BCC'ing AI: Using Modern Natural Language Processing to Detect Micro and Macro E-ggressions in Workplace Emails

Kelsi E. Cornett

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science

In

Psychology

Ivan Hernandez, Chair

Anna-Katherine Ward- Bartlett

Charles Calderwood

May 10th, 2024

Blacksburg, Virginia

Keywords: Microaggressions, Workplace Mistreatment, NLP, Diversity

BCC’ing AI: Using Modern Natural Language Processing to Detect Micro and Macro E-ggressions in Workplace Emails

Kelsi E. Cornett

ABSTRACT

Subtle offensive statements in workplace emails, which I term "Micro E-ggressions," can significantly impact the psychological safety and subsequent productivity of work environments despite their often-ambiguous intent. This thesis investigates the prevalence and nature of both micro and macro e-ggressions within workplace email communications, utilizing state-of-the-art natural language processing (NLP) techniques. Leveraging a large dataset of workplace emails, the study aims to detect and analyze these subtle offenses, exploring their themes and the contextual factors that facilitate their occurrence. The research identifies common types of micro e-ggressions, such as questioning competence and work ethic, and examines the responses to these offenses. Results indicate a high prevalence of offensive content in workplace emails and reveal distinct thematic elements that contribute to the perpetuation of workplace incivility. The findings underscore the potential for NLP tools to bridge gaps in awareness and sensitivity, ultimately contributing to more inclusive and respectful workplace cultures.

BCC’ing AI: Using Modern Natural Language Processing to Detect Micro and Macro E-ggressions in Workplace Emails

Kelsi E. Cornett

GENERAL AUDIENCE ABSTRACT

Subtle offensive statements in workplace emails, which I term "Micro E-ggressions," can significantly impact the psychological safety and subsequent productivity of work environments despite their often-ambiguous intent. This thesis investigates the prevalence and nature of both micro and macro e-ggressions within workplace email communications, utilizing state-of-the-art natural language processing (NLP) techniques. Leveraging a large dataset of workplace emails, the study aims to detect and analyze these subtle offenses, exploring their themes and the contextual factors that facilitate their occurrence. The research identifies common types of micro e-ggressions, such as questioning competence and work ethic, and examines the responses to these offenses. The results show a high occurrence of offensive content in workplace emails and highlight patterns that help maintain a negative work environment. The study demonstrates that advanced language analysis tools can help raise awareness and sensitivity, ultimately fostering more inclusive and respectful workplace cultures.

Table of Contents

Introduction

Subtly offensive statements are commonplace, and there is a fine line between complacency and complicity when they are tolerated in the workplace (Ruggs et al., 2011). Even if these comments are not approved of, they are often accepted due to ambiguity of intent (Marshburn et al., 2017). These statements may highlight a characteristic or feature that the recipient is sensitive about but not everyone shares. Therefore, the subtlety of the statements does not detract from their potency to cause offense and harm.

These offenses are similar to those coined by the term "microaggressions" in that they are subtle, often automatic, unintentional, or intentional statements that could communicate negative, hostile or derogatory slights toward an individual (Pierce et al., 1978). However, microaggressions target an individual's social identity such as gender, ability-status, ethnicity, sexual orientation, or other identities that the perpetrator ascribes to the target (Ross-Sheriff, 2012; Sue et al., 2007). For this paper, I will be looking at a broader type of subtle offenses, including but not limited to offenses tied to individuals' social identity.

I am looking at the broad conceptualization of subtle offenses due to the complexity of contexts that facilitate these offenses. A subtle offense, due to its subtle nature, can target many aspects of a person and make it hard for offended parties to label and communicate why it was offensive for one reason. Additionally, the statements may be so subtle that correcting the deliverer may be difficult due to their ambiguous motivation or the possibility of being denied. In order to aid in a more aligned understanding, it is paramount to understand the nature of these offenses, how targets perceive them, and how well people infer them. The current paper seeks to address this issue by leveraging modern text analysis tools to understand under what contexts these offenses persist. Utilizing big data, I gain a realistic and holistic preview into aggressive or

hostile workplace behaviors committed either through or with information and communication technologies (ICTs) such as email (Weatherbee & Kelloway, 2006). I will be coining the term "Micro E-ggressions" in investigating these subtle, ambiguous email offenses. I am interested in detecting all forms of dyadic cyber offenses, but using my methodology, I will be able to capture previously undetected, subtle, Micro E-ggressive offenses. By illustrating the nature of these offenses and finding better ways to detect them, this research can potentially bridge the current awareness gap by creating interventions for minimizing future occurrences. Ultimately, the implications of my findings can be applied to workplace communication monitoring systems to facilitate more sensitive and inclusive cultures.

Literature Review

Current Understanding of the Problem

The study of offensive speech is largely interested in understanding the effects of offense (Aquino & Douglas, 2003; Kim et al., 2008; Cortina et al., 2001; Kessler et al., 2013; Bruk-Lee & Spector, 2006). In addition to its behavioral impacts, research is revealing long-term occupational stressors, like workplace offenses or incivility, impact psychological and physical health as well (Vance et al., 2004; Young et al., 2003; Lim et al., 2008; Holm et al., 2015; De Dreu et al., 2004). As the use of email is ubiquitous in today's workforce in facilitating the sharing of information and connectivity, it can also be a powerful tool to facilitate mistreatment and offensive communication. Employers are increasingly interested in monitoring email to curb mistreatment and their potential liability for providing a platform for these offenses (Friedman & Reed, 2007).

More work in investigating the content and context of the experiences themselves that permit these offenses is needed. A better conceptualization of the commonalities in email

communications can add to our understanding of people's experiences of Micro E-ggressions and Macro E-ggressions and improve our ability to detect and anticipate the possibility of offense.

Effects of Subtly Offensive Speech in the Workplace

While direct hate speech has been studied more extensively, one less studied form of offensive speech in the workplace is subtle, often unintentional remarks. Hate speech targets an individual or group based on some characteristic such as their ethnicity, gender, race, sexual orientation, religion, nationality or other characteristics and is more overt in its potential to cause offense and harm (Nockleby, 2000). However, because of norms and standards guiding direct language, offensive statements in the workplace are likely to be more covert and perhaps unintentional. This subtlety, however, could predispose them to be more commonplace and substantially harmful to the recipient and the work climate. These statements, while less explicitly offensive, have the same disempowering role as the other forms. A disempowering act is any verbal or nonverbal and intentional or unintentional behavior expressed in the workplace which can be construed by the disempowered employee as hostile, intimidating, demeaning, threatening, or offensive and can diminish performance or impede productivity (Young et al., 2003).

As a result of feeling disempowered, recipients of workplace offenses could engage in antisocial workplace behavior or counterproductive work behavior (CWB; Richard et al., 2020). These are actions towards employees or the organization that can produce physical, economic, or emotional harm. These behaviors can include but are not limited to employees engaging in insubordination, sabotage, spreading rumors, lying, withholding effort, and absenteeism (Penney & Spector, 2002; Robinson & O'Leary-Kelly, 1998). These behaviors, in addition to being a

problem for organizational productivity, are a symptom of a larger problem. CWB or antisocial workplace behaviors as a response to interpersonal workplace offenses reveal the long-lasting psychological impact of offensive remarks and the insensitivity of individuals to recognize potentially harmful remarks. In the workplace, sensitivity should be seen as a strength, reflecting one's cognitive and emotional responsiveness to interactions with others (Bunk & Magley, 2011).

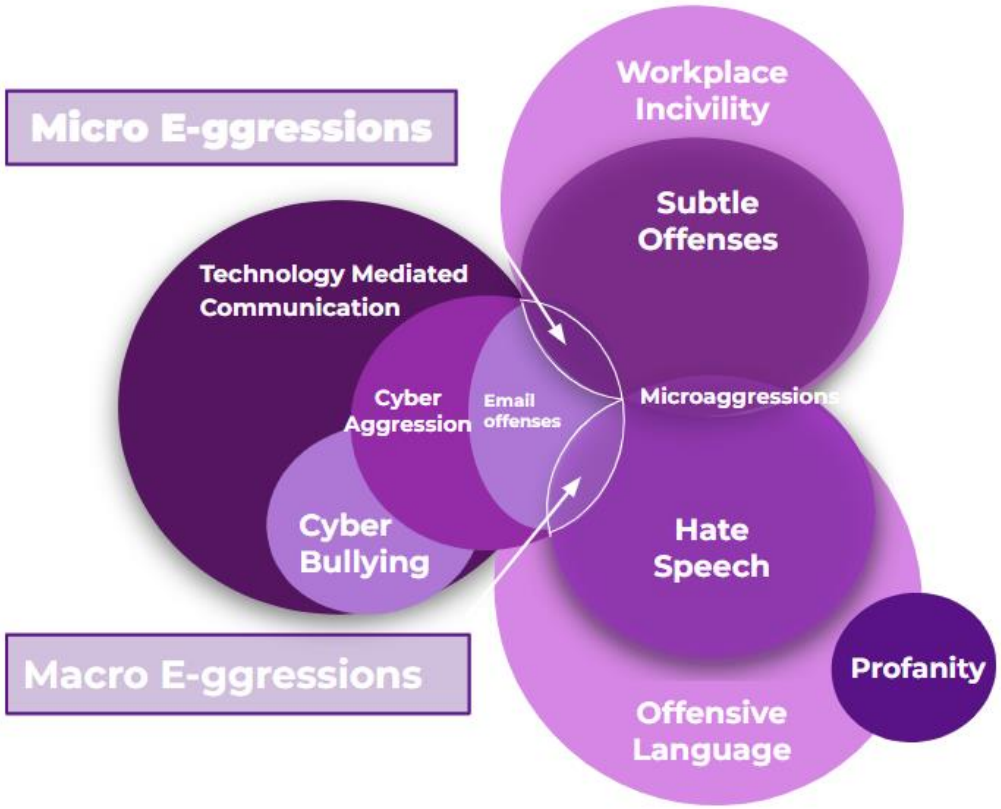
Defining Micro E-ggressions

Workplace offenses are incidents where an employee has been offended, hurt or unjustly treated by a coworker (Kim et al., 2008). The content of offensive language varies significantly, ranging from simple profanity to much more severe and targeted forms of expression (e.g., hate speech and insults; Nockleby, 2000; Sigurbergsson & Derczynski, 2019). Offensive language in the workplace can lead to the recipient feeling distressed, isolated, or harmed (Kocoń et al., 2021). The intention behind a statement is typically irrelevant to its classification as "offensive." Researchers consider offensive statements in the workplace as examples of "incivility" or low-intensity deviant (e.g., rude, discourteous) behavior that's intentionality to harm the recipient is ambiguous and is in violation of workplace norms for mutual respect (Andersson & Pearson, 1999; Pearson et al., 2005). Examples of incivility include being ignored, interrupted, ridiculed, or treated disrespectfully. Therefore, the current project focuses on a more specific and modern type of incivility: written cyber communications that subtly offend employees but lack overt derision and intent, which I call "Micro E-ggressions." These are different from cyberbullying, such that they do not require intentional harm, repetition of the offense, and an imbalance of power (Grigg, 2010). They are also different from cyber-aggression, as they are not always intentionally harmful. Similar to microaggressions, Micro E-ggression statements can be difficult

for the target to respond to (Sue, 2010). In addition, bystanders are less likely to support victims of these cyber microaggressions and subtle offenses (Coyne et al., 2019).

Figure 1

Defining the Problem



Existing Solutions to the Problem

Human Approaches to Improve Understanding

The illumination of the negative legal, behavioral, and health impacts of offensive interpersonal conflict has increased organizations' attention toward identifying and preventing potential workplace offenses (Ruggs et al., 2011). There is vast research into the effects of these offenses, despite little research existing on the nature of subtly offensive speech. Researchers

have attempted to study the negative interpersonal experiences people have had in a variety of ways, and those ways have various limitations.

Approaches to investigating and intervening in these offenses through prejudice reduction have been labeled T-group, diversity training, cultural competence training, sensitivity training, anti-bias training, and cognitive and emotional training (Dimock, 1971; Paluck et al., 2021). These training methods assume that simply understanding the historical context that shapes how different identities relate to social status and power can better equip people to identify offensive behavior (Marshburn et al., 2017). These interventions are expensive, time-consuming, hard to adapt, poorly structured, and often require several follow-ups. Most importantly, there is not enough evidence that practitioners should use these training methods as an efficacious method to reduce biases and offensive behaviors (Paluck et al., 2021). The highly inconclusive efficacy of these approaches has led some organizations to abandon them and encouraged research to develop more evidence-based solutions to detecting and preventing workplace offenses.

More recent research has investigated the role situational judgment tests (SJTs) could play in measuring our awareness of discriminatory speech and microaggressions (Sturdivant et al., 2017). Although this is a step to increasing awareness of the contents of offensive speech, microaggression-specific SJTs have not yet been implemented largely as aids in sensitivity training in workplaces. Instead, SJTs implementation has dominated selection programs when measuring academic or non-job-related measures like empathy, reliability, and dependability (Patterson et al., 2012; Tiffin et al., 2020). To develop effective SJTs, especially for less-studied subtle offenses, more research into the context and content is required to understand the nature of the problem.

