.3 Human-Annotation

Some emails are, by construction, not intended to get a response. When analyzing what linguistic and network factors influence responses, we want to control for the confounding effects of the sender's communicative intent and expectation of receiving a reply. To classify an author's intent and expectations, we design an annotation pipeline to collect labels from expert annotators.

Data Sampling and Annotation Schema By nature, the GMANE dataset does not distribute uniformly over characteristics that are likely to influence communicative intent and expectations (e.g., discussed topics, length of emails). Therefore, we use stratified sampling (Neyman, 1992) in the human annotation, to increase the diversity of our labeled data. The strata are defined using four email characteristics, including Topic, Length, Depth and Reply-Expectation (see Appendix A.1).

Nine expert annotators decided the annotation schema for this study, through three rounds of annotation on a small sample of data. The schema was grounded in existing schema (Wang et al., 2019; Sappelli et al., 2016)[3], but adapted to the GMANE dataset. Annotation guidelines were refined by discussing disagreements and removing low-frequency categories (e.g., scheduling). Our final annotation schema contains six email intents (*Sharing*, *Requesting*, *Promising*, *Personal Communication*, *Formalities*, *Auto-generated or Spam*) and three expectations (*Response Expected*, *Response Non-Expected*, and *Not Applicable*). A detailed description of the categories is in Appendix Table 3.

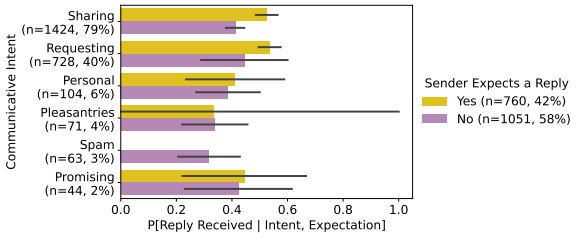


Figure 3: Probability of an email in the labeled data receiving a reply, stratified by intent and reply-expectation. Error bars represent 95% bootstrap confidence intervals. There is no bar in the (Spam,Yes) stratum because annotators were instructed that these categories are mutually exclusive. The error bar in the (Pleasantries,Yes) stratum is large because it contains only 3 observations.

Multiple rounds of alignment were required because many emails had multiple interpretations. Consider, for example, an email that reads "Here's an encrypted password: go decrypt it: xx/XxXXxxxxxXXX." During the pilot, annotators saw two interpretations of this email: 1) the sender giving the recipient a sample password to test their decryption algorithm on (Intent: sharing information, Reply Expected: no) or 2) the sender asking the recipient to decrypt a password for them (Intent: requesting help, Reply Expected: yes). Since the sender explicitly requests an action ("go decrypt it"), we chose the second interpretation.

Annotation Process Annotators included 9 students from a public university in the United States, who are all authors of the paper. For each example, annotators are shown the subject and body of one post and asked to rate the sender's intents and expectations. Annotation occurred in two phases. In the first phase, all annotators are asked to independently label the same 30 emails as a pilot study in order to calculate interannotator agreement. After this phase ended, the authors discussed cases with high disagreement to ensure better alignment in the final labeled data. In the second phase, each annotator is assigned around 200 instances to label. Since larger, noisier sets of labeled data may lead to higher quality models than fewer, cleaner labels (Song et al., 2022), emails in this second phase were each annotated by a single rater. As

[3]Wang et al. (2019) defined four distinct categories for email intents, including information exchange, task management, scheduling and planning, and social communication. Sappelli et al. (2016) and Carvalho and Cohen (2005) based their classification of intent on the theory of speech acts, categorizing emails into request, propose, commit, deliver, amend, refuse, greet, and remind.

a result, the second phase produced 1,811 labeled emails that were used in modeling. We use the POTATO (Pei et al., 2022) annotation platform for both phases (Appendix Figure 6).

2.4 SIZZLER Dataset Analysis

Labeled Data We use different agreement measures to validate our annotation process. Our average pair-wise agreement IAA and Krippendorff's α (Hayes and Krippendorff, 2007) were moderate, in part due to the class imbalance and inherent subjectivity in understanding email intention. The Expect-Reply category, which is the main focus of our analysis, had a sufficiently α high value of 0.58 to support our causal analysis. We report and discuss full details of the agreement in Appendix A.

Distribution of Email Intents Sharing was the most commonly occurring category in the labeled data (79%), followed by Reply-Expectation (42%) and Requesting (40%). The remaining categories occurred in under 10% of labeled emails. We analyze the distribution of email intents with respect to sender expectations. Figure 3 shows the probability of receiving a reply, stratified by the labels for the 5 intent categories and reply-expectation. In general, the sender's communicative intent is associated with both expectation and receipt of reply. Senders are more likely to receive a reply when they expect one (53%) than when they don't (40%). Messages that share or request information are most likely to receive replies. Senders are most likely to expect replies when requesting information (94%).

3 Classifying Intents and Expectations

We use our labeled data to model the sender's communicative intent and reply expectation across the entire dataset, in a multi-label classification task. To understand why this task is important, consider an email simply saying "Thank you!" where the sender probably not expecting a reply. This type of message would confound our analysis, because it is both less likely to receive a reply and has systematic differences in linguistic features (e.g., poor argumentation). The GMANE dataset is not labeled with the sender's communicative intent or reply expectation, so we modeled these features. Having access to these features at scale allows us to then match replied and unreplied messages on the sender's reply expectation and intent using propensity score matching, and analyze the effects of factors like network position, argumentation quality, etc. on reply behavior.

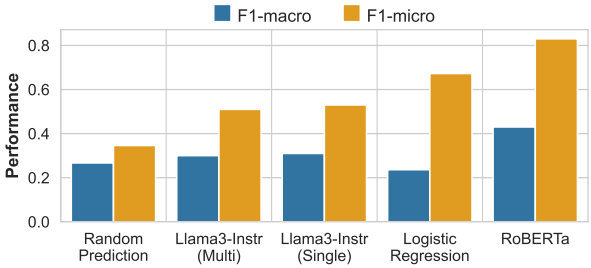


Figure 4: F1 scores of different models at identifying email intents and sender expectations. Llama3-Instruct (Multi) and Llama3-Instruct (Single) are zero-shot models with multi- and single-label settings, respectively.

3.1 Experimental Setup

The dataset is split into train, validation, and test using an 8:1:1 ratio. Detailed training and inference parameters are shown in Appendix A.

Masked Language Models We fine-tune a masked language model, RoBERTa (Liu et al., 2019), trained for three epochs to classify email intents and expectations from the labeled data.

Generative Language Models Additionally, we use Llama-3-Instruct(8B) (AI@Meta, 2024) model for zero-shot inference on classification tasks, applying vLLM (Kwon et al., 2023) technique to accelerate the inference process. We use two types of classification prompts to accomplish zero-shot inference tasks: "Multi-Options" and "Single-Option". For "Multi-Options", we ask the model to multi-select from all email intents in a single prompt. For "Single-Option", we ask the model about its tendencies on each email intent, across 7 prompts. Prompts are shown in Appendix Table 5.

Baselines For comparison, we develop two baseline models for the tasks. Our first baseline model randomly predicts each class. The second baseline is a logistic regression trained to output each of the 7 labels, where the input is the frequency of the most common unigrams and bigrams in the emails after removing stopwords. A penalty term of $C = 1$ is best over our validation set, with all other hyperparameters set to default values.

3.2 Prediction Results

Figure 4 depicts the comparative performance of each model. A detailed comparison, including each model's recall, precision, and F1 scores, is

provided in the Appendix Table 4. The generative and masked language models achieved strong performance in predicting communicative intents and expectations. All models outperform the random baseline, except the logistic regression F1-macro score—suggesting that email intents and expectations are not well-modeled by unigrams and bigrams alone, instead requiring models with stronger contextual and semantic representations. Overall the RoBERTa classifier outperforms both versions of the Llama3 zero-shot classifiers (Multi and Single) by a significant margin, underscoring the limitations of zero-shot classifiers in comparison to fine-tuned encoder models like the RoBERTa classifier in this study [4]. A notable exception are categories where our training data had very few positive labels (e.g., Promising, Personal, and Spam) and the model generated few to no positive labels.[5] The Llama3 classifier performs better in these categories, suggesting the utility of zero-shot approaches when getting enough labeled data is a challenge. The discrepancy between micro and macro F1 scores across all classifiers is the effect of class imbalance in most categories.

4 Factors Influencing Getting A Response

Our analysis studies how the likelihood of receiving a reply is influenced by several factors that relate to *social connection*, *social status*, and use of *argumentation* to engage on GMANE. This includes the style or tone of their writing and how the author is positioned in the social network. We propose 14 factors (eight linguistic, six network properties) to quantify different aspects of connection, status, and argumentation. We use propensity score matching to estimate how these 14 factors affect reply behavior, on a subset of emails that are matched to control for other confounding properties. We find that the effect of these social and linguistic factors is context-sensitive and heterogeneous.

4.1 Motivation and Causal Factors

Online forums allow users to seek **social connection**, *i.e.,* deepening relationships or building strong ties; **social status**, *i.e.,* increasing actual or perceived position in the social hierarchy; and engagement in **argumentation**, *i.e.,* explaining their ideas or experiences in a compelling way (Steinfield et al., 2009; Baek et al., 2011; Steinfield et al., 2013; Ryan et al., 2017; Lampel and Bhalla, 2007). Apart from the content they contribute to GMANE (e.g., what they post, what group it's posted in), users have two main levers for achieving these objectives via computer-mediated communication: how they write (**linguistic factors** like tone or style) (Zhang et al., 2018; Danish et al., 2021; Irani et al., 2024; Peterson et al., 2011; Bhat et al., 2021) and who they choose to engage with (reply behavior and the resulting **network factors**) (Ball and Newman, 2013; Shah et al., 2023).

Our analysis considers a range of linguistic and network factors that relate to social connection, social status, and argumentation. While prior work has often studied these factors independently, our study analyzes both linguistics and network factors together to determine their respective roles in reply behavior. For instance, linguistic cues can be used to build solidarity and connection with recipients (e.g., intimacy or, as a negative example, toxicity) (Koudenburg et al., 2017) or affirm social status (e.g., formality, politeness) (Peterson et al., 2011; Zhang et al., 2018). Similarly, a user's position in the network may indicate that they have high status (e.g., measures of node importance like PageRank) or strong connections (e.g., measures of tie strength like clustering and reciprocity) (Granovetter, 1973; Gupte et al., 2011). Since GMANE is primarily used to seek expertise and input, strong argumentation and information-sharing are also important markers of status (Lampel and Bhalla, 2007). Therefore, our analysis considers linguistic factors related to argumentation (cogency, effectiveness, quality, and clarity) as well as network factors related to expertise and engagement (Hub/Authority HITS scores and out-degree). Appendix Table 6 summarize all factors and Figures 14-15 show that these factors are only moderately correlated, suggesting that they measure distinct social and linguistic aspects of communication.

