

Computational Politeness in Natural Language Processing: A Survey

PRIYANSHU PRIYA*, Indian Institute of Technology Patna, India

MAUJAMA FIRDAUS, University of Alberta, Canada

ASIF EKBAL*, Indian Institute of Technology Patna, India

Computational approach to politeness is the task of automatically predicting and/or generating politeness in text. This is a pivotal task for conversational analysis, given the ubiquity and challenges of politeness in interactions. The computational approach to politeness has witnessed great interest from the conversational analysis community. This article is a compilation of past works in computational politeness in natural language processing. We view four milestones in the research so far, *viz. supervised and weakly-supervised feature extraction to identify and induce politeness in a given text, incorporation of context beyond the target text, study of politeness across different social factors, and study the relationship between politeness and various socio-linguistic cues*. In this article, we describe the datasets, approaches, trends, and issues in computational politeness research. We also discuss representative performance values and provide pointers to future works, as given in the prior works. In terms of resources to understand the state-of-the-art, this survey presents several valuable illustrations — most prominently, a table summarizing the past papers along different dimensions, such as the types of features, annotation techniques, and datasets used.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; • **Human-centered computing**;

Additional Key Words and Phrases: Politeness, (im)politeness, computational politeness, linguistic variations, politeness analysis

ACM Reference Format:

Priyanshu Priya, Mauajama Firdaus, and Asif Ekbal. 2024. Computational Politeness in Natural Language Processing: A Survey. In *Proceedings of (ACM Comput. Surv.)*. ACM, New York, NY, USA, 41 pages. <https://doi.org/10.1145/3654660>

1 INTRODUCTION

The Free Dictionary¹ defines the term polite as being “marked by or showing consideration for others and observance of accepted social usage”. The Oxford Learner’s Dictionary² defines the term “polite” as “having or showing good manners and respect for the feelings of others”. The term polite has been defined in multiple ways by the Merriam-Webster³ dictionary like “showing or characterized by correct social usage”, “marked by an appearance of consideration, tact, deference, or courtesy” and “marked by a lack of roughness or crudities”. All the definitions of the term *polite* encompass behavior that is socially correct and acceptable. According to [17, 80], politeness is a ubiquitous quality of human communication. In almost every situation, the speaker can be more or less polite to the listener, which might affect

*Corresponding authors

¹<https://www.thefreedictionary.com/>

²<https://www.oxfordlearnersdictionaries.com/>

³<https://www.merriam-webster.com/>

Authors’ addresses: Priyanshu Priya, priyanshu_2021cs26@iitp.ac.in, Indian Institute of Technology Patna, Bihta, Patna, Bihar, India, 801106; Mauajama Firdaus, mauzama.03@gmail.com, University of Alberta, Edmonton, Alberta, Canada; Asif Ekbal, asif@iitp.ac.in, Indian Institute of Technology Patna, Bihta, Patna, Bihar, India, 801106.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