Computer Approaches to Natural Language Understanding

Natural Language Processing (NLP) or Computer-Assisted Text Analysis (CATA) afford the opportunity for computers to understand complex narratives (Campion et al., 2016). Natural Language Processing (NLP) dates back to the 1950s with the goal that deep NLP will lead to complete natural language understanding (Allen, 1987). NLP requires high-level symbolic capabilities to shift to natural language understanding. These capabilities encompass the propagation and creation of dynamic bindings; the acquisition and retrieval of lexical, semantic, and episodic memories; the manipulation of recursive structures; the management of multiple learning modules and the transfer of information between them; and the grounding of basic language constructs (such as objects and actions) in perceptual and motor experiences (Dyer, 1994). This computational and statistical decomposition of text into its fundamental ideas can be utilized for indexing and searching extensive texts, categorizing text, retrieving information, extracting information, translating languages, acquiring knowledge, summarizing texts, answering questions, and generating texts or dialogues (Chowdhary, 2020).When coupled with large datasets, text analysis is capable of building robust models that can learn from the lived experiences of humans and form a generalizable understanding of the larger world in ways that may not be apparent to human annotators with finite time and attention.

Historically, text analysis methods used a "bag of words" approach, which treats each word as independent from other words when deriving the text's overall meaning (Schwarz and Ungar, 2015). Today, our cutting-edge methods and deep neural network architectures like BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019) (collectively known as "transformers") not only take into account word order and surrounding sentence context when inferring meaning, but they also leverage billions of prior examples to

human writing as the foundation of their language understanding (Vaswani et al., 2017). These new methods have brought us the closest we have ever been to using AI to understand natural language. When applying these modern models to new data, they use their prior understanding of the relationships between words and their interchangeability (i.e., synonyms) to provide greater predictive accuracy, efficiency, and consistency into the structures of language (Bender & Friedman, 2018; Malte & Ratadiya, 2019). Computer-based models provide a unique insight that can surpass the shallow understanding humans may have due to our limited processing capacities, which would have difficulty understanding regularities and patterns across a large corpus of diverse statements.

Some analyses facilitated by modern text analysis methods are topic modeling, text classification, and prediction explanation. Topic modeling involves reducing statements/documents into coherent and meaningful groupings (Blei et al., 2003). These methods serve as a dimension reduction technique, in which, similar to exploratory factor analysis, a collection of texts' underlying themes can be explored to label the meaning of these common themes, and the frequency of texts belonging to these themes (Kobayashi et al., 2018). Transforming and embedding sentences can use more simple methods to determine clusters, such as hierarchical clustering or K Means.

Another application of modern natural language processing models includes text classification/regression, in which a nominal/continuous valued property about the text (e.g, sentiment, clarity, tone) is inferred using only the content (Speer, 2021). As of 2022, neural network models currently demonstrate the highest predictive performance on 96 of 98 benchmarks as tracked by the PapersWithCode leaderboard[1]. This predictive performance can

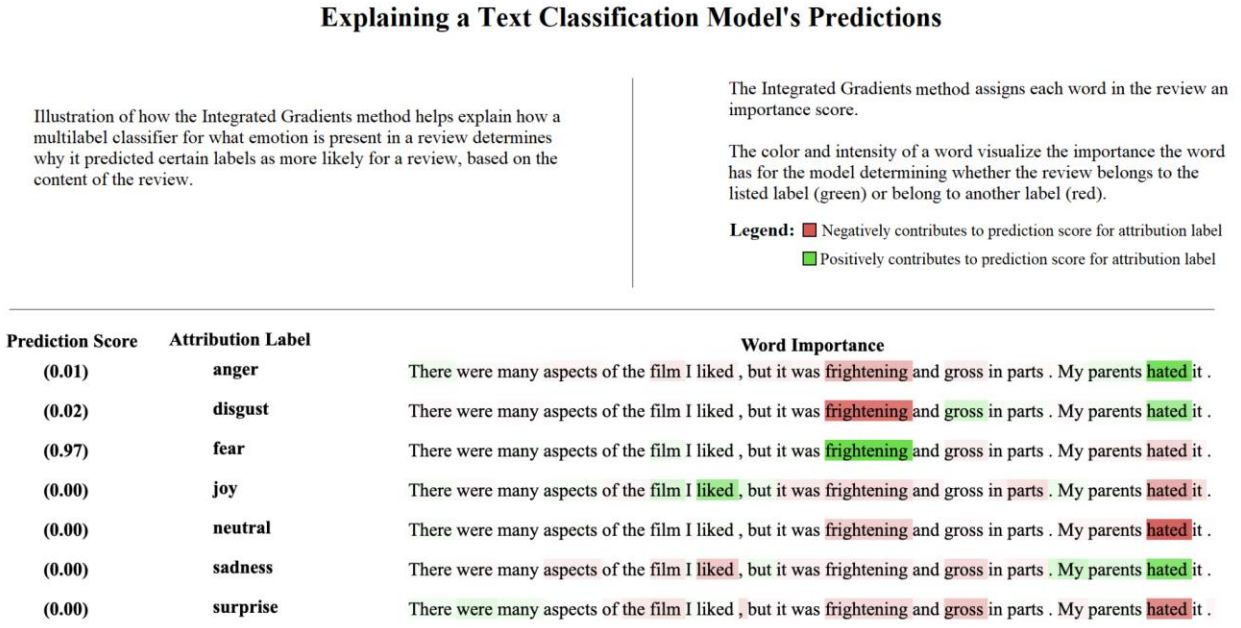
[1] https://paperswithcode.com/task/text-classification

offer societal benefits when tasks are difficult for humans. Similar to a calculator for an accountant or spell-check for a writer, these accurate tools can supplement knowledge deficiencies or provide greater consistency for tasks of great importance. Therefore, highly accurate models can potentially bridge the gap between intention and offense in the workplace.

Explainability/interpretability algorithms for machine learning models are a new tool for facilitating theory development. Historically, the complexity of machine learning models has constrained analysts' ability to comprehend them (Caruana et al., 2015). With no understanding of the internal workings, researchers using these complex black-box models have difficulty developing theory. Further, the inability to understand *which* features have *what* effects on the output of the model may decrease trust in these methods in both applied and research settings. Many algorithms, such as Gradient SHAP (Kokalj et al., 2021; Lundberg & Lee, 2017), DeepLIFT (Shrikumar et al., 2017), LIME (Ribeiro et al., 2016), and Integrated Gradients (Sundararajan et al., 2017) solve this explainability limitation by examining how changing the inputs affect the predicted values. By applying local changes (i.e., words) to contextually aware models, each local feature receives an importance score that accounts for its interdependence with other features (Figure 2). Therefore, we researchers are able to understand the positive and negative relationship that words have with outcomes of interests without separating those terms from the context in which they were used.

Figure 2.

Illustration of how model explanation methods can describe the relationship between input features and outcomes of interest.



Note. This figure demonstrates how the inferences of neural network natural language models can be understood in terms of their positive and negative relationships to the outcome variables.

Prior Computer Approaches to Understanding Offense

Recently, researchers have started applying machine learning and natural language processing models to the topic of offensive speech. Typically, these models are applied to detect hate speech on social media (Davidson et al., 2017; Kocoń et al., 2021; Mansourifar et al., 2021; Schmidt & Wiegand, 2017; Sigurbergsson & Derczynski, 2019; Waseem & Hovy, 2016; Zampieri et al., 2019). One example of overt offense research compiled tweets identified as offensive by internet users on hatebase.org, and then the Twitter API was used to search for tweets containing terms from the lexicon. This model looked to differentiate hate speech from offensive language (Davidson et al., 2017). Similar to Davidson's data collection method, researchers created a model using tweets containing phrases often found in offensive text (e.g., "you are") using Twitter's API and labeled them using a crowdsourcing platform (Zampieri et al., 2019). This model utilized a three-tier hierarchical annotation framework to ascertain (1) if

the text is offensive, (2) the nature of the offense (targeted or untargeted), and (3) its intended target (group, individual, or other). Another Twitter n-gram-based classification model with two human annotators was developed to group hate speech as race-based, sex-based, or neither (Waseem & Hovy, 2016). Looking outside of simple word content, they investigated how factors like geographic location, demographics, and word length distribution influenced the predictive model's performance in detecting hate speech but found that only knowledge of gender improved performance (Waseem & Hovy, 2016). More recent work eliminated the use of human annotators by the performance of DT, SVM, KNN, and RF models to group text by the categories toxicity, severe toxicity, identity attack, insult, and profanity (Mansourifar et al., 2021). Hate speech research is not limited to the English language but has included Danish (Sigurbergsson & Derczynski, 2019), Indonesian (Ibrohim & Budi, 2019), Arabic (Mulki et al., 2019), Italian (Bosco et al., 2018), Spanish (Pereira-Kohatsu et al., 2019), Polish (Ptaszyński et al., 2019) and multilingual data sets (Vidgen & Derczynski, 2020; Aluru et al., 2020).

Prior work in detecting microaggressions (Ali et al., 2020; Breitfeller et al., 2019) have used self-reported accounts of microaggressions using the crow-sourcing website www.microaggressions.com to train their models. This website has Tumblr users provide the context of the microaggression and how it made them feel. The researchers classified these as SelfMA (self-reported microaggressions), and they trained a model using these SelfMA with non-offensive, offensive (hate speech), and random posts manually selected from Reddit. They trained their model to classify types of microaggressions into themes of attributive, institutionalized, teaming, and othering (Breitfeller et al., 2019). More research using the same source on Tumblr focused on dichotomizing racial microaggressions from non-racial microaggressions (Ali et al., 2020). However, they did not include hate speech or non-offensive

statements in training their model. Rather, their focus was on whether different types of microaggressions could be distinguished.

Researchers have also attempted to compile microaggressive experiences to facilitate subsequent training models. The most recent work in building a more comprehensive dataset for microaggression detection has collected microaggressions from annotating popular American television and text mining on the same Tumblr microaggressions website as previous works (Washington, 2021). This multimodal dataset can potentially provide additional information used to increase the predictive validity of microaggression prediction models.

Limitations of Prior Approaches

Current solutions to understand and detect offensive language (1) are not workplace oriented, (2) do not include subtle offense, (3) lack exploration of the nature of subtle offense (4) have generalizability issues, and (5) neglect the predictive capabilities of state-of-the-art models. Workplace-oriented studies focus on the symptoms of offense and do not understand the context and content of the offense in order to facilitate awareness and prevention. In addition, these studies do not focus on workplace information and communication technologies. Furthermore, previous work largely neglected subtle offensive statements and instead focused on more overt offenses on social media. If the subtle offense is investigated, it is not workplace-oriented or does not integrate the predictive capabilities of AI. Research is needed to explore how offenses manifest and how well people can infer offense to justify examining watts to anticipate it. Previous NLP models detecting hate speech have a generalization deficiency, as a classifier trained on a hate speech dataset cannot perform well on another dataset. Generalizability to other types of offense (i.e., microaggressions) or for other platforms for speech (i.e., formal speech as opposed to social media-derived informal speech) is low. Current models also fail to include

actual perceptions of offense but instead look at something the general consensus finds offensive. This leaves out information that could be gained through subjective offense as well as the role individual or group differences could play in interpretation. Another issue that computer models lack is a human baseline. The accuracy of models is compared to chance or other models. Failing to compare to human annotators fails to answer the question of how computer-centered approaches can improve our human understanding. Not only do most models fail to have a human baseline to compare to, but they also do not include control variables in training their models (i.e., they only train a model to know hate speech and never neutral or positive text, failing to understand the nuances of language). Finally, most of the models created to detect offensive language do not utilize the most up-to-date models due to technology advancing faster than practical implementation is possible. The subtle nature of the offense may benefit from the predictive power of modern NLP methods to ameliorate the gap.

Overview of the Current Approach

I propose applying state-of-the-art natural language models to aid in the understanding of subtly offensive speech and highlight the role it can serve in minimizing the gap in human sensitivity and awareness. I describe the various gaps in understanding subtly offensive speech that can be bridged through collecting these unique workplace experiences and applying natural language processing models in ways not yet applied to this context.

Benefits Offered by Approach

The current approach suggests addressing the limitations of prior studies by emphasizing the more subtle and commonplace types of offensive statements. Specifically, I seek to detect incivility experiences that are not explicit hate speech, direct insults, or unambiguous harassment. I am not only looking at speech which overtly targets an individual's social identity

(e.g., ethnicity, gender, etc.), but all possible communications that could offend someone. My model can capture the nuances in offenses that prior work could not. These types of subtle statements, while potentially more common, are also potentially more difficult to detect; therefore, the nature of the topic lends itself to methods that optimize detection performance.