[4] We experimented with a two-shot setting for the Single-Choice task, but performance declined for some intent types, with the micro-averaged F1 score dropping from 0.53 to 0.47. Adding two-shot examples made the model more likely to predict the positive label, decreasing precision while increasing recall. These results highlight Llama3-8B-Instruct's limited ability to judge email intent, and that RoBERTa is superior both in performance and speed for constructing our dataset.

[5] In our analysis, the modeled intentions and expectations were used as control variables in our propensity score matching. We suspect there was little effect of these three classes on the PSM because (1) these classes are rare in practice so the matching is less likely to place significant emphasis on them, (2) because of their infrequency, any distortion to the matching expected to be low, and (3) although classification performance is low, matching is performed on the P(class|text) value, so the matching model itself may still make use of relatively weak signals of these classes, even if the classifier's absolute prediction is off.

Prior works have examined the effects of text and network factors on engagement online, often finding that they result in more likes, shares, and replies (cf. Section 5). Often, this work suggests that posts should get greater engagement, i.e., higher likelihood of receiving a reply, when they contain **[H1]** stronger interpersonal connection (higher reciprocity, clustering, intimacy; and lower toxicity), **[H2]** markers of higher status (PageRank, HITS, and outdegree), **[H3]** affirmation of higher status (formality, politeness), and **[H4]** stronger argumentation. However, this prior work was not conducted in email forums like GMANE, so it is unclear whether these findings will generalize. For instance, engagement in other online platforms is often driven by curated feed algorithms, which allow users to imagine they are talking to only their strongest ties (Kaplan, 2021). By contrast, public intimacy does not necessarily facilitate social connection on sites like GMANE, since initial emails are broadcast to the whole list.

4.2 Propensity Score Matching

We use stratified propensity score matching (PSM) to estimate whether the likelihood of a response is associated with linguistic and network factors. This approach involves analyzing emails that have received a reply against a comparison group of unresponded emails with similar covariates, as described by Imbens and Rubin (2015).

Subsetting We separately model what factors affect response emails (i.e., emails that reply to an existing post or thread in a mailing list) and initial emails (i.e., a root post). A reply to an initial email allows a conversation to begin while replying to a response email allows an existing conversation to continue. Since conversations begin and continue under very different contexts (e.g., initial emails are broadcast to the whole list and response emails are visible to the whole list but only go to the people on the thread), we build separate models rather than a joint model of reply behavior.

Matching Following a standard PSM approach, a Random Forest classifier is used to predict the likelihood of each email receiving a reply from a set of relevant covariates: (1) Discussion tree features (6 covariates): tree depth, number of branches, number of nodes, number of leaves, depth of mail in tree; (2) Linguistic features (11 covariates): number of characters per email section (e.g., main body), and the predicted likelihood of our RoBERTa model of the seven email intents and expectations. and (3) Time features (3 covariates): the hour of the day (standardized to UTC), the day of the week (Mon, Tues, etc.), and the month.

This classifier is trained using 5-fold cross-validation on all of the data to estimate the probability of receiving a reply for the held-out fold.[6] Each email that received a reply is matched with an email that did not receive a reply with similar probabilities (within 0.01), effectively equalizing the covariate distribution in the matched sample.

Modeling Finally, we run a logistic regression on the matched data to model the odds of a post receiving a reply (dependent variable) as a function of all network and linguistic factors in Table 6. For double robustness, we control for a subset of matching covariates with low multicollinearity. All independent variables in the regression had variance inflation factors < 4. We report standardized coefficients, and 95% confidence intervals are adjusted for multiple comparisons using a Bonferroni correction. We calculate all network factors from the network of the month *prior* to the email date, in order to avoid confounding the outcome (whether the email received a reply) with network covariates (based on edges representing email replies).

4.3 Results

Figure 5a shows factors that influence whether a conversation begins, from 862,820 matched pairs of initial emails, while Figure 5b shows factors that influence whether a conversation continues, from 816,902 matched pairs of response emails. Notably, these associations are not confounded by any of the factors we matched on, including email intent (share, request, etc.) and expectation of reply. The regression results exhibit minimal temporal variation within 2009 (cf. Appendix Figure 17 and Figure 18).

We find both network and text factors are essential in predicting whether an email gets a response. Overall, social connection factors (i.e., strategies to build solidarity and strong relationships), social status factors (i.e., strategies to assert or improve their position in the social structure), and argumentation (i.e., strategies to communicate logically and compellingly) elicit responses at different times and under different contexts in email conversations. While

[6] We use the default parameters of the Random-Forest classifier in Scikit-Learn (Pedregosa et al., 2011) but set the maximum number of trees to 50 to avoid overfitting.

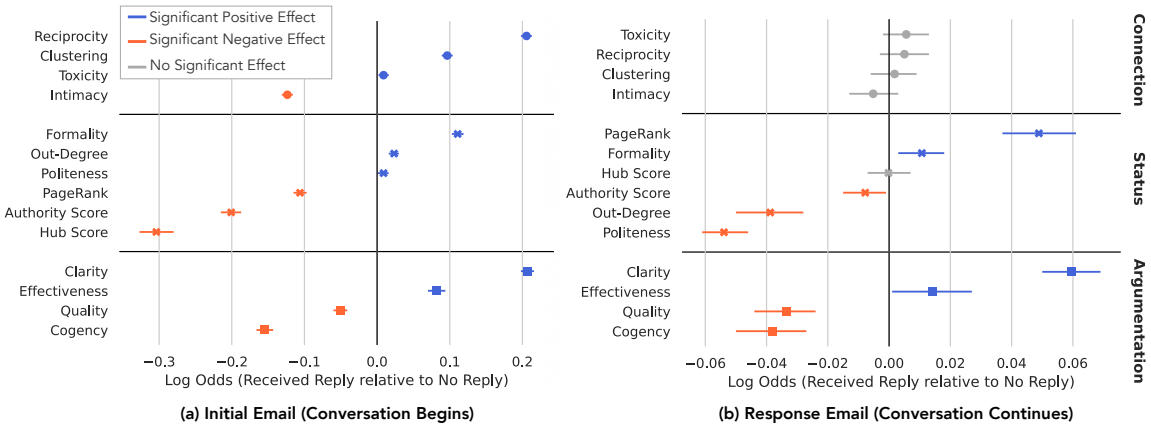


Figure 5: An analysis of what linguistic and network features are associated with the likelihood that an email gets a reply, stratified by whether it is the first email in a chain or a later response. Panel A examines initial emails where conversations may begin, while Panel B shows associations for response emails where conversations may continue. We see that social connection affects response rates in initial emails but not response emails, text-based social status factors play a bigger role in initial emails while network factors play a bigger role in response emails, and affective argumentation is associated with more replies while cognitive argumentation is not. Comparisons are made using Propensity Score Matching, to account for the confounding effects that the email intent, expectation, and structural factors (e.g., time, community, tree depth) may have on reply behavior.

consistent with prior literature suggesting that these social and linguistic factors influence how people choose to interact with each other (Koudenburg et al., 2017), our findings also depart from prior theory in some cases, pointing to some differences between GMANE and other online forums likely related to the platform's structure.

Social Connection In testing **H1**, we find that the relationship between social connection and reply behavior is very context-sensitive and not always as hypothesized. In many cases, social connection does not affect response rates at all. This variation is likely attributable to audience design, or the ways in which people shift their style to match the inferred expectation of their audience (here, email readers) (Clark and Murphy, 1982).

First, conversations are more likely to begin when the sender seeks social connection through the *network* rather than via *text*. Initial emails tend to receive a response when senders have strong social ties (higher levels of clustering and reciprocity in the network) and are less likely to receive a response when written more intimately. Intimate posts on SNS are often understood as encouraging the user's strongest ties to engage (Imlawi and Gregg, 2014). However, sites like GMANE do not allow users to target initial emails to a specific audience (Kaplan, 2021), which may make intimacy less effective and shift social norms around self-disclosure. Instead, individuals with stronger ties in the network, are more likely to receive replies.

Second, social connection factors differently affect whether conversations begin vs. whether they continue. Although replies are accessible to everyone, only those participants in the conversation are directly notified of them, so conversations can be imagined as having a narrower audience. In this context, social connection factors are mostly not associated with whether conversations continue. Although strong social ties may facilitate replies to initial emails, they may not be as useful once a conversation has already begun.

Third, more toxic initial posts were slightly more likely to receive a response, possibly because such emails capture the attention of community moderators (Hanrahan et al., 2016a). This finding is consistent with prior work suggesting that users in online forums tend to expect and receive stronger engagement when content elicits stronger (positive or negative) feelings (Dabbish et al., 2005).

Social Status In testing **H2**, we find that only some of the strategies used to seek or signal social status result in more replies. First, reply behaviors vary based on the types and timing of network strategies. High-status (PageRank and HITS) individuals are less likely to receive replies to initial emails, but high-PageRank individuals are more likely to receive a reply to a response email. These differences between initial and response emails may relate to audience design: Per the discussion

above, in initial emails, senders with higher PageRank may receive replies less often because they tend to have more weak ties (Granovetter, 1973); by contrast, once a conversation has begun with a high-PageRank individual, users may be more likely to keep these conversations going to connect with a high-status individual (Diesner and Carley, 2005; Ball and Newman, 2013).

Second, conversations are more likely to continue when the sender has high PageRank but low out-degree and Authority scores. Whereas the outdegree and Authority score is driven by a user's *level* of engagement, PageRank is more related to the *quality* of engagement (Ding et al., 2002). This finding suggests that status-seeking is more likely to result in a reply when focused on *who* you talk to, not *how many* people you engage with. Additionally, in some contexts, not responding may be seen as status-affirming. Since users with higher Authority scores and outdegree may reply to other people's posts and be perceived as more knowledgeable (Zhang et al., 2007), a lack of follow-up questions or responses may be seen as deference to their status and subject matter expertise (Koudenburg et al., 2014). As experts, these users may also get fewer responses because they ask harder questions, or share better information.

As expected from **H3**, writing in a formal tone, which indicates deference to social hierarchy and affirms the (higher) social status of the recipient (Peterson et al., 2011), is associated with higher response rates. However, polite emails are more likely to start new conversations and end existing conversations – even after controlling for potential confounders like communicative intent. This may occur because, by following social conventions, politeness allows initial emails to be acceptable to a broader audience and facilitates new connection; but once the conversation has begun, it becomes a barrier to smooth communication (Brown and Levinson, 1987; Stephan et al., 2010).

Argumentation Emails tend to receive replies when they are written with affective rather than cognitive argumentation. Affective arguments tend to be simple and emotional, while cognitive arguments are more intellectual in nature (Lai et al., 2012). When testing **H4**, we find that emails tend to receive replies more often when written with greater clarity (ease of understanding) and rhetorical effectiveness (emotional appeals and social appropriateness), but tend to receive replies less often when written with greater argument quality (better support for claims) and cogency (logical completeness). Although affective and cognitive argumentation strategies are known to influence persuasion (Schwarz et al., 1991), this is the first study to our knowledge that links them to reply behavior in emails. Affective arguments tend to be more persuasive and easier to process (Greifeneder et al., 2011; Minton et al., 2017). Similar mechanisms could drive the increase in replies on GMANE, since this is an online forum with many posts competing for a user's attention.