Manuscript submitted to ACM

1

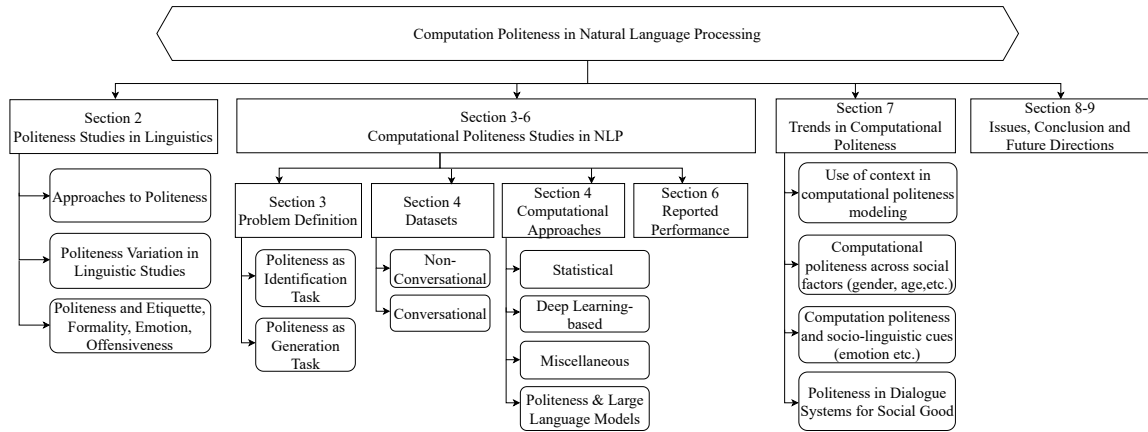


Fig. 1. Organization of the present survey

the speaker's social goals. Politeness is conveyed by a specific and extensive collection of linguistic markers that can alter the information content of an utterance. Occasionally, politeness is an act of commission such as saying “thank you”, “you’re welcome” and “please”, while the other times it is an act of omission, for instance, declining to be contradictory. Moreover, the mapping of politeness indicators is frequently context-specific (workplace vs. family vs. friends), speaker-specific (male vs. female), goal-specific (buyer vs. seller), or culture-specific. Although politeness often relies on a shared pool of linguistic markers for social coordination between the speaker and the listener, the importance and valence of each marker are often determined by the context, which poses a challenge for computational politeness studies.

The growth and development of Artificial Intelligence (AI) and Natural Language Processing (NLP) have led to a rise in the popularity of human-computer interaction. Consequently, the need to emulate human behavior has increased. Politeness is one of the most significant characteristics of human beings, and we often make decisions about when, where, and how to use politeness in communication to promote social interactions. Since politeness exemplifies the socially desirable behavior that helps regulate the essential aspects of social interaction, computational approaches to politeness are crucial for analyzing, simulating, and facilitating such interactions.

The challenges and benefits of politeness to human-human and human-machine interactions have led to an interest in computational approaches to politeness as a research problem. The computational approach to politeness involves developing tools that automatically identify, evaluate, adapt, interpret, and generate linguistic cues conveying polite behavior in human-human and human-machine interactions. Developing a computational model of politeness and applying and evaluating it in different contexts is a challenging task due to the nuanced ways in which politeness may be expressed. The initial work by Danescu-Niculescu-Mizil et al. [33] on politeness leveraged foundational research in linguistic politeness [17, 80] to discern linguistic markers of politeness in online requests. This was achieved through a computational framework based on Support Vector Machines (SVMs), which has since garnered significant attention within the Natural Language Processing (NLP) community. The computational approach to politeness encompasses a variety of data types and methodologies. This combination has brought in some intriguing advancements in the field of computational politeness.

As shown in Figure 1, we organize the remainder of the article as follows: Section 2 describes the linguistic studies of politeness. To have a better understanding of the different aspects of previous work in computational politeness, this survey article examines politeness in six phases. Sections 3, 4, and 5 successively discuss the different problem definitions, datasets and approaches for computational politeness studies. In Section 6, we present the reported performance values. Section 7 covers the underlying trends in computational politeness followed by the issues in Section 8. Eventually, Section 9 concludes the article and makes recommendations for future research work. We have included tables and examples in this survey paper that may be used to gain a better understanding of computational politeness studies. The primary objective of this survey is to provide a comprehensive overview of previous work in the field of computational politeness, with insights into future research that may prove valuable to researchers interested in computational politeness and related fields.

2 POLITENESS STUDIES IN LINGUISTICS

Before discussing computational approaches to politeness, we outline a brief overview of linguistic studies about politeness in this section. Politeness is amongst the most crucial components of human communication, and it has the potential to determine the success or failure of any human interaction [75]. Politeness studies have taken center stage in pragmatics research during the past few decades. These studies have mainly focused on applying various communicative tactics to preserve or promote social harmony. Politeness has been defined in different ways by different theoreticians. Some of the most notable definitions are as follows:

- (1) Leech [83] defines politeness as “maintain[ing] the social equilibrium and the friendly relations which enable us to assume that our interlocutors are being cooperative in the first place”.
- (2) Another notable definition of politeness was proposed by Brown et al. [17] as “politeness, like formal diplomatic protocol (for which it must surely be the model), presupposes that potential for aggression as it seeks to disarm it, and makes possible communication between potentially aggressive parties”.
- (3) Lakoff [81] defines politeness as “a means of minimizing confrontation in discourse - both the possibility of confrontation occurring at all, and the possibility that a confrontation will be perceived as threatening”.
- (4) Thomas [158] summarizes the research agendas of linguists such as those mentioned above who are concerned with the study of pragmatic politeness as “All that is really being claimed is that people employ certain strategies (including the 50+ strategies described by Leech, Brown and Levinson, and others) for reasons of expediency – experience has taught us that particular strategies are likely to succeed in given circumstances, so we use them”.
- (5) Watts [170] holds that politeness should be defined through a discursive approach. He believes that such struggle defines “the ways in which (im)polite behaviour is evaluated and commented on by lay members and not with ways in which social scientists lift the term (im)politeness out of the realm of everyday discourse and evaluate it to the status of a theoretical concept in what is frequently called Politeness Theory”.
- (6) Yule [177] in his study asserts that individuals can be polite through “being tactful, modest and nice to other people”.