Prior research emphasized obtaining offensive experiences through social media and rating those statements using third-party coders. Rather than have raters collectively determine offense, I used the real-life instances and ratings of offenders to train my model. Thus, the predictive model trained on this data is helpful for understanding the full weight that words can carry. My model was trained using a robust data set with uncensored, real-lived work email communications. This approach does not rely on subjects remembering or reliving experiences, but instead the objective, real occurrences.

The final novel aspect of this approach is using state-of-art machine learning models to detect patterns that distinguish offensive from non-offensive statements. Historically, text analysis within I-O psychology is addressed using dictionary-based approaches that examine the presence of words known to be positive or negative (Pennebaker et al., 2015). These approaches, while computationally simple, ignore word sequences and positions. Currently, in computer science, neural network approaches offer the highest performance for text classification. The most performant class of models are “transformer models” which are pre-trained to have a general ability to process text to draw inferences based on the combination and order of words present in the text. Specifically, these models employ a technique called Masked Language Modeling to get a better understanding of the general structure of a language In this task, a large corpus (e.g., Wikipedia, Twitter) is provided to the model with some words hidden (i.e., masked). The model must infer the masked words based on the context provided by the

surrounding unmasked words. This task optimizes the weights in the attention layers to place emphasis on terms and sentence positions that are indicative of semantic meaning. My approach takes into account word order and meaning and can distinguish between statements that are semantically and syntactically similar yet differ in meaning and offense. My model was trained and tested against human raters and can provide a more reliable rating of offense than their counterparts (i.e., humans) when scoring workplace emails.

Research Questions

My first research goal for my study is to explore how prevalent Micro and Macro E-ggressions are in workplace email chains. I will be looking at what proportion of communications include offensive content (i.e., scored above the midpoint from the subtle offense classification model). To contextualize the distribution, I will also be looking at the variability and modality within the distribution.

Research Question 1: How prevalent are workplace "Micro E-ggressions" and "Macro e-ggressions"?

My next research goal for my study is to explore the underlying themes in the subtly offensive statements sent to others and compare them to the general themes found within the emails. This analysis leverages topic modeling to discover underlying themes found within offenses that people experience. I will explore what these themes are by using a topic clustering model to find groupings of topics within the most offensive text that are similar within that group and distinct between groups.

Research Question 2: What types/themes of offenses are most common?

My last research question will involve using the same exploratory clustering method as before, but I will look at the commonalities in how recipients of offensive emails respond. This

analysis will search for underlying themes. I will also look at how offensive the responses are scored in comparison to the original messages.

Research Question 3: How do recipients of offensive emails respond?

Hypotheses

I also have specific hypotheses regarding the interpersonal characteristics of the offender and the target of the offense that may be associated with stronger perceptions of the offense. I expect the identities of the perpetrators and their dissimilarity to the target to affect how the offense is experienced. Power and status differences between majority and minority (non-White, non-male) groups are ingrained in American systems and awareness. These ingrained prejudices can explain why social category diversity increases relationship conflict in workgroups (Jehn et al., 1999). Prior research has found that those of majority status are more discriminatory, more influenced by differentiation, and less parity-oriented (equality) when in a position of high status, and therefore, status (Sachdev & Bourhis, 1991). This same research found that minorities were less discriminatory when they were of lower status and power. Moreover, research into perpetrators of offensive speech online found that most were perpetrated by males (Waseem & Hovy, 2016).

I expect that women will be more successful at inferring what would be offensive to others and therefore, less likely to send offensive emails. Prior research found that there are gender-based sensitivity differences and that women are of greater vulnerability to disempowering behavior in organizations (Vance et al., 2004). This heightened vulnerability and sensitivity to offense is further substantiated by the fact that men apologize less frequently than women, as they are perceived to possess a higher threshold for determining what constitutes offensive behavior. The same research found that when viewing the same workplace interaction,

women rated offensive interactions as more severe than men did (Schumann & Ross, 2010). Additionally, women exhibit greater sensitivity to disgust compared to men (Druschel & Sherman, 1999). A possible explanation for the gender difference in perceptions of severity is that women are more conditioned to focus on the experiences of others and prioritize maintaining amity in their relationships (Schumann & Ross, 2010). It has also been found that women experience more guilt after transgressions, greater empathy for victims, and a greater willingness to forgive transgressors. An alternative explanation is that men possess a higher tolerance for both physical and social discomfort/pain. Physical and social pain share common physiological mechanisms and could account for why men are more resilient to offense (MacDonald & Leary, 2005). Women are better than men at perceiving all forms of mistreatment, but more specifically, this trend especially applies in cases of sex-based offenses in the workplace (McCord et al., 2018).

Similar to women, racial minorities are chosen in a systematic way to be likely targets for workplace incivility (i.e. "selective incivility") because of their social group membership and the power differentials associated with that status (Cortina, 2008; Cortina et al., 2013). Specific to email communication, it has been found that non-White employees experience subtle forms of discrimination through the use of email more on average than their White counterparts (Daniels & Thornton, 2019). Prior research suggests that, due to experiencing negative racial experiences in their past, some ethnic minority individuals may be particularly sensitive to subtle slights like microaggressions (Lilienfeld, 2017; Watson & Clark, 1984). Because of their personal experience with prejudice, minority-group members are better judges of discrimination and are more likely to perceive race-based offenses than ethnic majority (i.e., White) members and are therefore less likely to perpetuate them (McCord et al., 2018). In addition, the intersectionality of

identities can influence mistreatment. Gender and race can interact to affect experiences of incivility such that holding multiple low-status social identities (i.e., women of color) can make someone more susceptible (i.e. double jeopardy) and more likely to report worse mistreatment (Cortina et al., 2013).

Hypothesis 1a: Offensive emails are more likely to come from men.

Hypothesis 1b: Offensive emails are more likely to be received by women.

Hypothesis 2a: Offensive emails are more likely to come from White employees.

Hypothesis 2b: Offensive emails are more likely to be received by non-White employees.

Hypothesis 2c: Offensive emails are most likely to be received by non-White, women employees.

Hypothesis 2d: Offensive emails are most likely to come from White, male employees.

A grounded theory of social psychology is that we tend to like people who are more similar, belonging to an in-group, than dissimilar to us (Byrne, 1971). We tend to not punish those who we like more and are more forgiving of transgressions in order to preserve relationships (Bradfield & Aquino, 1999). Due to our tendency to assign greater value to relationships we like and are similar to, we are more motivated to preserve the relationship with a transgressor who is perceived as more similar to oneself (Kim et al., 2008). As mentioned before these biases can explain why social category diversity has been linked to increased relationship conflict in workgroups (Jehn et al., 1999). Therefore, I expect that more offenses are likely to transpire between those of dissimilar social identities and status (i.e., race and/or gender).

Hypothesis 3: Offensive emails are more likely to occur between dissimilar races and/or genders.

My next hypothesis centers around the temporal context of these interactions: more specifically, how email offenses that are sent outside of traditional work hours differ from those sent inside work hours. I expect that emails sent outside of the hours of 9:00 am to 5:00 pm will be, on average, more aggressive than those sent inside of these work hours. Prior research found that after-hour work emails have detrimental effects on employees. Constant connectivity to work is amplified by email communication, and this has been associated with diminished well-being and negative emotions (Büchler et al., 2020).

Organizational expectations for extensive email monitoring are correlated with decreased detachment, heightened emotional exhaustion, diminished work-life balance, and increased turnover intentions (Belkin et al., 2020). In general, work-related smartphone use after work is a predictor of job burnout (Park et al., 2020). In particular, the characteristics of these emails such as frequency (i.e., the number of emails), duration (i.e., the time spent on emails), and emotional valence (i.e., the tone of emails), uniquely influence vigor and fatigue through rumination and problem-solving pondering (Minnen et al., 2021). Unpleasant emotional valence, like email incivility (rude or aggressively toned emails), has been found to impair well-being and increase negative affect (Giumetti et al., 2013; Park, Fritz, & Jex, 2018; Pearson & Porath, 2005). In addition, technostress has been linked to deficiencies in self-control and increased burnout (Li & Liu, 2022). This constant stress and later burnout results is characterized by emotional exhaustion, which debilitates emotional regulation and self-monitoring (Hülsheger et al., 2013). Given the negative emotional outcomes of after-hour email communication, I expect that those who engage in this communication outside of normal temporal work demands will be more likely to curate less emotionally positive emails.

Hypothesis 4: Emails sent after normative work hours are more likely to be more offensive than emails sent within normative work hours.

My final hypothesis builds off of the temporal aspect of the previous hypothesis but focuses on the temporal lag in responses. I will examine how response time differs given the offense severity of the original message. Compared to non-offensive emails, I expect the distribution of lag time for receivers' responses to offensive emails will be characterized by two latent profiles: either the receivers will respond quickly and reactively or not at all or much later, after ruminating about the offense. Prior literature found that the experience of microaggressions can lead to feeling isolated and threatened (Sue, 2010). Race-related coping literature suggests the first rule of thumb for the target is to take care of oneself (Holder, Jackson, & Pontelertto, 2015; Mellor, 2004). The experience of offense can cause a "freeze effect," in which the target does not know how to respond and therefore does not respond, later ruminating on the situation and experiencing negative self-evaluations (Goodman, 2011; Sue et al., 2019). An alternative reaction to these offenses, especially under conditions of situational ambiguity, is reactive aggression and more specifically, impulsive-reactive aggression. Reactive aggression is motivated by the reaction to aversive emotions, and impulsive reactive aggression is a quick, hostile reaction to provocation in which the goal is to block a threat to one's self (Madan, 2014). A core feature of ICT-mediated communication, such as email, is the paucity of semantic cues due to text-only communication (Runions et al., 2013). The paucity of social cues may heighten perceived aggression and therefore, an impulsive, reciprocal, and spontaneous reaction. In an organization, perceiving the tone of an email as hostile may lead to a trail of communications that violate norms of civility and affect work relationships and performance. Given the two schools of thought, I propose the hypothesis below:

Hypothesis 5: The distribution of lag time for responding to offensive emails will be better characterized by two latent profiles, each with their own underlying means and variance. This two-profile solution should fit the distribution of data better than alternative profile solutions (e.g., 1 underlying mean, 3 underlying means, etc.).

Method

Sample Collection

I utilized a large dataset that includes over 500,000 emails between 158 employees, mostly senior management of Enron. This dataset was generated from Enron email servers by the Federal Energy Regulatory Commission (FERC) during its investigation in 2015. This dataset allows seeing how dyadic workplace email communications actually transpire and to whom and from whom they most occur. The dataset includes the names of the senders and receivers, the time it was sent, the subject of the email, and the content of the email. Knowing the time stamps can help better understand temporal influences in offense severity and lag time differences. The context of the emails can lend to measures of the offensiveness of the emails using a pre-trained RoBERTa-based transformer model (Cornett & Hernandez, 2022). Using the names of senders and receivers, I predicted their race and gender using a pre-trained NLP demographic inference model in order to understand the contextual identity dynamics that influence communication (Chekili & Hernandez, 2022). These measures are described below.

Measures

Gender

Gender was indirectly inferred by a gender prediction neural network model, and was predicted as either male or female using a convolutional neural network trained to classify gender from a name's spelling (Chekili & Hernandez, 2022). The model's phi correlation with actual gender (r=0.93), exceeds conventional reliability thresholds (r=*0.*70) for the correlation between a scale and itself.

Race

Race was indirectly inferred using a race prediction neural network model and was predicted as White, Hispanic, Black, Asian, or Middle Eastern using a convolutional neural network trained to classify race from a name's spelling (Chekili & Hernandez, 2022). The classification model's phi correlation with actual race (r=0.80) exceeds conventional reliability thresholds.

Statement Ratings

Statement offensiveness was derived from a pre-trained "subtle offensive" transformer-based model that fine-tuned the RoBERTa architecture to people's perceived offense from a variety of personally experienced statements (Liu et al.,2019). Participants' real lived experiences of interpersonal offenses in the workplace was used as the the training data. They reported their offenses, the rating of the offense, and a rationale. Ratings were made on a 7-point semantic differential scale (1=Extremely Positive; 4=Neutral; 7=Extremely Offensive). The model's cross-validated correlation with rated offense was (r=0.73, 95% C.I.=[0.69, 0.77]), which exceeded the correlation of independent raters evaluating another person's perceived offense (r=0.61, p<0.0001, 95% CI=[0.47, 0.72]).

Normative Time

Normative time was derived from collecting the send times of both original and reply emails and then dichotomizing the range between 8am and 5pm as “normative work time” and all other times as “non-normative work time” (i.e. df$normativehour_reply <-(df$hour_of_day_reply >= 8 | df2$hour_of_day_reply <= 17 , 1, 0), as these hours are typically the window for an 8-hour work day (with 1 hour lunch break).