5 Related Work

Speech Acts and Email Intents Emails are often analyzed based on the author's communicative intent or (inferred) purpose in sending the email. Multiple taxonomies of email communication have been developed based on Searle (1975)'s seminal theory on speech acts, including classifications of sender intentions and expectations and recipient actions (Cohen et al., 2004; Carvalho and Cohen, 2005; Dabbish et al., 2005; Sappelli et al., 2016; Lin et al., 2018). Recent work adds context to better predict email intentions, including email metadata, message body (Wang et al., 2019), and action logs (Shu et al., 2020b). Our classification schema is grounded in speech act theory, and extends existing classification schemes on speaker intentions (Cohen et al., 2004; Carvalho and Cohen, 2005) and response expectation (Sappelli et al., 2016). We also extend this work by applying email intent classifiers to study social status and connection.

Email Response. While most users read emails, only 29% reply (Di Castro et al., 2016). Structural factors including information about the email's content, thread, and sender can predict whether and in what timeframe an email will receive a reply (Kooti et al., 2015). Content features that affect response rates include personalizing the message (e.g., including a self-disclosing introduction), explicitly communicated reply expectation (e.g., making requests), and the topic of an email (e.g., male-dominated topics tend to get more replies) (Burke et al., 2007; Wang et al., 2013). However, a recipient's perception of and response to an email is affected by factors beyond content and structure, including linguistic and network factors. Linguistic cues like formality, politeness, clarity, positivity, frustration, and toxicity inform the ways in which users understand email and evaluate automatic re-

ply suggestions (Peterson et al., 2011; Chhaya et al., 2018; Robertson et al., 2021a; Bhat et al., 2021). The locations of the sender and recipient in the email network also influence the conversation by shaping their roles and social status (Diesner and Carley, 2005; Ball and Newman, 2013). Shah et al. (2023) integrated email content with the sender's social network to predict the probability of receiving a response in a corporate setting.

Our work builds on this literature to study how linguistic and network factors are associated with reply behavior. We use a larger dataset, more nuanced annotations, and an interpretable, causal (rather than predictive) analysis of these factors. Through our causal inference setup, we are able to control for factors like topic and whether a request or introduction was made, and test whether replies are associated with the position of the user in the GMANE network and linguistic features pertaining to social connection, social status, and argumentation (e.g., intimacy, formality, and cogency). This new setup allows our study to look beyond the content of the email and ask how the style of writing and network features of the sender affect replies. For applications in automatic email reply suggestions (Kannan et al., 2016), our framework for modeling intent and reply expectations could help generate more contextually appropriate responses and enhance communication.

Social Factors in Engagement. Our work also contributes to literature examining the effects of social status and connection on engagement online. These works show that stylistically normative posts are more likely to receive replies (Robertson et al., 2021a). For instance, politeness is associated with more replies on Wikipedia but not on Reddit (Zhang et al., 2018; Danish et al., 2021), because of different expectations on tone. On forums like Reddit, argumentation plays an important role in determining an author's persuasuion and likelihood of receiving a reply (Tan et al., 2016; Irani et al., 2024). Factors like reciprocity (Guadagno et al., 2024), formality (Peterson et al., 2011), politeness, and rhetorical prompts (Zhang et al., 2018) are associated with both higher engagement and the quality of social interaction, while toxicity detracts from these outcomes (Bhat et al., 2021).

6 Discussion

This research presents an integrated view of factors that are often considered separately in analyses of message reply behaviors: social connection, social status, and argumentation. Social connection and social status play complementary roles in email reply behaviors, in ways that are heavily moderated by the conversational setting. For instance, attributes related to social status and connection that promote replies in initial emails (when a sender is addressing a broader audience) are unimportant or even anticorrelated to response rates in response emails. Instead, network social status markers are the most predictive of replies in response emails. In addition to various network and linguistic markers of social status and connection, we also explored the role of argumentation strategy. For instance, affective argumentation consistently produces higher response rates than cognitive argumentation.

This work makes three contributions. First, we offer a new *dataset* labeled with each post's email intentions and reply expectation. We used these labels as control variables in analyses on the dynamics of communication on GMANE. Second, we propose a *methodological* framework for causally modeling the effects of network and linguistic factors on reply behavior using propensity score matching. And third, we examine the role of linguistic and network factors in reply behavior, presenting a *theoretically* integrated view of social connection, social status, and argumentation factors in message reply behavior on online forums.

7 Conclusion

Despite the intention to elicit a response, some emails remain unanswered. In this work, we investigate why some emails don't get a response; we analyze more than 11 million emails from over 1.78 million users in the GMANE dataset to identify linguistic and social determinants of receiving a reply. We model the email's intention and expectation of getting a reply with models trained on our new SIZZLER dataset. We then conduct a causal analysis to assess the impact of social connection, status, and argumentation on the likelihood of receiving a response. We find email response rates are impacted, at different times and in different contexts, by these pragmatic factors. Future work could investigate if the factors influencing email responses generalize across different platforms, languages, and cultural contexts. Applying our framework to observe changes in email patterns over time could also reveal insights into the evolution of digital communication and social interaction norms.

8 Limitations

In this work, we focus on analyzing and modeling email conversations. We decided to exclusively work with English-content emails due to the lack of analogous tools for text analysis in languages other than English. It should be noted that some replies to English emails can be written in other languages rather than English. We do not take these into consideration within our work.

The GMANE corpus primarily consists of technical issues, where individuals reach out to recipients in hopes for a solution. As such, much of the sampled data used for final annotations fell under the technical taxonomy. The scope of this work is hence limited to the GMANE corpus only due to the class-imbalanced nature of the dataset.

The corpus also spans several decades of email threads. Due to the sheer volume, we decided to focus on 2009 as it had the largest amount of data; years 2007-2008 and post-2010 had a sharp decline in email volumes for reasons that were unclear. It is hence notable to consider that there may be instances where this analysis does not apply to the full dataset.

In addition, our *Sender Expectation* (the main category of interest) and *Requesting* labels have high IAA (Krippendorff's α) but other labels have lower IAA. This is likely the result of class imbalance and inherent challenges of interpreting intention in email. We attempted to address these challenges by running several pilot studies to iteratively refine the annotation guidelines and improve agreement. We also do not believe that the low IAA scores negatively affected our results, since the predicted probabilities from the intention classifiers were simply used in propensity score matching to create a balanced analytic sample.

The main focus of this research is not on optimizing the performance of the classifiers, but instead on using their prediction for further analysis. As a result, we did not consider different prompt structures, which could lead to more optimal performance. Our final structure for zero-shot prompting was designed to mimic the same setup that human-annotators saw. We may see beneficial increase in changing and leveraging prompt structure as suggested by (Muktadir, 2023), but this is a limitation that is presented as beyond the scope of this research. Finally, while our propensity score matching inference controlled for several covariates that are likely to be related to email reply behavior, there are undoubtedly confounders that we were not able to control for. However, our anaylsis includes the confounders that are most often discussed in the literature suggesting that our findings are likely to be robust.

9 Ethical Considerations

Analyzing and modeling email conversations between users is of great importance. However, as humans generate the data we deal with, careful caution should be taken while working with this data type. Since all emails in our dataset come from the GMANE corpus (i.e., we are not releasing any of the emails in the dataset), we follow the GMANE terms of use and keep all usernames anonymous throughout analysis.

The GMANE corpus was taken from a public email archive and is also available for public use. The data contains some instances of hate speech, toxicity, offensive language, and manifesto exchange. As a result, it is highly likely that some of this harmful content is present within the SIZZLER dataset.

In addition, human personas are defined by attributes such as personality, values, and demographics (Shu et al., 2024). Therefore, it is essential to consider that biases within our data and threaded conversations may influence the outcomes and insights derived within these models. It may be the case that our data over represents certain personas, leading to bias in model annotations. This is of particular note, since there was some subjective variation in how email intents were perceived and since annotators were students at the same school.

Acknowledgments

This work was supported by the National Science Foundation under Grant No. IIS-2143529.

References

AI@Meta. 2024. Llama 3 model card.

Nikolay Babakov, David Dale, Ilya Gusev, Irina Krotova, and Alexander Panchenko. 2023. Don't lose the message while paraphrasing: A study on content preserving style transfer. In *Natural Language Processing and Information Systems*, pages 47–61, Cham. Springer Nature Switzerland.

Kanghui Baek, Avery Holton, Dustin Harp, and Carolyn Yaschur. 2011. The links that bind: Uncovering novel motivations for linking on facebook. *Computers in human behavior*, 27(6):2243–2248.

Brian Ball and Mark EJ Newman. 2013. Friendship networks and social status. *Network Science*, 1(1):16–30.

Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *Proceedings of the Web Conference 2021*, WWW '21, page 1134–1145, New York, NY, USA. Association for Computing Machinery.

Pablo Barberá, Amber E Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. Automated text classification of news articles: A practical guide. *Political Analysis*, 29(1):19–42.

Naomi S Baron. 2002. *Alphabet to email: How written English evolved and where it's heading*. Routledge.

Janek Bevendorff, Khalid Al Khatib, Martin Potthast, and Benno Stein. 2020. Crawling and preprocessing mailing lists at scale for dialog analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1151–1158.

Meghana Moorthy Bhat, Saghar Hosseini, Ahmed Hassan Awadallah, Paul Bennett, and Weisheng Li. 2021. Say 'YES' to positivity: Detecting toxic language in workplace communications. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2017–2029, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge University Press, Cambridge, England.

Moira Burke, Elisabeth Joyce, Tackjin Kim, Vivek Anand, and Robert Kraut. 2007. Introductions and requests: Rhetorical strategies that elicit response in online communities. In *Communities and Technologies 2007*, pages 21–39, London. Springer London.

Vitor R. Carvalho and William W. Cohen. 2005. On the collective classification of email "speech acts". In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, page 345–352, New York, NY, USA. Association for Computing Machinery.

Niyati Chhaya, Kushal Chawla, Tanya Goyal, Projjal Chanda, and Jaya Singh. 2018. Frustrated, polite, or formal: Quantifying feelings and tone in email. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 76–86.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Herbert H. Clark and Gregory L. Murphy. 1982. Audience design in meaning and reference. In Jean-François Le Ny and Walter Kintsch, editors, *Advances in Psychology*, volume 9, pages 287–299. North-Holland.

William Cohen, Vitor Carvalho, and Tom Mitchell. 2004. Learning to classify email into "speech acts". In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 309–316.

Laura A Dabbish, Robert E Kraut, Susan Fussell, and Sara Kiesler. 2005. Understanding email use: predicting action on a message. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 691–700.

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 (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Danish, Yogesh Dahiya, and Partha Talukdar. 2021. Discovering response-eliciting factors in social question answering: A reddit inspired study. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):82–91.