2.1 Approaches to Politeness

Several linguistic studies describe different approaches to politeness. Starting from the classic view of politeness that is based on traditional pragmatic theories, most notably Conversational Implicature [51] and Speech Act Theory [143] to post-modern view, politeness has been conceptualized as maxim-based, face-based, discursive, relational and frame-based approaches.

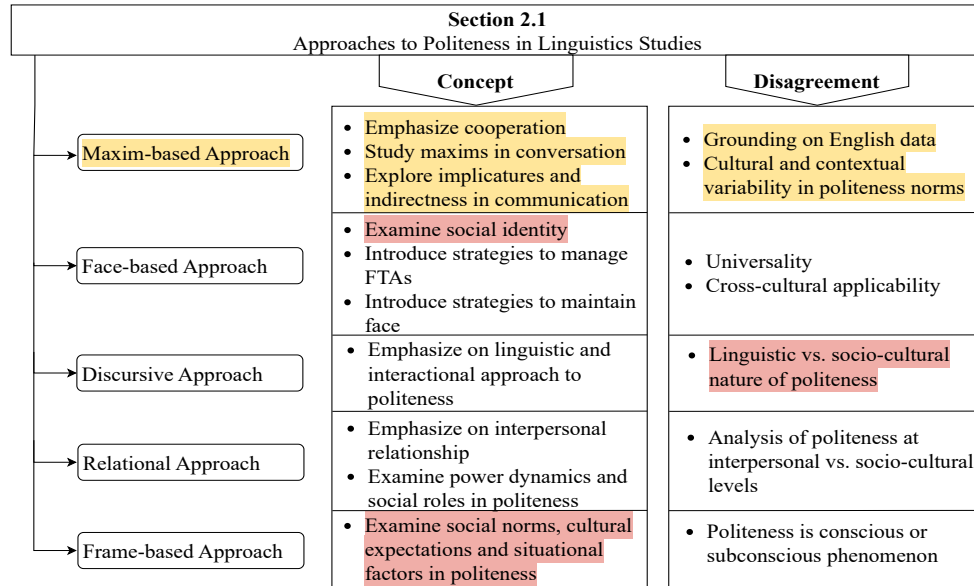


Fig. 2. Overview of various approaches to politeness in linguistic studies

The overview of the different facets of politeness considered in these approaches and their significant disagreements/criticisms in the linguistic studies are depicted in Figure 2. These different approaches to politeness offer distinct perspectives on how people manage politeness in communication. They differ in their emphasis on linguistic maxims, face, discourse, relationships, and social frames.

Maxim-based Approach to Politeness. The language philosophers like Lakoff and Leech utilize maxims to describe politeness. Lakoff identifies two norms of pragmatic competence: (i) ‘be clear’ and (ii) ‘be polite’. The former is formalized using Grice [51] Cooperative Principle (CP), while the latter is formalized in terms of Leech [83] Politeness Principle (PP). The Cooperative Principle is stated as “*Make your conversational contribution such as is required, at the state at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged*”. It comprises of four maxims of conversation, viz. “maxim of quantity” (informative), “maxim of quality” (truthful), “maxim of relation” (relevant), and “maxim of manner” (clear), collectively known as the Gricean maxims. The Politeness Principle, on the other hand, is defined as “*in its negative form, minimize (other things being equal) the expression of impolite beliefs and in the corresponding positive form, maximize (other things being equal) the expression of polite beliefs*”. It comprises of six maxims, viz. “Tact maxim” (Minimize cost to other; Maximize benefit to other), “Generosity maxim” (Minimize benefit to self; Maximize cost to self), “Approbation maxim” (Minimize dispraise of other; Maximize praise of other), “Modesty maxim” (Minimize praise of self; Maximize dispraise of self), “Agreement maxim” (Minimize disagreement between self and other; Maximize agreement between self and other) and “Sympathy maxim” (Minimize antipathy between self and other; Maximize sympathy between self). These maxims are popularly known as politeness maxims. The PP trades off with the CP; indeed, it offers considerable explanatory power to the CP, in the sense that the CP explains *how* individuals express indirect meanings, whereas the PP explains *why* individuals express indirect meanings. To illustrate this, Leech [84] analyzed an excerpt from the G.B. Shaw’s play “*You Never Can Tell*” which is as follows:

[Context: The waiter, as the most tactful communicator, has been chosen to convey some bad news to Crampton, namely that he is Philips' father.]