Lag Time

Lag time was calculated by converting time stamps into a floating point format called Unix time, which represents the number of seconds that have elapsed since the Unix operating system was created. This is a common convention within programming languages for converting time into an interval-scaled variable that allows mathematical operations like addition and subtraction. With the Unix formatted time, I subtracted the reply time from the original send time (i.e. df2$lag <- (df2$origionaltime - df2$replytime). This, however, stayed in Unix time to detect incremental differences caused by time.

Results

Prevalence of Offenses in Workplace Emails (RQ1)

RQ1 is concerned with the prevalence of these offenses in workplace emails. The analysis of the prevalence of offensive behaviors in workplace emails revealed that a majority of the emails examined scored above the midpoint (i.e. 4) for offensive content. Specifically, the emails sent had a mean offense score of 4.314, with a standard deviation of 0.717, highlighting the variability in the level of offense across different communications (Table 1). Additionally, 65% of the emails sent had at least one sentence with an offense score greater than neutral (i.e., 4). The distribution of offense scores was visualized using a kernel density estimate plot, which illustrated a continuous distribution of offense levels within the analyzed emails (Figure 3).

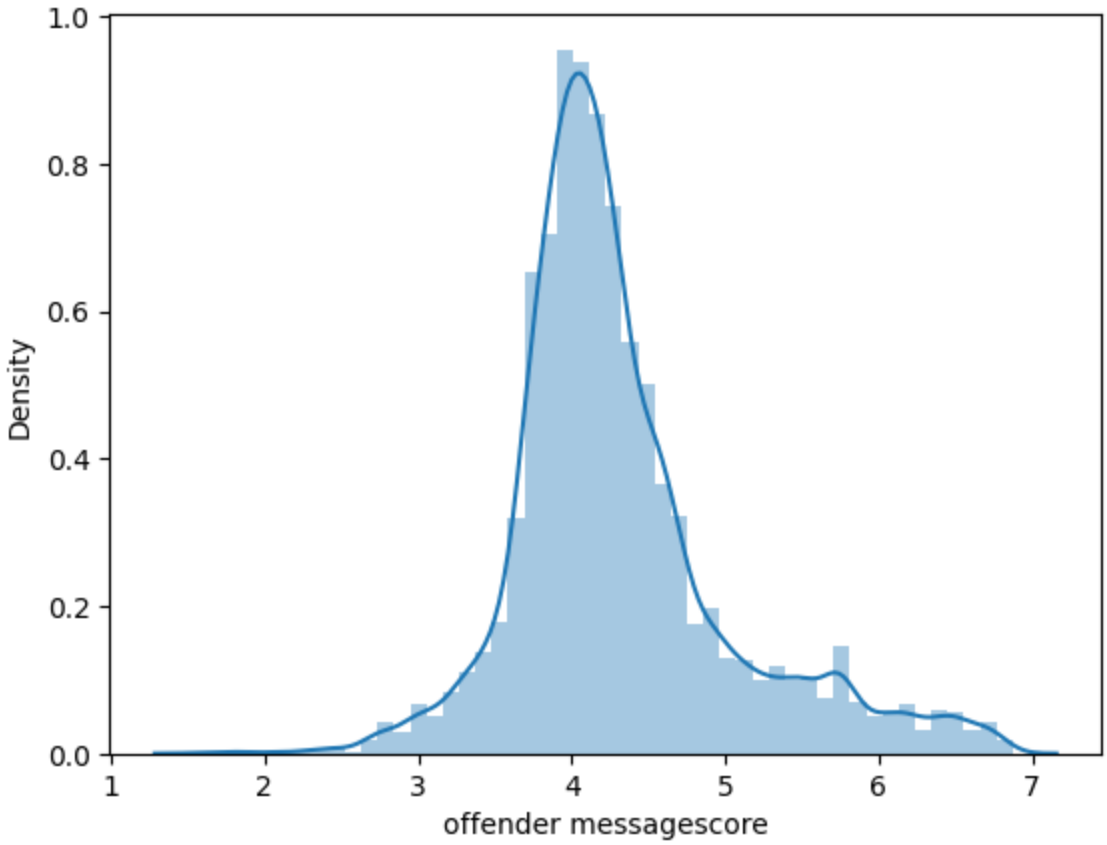
Table 1

Descriptive Statistics of Offender and Offendee Message Scores

	Offender Message Score	Offendee Reply Message Score
Count	24281.000	24281.000
Mean	4.314	4.540
Std	0.717	0.848
Min	1.570	1.207
25%	3.896	4.031
50%	4.162	4.391
75%	4.572	4.937
Max	6.871	6.946

Figure 3

Kernel Density Plot of Original Message Scores



Thematic Elements in Offensive Emails (RQ2)

For RQ2, the clustering of offensive emails revealed distinct thematic elements among micro and macro e-aggressions. The application of k-means clustering on sentence embeddings isolated from offensive emails resulted in 7 primary clusters, with themes ranging from explicit derogatory remarks to subtler forms of exclusionary language. I found these themes by isolating statements with an offense score greater than or equal to 5.5 indicating they are slightly more offensive than neutral. Although this cutoff is arbitrary, it provides a threshold that retains a large amount of emails, while also minimizing the inclusion of marginal cases of offense. This resulted in 2044 emails. From those, I excluded spam or any unintentional original messages which produced a dataset of 287 clusterable emails. Using an embedding model that converts each email as a vector (numeric representation of sentences), I converted each email into a

quantitative representation of its meaning (Reimers & Gurevych, 2019). Vectorizing encodes word meanings and positions to represent meaningful and contextually complex semantic information. My embedding shape was (287, 768) indicating each of the 287 emails is represented by a vector length of 768 (i.e. the number of semantic dimensions in the embedding). After embedding, I performed a k-means cluster analysis using the Scikit-learn library and calculated the silhouette score, a metric used to determine the optimal number of clusters, for each clustering configuration (Figure 4; Shahapure & Nicholas, 2020).The highest silhouette score indicated a well-defined clustering solution of 5.

Figure 4

Plot of Reply Clusters and their Respective Silhouette Scores

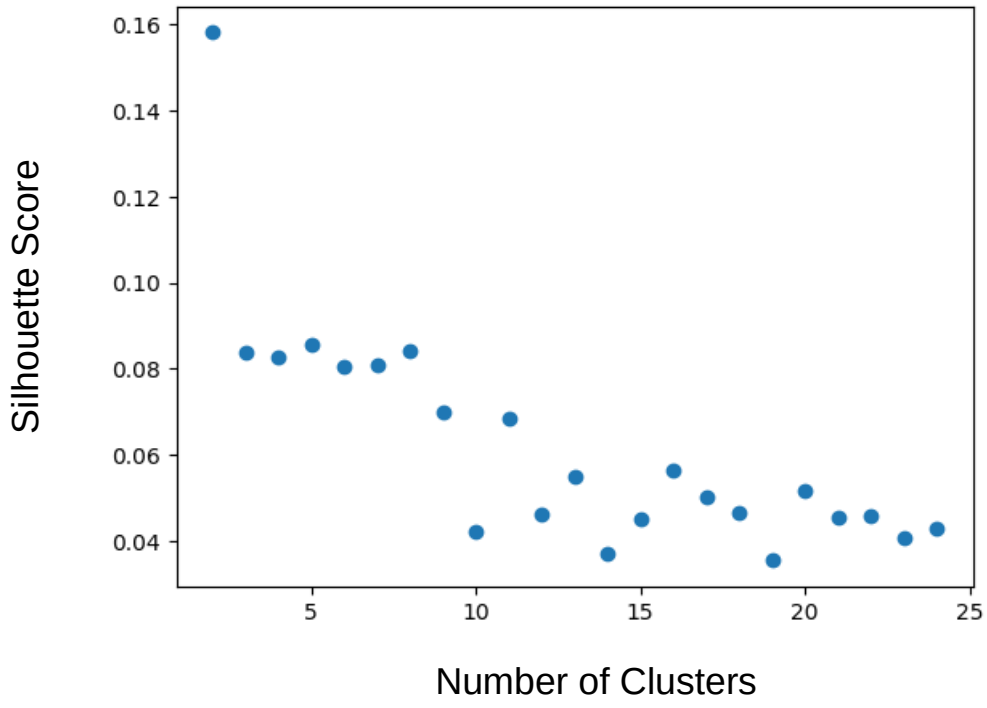


Table 2

List of Common Themes in Offender's Emails

Theme	Description	Category	Example	Proportion
Doubting/Questioning competence	Statements that imply a person's is not capable of doing their work.	Micro E-ggression	" I will want Berney to have some oversight to ensure quality and consistency" "I think most folks are just too damn busy" "Everyone is having problems" "I believe that your approach of going on a series of trips is not the most effective. Not only do I think it will be relatively unproductive… it will be counterproductive. " " I question whether you are the best qualified from a skill perspective."	25.436%
Disbelief/Questioning Work Ethics	Statements that imply that the person or their work is unfair/ethical.	Micro E-ggression	"I would think that Credit would have no biased interest in the outcome of this race. Could such slanderous allegations be true?" "Sorry about the P&L and Empower being hairy" "You changed the rate on this deal…What's up with that girl?"	22.996%
Passing Judgement and Coercion on/using Rumors/Secrets/News	Statements that criticize how or what work is being done more overtly.	Macro E-ggression	"The more i read and hear you should be ashamed" "I though this was really cute-but I have asick mind…" "Can you sneak the clubs past your wife?" "Congratulations and you suck" "you said you were going to return those items.Given what has happened and what those items represent, I do not see how you could now feel it is appropriate to keep them" "... she & her husband fight a lot but always work things out." "This might be a good time to remind you that this week is Secretary's week (since I am holding all the goodies) Just joking Brenda."	18.467%
Frustration with Work (Corruption)	Statements that communicate disbeliefs and frustrations with work systems and.	Micro E-ggression	"apparently this business is a disaster!!! " "Its a crying shame what happening to enron, Who would have ever though this?"	16.376%
Political	Statements that	Micro E-	"Here's an attempt at being frank about the	16.735%

Commentary/ Criticism	communicate disapproval of political systems or decisions	ggression	legislation without jeopardizing our relationships …" "Wouldn't this fix everything? What a great idea- a government-owned utility. Who in their right mind would want to give the money back to taxpayers?"

The common themes identified in the offender's emails, including passing judgment, doubting competence, questioning work ethics, expressing frustration with work or corruption, and engaging in political commentary. This highlights the prevalence and nature in which workplace environments and communication modalities influence what offenses transpire and warrant a reply.

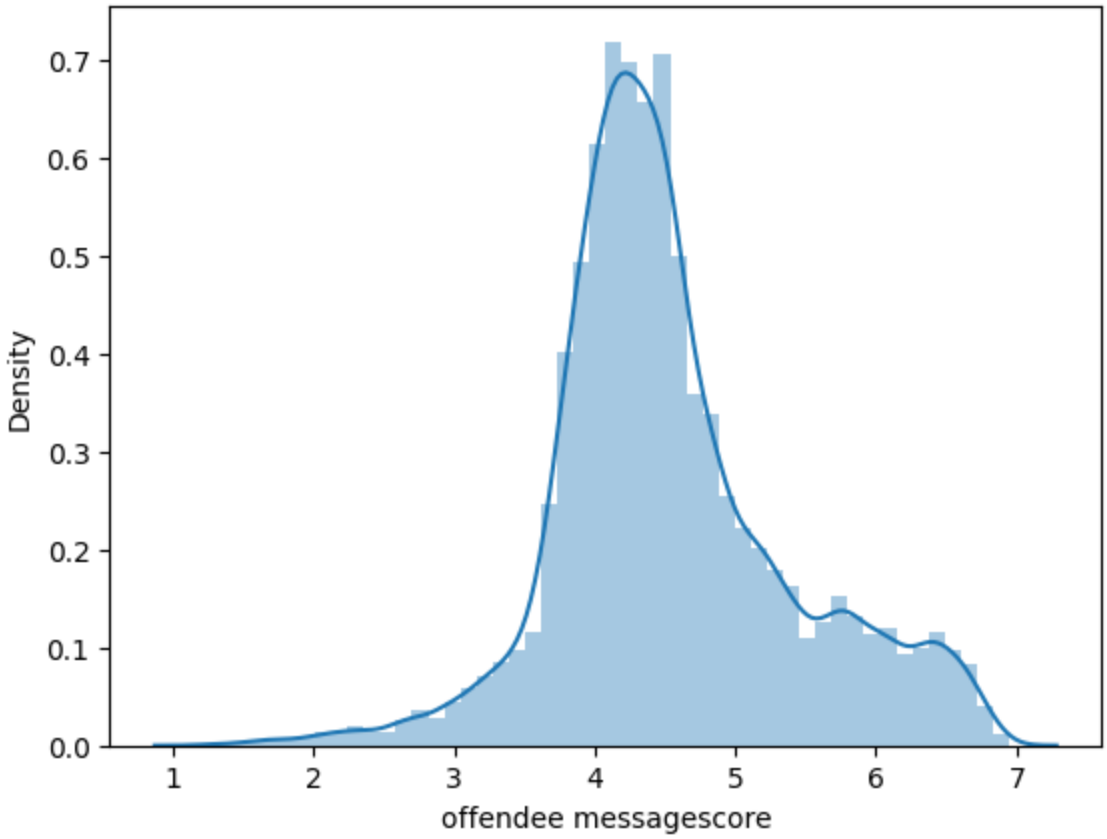
Dyadic Interactions and Responses to Offensive Emails (RQ3)

RQ3 is interested in examining the dyadic aspect of email communication and how people respond to offensive emails. I looked at this by first isolating offense emails and obtaining the responses to those emails. Responses were identified by looking at the subject of the email, finding replies to that subject within a given date range, and examining the temporally first response. Then, I performed a multilevel analysis to see if the offense level of the original message, influence the offensiveness of the reply (i.e. ('offendee.messagescore ~ offender.messagescore + (1|offender) + (1|offendee)). Examining the dyadic aspect of email communication, the study found that the original message offensiveness significantly and positively (*B*= 0.308, *p* =2e-16, 95% CI: [0.293, 0.322]) predicted the reply. Such that, the intercept (average message offense) was 3.174 (slightly more positive), and every one unit increase in offense of the original message, would, on average, result in a 0.308 increase in offensiveness of the reply message. This supports the notion that workplace incivility breeds more incivility. The mean offense score of responses was 4.540 (i.e. most replies were negatively

coded) which is greater than the mean of the offender's message score 4.324. Additionally, 77.011% of all reply messages were coded as more offensive than neutral (i.e., 4) compared to the original messages in which 65% scored above neutral. The distribution of reply offense scores was visualized using a kernel density estimate plot (Figure 5).