Dotan Di Castro, Zohar Karnin, Liane Lewin-Eytan, and Yoelle Maarek. 2016. You've got mail, and here is what you could do with it! analyzing and predicting actions on email messages. In *Proceedings of the ninth acm international conference on web search and data mining*, pages 307–316.

Jana Diesner and Kathleen M Carley. 2005. Exploration of communication networks from the enron email corpus. In *SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA*, pages 3–14.

Chris Ding, Xiaofeng He, Parry Husbands, Hongyuan Zha, and Horst D. Simon. 2002. Pagerank, hits and a unified framework for link analysis. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, page 353–354, New York, NY, USA. Association for Computing Machinery.

Christa Dürscheid, Carmen Frehner, Susan C Herring, Dieter Stein, and Tuija Virtanen. 2013. Email communication. *Handbooks of Pragmatics [HOPS]*, (9):35–54.

Neele Falk and Gabriella Lapesa. 2023. Bridging argument quality and deliberative quality annotations with adapters. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2469–2488, Dubrovnik, Croatia. Association for Computational Linguistics.

Diego Garlaschelli and Maria I Loffredo. 2004. Patterns of link reciprocity in directed networks. *Physical review letters*, 93(26):268701.

Mark S Granovetter. 1973. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380.

Rainer Greifeneder, Herbert Bless, and Michel Tuan Pham. 2011. When do people rely on affective and cognitive feelings in judgment? a review. *Personality and Social Psychology Review*, 15(2):107–141.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Rosanna E Guadagno, Amanda Sardos, and Amanda M Kimbrough. 2024. If you reply to me, i will buy from you: A social influence examination of reciprocity on twitter. In *Proceedings of the 12th International Workshop on Behavior Change Support Systems (BCSS 2024)*. R. Piskac c/o Redaktion Sun SITE, Informatik V, RWTH Aachen.

Mangesh Gupte, Pravin Shankar, Jing Li, Shanmugauelayut Muthukrishnan, and Liviu Iftode. 2011. Finding hierarchy in directed online social networks. In *Proceedings of the 20th international conference on World wide web*, pages 557–566.

Benjamin V Hanrahan, Manuel A Pérez-Quiñones, and David Martin. 2016a. Attending to email. *Interacting with Computers*, 28(3):253–272.

Benjamin V. Hanrahan, Manuel A. Pérez-Quiñones, and David Martin. 2016b. Attending to email. *Interacting with Computers*, 28(3):253–272.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Jehad Imlawi and Dawn Gregg. 2014. Engagement in online social networks: The impact of self-disclosure and humor. *International Journal of Human-Computer Interaction*, 30(2):106–125.

Arman Irani, Michalis Faloutsos, and Kevin Esterling. 2024. Argusense: Argument-centric analysis of online discourse. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):663–675.

Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufman, Balint Miklos, Greg Corrado, Andrew Tomkins, Laszlo Lukacs, Marina Ganea, Peter Young, and Vivek Ramavajjala. 2016. Smart reply: Automated response suggestion for email. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (2016).*

Danny Kaplan. 2021. Public intimacy in social media: The mass audience as a third party. *Media, Culture & Society*, 43(4):595–612.

Jon M Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.

Farshad Kooti, Luca Maria Aiello, Mihajlo Grbovic, Kristina Lerman, and Amin Mantrach. 2015. Evolution of conversations in the age of email overload. In *Proceedings of the 24th international conference on world wide web*, pages 603–613.

Namkje Koudenburg, Tom Postmes, and Ernestine H Gordijn. 2014. Conversational flow and entitativity: The role of status. *British Journal of Social Psychology*, 53(2):350–366.

Namkje Koudenburg, Tom Postmes, and Ernestine H Gordijn. 2017. Beyond content of conversation: The role of conversational form in the emergence and regulation of social structure. *Personality and Social Psychology Review*, 21(1):50–71.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement*, 30(1):61–70.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Vicky T. Lai, Peter Hagoort, and Daniel Casasanto. 2012. Affective primacy vs. cognitive primacy: Dissolving the debate. *Frontiers in Psychology*, 3.

Joseph Lampel and Ajay Bhalla. 2007. The role of status seeking in online communities: Giving the gift of experience. *Journal of computer-mediated communication*, 12(2):434–455.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995*, pages 331–339. Elsevier.

Chu-Cheng Lin, Dongyeop Kang, Michael Gamon, and Patrick Pantel. 2018. Actionable email intent modeling with reparametrized rnns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808.

Maitrey Mehta and Vivek Srikumar. 2023. Verifying annotation agreement without multiple experts: A case study with gujarati snacs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10941–10958.

Elizabeth A Minton, T Bettina Cornwell, and Lynn R Kahle. 2017. A theoretical review of consumer priming: Prospective theory, retrospective theory, and the affective–behavioral–cognitive model. *Journal of Consumer Behaviour*, 16(4):309–321.

Golam Md Muktadir. 2023. A brief history of prompt: Leveraging language models. (through advanced prompting).

Mark EJ Newman. 2003. Properties of highly clustered networks. *Physical Review E*, 68(2):026121.

Jerzy Neyman. 1992. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. In *Breakthroughs in statistics: Methodology and distribution*, pages 123–150. Springer.

Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining*, pages 117–126, Online. Association for Computational Linguistics.

Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, et al. 1999. The pagerank citation ranking: Bringing order to the web.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Jackson Sargent, Apostolos Dedeloudis, and David Jurgens. 2022. Potato: The portable text annotation tool. *arXiv preprint arXiv:2212.08620*.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326, Online. Association for Computational Linguistics.

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.

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

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic proto-roles. *Transactions of the Association for Computational Linguistics*, 3:475–488.

Ronald E. Robertson, Alexandra Olteanu, Fernando Diaz, Milad Shokouhi, and Peter Bailey. 2021a. "i can't reply with that": Characterizing problematic email reply suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18, New York, NY, USA. Association for Computing Machinery.

Ronald E Robertson, Alexandra Olteanu, Fernando Diaz, Milad Shokouhi, and Peter Bailey. 2021b. "i can't reply with that": Characterizing problematic email reply suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Tracii Ryan, Kelly A Allen, DeLeon L Gray, and Dennis M McInerney. 2017. How social are social media? a review of online social behaviour and connectedness. *Journal of Relationships Research*, 8:e8.

Maya Sappelli, Gabriella Pasi, Suzan Verberne, Maaike de Boer, and Wessel Kraaij. 2016. Assessing e-mail intent and tasks in e-mail messages. *Information Sciences*, 358:1–17.

Norbert Schwarz, Herbert Bless, and Gerd Bohner. 1991. Mood and persuasion: Affective states influence the processing of persuasive communications. In Mark P. Zanna, editor, *Advances in Experimental Social Psychology*, volume 24, pages 161–199. Academic Press.

John R. Searle. 1975. A taxonomy of illocutionary acts. Retrieved from the University Digital Conservancy.

Harsh Shah, Kokil Jaidka, Lyle Ungar, Jesse Fagan, and Travis Grosser. 2023. Building a multimodal classifier of email behavior: Towards a social network understanding of organizational communication. *Information*, 14(12):661.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments.

Kai Shu, Subhabrata Mukherjee, Guoqing Zheng, Ahmed Hassan Awadallah, Milad Shokouhi, and Susan Dumais. 2020a. Learning with weak supervision for email intent detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research*

and Development in Information Retrieval, pages 1051–1060.

Kai Shu, Subhabrata Mukherjee, Guoqing Zheng, Ahmed Hassan Awadallah, Milad Shokouhi, and Susan T. Dumais. 2020b. Learning with weak supervision for email intent detection. *CoRR*, abs/2005.13084.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey.

Charles Steinfield, Joan M DiMicco, Nicole B Ellison, and Cliff Lampe. 2009. Bowling online: social networking and social capital within the organization. In *Proceedings of the fourth international conference on Communities and technologies*, pages 245–254.

Charles Steinfield, Nicole B Ellison, Cliff Lampe, and Jessica Vitak. 2013. Online social network sites and the concept of social capital. *Frontiers in new media research*, pages 115–131.

Elke Stephan, Nira Liberman, and Yaacov Trope. 2010. Politeness and psychological distance: a construal level perspective. *Journal of Personality and Social Psychology*, 98(2):268–280.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Wei Wang, Saghar Hosseini, Ahmed Hassan Awadallah, Paul N Bennett, and Chris Quirk. 2019. Context-aware intent identification in email conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 585–594.

Yi-Chia Wang, Moira Burke, and Robert E Kraut. 2013. Gender, topic, and audience response: An analysis of user-generated content on facebook. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 31–34.

Jun Zhang, Mark S. Ackerman, and Lada Adamic. 2007. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 221–230, New York, NY, USA. Association for Computing Machinery.

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. *arXiv preprint arXiv:1805.05345*.

Justine Zhang, James Pennebaker, Susan Dumais, and Eric Horvitz. 2020. Configuring audiences: A case study of email communication. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–26.

A Appendix

Inference and Training Details

All deep learning experiments are conducted on single NVIDIA RTX A6000 GPU using HuggingFace Transformers 4.36.2 and Pytorch 2.1.2 on a CUDA 12.1 environment.

For zero-shot inference task on Llama3-Instruct, we used the following parameters: temperature = 1.0, max_len = 2048 tokens, max_gen_len = 16(single)/32(multi) tokens. Accelerated by vLLM, the model requires ∼10h(multi)/∼60h(single) to annotate all Emails in one month (∼600K).

For fine-tuning task on RoBERTa classifier, we trained 3 epoches with 500 warmup steps, and saved the best model according to the validation loss. The training process took only a few minutes on ∼1,800 annotated data. We then apply the best model on unlabeled data, which takes ∼2h to annotate all emails in one month (∼600K).

A.1 Data Sampling

The strata are defined of the email are defined using four email characteristics (i) *Topic* of the built-in "group" tag provided in the GMANE corpus (e.g., Linux, music); (ii) *Length* in characters of the the main body of the email; (iii) *Depth* of the email in the discussion tree associated with each mail thread, ensuring we include emails that had no response at all; and (iv) the *Sender's Expectation* of a reply, inferred using a model we trained from a set of 1,640 pilot annotations (F1 = 0.79). To increase diversity, the data for annotation includes at most one email from the same discussion thread and includes emails from all years.

Annotation Details

Annotation Schema. We show the full annotation schema used in our study in Table 3. The annotation interface is pictured in Figure 6.

To decide the annotation schema, we conducted two rounds of annotations. Our categories are *Shar-*

ing, *Requesting*, *Promising*, *Personal Communication*, and *Formalities* (see Appendix Table 3). We only annotate explicit intents. *Sharing*, *Requesting*, *Promising*, and *Personal Communication* are not mutually exclusive (an email might contain any or all intents), while *Formalities* are mutually exclusive from the rest (we only mark an email as formalities if it only contains pleasantries, formalities, or acknowledgment of receiving the email). In addition to email intent, we also include *Sender Expectation* dimension into the schema to analyze the email sender's expectation of response, which will support our analysis of email conversational-ending phenomena.