Waiter: [smoothly melodious] Yes, sir. Great flow of spirits, sir. A vein of pleasantry, as you might say sir ... The young gentleman's latest is that you're his father.

Crampton: What!

Waiter: Only his joke, sir, his favourite joke. Yesterday I was to be his father ...

This excerpt demonstrates that the waiter is successful in communicating Crampton's parentage. Meanwhile, to retain politeness (i.e., to avoid upsetting), he further claims that this information is false and just a joke; in that way, he compromises the "maxim of quality".

Both CP and PP are exposed to criticism. The CP receives criticism owing to the fact that Gricean maxims do not apply universally, and PP is disputed by arguing that it has a large number of maxims which results in the overlap. Brown and Levinson, among others, have emphasized this aspect and formulated their theory of politeness, popularly known as Brown and Levinson's (B&L) Politeness Theory [16, 17].

Face-based Approach to Politeness. Brown and Levinson [16] (B&L) Politeness Theory is the most well-known and researched in the field of politeness. B&L studied politeness by utilizing Goffman [48] idea of face, which defines face as "the positive social value a person effectively claims for himself by the line others assume s/he has taken during a particular contact. Face is an image of self-delineated in terms of approved social attributes". B&L treated face as **basic wants**, which each member is aware that every other member wants, which are generally in each member's best interests to satisfy partially. They introduced two notions of face, namely *positive face* and *negative face*. Positive face reflects an individual's wants to be desirable to at least some others. Negative face, on the other hand, is the individual's desire that his actions be unimpeded by others. This theory is dependent on the hypothesis that certain kinds of speech acts threaten either the speaker's (S) face or the addressee's (H) face. Consequently, they define politeness as the necessary element to mitigate such face-threatening acts (FTAs). They claimed that different sociological variables, namely Relative Power (P, power dynamic between S and H, whether H is superior, subordinate, or at the same level as S), Distance (D, social distance between S and H, whether S and H are close friends, family members or colleague) and Absolute Ranking (R, degree of sensitivity of the topic in a particular situation or culture) determine the amount of face threat of a particular act. Eventually, Brown and Levinson [16] formulated five sets of super-strategies for doing FTAs that are associated with the degree of face threat, namely *bald on-record*, *positive politeness*, *negative politeness*, *Off-record/indirect* and *withhold the FTA*. They mentioned in their politeness theory that bald on-record as a politeness strategy is a direct way of saying things to the hearer without making an effort to reduce the potential threat to the hearer's self-esteem. However, there are some methods to employ this strategy that implicitly mitigate FTAs, such as straightforwardly offering advice (e.g., *Leave it, I will clean it up later*). They stated that the frequent use of such a strategy may cause discomfort to the hearer; hence, it is typically employed in contexts where the speaker shares a strong bond with the hearer, such as within family or close friendships. They enumerated several scenarios in which one might opt for this strategy, including urgent situations (e.g. *Watch out!*), situations when efficiency is necessary (e.g., *Hear me out*), and so on. Positive politeness strategies aim to mitigate the potential harm to the hearer's positive self-image. These strategies are employed to enhance the hearer's feeling of self-esteem, their interests, or their possessions. These are typically applied when there is a level of familiarity among the participants or when it is essential to address an individual's need for positive self-esteem. Negative politeness strategies, on the other hand, are directed towards the hearer's negative face and emphasize avoidance of imposition on the hearer. The off-record (indirect) strategy involves the use of indirect language and removes the speaker from the potential to be imposing. A few example sentences for each of these strategies are shown in Table 1.

B&L work primarily focuses on positive and negative politeness, which in turn, consists of 15 output strategies (i. Notice, Attend to H, ii. Exaggerate, iii. Intensify interest to H, iv. Use in-group identity markers, v. Seek agreement, vi. Avoid disagreement, vii. Presuppose/raise/assert common ground, viii. Joke, ix. Assert/presuppose S's knowledge of and concern for H's wants, x. Offer, promise, xi. Be optimistic, xii. Include both S and H in the activity, xiii. Give (or ask for) reasons, xiv. Assume or assert reciprocity, xv. Give gifts to H) and 10 output strategies (i. Be conventionally indirect, ii. Question, hedge, iii. Be pessimistic, iv. Minimize the imposition, v. Give deference, vi. Apologise, vii. Impersonalise S and H, viii. State the FTA as a general rule, ix. Nominalise, x. Go on-record as incurring a debt or as not incurring a debt), respectively [16]. Each output strategy serves as a method for fulfilling the strategic objectives of a higher-level super-strategy.