Figure 5

Kernel Density Plot of Offendee's Reply Message Scores



Cluster analysis on the text of responses highlighted several response types, ranging from offensive retaliation to self-defensive humor, illustrating the diverse ways individuals react to receiving offensive content. The application of k-means clustering on sentence embeddings isolated from offensive emails resulted in 11 primary clusters, with themes ranging from explicit derogatory remarks to subtler forms of exclusionary language. Similar to the cluster analyses of the original offenses, I isolated statements with an offense score greater than or equal to 5 indicating. This culminated in 5,721 reply emails which was reduced to 437 after excluding spam and forwarded messages. My embedding shape was (437, 768). After embedding, I performed a k-means cluster analysis again and calculated the silhouette score for each clustering configuration (Figure 5). The highest silhouette score indicated a well-defined clustering solution of 13, however, after excluding two of the clusters identified as non-offensive for sports and

trade-related commentary, I found 10 distinct offensive clusters. The qualitative investigation of offensive statements found that more overt offenses largely targeted identity, whereas less overt statements attempted to dilute the potency of a message with sarcasm/joking humor, cheerful introductory/concluding remarks, or framing/embedding it as work-related (Table 3).

Figure 6

Plot of Reply Clusters and their Respective Silhouette Scores

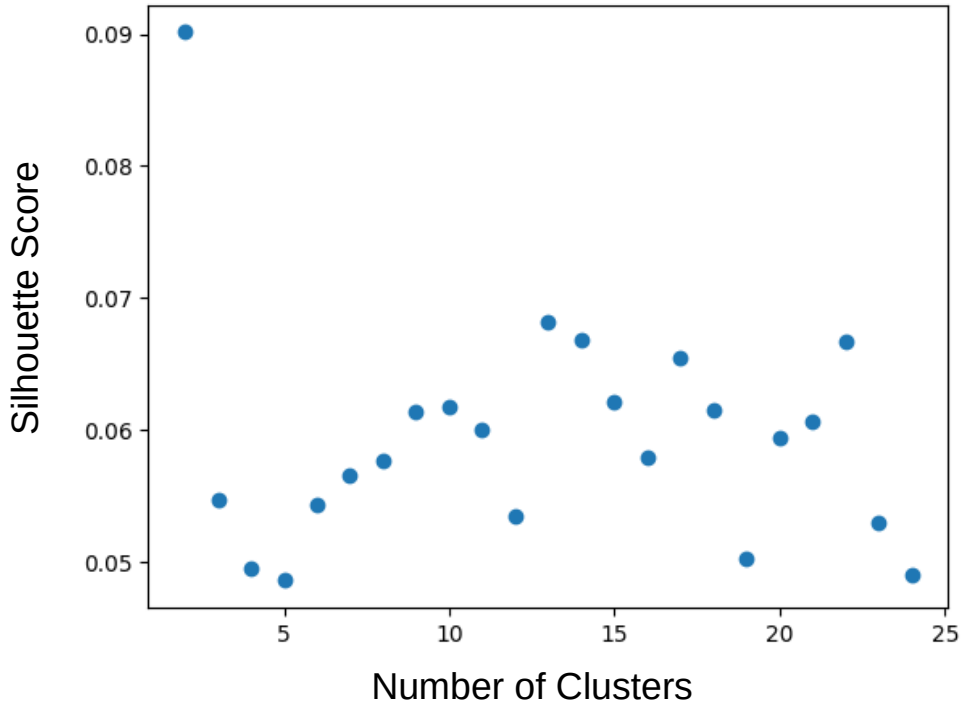


Table 3

List of Common Themes in People's Reply Emails

Theme	Description	Category	Example	Proportion

Work/Task Citicism	Statements that criticize how or what work is being done more overtly.	Macro E-ggression	“The creation of this contract is the most ridiculous waste of time.” “Got me. Don't get old Gerald, the old rumor about the first thing to go is false.”	12.815%
Joking Interpersonal Criticism	Statements that comment on the personality/behavior of another with jugement and dissent	Micro E-ggression	“Yeah, you are way too excitable in the mornings. I get dizzy just reading your emails.” “Are you experiencing some sort of split personality today?”	10.984%
Confronting Offense with impenitence	Statements that address a possible offense caused to another person without being apologetic	Micro E-ggression	I'm glad you appreciated my observation rather than taking offense. I was afraid after two weeks of silence I may have pissed you off..“ “ I know you're probably mad at me but I took vacation …Just because I didn't make your party does not mean you are out of sight - out of mind!”	10.068%
Joking Sexism/Misogyny	Statements that target, objectify, or diminish women or feminine presenting people	Macro E-ggression	“the girls are not that bad if we bring some clearisil for the blemishes and a couple of cases of beer.” “how can we get that girl cassie to go? do i need to invite her roomate? hull could keep her occupied.” “What Chics will be there - and you don't count.”	9.611%
Flirtatious and derogatory Communication	Statements that are flirtatious/playful in nature but condescending and gendered towards women/feminine presenting people	Micro E-ggression	“Easy,babe” “Oh, you little sweet talker…!” “ I bet your boyfriend was upset.”	9.382%
Work-Related Self Criticism	Statements that imply a person is not capable or in control of their work	Micro E-ggression	“ I feel so useless. Thanks so much for taking charge of this” “They want it to stay in. Can you live with that?”	9.382%

Political Commentary/Criticism	Statements that communicate disapproval of political systems or decisions	Micro E-ggression	"If you do your job of protecting consumers…. I cannot in good conscience, however, forego any measure that would serve to protect the people" "U.S. liberal economists and media sophisticates for being so naive"	7.551%
Work/Task Frustration	Statements that communicate slight frustrations with work systems	Micro E-ggression	"Remember the thorn in Bill's side and mine over the last years... I have no idea how far Dick or Tom may let this go, but it is bad form.." "I would be real hard-pressed to do a respectable job today.I do not think our department has anyone who is up to speed on how to negotiate an..."	6.865%
Task Delegation/Command	Statements that are related to work and dominating	Micro E-ggression	"Pay it" "Let me know"	4.805%
Cheerful Sarcastic Work Criticism	Statements that communicate frustration with work processes with sarcasm.	Micro E-ggression	"The model looks great except when I changed an input and hit go it blew up." "Everything is just peachy up here…he'll just switch me with an analyst in the risk group that works on our books, sooooooooo....I'm curious how this "track" will actually affect me. Happy hour tomorrow?"	3.204%

Note. These percentages do not add up to 100% because not all clusters were found to be offensive/negative

Of the themes assigned, the most common theme was work or task-related criticism. Almost 20% of email experiences involved either work/task-related criticism or frustration. Frustration, however, was less overt than criticism in its intention of offensiveness (i.e. was a Micro E-ggression). Next were statements that were judgemental or communicated annoyance/disapproval of behavior that were masked as jokes, and 11% of emails included these. The third most common theme found was "confronting offense with impenitence" which were

statements that address offenses caused to another person without being apologetic. This theme reveals the cyclical nature of workplace incivility (i.e. people who "received" offensive emails and replied may also have been perpetrators previously which warranted the "original" message). Subsequent themes include "joking sexism/misogyny," "flirtatious and derogatory communication," "work-related self-criticism," "political commentary/criticism," "work/task frustration," "task delegation/command, " and "cheerful sarcastic work criticism."

Gender Differences in Sending and Receiving Offensive Emails (H1a & H1b)

H1a is interested in which gender is sending more offensive emails, and I expect it is males. To examine whether men send more offensive emails than women, I applied a multilevel model to the data, in which max offense within an email is predicted by the gender of the sender, with a random intercept for each unique person, as the emails are nested within people (i.e., offender.messagescore ~ offender.gender + (1|offender)).Multilevel modeling revealed that males were not statistically significantly (B=0.064, t(1176) = 1.891, p = .059, 95% CI:[-0.002, 0.131]) more likely to send offensive emails compared to females. Which was not in support of H1a.

H1b is interested in what gender receives offensive emails the most, and I hypothesized it is females. To examine whether women receive more offensive emails than men, applied a similar multilevel model to the data, in which max offense within an email is being predicted by the gender of the receiver, with a random intercept for each unique person, as the emails are nested within people (i.e., offender.messagescore ~ offendee.gender + (1|offendee)). H1b was also not supported, as females were not found to receive offensive emails more statistically significantly in frequency than males (B = 0.003, t(859.500) = 0.090, p=0.929, 95% CI:[-0.060,

0.065]). This suggests that in our data there is not a significant gender pattern in the directionality of offensive communications within the workplace.

Racial Dynamics in Offensive Email Communication (H2a, H2b, H2c, H2d)

H2a suggested that White employees were more likely to send offensive emails compared to non-White employees. To examine whether White employees send more offensive emails than non-White employees, I applied a multilevel model to the data, in which max offense within an email is being predicted by the race of the sender, with a random intercept for each unique person, as the emails are nested within people (i.e., offender.messagescore ~ offender.race + (1|offender)). The analysis partially supported H2a which indicates that Hispanic employees send less offensive emails compared to White employees (B = -0.184 t(1215) = -2.912, p=0.004, 95% CI: [-0.308, -0.060]). All other racial groups (i.e. Black, Asian, and Middle Eastern) were not statistically significantly less offensive than White employees.

H2b suggested that non-White employees are more likely targets of offensive emails than White employees . To examine whether non-White employees receive more offensive emails than White employees receive, I applied a similar multilevel model to the data, in which max offense within an email is being predicted by the race of the receiver, with a random intercept for each unique person, as the emails are nested within people (i.e., offender.messagescore ~ offendee.race + (1|offendee). None of the non-White employees' Beta coefficients were statistically significant, therefore H2b was not supported (Table 4).

Table 4

Hypothesis 2b Racial Targets/Recipients of Offense Regression Results

Race	Unstandardized Beta Coefficients	Degrees of Freedom	T Statistic	P-Value
Reference Category (White)	4.371	887.576	260.604	<2e-16***
Asian	-0.059	831.925	-0.708	0.479
Black	-0.062	789.331	-1.406	0.160
Hispanic	-0.068	970.721	-1.109	0.268
Middle Eastern	-0.047	1401.214	-0.247	0.805

Further analysis for H2c and H2d investigated the role of intersectionality in workplace micro and macro E-ggressions and showed that no interaction of race and gender were significantly more likely to receive and send offensive emails (Tables 5 and 6). H2c expected that offensive emails are more likely to be received by marginalized employees (i.e., non-White women). To examine this, I applied a multilevel model to the data, in which max offense within an email is being predicted by the gender and race of the receiver, with a random intercept for each unique person, as the emails are nested within people (i.e., offender.messagescore ~ offendee.race + offendee.gender + offendee.race*offendee.gender + (1|offendee)).

H2d expected that offensive emails are more likely to be sent by non-marginalized employees (i.e., White men). To examine whether White men send more offensive emails, I applied a similar multilevel model to the data, in which max offense within an email is predicted by the gender and race of the sender, with a random intercept for each unique person, as the

emails are nested within people (i.e., offender.messagescore ~ offender.race + offender.gender + offender.race*offender.gender + (1|offender)).