Annotated Sample. The inter-annotator agreement (IAA) is calculated on a set of 30 documents labeled by all annotators, after multiple rounds of pilot annotation and discussion. After calculating IAA, we generate a singly annotated corpus of 1,811 emails. The number of emails labeled positive in each category is shown in Table 2. The choice to singly annotate is motivated by where the expected source of variance would come from: data or annotator. Based on our discussions during pilot annotation, we view the data as containing more variance. By annotating more instances, we are able to capture more variance from different emails than would be possible with multiply-labeled data. Prior work on maximizing performance on fixed annotation sizes has shown that this approach leads to better model performance (Barberá et al., 2021; Mehta and Srikumar, 2023). This annotation approach has also been taken for other NLP papers. For example Reisinger et al. (2015) used just a single annotator for the majority of their dataset after pilot work showed this annotator had high IAA. Similarly, after showing that their annotators had high IAA on a small set of documents, Clark et al. (2019) used singly-annotated data. More broadly, Mathur et al. (2019) trained models on data from the Conference on Machine Translation that were mostly singly-annotated.

Model Performance. We show the detailed model performance on the task of identifying the email intents in the Table 4.

Statistics of the Labeled Data. Table 1 shows a summary of the annotated dataset. This includes the annotator's agreement in the first phase before discussion, measured using Krippendorff's α and average pairwise agreement (Krippendorff, 1970),

Dimension	Category	% Agreement	α
Sender Expectation	Expect Reply	0.78	0.58
Communicative Intent	Sharing	0.77	0.29
	Requesting	0.83	0.64
	Personal Communication	0.88	0.01
	Promising	0.94	0.12
	Formalities Only	0.95	0.30
Spam	Spam	0.98	0.09

Table 1: Summary of statistics of annotated data, including the inter-annotator agreement – measured using average pairwise agreement and Krippendorff's α . As expected, values of α are low for categories with large class imbalances, although the average pairwise agreement is high in all cases.

Dimension	Category	# Emails	# in Pilot
Sender Expectation	Expect Reply	760	11
Communicative Intent	Sharing	1,424	28
	Requesting	728	12
	Personal Communication	104	0
	Promising	44	1
	Formalities Only	71	0
Spam	Spam	63	0

Table 2: Number of emails labeled positive in each category; both the number of agreed-upon positive labels (i.e., majority of annotators selected the label) out of the 30 multiply-labeled pilot documents; and the number of positive labels out of the 1,811 singly labeled documents annotated.

and the number of emails falling into each category in the second phase. The top 3 intent categories are *Sharing*, *Requesting*, and *Personal Communication*.

Requesting and *Sender Expectation* had moderate interannotator agreement, but other categories have relatively low values of Krippendorff's α. The low agreement is partially explained by class imbalance, as these categories are either very frequent or infrequently occurring in the data; as a result, chance-adjusted agreement measures become very low, in spite of high mean pairwise agreement. However, the low agreement also reflects genuine differences in how annotators interpreted the sender's intents. For instance, annotators were split on whether the following email constituted an instance of *Sharing* or *Requesting*: "Tested? I thought he just said they were patching them in Ubuntu." The disagreement was over two equally reasonable interpretations of the sender's intent: sharing information (asserting that the patch is occurring) and requesting information (asking for confirmation of the patch).

Therefore, we argue that this variation in label-

Type	Category	Description
Email Intent	Sharing	The sender is sharing information, opinions, updates, or other content with the recipient(s). This includes (but is not limited to) answering a question posed by the recipient; sharing background information to ask a question; stating opinions or takes; sharing documents/external links/resources/references/sample code; or proactively sharing FYI messages and updates.
	Requesting	The sender is asking the recipient(s) for something. This includes (but is not limited to) asking a question, asking for more information, or requesting a task or action be completed. Only EXPLICIT requests should be annotated.
	Promising	The sender writes that they will perform some future action. This includes (but is not limited to) saying they will look into a problem, rewrite some code, fix a bug, give an updated in the future, or do some joint activity with the recipient.
	Personal Communication	The sender makes some interpersonal investment in the recipient(s), beyond the subject of the email. Includes indicators that they have a relationship outside the email thread, personal self-disclosures, engagement with the recipient's personal feelings/worldviews, or other communication that goes beyond the subject of the email and addresses the recipient as a person. Does not include pleasantries like please, thank you, expressions of gratitude, etc.
	Pleasantries/Acknowledgment Only	The email contains only pleasantries, formalities, or acknowledgement of the prior email (e.g., the whole email is a 'Thank you very much' or a 'Will try that!' or 'Got it!') – or if pleasantries/formalities/acknowledgements like these are the main point of the message and all other content is unimportant or secondary.
	Auto-generated or Spam	Email digests, auto-reminders, github or auto-change notification messages, do not reply emails, scam emails, etc.
Sender Expectation	Response Not-Expected	There is no explicit indication in the email that the sender expects the recipient(s) to respond to the email (including asking a question that doesn't require a response).
	Response Expected	Sender likely expects the recipient to respond to the message — e.g., to answer a question, engage with content/attachments and send a follow up email, take some action outside the email chain and send a confirmation that the action was taken, etc. Only annotate if the sender explicitly expects a response.
	Not Applicable	System generated or spam content.

Table 3: Annotation schema used in our study

ing data is likely to lead to an analytically useful model. While some of these disagreements may be due to a lack of context (e.g., annotators did not see the entire email thread), the different interpretations also likely reflect heterogeneity in how recipients would perceive the sender's intents on GMANE. One option to improve IAA in cases like these are to impose stricter rules guiding interpretation, therefore relying less on potentially idiosyncratic annotator judgments in cases like these; we chose not to do this, to ensure our models captured genuine variation in how intent is interpreted. Ultimately, the goal in developing these models is to generate predicted probabilities for each label, which we will use to control for communicative intent and expectation in our final analysis. A labeled dataset with some inconsistent labels would tend to have less certain predicted probabilities on emails like the one shown above. Arguably, these less certain probabilities are a better indication of the factors influencing reply behavior, since they reflect the uncertainty or differences in how the person replying perceived the email.

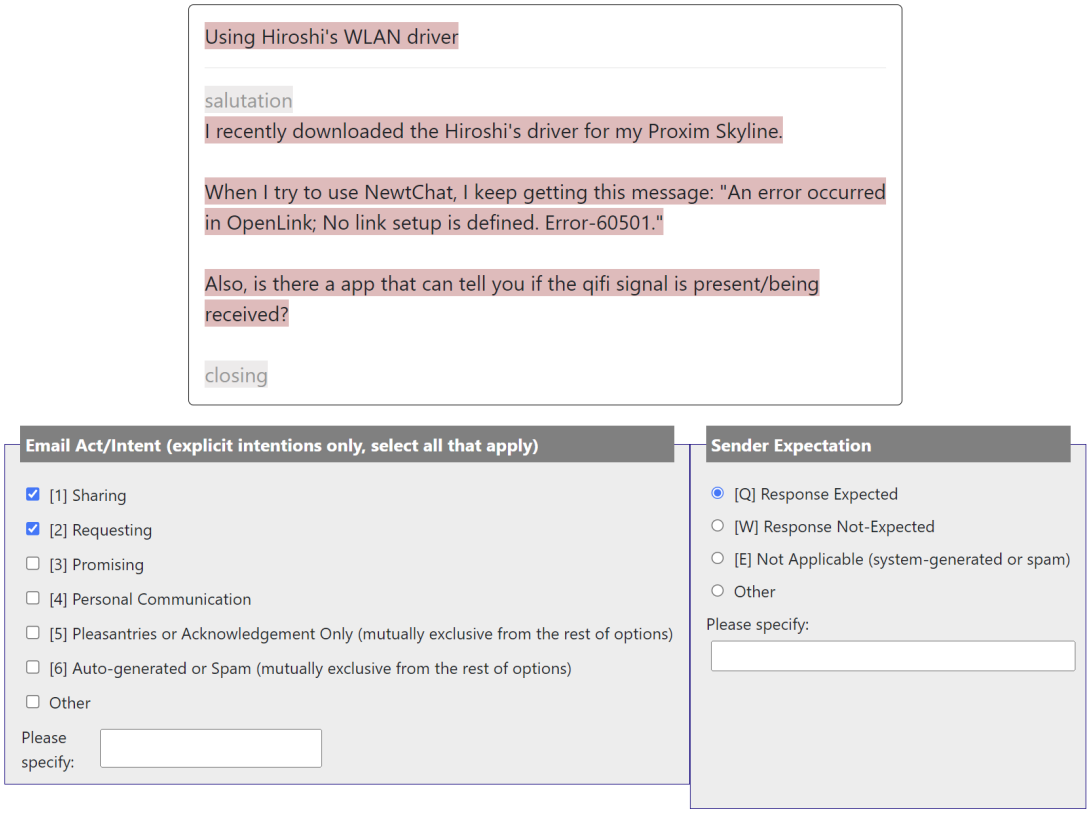


Figure 6: Annotation interface for email understanding

Settings		Sharing	Requesting	Promising	Personal	Pleasantries	Spam	Expectation	Micro_Avg	Macro_Avg
Random Prediction	**P**	0.777	0.390	0.068	0.013	0.037	**0.064**	0.470	0.262	0.260
	R	0.440	0.487	**0.875**	0.200	0.500	**0.750**	0.627	0.509	**0.554**
	F1	0.562	0.433	0.126	0.024	0.069	**0.118**	0.537	0.346	0.267
Logistic-Regression	**P**	0.824	0.810	0.000	0.000	0.000	0.000	**0.800**	**0.820**	0.348
	R	**1.000**	0.224	0.000	0.000	0.000	0.000	0.267	0.570	0.213
	F1	**0.904**	0.351	0.000	0.000	0.000	0.000	0.400	0.673	0.236
RoBERTa-Base	**P**	**0.880**	0.810	0.000	0.000	0.430	0.000	0.780	**0.820**	**0.410**
	R	0.910	**0.890**	0.000	0.000	0.500	0.000	**0.880**	**0.830**	0.450
	F1	0.890	**0.850**	0.000	0.000	0.460	0.000	**0.820**	**0.830**	**0.430**
Llama3-Instruct (Multi-Options)	**P**	0.850	0.500	0.090	**0.110**	0.090	0.000	0.710	0.490	0.340
	R	0.710	0.680	0.500	**0.800**	0.500	0.000	0.130	0.550	0.470
	F1	0.770	0.580	0.150	**0.200**	0.150	0.000	0.220	0.510	0.300
Llama3-Instruct (Single-Option)	**P**	0.830	**0.870**	**0.230**	0.000	**0.240**	0.000	0.670	0.700	0.400
	R	0.540	0.530	0.380	0.000	**0.670**	0.000	0.160	0.430	0.320
	F1	0.650	0.660	**0.290**	0.000	**0.350**	0.000	0.260	0.530	0.310

Table 4: The models' F1 score performance on identifying all the email intents.

B Propensity Score Matching Details

Distribution of Covariates and Outcomes. To make fine-grained analysis of propensity scores of all factors, we also compute all the propensity scores and visualize their distributions. We visualize the Kernel Density Estimate plots of various features as normalized scores for Politeness, Certainty, and Intimacy factors in Figure 10. More details of the social factors' score distributions can be found in Figure 13,11, and linguistic factors' score distributions can be found in Figure 8, 9, 12, 10. We also visualize the correlations between the independent variables, which are generally fairly low (Figures 14-15).

Prompt Type	Prompt
Multi-Options	For the following email:{Email}. Please classify the email's intent into the following categories: 1. Sharing: The sender is sharing information, opinions, updates, or other content with the recipient(s). This includes (but is not limited to) answering a question posed by the recipient; sharing background information to ask a question; stating opinions or takes; sharing documents/external links/resources/references/sample code; or proactively sharing FYI messages and updates. 2. Requesting: The sender is asking the recipient(s) for something. This includes (but is not limited to) asking a question, asking for more information, or requesting a task or action be completed. Only EXPLICIT requests should be annotated. 3. Promising: The sender writes that they will perform some future action. This includes (but is not limited to) saying they will look into a problem, rewrite some code, fix a bug, give an updated in the future, or do some joint activity with the recipient. 4. Personal Communication: The sender makes some interpersonal investment in the recipient(s), beyond the subject of the email. Includes indicators that they have a relationship outside the email thread, personal self-disclosures, engagement with the recipient's personal feelings/worldviews, or other communication that goes beyond the subject of the email and addresses the recipient as a person. Does not include pleasantries like please, thank you, expressions of gratitude, etc. 5. Pleasantries or Acknowledgement Only: The email contains only pleasantries, formalities, or acknowledgement of the prior email (e.g., the whole email is a 'Thank you very much' or a 'Will try that!' or 'Got it!') – or if pleasantries/formalities/acknowledgements like these are the main point of the message and all other content is unimportant or secondary. This is mutually exclusive from the rest of options. 6. Auto-generated or Spam: Email digests, auto-reminders, github or auto-change notification messages, do not reply emails, scam emails, etc. This is mutually exclusive from the rest of options. Respond with all the applicable categories without explanation.
Sharing	For the following email:{Email}. Does the email's intent align with "Sharing"? "Sharing" means the sender is sharing information, opinions, updates, or other content with the recipient(s). This includes (but is not limited to) answering a question posed by the recipient; sharing background information to ask a question; stating opinions or takes; sharing documents/external links/resources/references/sample code; or proactively sharing FYI messages and updates. Respond with only yes or no.
Requesting	For the following email:{Email}. Does the email's intent align with "Requesting"? "Requesting" means the sender is asking the recipient(s) for something. This includes (but is not limited to) asking a question, asking for more information, or requesting a task or action be completed. Only EXPLICIT requests should be considered. Respond with only yes or no.
Promising	For the following email:{Email}. Does the email's intent align with "Promising"? "Promising" means the sender writes that they will perform some future action. This includes (but is not limited to) saying they will look into a problem, rewrite some code, fix a bug, give an updated in the future, or do some joint activity with the recipient. Respond with only yes or no.
Personal	For the following email:{Email}. Does the email's intent align with "Personal Communication"? "Personal Communication" means the sender makes some interpersonal investment in the recipient(s), beyond the subject of the email. Includes indicators that they have a relationship outside the email thread, personal self-disclosures, engagement with the recipient's personal feelings/worldviews, or other communication that goes beyond the subject of the email and addresses the recipient as a person. Does not include pleasantries like please, thank you, expressions of gratitude, etc. Respond with only yes or no.
Pleasantries	For the following email:{Email}. Does the email's intent align with "Pleasantries or Acknowledgement Only"? "Pleasantries or Acknowledgement Only" means the email contains only pleasantries, formalities, or acknowledgement of the prior email (e.g., the whole email is a 'Thank you very much' or a 'Will try that!' or 'Got it!') – or if pleasantries/formalities/acknowledgements like these are the main point of the message and all other content is unimportant or secondary. Respond with only yes or no.
Spam	For the following email:{Email}. Does the email's intent align with "Auto-generated or Spam"? For example, email digests, auto-reminders, github or auto-change notification messages, do not reply emails, scam emails, etc. Respond with only yes or no.
ResponseExpected	For the following email:{Email}. Does email sender explicitly expects an email in reply? E.g., to answer a question, engage with content/attachments and send a follow up email, take some action outside the email chain and send a confirmation that the action was taken, etc. Only annotate if the sender explicitly expects a response. Please respond with only yes or no.

Table 5: Prompts for Generative Language Model Zero-shot Inference.

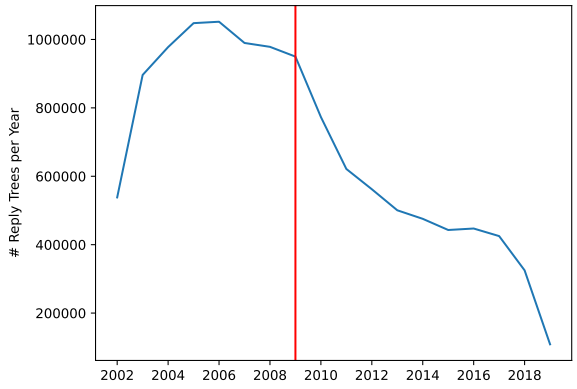


Figure 7: The number of conversations (reply trees) on GMAME each year from 2001 - 2019. The volume of conversations is relatively stable from 2007-2009 but then sharply drops off from 2009 to 2010. Even though the platform contains data through 2019, we chose to analyze data from 2009 as we are unaware of why these declines occur. Utilizing 2009 was the best way to ensure the most recent and reliable data.

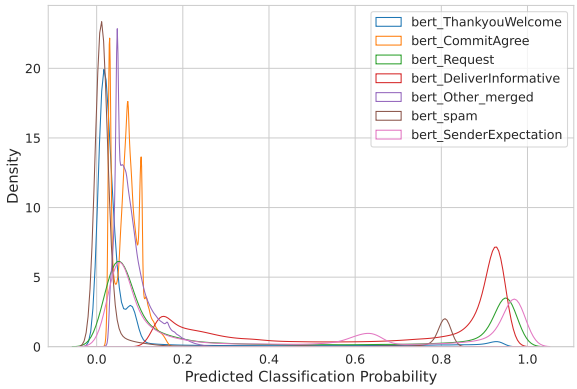


Figure 8: KDE plots for BERT predictions

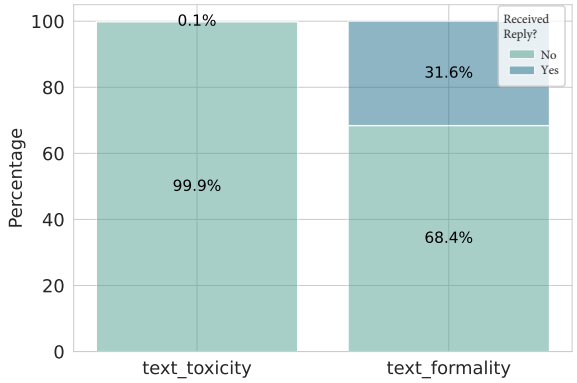


Figure 9: Distributions of toxicity and formality scores in main body texts.

Propensity Score Distributions. The distribution of propensity scores in the full and matched samples is given in Figure 16.

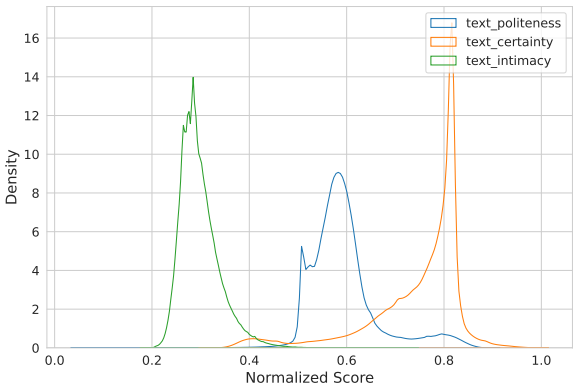


Figure 10: KDE plots for politeness, certainty, and intimacy scores of main body texts.

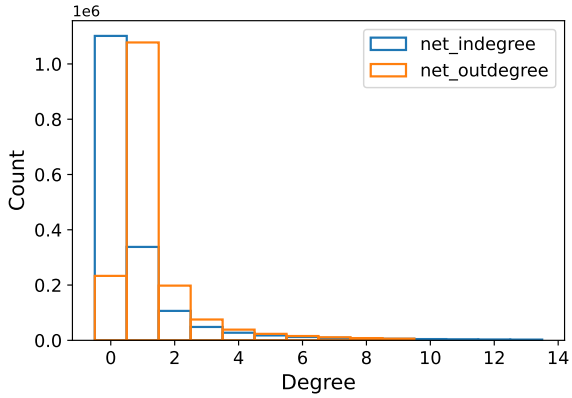


Figure 11: Counts of in-degree and out-degree of network nodes.

Outcomes of Interest The 14 outcomes of interest are described in Table 6.

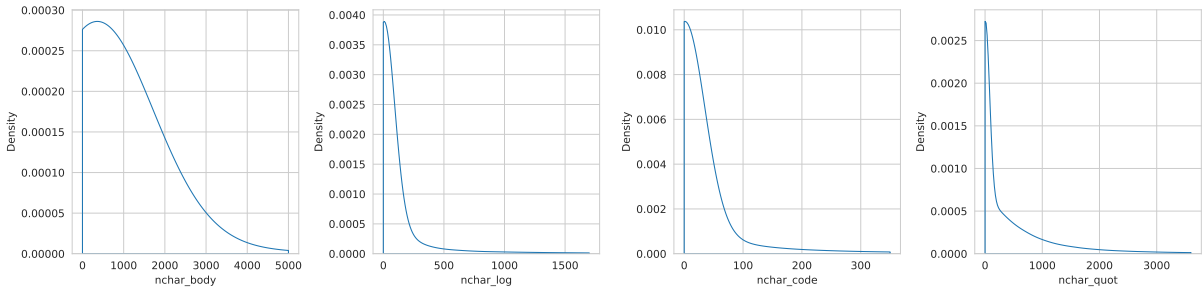


Figure 12: KDE plots for emails with respect to the number of characters of the main body texts.

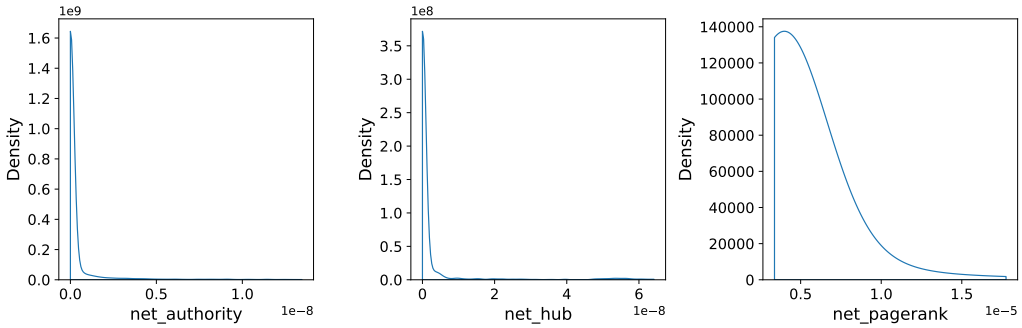


Figure 13: KDE plots for hub, authority and PageRank scores of users.