Table 5

Hypothesis 2c Racial and Gender Targets/Recipients of Offense Regression Results

Race & Gender	Unstandardized Beta Coefficients	Degrees of Freedom	T Statistic	P-Value
White Male (Reference Category)	4.363	839.896	142.975	<2e-16***
Asian Male	-0.298	1009.127	-1.075	0.283
Black Male	-0.035	781.872	-0.360	0.719
Hispanic Male	-0.021	983.82	0.873	0.873
Middle Eastern Male	0.095	644.471	0.200	0.842

Table 6

Hypothesis 2d Racial and Gender Senders of Offense Regression Results

Race & Gender	Unstandardized Beta Coefficients	Degrees of Freedom	T Statistic	P-Value
White Male (Reference Category)	4.263	1160	131.636	<2e-16***
Asian Male	-0.014	1245	-0.056	0.966
Black Male	-0.023	1126	-0.239	0.811
Hispanic Male	-0.093	1209	-0.724	0.469
Middle Eastern Male	-0.056	1140	-0.134	0.893

Offense Emails and Dissimilarity in Sender-Receiver Dynamics (H3a & H3b)

H3a expects that offensive emails are more likely to occur between people who are dissimilar by race or gender. To test if offensive emails are more likely to occur for dissimilar genders of the sender and the receiver, I used a model that predicts max offense within an email by the difference between sender and receiver gender (i.e.,offender.messagescore ~ same.gender + offender.gender + offendee.gender + (1|offender)). To test if offense emails are more likely to occur for dissimilar race of sender and race of receiver, I will use a similar multilevel model that predicts max offense within an email by the difference between sender and receiver race (i.e., offender.messagescore ~ same.race offender.race + offendee.race + (1|offender)).

The study found that offensive emails were more likely to occur between individuals of similar gender (B=0.023, t(24270)= 2.130, p =0.033, CI 95%: [-0.002, 0.044]), but no significant effect for race similarity (B=0.008, t(24230)= 0.643, p =0.520, CI 95%: [-0.018, 0.0436]), not supporting H3a or H3b. However, this suggests that similarities in gender between sender and receiver contribute to the likelihood of offensive email exchanges.

Offense Levels and Work Hours (H4)

H4 predicted that emails sent outside of normative work hours will be more offensive than those sent inside work hours (i.e., 8:00 am to 5:00 pm). To examine the effect of after-normative work hours and during-normative work hours, I ran a multilevel analysis predicting maximum email offense using a dichotomous variable representing the time (normative vs. non-normative) the email was sent. To address the nesting of emails from the same person, I included a random intercept for each unique sender. (offendee.messagescore ~ normativehour_reply + (1|offendee))

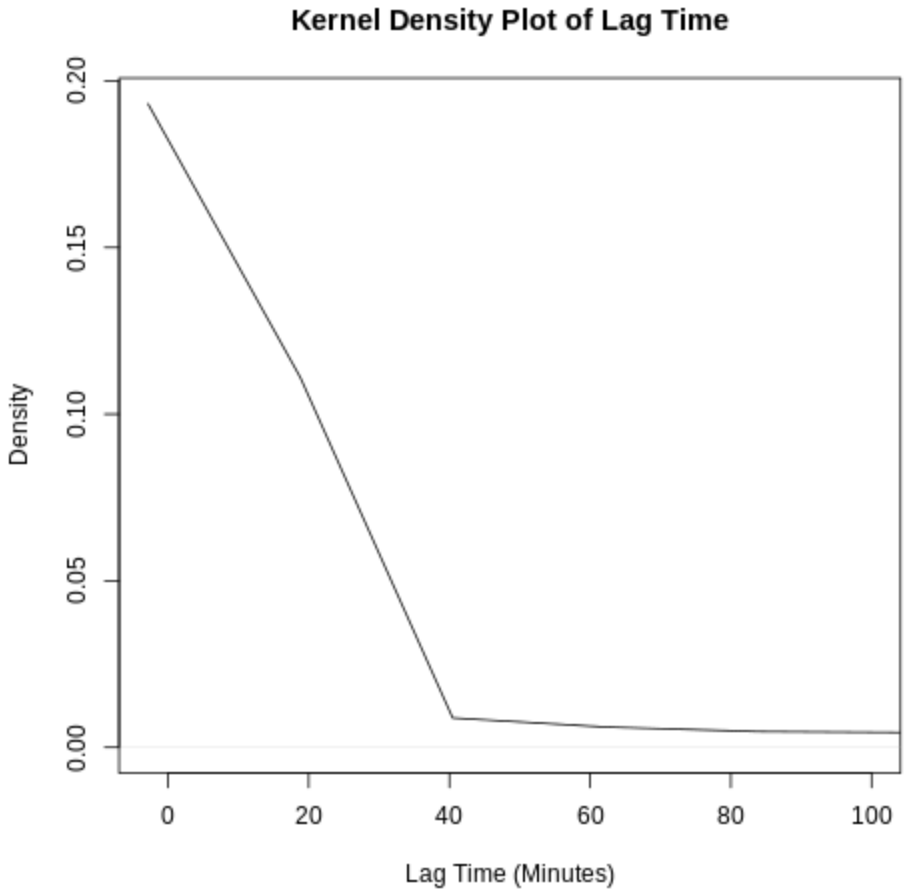
Multilevel analysis for H4 revealed that emails sent outside of normative work hours were not significantly more offensive than those sent during work hours (B=-0.017, t(24220)= -0.899, p =0.369, CI 95%: [-0.053, 0.020]), not supporting H4. Suggesting that the temporal context of email communication does not play a role in the level of offense conveyed for my sample.

Response Times to Offensive Emails (H5)

H5 expected that there will be two distinct lag times in responses to offensive emails. Looking at the kernel density estimation (KDE) plot, I can see the data is distributed negatively and linearly, not supporting H5 (Figure 7)

Figure 7

Distribution of Lag Times for Replies to Offenses.



Multilevel analysis found that the offensiveness of a message statistically significantly influenced the time it took the target to reply, such that the more offensive a message was, the quicker the reply was (*B*= -41.078, t(20617.489)= -5.295, *p* = 1.20e-07***, CI 95%: [-56.283, -25.874]). This suggests every 1 unit increase in the offensiveness of an email, the expected reply time decreases by 41.078 minutes.

Discussion

Utilizing state-of-the-art natural language processing and a robust dataset we are better able to understand the prevalence, content, and contexts of offenses in the workplace. Results revealed that workplace Micro and Macro E-ggressions are plentiful and complex and the application of AI provides fertile opportunities for future investigation.

The Prevalence of Micro and Macro E-ggressions (RQ1)

The analysis of in workplace emails highlights a concerningly high prevalence of such content. Results indicate that a majority of all emails in the Enron corpus (65%) include some form of communication above the midpoint (neural) of offense perception. Implications of the findings reveal a need for further examination into the underlying factors which contribute to the manifestation of offensive language in professional technology-mediated correspondence.

Thematic Elements in Offensive Emails (RQ2)

Cluster analysis of the offensive emails, revealed distinct thematic elements. These themes ranged in severity from "Macro E-ggressions" like passing judgment and coercion to "Micro E-ggressions" such as doubting competence and expressing subtle work dissatisfaction. Identifying these themes helps provide insight into underlying motivations and dynamics driving offensive communication in the workplace. The most common theme was doubting/questioning competence followed by disbelief/questioning work ethic, both of which are means to disempower others and discredit their abilities. Findings suggest that individuals employ technology-mediated work communication modalities such as email to express their discontent or exert power dynamics within organizational settings.

Dyadic Interactions and Responses to Offensive Emails (RQ3)

Examining the dyadic aspect of email communication revealed a significant relationship between the offensiveness of the original message and the subsequent response. This finding, though correlational, supports the notion that workplace incivility tends to perpetuate itself. Interestingly, replies included similar themes around work ethic and competence criticism, however, there was a greater diversity of response themes. One explanation for this finding is an

attribution effect (i.e., offendees construal of why they are being targeted may influence how they respond differently). Evidence of possible attribution effects can be seen in the prevalence of the new themes such as “flirtatious and derogatory communication” in which such recipients might have interpreted a colleague's communication as amorous. More escalatory, yet still tolerant themes like “joking interpersonal criticism” and “joking sexim/mysogeogy” were also new and revealed the power of humor in disempowering behavior. Most interesting was the theme, “confronting offense with impenitence” which revealed that some of the “offendees” may have been the original “offenders,” therefore, supporting the notion that incivility breeds. Replies less intended to escalate incivility include “cheerful sarcastic work criticism,” to serve as a distraction from subsequent offenses in the text, “work-related self-criticism” which takes power away from a possible perpetrator by diminishing oneself, “work/task frustration” which puts blame on the situation instead of oneself, and “task delegation/command” which shifts focus from interpersonal to task-related matters. The diverse range of response types, from offensive retaliation to self-defensive humor, underscores the complex nature of interpersonal dynamics within professional contexts due to power, identity, and temporal differences. Excluding these contextual factors, however, we are able to postulate that in general, incivility is not always escalatory, but is reactive and cyclical.

Gender Differences in Sending and Receiving Offensive Emails (H1a & H1b)

Contrary to the initial hypotheses, my analysis did not reveal any significant gender differences in the propensity to send or receive offensive emails. This challenges traditional assumptions regarding gendered patterns of social dominance-coded communication and suggests a more nuanced understanding of workplace dynamics concerning offensive behavior in the workplace.

Racial Dynamics in Offensive Email Communication (H2a, H2b, H2c, H2d)

Although differential behavior of Hispanic employees was observed, supporting the notion that racial dynamics influence workplace communication. Intersectionality of race and gender did not emerge as a statistically significant factor in predicting email offensiveness. These results emphasize the complexity of understanding and addressing workplace incivility. Additionally, including more intersectional identities (eg., religion, political identity, incarceration status, LGBTQ+ status, etc) and contextual information could lend nuance to the understanding of workplace mistreatment in future research.

Offense Emails and Dissimilarity in Sender-Receiver Dynamics (H3a & H3b)

Analyses found a significant relationship between the similarity in gender between sender and receiver and the likelihood of offensive email exchanges. However, the effect of race similarity was not found. Although contrary to my hypothesis, this highlights the role of gender dynamics in shaping interpersonal interactions. Alternative theories suggest that identity dissimilarity may make us more perceptive of interpersonal conflict, and therefore, more conscious of what we say (Philips, 2014). Additionally, men may feel more comfortable expressing dissenting or controversial opinions around other men, masked as corporate"locker room" talk (Dellinger & Williams, 2002). Further, cross-cultural gender-focused research has found that women are encouraged and incentivized to uphold patriarchal norms by continuing the oppression and bullying of other women in the workplace (Agarwal, 2016).

Offense Levels and Work Hours (H4)

Contrary to expectations, emails sent outside of normative work hours were not found to be statistically significantly more offense were not found to be significantly more offensive or more offensively replied to than those sent during work hours. This suggests that the temporal

contexts of work may be more complex than previously assumed and could be a result of "normative work hours" being dependent on the workplace and culture. In addition, it could suggest that the notion of "work hours" may be dissolving in a post-COVID workforce all altogether as work hours previously began and ended with arriving at and departing from a physical workplace (Vyas, 2022).

Response Times to Offensive Emails (H5)

Further temporal analysis revealed a significant relationship between the offensiveness of a message and the time it took for the recipient to reply. This suggests that individuals may respond more promptly to offensive content, possibly indicating heightened emotional arousal or a desire to address the issue promptly. This is supported by the previous notion of "reactive aggression" in self-defense literature (Madan, 2014). Future research could investigate how offense types/themes, intersectionality, and workplace norms influence lag times.

Limitations

In my study, I was able to use the technological and statistical power of artificial intelligence and big data with a psychological, theory-backed lens that lends humanity to data. However, some notable limitations include (a) generalizability, (b) restriction of range, (c) uninvestigated contextual factors, and (d) model overspecialization.

First, like all case studies, the study findings may not be generalizable beyond the context of the Enron corporation and its specific employees and organizational culture. Enron, specifically was under investigation for fraud, and therefore, themes such as "frustration with work (corruption)" may be specific to this company due to heightened employee stress.

Next, the issue of range restriction may influence the validity of results. Since emails that were not replied to were excluded from the dataset, this restricted the range of it, and therefore,

could potentially limit the representativeness of the sample and affect the comprehensiveness of the analysis. Further analyses could investigate what type of emails were not replied to and why. In addition, while investigating qualitative data, statements such as: “I am sending this notification to your department to seek a solution of halting and stopping inappropriate email materials which I am receiving…The content of materials received, I consider offensive, and am sure it would also be considered objectionable and inappropriate by the standards of Enron Corporation.” And, “This was the worst we have seen at Enron and I don't even want to put in this email what I heard was going on there" indicates possible behavior shifts to conceal offensive statements. This concealment would restrict the dataset from what communications would naturally transpire and pose a challenge in accurately identifying and analyzing offensive language. Additionally, it could lead to the underestimation of the prevalence of offensive communication.

Next, the lack of emphasis on possible contextual factors could leave important details out regarding the specifics of the circumstances that produce these finds. Future analyses could look into thematic co-occurrences to provide a nuanced understanding of studying offensive statements that are offensive for a multitude of reasons due to complexities in interpersonal dynamics (e.g., power differentials and intersectionality). Looking at co-occurrences could result in a partial representation of the complexity of offensive communication patterns.

Finally, although the developed model can accurately predict the offensiveness of statements, it may be overspecified to its training set. The model primarily analyzes short text segments, potentially overlooking subtle offensive statements embedded within longer dialogues like emails. Moreover, text-based analysis may exclude contextual factors, such as tone, body

language, and micro-expressions, which can influence the interpretation of offense in interpersonal interactions.

Conclusion

The findings of this study shed light on the prevalence, thematic elements, and interpersonal dynamics of offensive email communication within workplace environments. While some hypotheses were supported, others yielded unexpected results, challenging conventional assumptions regarding gender and racial dynamics in workplace communication. These findings underscore the importance of understanding and addressing workplace incivility to champion professional and respectful communication and subsequent healthy organizational cultures. The themes identified can serve as reminders of the potential topics people may be sensitive about and others may be privy to. By fostering a more aware culture that emphasizes respectful communication, organizations can cultivate work environments in which individuals feel understood, valued, and motivated to collaborate towards shared goals. Including (Bcc'ing) artificial intelligence in the study and addressing of these Micro and Macro E-ggressions whole exposes a whole new frontier for workplace offense research methods and attitudes towards intolerance.

References

Ali, O., Scheidt, N., Gegov, A., Haig, E., Adda, M., & Aziz, B. (2020). Automated Detection of Racial Microaggressions using Machine Learning. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2477–2484. https://doi.org/10.1109/SSCI47803.2020.9308569

Agarwal, S. (2016). Women Bullying Women (WBW) at Workplace:A Literature Review. Journal of Applied Management- Jidnyasa, 8(1), 57–65. Retrieved from http://simsjam.net/index.php/Jidnyasa/article/view/121052

Allen, J. (1987). *Natural language understanding*. Benjamin/Cummings Pub. Co.

Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). *Deep Learning Models for Multilingual Hate Speech Detection*. https://doi.org/10.48550/ARXIV.2004.06465

Andersson, L. M., & Pearson, C. M. (1999). Tit for Tat? The Spiraling Effect of Incivility in the Workplace. *The Academy of Management Review*, *24*(3), 452. https://doi.org/10.2307/259136

Aquino, K., & Bommer, W. H. (2003). Preferential Mistreatment: How Victim Status Moderates the Relationship Between Organizational Citizenship Behavior and Workplace Victimization. *Organization Science*, *14*(4), 374–385. https://doi.org/10.1287/orsc.14.4.374.17489

Aquino, K., & Douglas, S. (2003). Identity threat and antisocial behavior in organizations: The moderating effects of individual differences, aggressive modeling, and hierarchical status.

Organizational Behavior and Human Decision Processes, *90*(1), 195–208. https://doi.org/10.1016/S0749-5978(02)00517-4

Aquino, K., Tripp, T. M., & Bies, R. J. (2001). How employees respond to personal offense: The effects of blame attribution, victim status, and offender status on revenge and reconciliation in the workplace. *Journal of Applied Psychology*, *86*(1), 52–59. https://doi.org/10.1037/0021-9010.86.1.52

Belkin, L. Y., Becker, W. J., & Conroy, S. A. (2020). The Invisible Leash: The Impact of Organizational Expectations for Email Monitoring After-Hours on Employee Resources, Well-Being, and Turnover Intentions. *Group & Organization Management*, *45*(5), 709–740. https://doi.org/10.1177/1059601120933143

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, *3*(null), 993–1022.

Bosco, C., Dell'Orletta, F., Poletto, F., Sanguinetti, M., & Tesconi, M. (2018). Overview of the EVALITA 2018 Hate Speech Detection Task. In T. Caselli, N. Novielli, V. Patti, & P. Rosso (Eds.), *EVALITA Evaluation of NLP and Speech Tools for Italian* (pp. 67–74). Accademia University Press. https://doi.org/10.4000/books.aaccademia.4503

Breitfeller, L., Ahn, E., Jurgens, D., & Tsvetkov, Y. (2019). Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 1664–1674. https://doi.org/10.18653/v1/D19-1176

Bruk-Lee, V., & Spector, P. E. (2006a). The social stressors-counterproductive work behaviors link: Are conflicts with supervisors and coworkers the same? *Journal of Occupational Health Psychology*, *11*(2), 145–156. https://doi.org/10.1037/1076-8998.11.2.145

Bruk-Lee, V., & Spector, P. E. (2006b). The social stressors-counterproductive work behaviors link: Are conflicts with supervisors and coworkers the same? *Journal of Occupational Health Psychology*, *11*(2), 145–156. https://doi.org/10.1037/1076-8998.11.2.145

Büchler, N., ter Hoeven, C. L., & van Zoonen, W. (2020). Understanding constant connectivity to work: How and for whom is constant connectivity related to employee well-being? *Information and Organization*, *30*(3), 100302. https://doi.org/10.1016/j.infoandorg.2020.100302

Bunk, J. A., & Magley, V. J. (2011). Sensitivity to interpersonal treatment in the workplace: Scale development and initial validation: Sensitivity to interpersonal treatment. *Journal of Occupational and Organizational Psychology*, *84*(2), 395–402. https://doi.org/10.1348/096317910X488626

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology*, *101*(7), 958–975. https://doi.org/10.1037/apl0000108

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730.

Chowdhary, K. R. (2020). Natural Language Processing. In K. R. Chowdhary (Ed.), *Fundamentals of Artificial Intelligence* (pp. 603–649). Springer India. https://doi.org/10.1007/978-81-322-3972-7_19

Collins, D. R., & Stukas, A. A. (2008). Narcissism and self-presentation: The moderating effects of accountability and contingencies of self-worth. *Journal of Research in Personality*, *42*(6), 1629–1634. https://doi.org/10.1016/j.jrp.2008.06.011

Cortina, L. M. (2008). Unseen Injustice: Incivility as Modern Discrimination in Organizations. *Academy of Management Review*, *33*(1), 55–75. https://doi.org/10.5465/amr.2008.27745097

Cortina, L. M., Kabat-Farr, D., Leskinen, E. A., Huerta, M., & Magley, V. J. (2013). Selective Incivility as Modern Discrimination in Organizations: Evidence and Impact. *Journal of Management*, *39*(6), 1579–1605. https://doi.org/10.1177/0149206311418835

Cortina, L. M., Magley, V. J., Williams, J. H., & Langhout, R. D. (2001). Incivility in the workplace: Incidence and impact. *Journal of Occupational Health Psychology*, *6*(1), 64–80. https://doi.org/10.1037/1076-8998.6.1.64

Coyne, I., Gopaul, A.-M., Campbell, M., Pankász, A., Garland, R., & Cousans, F. (2019). Bystander Responses to Bullying at Work: The Role of Mode, Type and Relationship to Target. *Journal of Business Ethics*, *157*(3), 813–827. https://doi.org/10.1007/s10551-017-3692-2

Daniels, S., & Thornton, L. M. (2019). Race and workplace discrimination: The mediating role of cyber incivility and interpersonal incivility. *Equality, Diversity and Inclusion: An International Journal*, *39*(3), 319–335. https://doi.org/10.1108/EDI-06-2018-0105

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated Hate Speech Detection and the Problem of Offensive Language*. 4.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674.

De Dreu, C. K. W., van Dierendonck, D., & Dijkstra, M. T. M. (2004). Conflict at Work and Individual Well-Being. *International Journal of Conflict Management*, *15*(1), 6–26. https://doi.org/10.1108/eb022905

Dellinger, K., & Williams, C. L. (2002). The Locker Room and the Dorm Room: Workplace Norms and the Boundaries of Sexual Harassment in Magazine Editing. Social Problems, 49(2), 242–257. https://doi.org/10.1525/sp.2002.49.2.242

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://doi.org/10.48550/ARXIV.1810.04805

Dimock, H. G. (1971). Sensitivity Training as a Method of Increasing On-The-Job Effectiveness. *Sociological Inquiry*, *41*(2), 227–231. https://doi.org/10.1111/j.1475-682X.1971.tb01144.x

Druschel, B. a, & Sherman, M. F. (1999). Disgust sensitivity as a function of the Big Five and gender. *Personality and Individual Differences*, *26*(4), 739–748. https://doi.org/10.1016/S0191-8869(98)00196-2

Fai, Y. (2010). *Conflict management and emotional intelligence* [Dissertation]. Southern Cross University.

Fariselli, L., Ghini, M., & Freedman, J. (2006). *Age and Emotional Intelligence*. 10.

Friedman, B. A., & Reed, L. J. (2007). Workplace Privacy: Employee Relations and Legal Implications of Monitoring Employee E-mail Use. *Employee Responsibilities and Rights Journal*, *19*(2), 75–83. https://doi.org/10.1007/s10672-007-9035-1

Giumetti, G. W., Hatfield, A. L., Scisco, J. L., Schroeder, A. N., Muth, E. R., & Kowalski, R. M. (2013). What a rude e-mail! Examining the differential effects of incivility versus support on mood, energy, engagement, and performance in an online context. *Journal of Occupational Health Psychology*, *18*(3), 297–309. https://doi.org/10.1037/a0032851

Goodman, D. J. (2011). *Promoting Diversity and Social Justice* (0 ed.). Routledge. https://doi.org/10.4324/9780203829738

Grigg, D. W. (2010). Cyber-Aggression: Definition and Concept of Cyberbullying. *Australian Journal of Guidance and Counselling*, *20*(2), 143–156. https://doi.org/10.1375/ajgc.20.2.143

He, P., Liu, X., Gao, J., & Chen, W. (2020). *DeBERTa: Decoding-enhanced BERT with Disentangled Attention*. arXiv. https://doi.org/10.48550/ARXIV.2006.03654

Hermans, A., Beyer, L., & Leibe, B. (2017). In Defense of the Triplet Loss for Person Re-Identification. *CoRR*, *abs/1703.07737*. http://arxiv.org/abs/1703.07737

Hodson, G., & Costello, K. (2007). Interpersonal Disgust, Ideological Orientations, and Dehumanization as Predictors of Intergroup Attitudes. *Psychological Science*, *18*(8), 691–698. https://doi.org/10.1111/j.1467-9280.2007.01962.x

Holder, A. M. B., Jackson, M. A., & Ponterotto, J. G. (2015). Racial microaggression experiences and coping strategies of Black women in corporate leadership. *Qualitative Psychology*, *2*(2), 164–180. https://doi.org/10.1037/qup0000024

Holm, K., Torkelson, E., & Bäckström, M. (2015). Models of Workplace Incivility: The Relationships to Instigated Incivility and Negative Outcomes. *BioMed Research International*, *2015*, 920239. https://doi.org/10.1155/2015/920239

Hülsheger, U. R., Alberts, H. J. E. M., Feinholdt, A., & Lang, J. W. B. (2013). Benefits of mindfulness at work: The role of mindfulness in emotion regulation, emotional exhaustion,

and job satisfaction. *Journal of Applied Psychology*, *98*(2), 310–325. https://doi.org/10.1037/a0031313

Ibrohim, M. O., & Budi, I. (2019). Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter. *Proceedings of the Third Workshop on Abusive Language Online*, 46–57. https://doi.org/10.18653/v1/W19-3506

Kessler, S. R., Bruursema, K., Rodopman, B., & Spector, P. E. (2013). Leadership, Interpersonal Conflict, and Counterproductive Work Behavior: An Examination of the Stressor-Strain Process. *Negotiation and Conflict Management Research*, *6*(3), 180–190. https://doi.org/10.1111/ncmr.12009

Kim, T.-Y., Shapiro, D. L., Aquino, K., Lim, V. K. G., & Bennett, R. J. (2008). Workplace offense and victims' reactions: The effects of victim-offender (dis)similarity, offense-type, and cultural differences. *Journal of Organizational Behavior*, *29*(3), 415–433. https://doi.org/10.1002/job.519

Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T., & Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, *58*(5), 102643. https://doi.org/10.1016/j.ipm.2021.102643

Kokalj, E., Škrlj, B., Lavrač, N., Pollak, S., & Robnik-Šikonja, M. (2021). BERT meets shapley: Extending SHAP explanations to transformer-based classifiers. *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, 16–21.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. arXiv. https://doi.org/10.48550/ARXIV.1909.11942

Li, M., Fu, B., Ma, J., Yu, H., & Bai, L. (2021). Sensitivity and emotional intelligence: An empirical study with mental health as a regulating variable. *Current Psychology*, *40*(6), 2581–2589. https://doi.org/10.1007/s12144-020-00669-5

Li, X., & Liu, D. (2022). The Influence of Technostress on Cyberslacking of College Students in Technology-Enhanced Learning: Mediating Effects of Deficient Self-Control and Burnout. *International Journal of Environmental Research and Public Health*, *19*(18), 11800. https://doi.org/10.3390/ijerph191811800

Lilienfeld, S. O. (2017). Microaggressions: Strong Claims, Inadequate Evidence. *Perspectives on Psychological Science*, *12*(1), 138–169. https://doi.org/10.1177/1745691616659391

Lim, S., Cortina, L. M., & Magley, V. J. (2008). Personal and workgroup incivility: Impact on work and health outcomes. *Journal of Applied Psychology*, *93*(1), 95–107. https://doi.org/10.1037/0021-9010.93.1.95

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. https://doi.org/10.48550/ARXIV.1907.11692

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

MacDonald, G., & Leary, M. R. (2005). Why Does Social Exclusion Hurt? The Relationship Between Social and Physical Pain. *Psychological Bulletin*, *131*(2), 202–223. https://doi.org/10.1037/0033-2909.131.2.202

Madan, A. O. (2014). *Cyber Aggression / Cyber Bullying and the Dark Triad: Effect on Workplace Behavior / Performance*. *8*(6), 7.