Social Features	Features	Definitions
Argumentation	Cogency	logical strength of argument (e.g., are justifications adequate) (Ng et al., 2020; Falk and Lapesa, 2023)
	Effective	rhetorical strength of argument (e.g., emotional, appropriate) (Ng et al., 2020; Falk and Lapesa, 2023)
	Quality	how well the argument makes its intended claim (Swanson et al., 2015; Falk and Lapesa, 2023)
	Clarify	argument's ease of interpretation (Gretz et al., 2020; Falk and Lapesa, 2023)
Social Status	Formality	formality of style, as judged by annotators (Babakov et al., 2023)
	Politeness	politeness of tone, as judged by annotators (Bao et al., 2021)
	Out-Degree	the number of emails sent
	Pagerank	the importance of a node, measure by the importance of nodes pointing to it (Page et al., 1999)
	Hub	a node's access to expertise, measured by the authority of nodes it points to (HITS) (Kleinberg, 1999)
	Authority	a node's expertise, measured by the hub score of nodes pointing to it (HITS) (Kleinberg, 1999)
Social Connection	Intimacy	how personal, deep, self-disclosing an email is (Pei and Jurgens, 2020)
	Toxicity	toxicity of tone and content, as judged by annotators[7]
	Clustering Coefficient	interconnectedness of an author's ego-network (Newman, 2003)
	Reciprocity	how often authors receive replies from people they reply to, where high reciprocity along a given edge suggests a strong tie (Garlaschelli and Loffredo, 2004)

Table 6: The linguistic and network factors that are studied in the causal analysis. Citations indicate which models were used to generate the features. Linguistic features are in orange while network features are in blue.

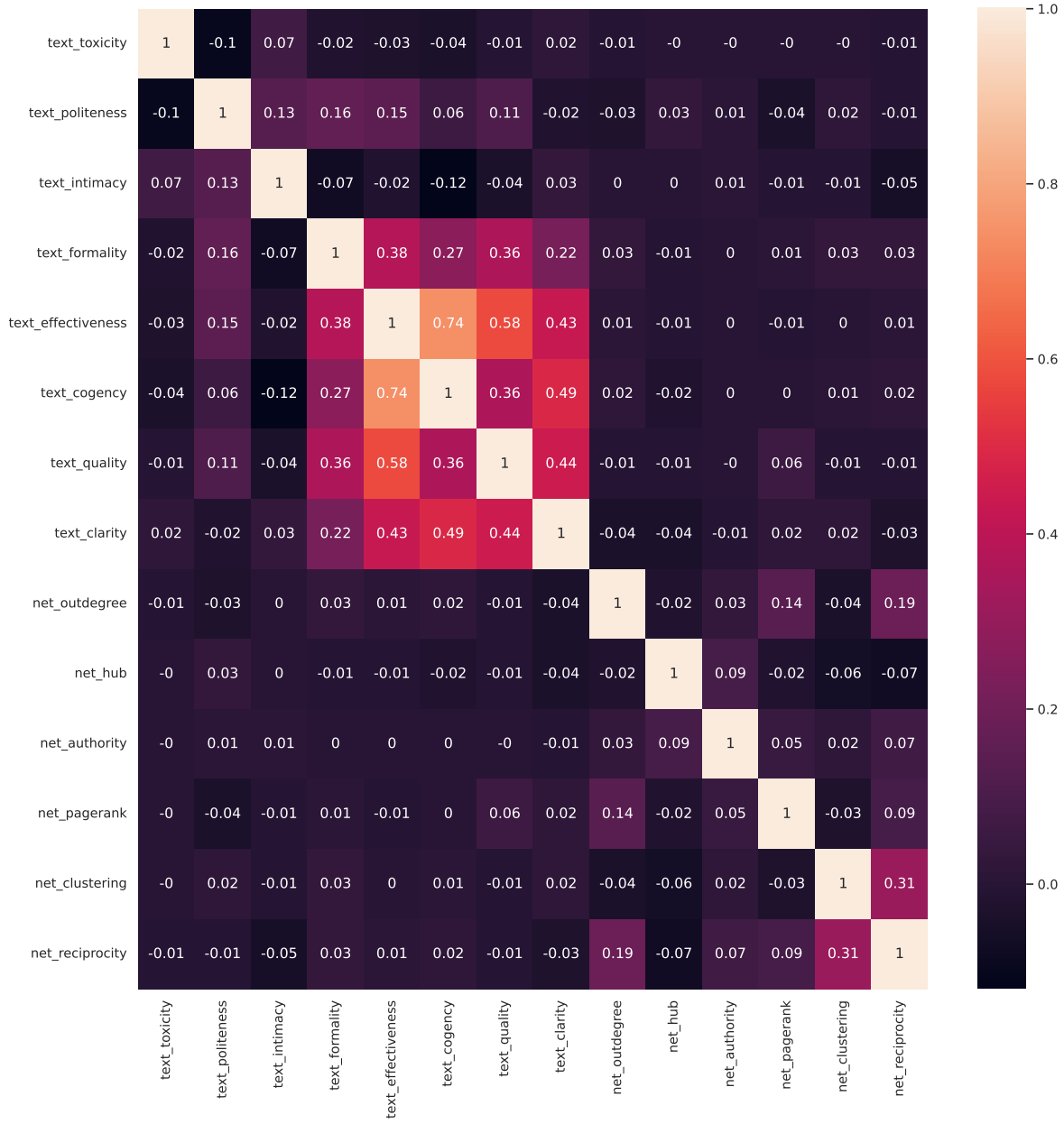


Figure 14: Correlation between features in the matched dataset of initial emails.

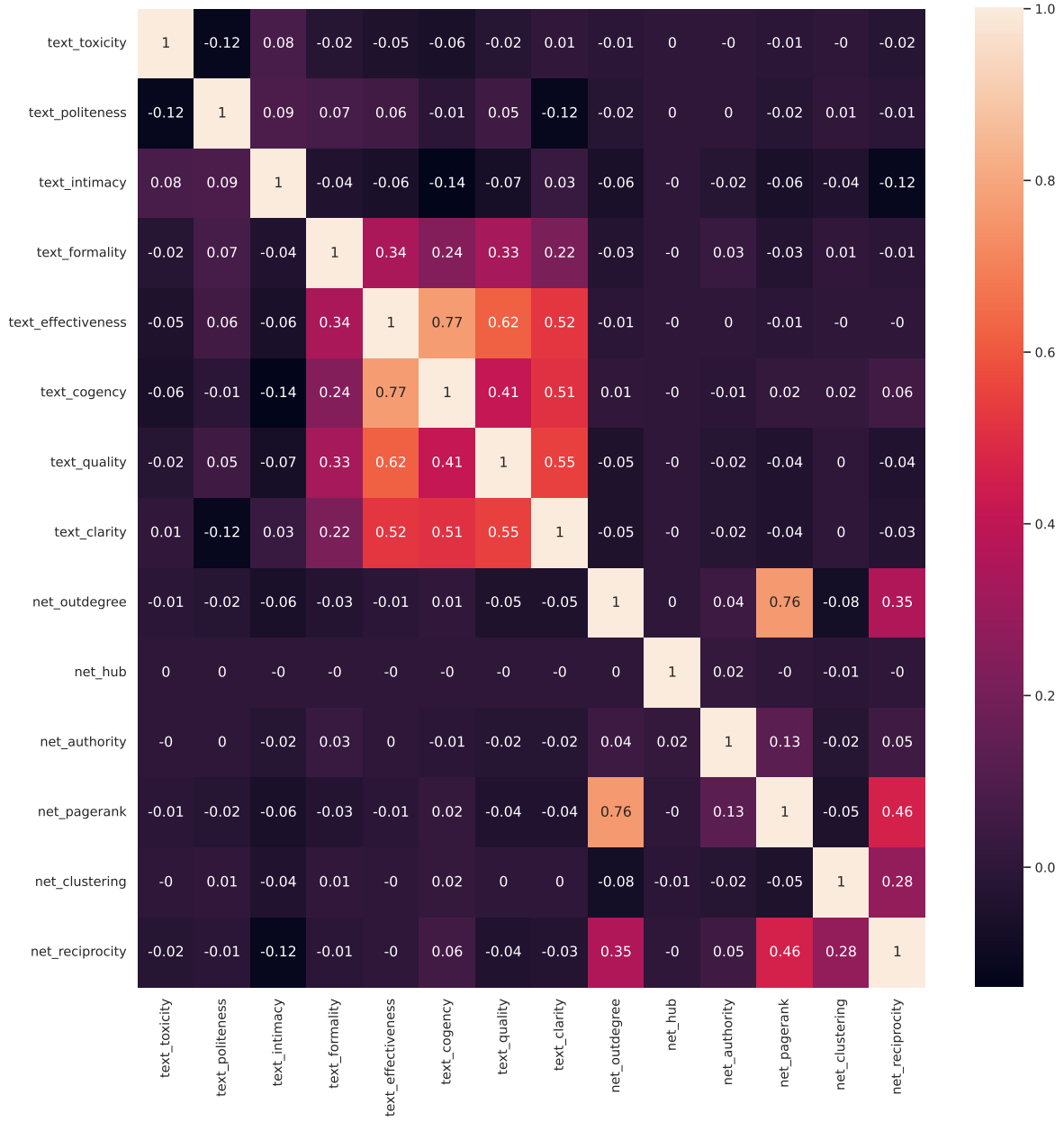


Figure 15: Correlation between features in the matched dataset of response emails.