Malte, A., & Ratadiya, P. (2019). *Evolution of transfer learning in natural language processing*. https://doi.org/10.48550/ARXIV.1910.07370

Mansourifar, H., Alsagheer, D., Shi, W., Ni, L., & Huang, Y. (2021). Statistical Analysis of Perspective Scores on Hate Speech Detection. *ArXiv:2107.02024 [Cs]*. http://arxiv.org/abs/2107.02024

Marshburn, C. K., Harrington, N. T., & Ruggs, E. N. (2017). Taking the Ambiguity Out of Subtle and Interpersonal Workplace Discrimination. *Industrial and Organizational Psychology*, *10*(1), 87–93. https://doi.org/10.1017/iop.2016.106

McCord, M. A., Joseph, D. L., Dhanani, L. Y., & Beus, J. M. (2018). A meta-analysis of sex and race differences in perceived workplace mistreatment. *Journal of Applied Psychology*, *103*(2), 137–163. https://doi.org/10.1037/apl0000250

Mellor, D. (2004). Responses to Racism: A Taxonomy of Coping Styles Used by Aboriginal Australians. *American Journal of Orthopsychiatry*, *74*(1), 56–71. https://doi.org/10.1037/0002-9432.74.1.56

Minnen, M. E., Mitropoulos, T., Rosenblatt, A. K., & Calderwood, C. (2021). The incessant inbox: Evaluating the relevance of after-hours e-mail characteristics for work-related rumination and well-being. *Stress and Health*, *37*(2), 341–352. https://doi.org/10.1002/smi.2999

Mulki, H., Haddad, H., Bechikh Ali, C., & Alshabani, H. (2019). L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language. *Proceedings of the Third Workshop on Abusive Language Online*, 111–118. https://doi.org/10.18653/v1/W19-3512

Nockleyby, J. T. (2000a). *Hate Speech in Encylopedia of the American Constitution* (2nd ed., pp. 1277–1279). Macmillan.

Nockleyby, J. T. (2000b). *Hate Speech in Encylopedia of the American Constitution* (2nd ed., pp. 1277–1279). Macmillan.

Onraet, E., Van Hiel, A., De Keersmaecker, J., & Fontaine, J. R. J. (2017). The relationship of trait emotional intelligence with right-wing attitudes and subtle racial prejudice. *Personality and Individual Differences*, *110*, 27–30. https://doi.org/10.1016/j.paid.2017.01.017

Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice Reduction: Progress and Challenges. *Annual Review of Psychology*, *72*(1), 533–560. https://doi.org/10.1146/annurev-psych-071620-030619

Park, J.-C., Kim, S., & Lee, H. (2020). Effect of work-related smartphone use after work on job burnout: Moderating effect of social support and organizational politics. *Computers in Human Behavior*, *105*, 106194. https://doi.org/10.1016/j.chb.2019.106194

Park, Y., Fritz, C., & Jex, S. M. (2018). Daily Cyber Incivility and Distress: The Moderating Roles of Resources at Work and Home. *Journal of Management*, *44*(7), 2535–2557. https://doi.org/10.1177/0149206315576796

Patterson, F., Ashworth, V., Zibarras, L., Coan, P., Kerrin, M., & O'Neill, P. (2012). Evaluations of situational judgement tests to assess non-academic attributes in selection: Situational judgement tests. *Medical Education*, *46*(9), 850–868. https://doi.org/10.1111/j.1365-2923.2012.04336.x

Pearson, C. M., Andersson, L. M., & Porath, C. L. (2005). Workplace incivility. In S. Fox & P. E. Spector (Eds.), *Counterproductive work behavior: Investigations of actors and targets.* (pp. 177–200). American Psychological Association. https://doi.org/10.1037/10893-008

Pearson, C. M., & Porath, C. L. (2005). On the nature, consequences and remedies of workplace incivility: No time for “nice”? Think again. *Academy of Management Perspectives*, *19*(1), 7–18. https://doi.org/10.5465/ame.2005.15841946

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. University of Texas at Austin.

Penney, L. M., & Spector, P. E. (2002). Narcissism and Counterproductive Work Behavior: Do Bigger Egos Mean Bigger Problems? *International Journal of Selection and Assessment*, *10*(1 & 2), 126–134. https://doi.org/10.1111/1468-2389.00199

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and Monitoring Hate Speech in Twitter. *Sensors*, *19*(21), 4654. https://doi.org/10.3390/s19214654

Petrovici, A., & Dobrescu, T. (2014). The Role of Emotional Intelligence in Building Interpersonal Communication Skills. *Procedia - Social and Behavioral Sciences*, *116*, 1405–1410. https://doi.org/10.1016/j.sbspro.2014.01.406

Phillips, K. W. (2014). How diversity makes us smarter. Scientific American, 311(4), 43-47.

Pierce, C. M., Carew, J. V., Pierce-Gonzalez, D., & Wills, D. (1977). An Experiment in Racism: TV Commercials. *Education and Urban Society*, *10*(1), 61–87. https://doi.org/10.1177/001312457701000105

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *CoRR, abs/1602.04938*. http://arxiv.org/abs/1602.04938

Richard, E. M., Young, S. F., Walsh, J. J., & Giumetti, G. W. (2020). Cyberaggression in Work-Related Email: Nomological Network and Links to Victims' Counterproductive Work Behavior. *Occupational Health Science*, *4*(1–2), 161–190. https://doi.org/10.1007/s41542-020-00056-3

Robinson, S. L., & O'Leary-Kelly, A. M. (1998). Monkey See, Monkey Do: The Influence of Work Groups on the Antisocial Behavior of Employees. *Academy of Management Journal*, *41*(6), 658–672. https://doi.org/10.5465/256963

Ross-Sheriff, F. (2012). *Microaggression, Women, and Social Work*. 4. https://doi.org/10.1177/0886109912454366

Ruggs, E. N., Martinez, L. R., & Hebl, M. R. (2011). How Individuals and Organizations Can Reduce Interpersonal Discrimination: Reduce Interpersonal Discrimination. *Social and Personality Psychology Compass*, *5*(1), 29–42. https://doi.org/10.1111/j.1751-9004.2010.00332.x

Runions, K., Shapka, J. D., Dooley, J., & Modecki, K. (2013). Cyber-aggression and victimization and social information processing: Integrating the medium and the message. *Psychology of Violence*, *3*(1), 9–26. https://doi.org/10.1037/a0030511

Sachdev, I., & Bourhis, R. Y. (1991). Power and status differentials in minority and majority group relations. *European Journal of Social Psychology*, *21*(1), 1–24. https://doi.org/10.1002/ejsp.2420210102

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, *abs/1910.01108*.

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. https://doi.org/10.18653/v1/W17-1101

Schumann, K., & Ross, M. (2010). Why Women Apologize More Than Men: Gender Differences in Thresholds for Perceiving Offensive Behavior. *Psychological Science*, *21*(11), 1649–1655. https://doi.org/10.1177/0956797610384150

Schwartz, B. (2016). Knock Yourselves Out: "Punching up" in American comedy. *The Baffler*, *31*, 134–146.

Schwartz, H. A., & Ungar, L. H. (2015). Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 78–94. https://doi.org/10.1177/0002716215569197

Shahapure, K. R., & Nicholas, C. (2020). Cluster Quality Analysis Using Silhouette Score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 747–748. https://doi.org/10.1109/DSAA49011.2020.00096

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. *CoRR*, *abs/1704.02685*. http://arxiv.org/abs/1704.02685

Sigurbergsson, G. I., & Derczynski, L. (2019). Offensive Language and Hate Speech Detection for Danish. *ArXiv:1908.04531 [Cs]*. http://arxiv.org/abs/1908.04531

Speer, A. B. (2021). Scoring dimension-level job performance from narrative comments: Validity and generalizability when using natural language processing. *Organizational Research Methods*, *24*(3), 572–594.

Sturdivant, M., Yibass, S., Abraham, E., & Hauenstein, N. M. A. (2017). Using Situational Judgment Tests To Study Subtle Discrimination. *Industrial and Organizational Psychology*, *10*(1), 94–97. https://doi.org/10.1017/iop.2016.107

Sue, D. W. (2010). *Microaggressions in everyday life: Race, gender, and sexual orientation*. John Wiley & Sons.

Sue, D. W., Alsaidi, S., Awad, M. N., Glaeser, E., Calle, C. Z., & Mendez, N. (2019). Disarming racial microaggressions: Microintervention strategies for targets, White allies, and bystanders. *American Psychologist*, *74*(1), 128–142. https://doi.org/10.1037/amp0000296

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A. M. B., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist*, *62*(4), 271–286. https://doi.org/10.1037/0003-066X.62.4.271

Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Zhou, D. (2020). *MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices*. arXiv. https://doi.org/10.48550/ARXIV.2004.02984

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. *CoRR*, *abs/1703.01365*. http://arxiv.org/abs/1703.01365

Tiffin, P. A., Paton, L. W., O'Mara, D., MacCann, C., Lang, J. W. B., & Lievens, F. (2020). Situational judgement tests for selection: Traditional vs construct-driven approaches. *Medical Education*, *54*(2), 105–115. https://doi.org/10.1111/medu.14011

Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Vance, C. M., Ensher, E. A., Hendricks, F. M., & Harris, C. (2004). Gender-Based Vicarious Sensitivity to Disempowering Behavior in Organizations: Exploring an Expanded Concept of Hostile Working Environment. *Employee Responsibilities and Rights Journal*, *16*(3), 135–147. https://doi.org/10.1023/B:ERRJ.0000038649.75806.28

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *CoRR*, *abs/1706.03762*. http://arxiv.org/abs/1706.03762

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, *15*(12), e0243300. https://doi.org/10.1371/journal.pone.0243300

Vyas, L. (2022). "New normal" at work in a post-COVID world: Work–life balance and labor markets. Policy and Society, 41(1), 155–167. https://doi.org/10.1093/polsoc/puab01

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, 88–93. https://doi.org/10.18653/v1/N16-2013

Washington, G., Mance, G., Aryal, S., Ngueajio, M., Salaam, C., & Alim, C. (2021). *ABL-MICRO: Opportunities for Affective AI Built Using a Multimodal Microaggression Dataset*. 7.

Woodley, H. J. R., Bourdage, J. S., Ogunfowora, B., & Nguyen, B. (2016). Examining Equity Sensitivity: An Investigation Using the Big Five and HEXACO Models of Personality. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.02000

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *CoRR, abs/1906.08237*. http://arxiv.org/abs/1906.08237

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. *CoRR, abs/1909.00161*. http://arxiv.org/abs/1909.00161

Young, A. M., Vance, C. M., & Ensher, E. A. (2003). Individual Differences in Sensitivity to Disempowering Acts: A Comparison of Gender and Identity-Based Explanations for Perceived Offensiveness. *Sex Roles*, 9.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the Type and Target of Offensive Posts in Social Media. *Proceedings of the 2019 Conference of the North*, 1415–1420. https://doi.org/10.18653/v1/N19-1144

Appendix

Example of Data Frame

sender_genders	sender_races	receiver_genders	receiver_races	max_email_offense	avg_email_offense	time_sent	Email text	lag
male	White	female	White	3.828602	2.63285		test successful. way to go!!!	
male	White	male	White	3.941368	3.941368		How about either next Tuesday or Thursday? Ph...	
male	White	male	Black	4.333849	4.333849		For delivered gas behind San Diego, Enron Ener…	
male	White	male	White	4.079687	4.079687		Here are the rentrolls: Open them and save...	
male	White	male	Black	3.941368	3.941368		Here are the names of the west desk members by...	
male	White	female	White	6.527519	5.591187		Hey woman!! How are things going? Happy with the job? Happy with the man? Be careful, something happened to my woman and she has a really big stomach	
female	White	female	White	6.61543	4.850307		Do I need to come slap some sense in you?	
female	White	female	White	6.61543	6.61543		What's and infactuation? You being a brat? That will never pass	