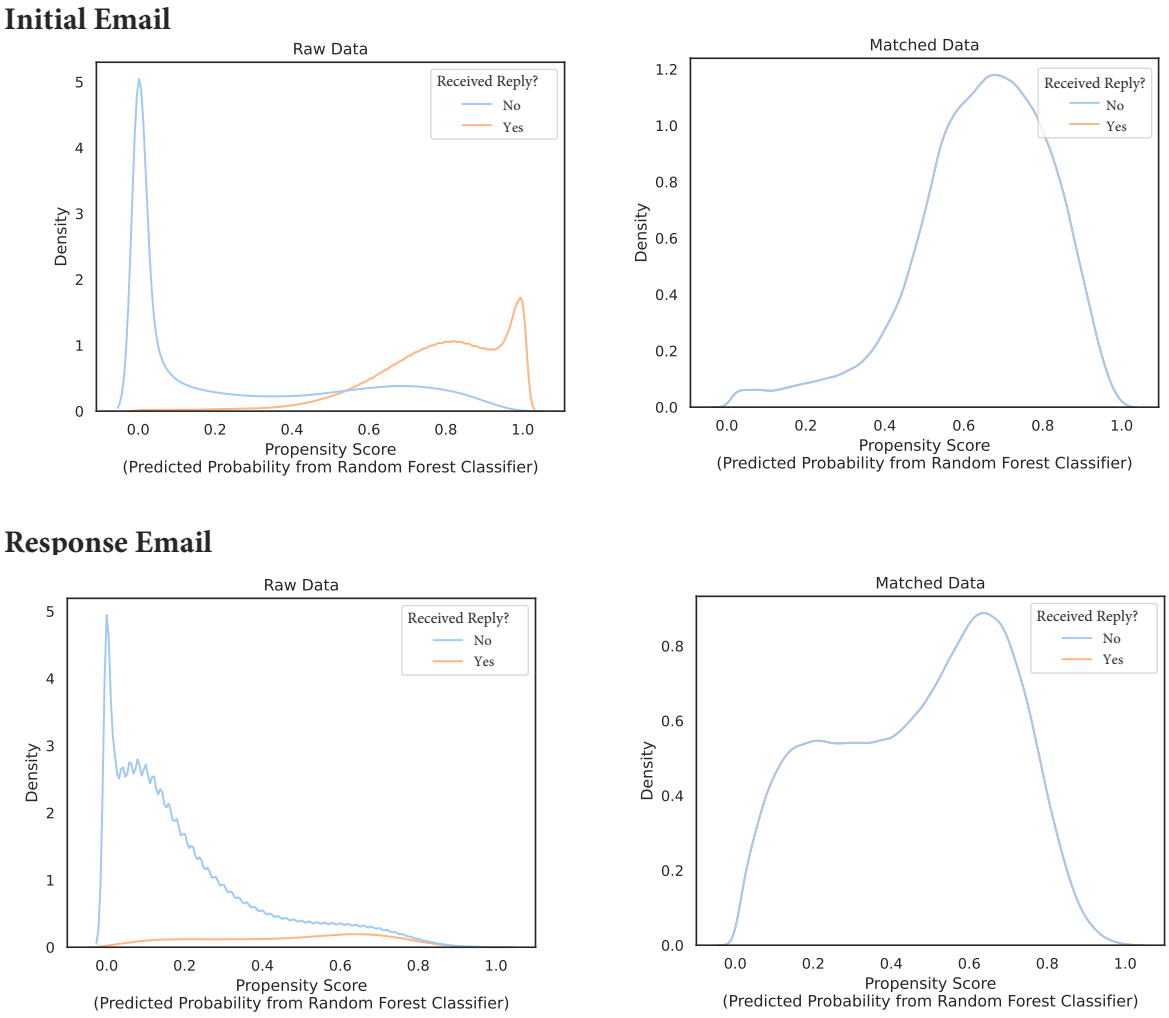
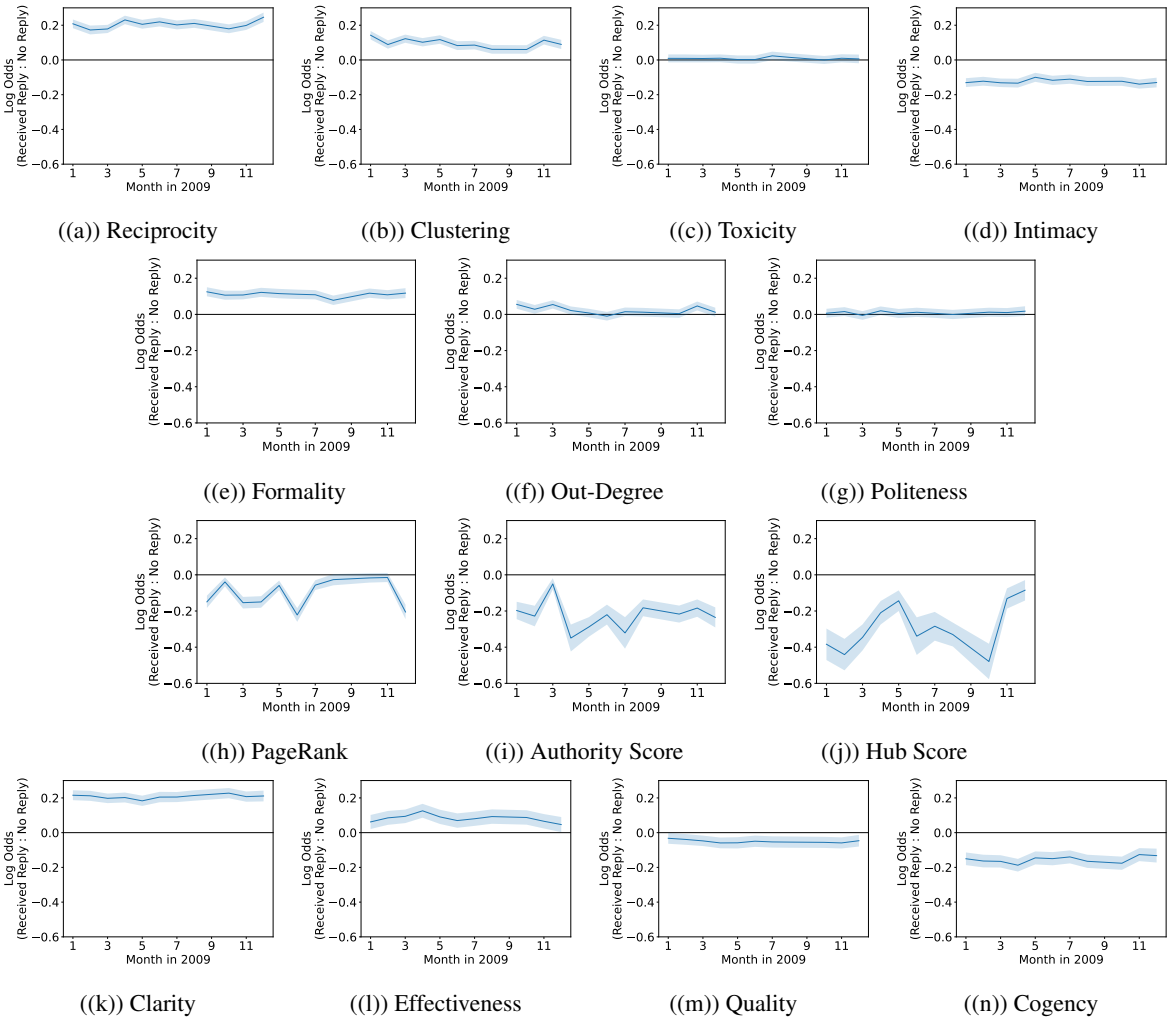


Figure 16: Propensity scores in full and matched samples. (The two density curves in the matched sample are very similar and, therefore, appear overlapping.)



((a)) Reciprocity ((b)) Clustering ((c)) Toxicity ((d)) Intimacy

((e)) Formality ((f)) Out-Degree ((g)) Politeness

((h)) PageRank ((i)) Authority Score ((j)) Hub Score

((k)) Clarity ((l)) Effectiveness ((m)) Quality ((n)) Cogency

Figure 17: Regression coefficients for initial emails exhibit minimal temporal variation. In this plot, the regression from Figure 5a is fit separately for matched data from each month. Each regression coefficient is plotted over time.

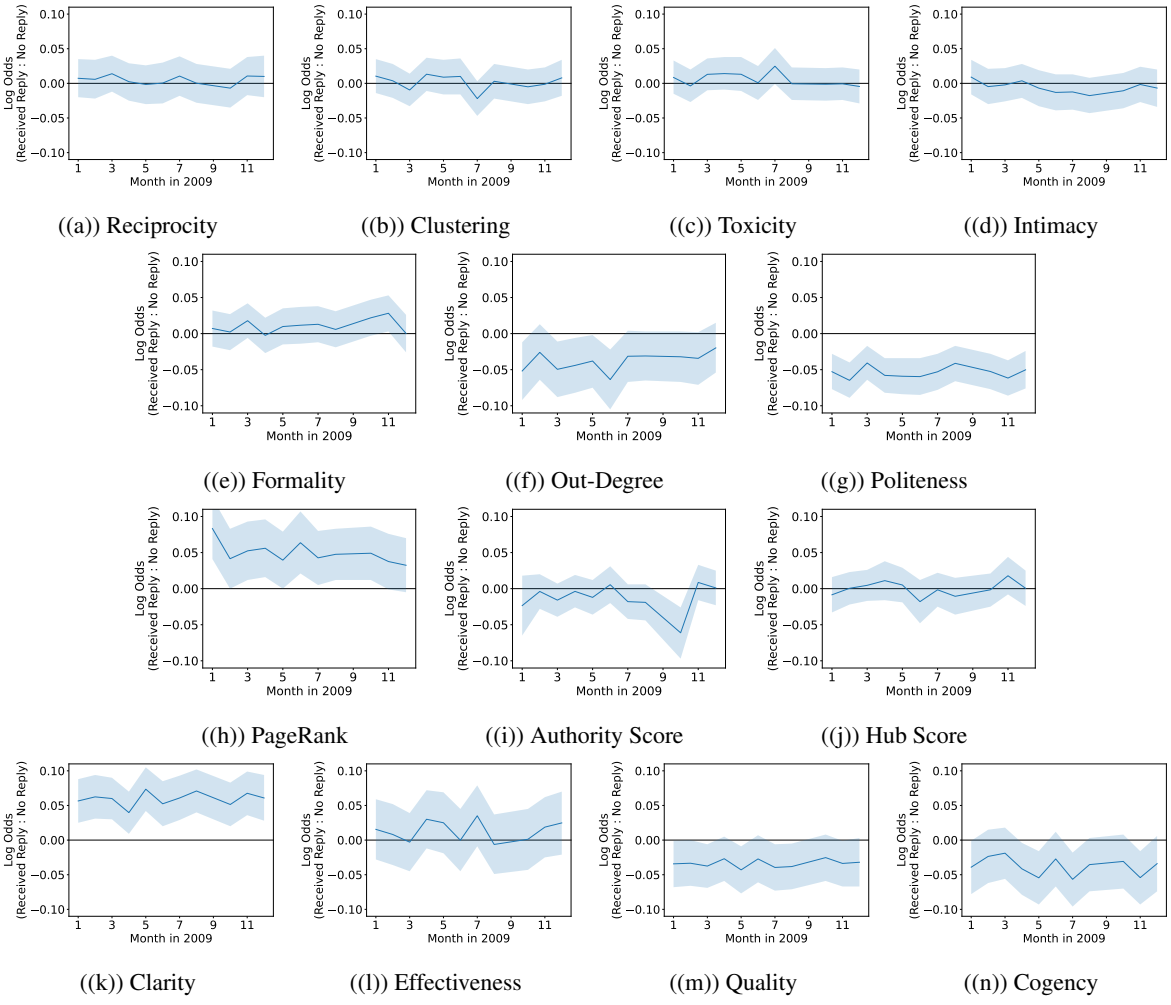


Figure 18: Regression coefficients for response emails exhibit minimal temporal variation. In this plot, the regression from Figure 5b is fit separately for matched data from each month. Each regression coefficient is plotted over time.