Table 1. Strategies for doing FTAs. S refers to the speaker, and H refers to the addressee.

| Strategy | Explanation | Example |
|-----------------------|---|--|
| Bald on-record | Perform the FTA in the most direct, clear, unambiguous and concise way possible | (i) Please make a cup of tea. (said to a close friend) (ii) Don't forget to clean the blinds! |
| Positive politeness | Minimizing the threat to H's positive face wants | (i) Yum! You make such a great sandwich! Would you make some? (ii) This piece of work is really fantastic. Well done! |
| Negative politeness | Redressal of H's negative face wants | (i) Could you please serve the meal? (ii) I hope offense will not be taken. |
| Off-record (indirect) | Perform the FTA such that there is more than one unambiguously attributable intention so that the actor cannot be held to have committed himself to one particular intent | (i) If S says "I have stomachache", the H may infer that S is asking for some medicines, however, if questioned, S may deny this. (ii) S: Are you going out? -H: Yes, but I'll come home early. |
| Withhold the FTA | The speaker makes a conscious decision to abstain from performing the FTA | - |

However, B&L's theory of politeness is also exposed to criticism mainly because of its universality claim [28]. They argued that their conception of "face" applied across diverse cultures and languages. Moreover, they failed to articulate the role of context in the judgment of politeness; they assumed that certain speech acts/strategies are inherently polite and will produce the same politeness effects. While addressing these issues among others, several linguists posit discursive approaches to politeness.

Discursive Approach to Politeness. The discursive approach to politeness is a noteworthy line of work in the field of linguistic politeness [47, 94–96, 105, 170, 171]. Locher and Watts [96] made a distinction between first-order politeness (politeness as is perceived by the speakers of the language, the lay person's understandings) and second-order politeness (theoretical constructs of politeness model proposed in the literature) and claimed that there is no one meaning of the term 'politeness', but rather a subject of discursive struggle. The discursive approach to politeness emphasizes the context and the situated and emergent meanings rather than pre-defined meanings. Furthermore, this approach acknowledges that politeness is evaluative in essence i.e., it is used to make judgments about people's behavior and is social norm-oriented, which provides insight into the concept of appropriateness. Locher [94] stated that the "discursive approach to politeness stresses that we first of all have to establish the kind of relational work the interactants in question employ to arrive at an understanding of the then-current norms of interaction".

Relational Approach to Politeness. The relational approach to politeness is based on the concept of second-order politeness. Christie [25] pointed out that politeness is some form of 'relational work'. The term 'relational work' relates to communication at the interpersonal level that helps in negotiating the relationships among people in interaction [96, 97]. It encompasses the whole spectrum of behavior like rude, impolite, normal, polite, etc. As such, 'relational work' cannot be simply limited to a distinction between polite and impolite behavior; in fact, there are some forms of 'relational work', which is neither polite nor impolite. Locher and Watts [96] addressed facework from a new perspective, namely 'relational work', by extending the research scope from "direct, impolite, rude, or aggressive interaction to polite interaction, encompassing both appropriate and inappropriate forms of social behavior" [95]. Additionally, they

proposed a lay interpretation of polite conduct as politic/appropriate verbal or non-verbal behavior in any social encounter and defined polite behavior as the surplus of politic behavior. This is illustrated in Culpeper [28] using an example from Watts [170] (Example1) and a rework of this example (Example2).

Example1: politic behaviour

S: Would you like some more coffee?

H: Yes, *please*

Example2: polite/positively marked behaviour

S: Would you like some more coffee?

H: Yes, *please, that's very kind, coffee would be wonderful.*

There are no separate distinctions between unmarked politic behavior and positively marked behavior, as they are of the same character and may overlap in specific contexts. This approach represents the premise that the understanding of polite and impolite behavior varies depending on contextual and social variables.

Frame-based Approach to Politeness. The authors in [154–157] introduced the frame-based approach to politeness. In these works, the authors asserted that the particular context of use and concrete linguistic realizations i.e., formulae, which together constitute “frames”, should be analyzed. This helps eliminate the problematic concept of directness. Furthermore, the authors in [28, 156, 157] stated that “[i]t is the regular co-occurrence of particular types of context and particular linguistic expressions as the unchallenged realizations of particular acts that create the perception of politeness”. The frame-based approach to politeness is data-driven, in contrast to the classic and post-modern views of politeness, which are theory-driven [154]. Data-driven means that it is predicated in the analysis of a large Cypriot Greek corpus, and it recognizes the norms to the level that these may be observed empirically. These norms were then evaluated quantitatively in order to identify regularities of co-occurrence between linguistic expressions and their extra-linguistic contexts of use. This approach provides a rich, coherent, and pragmatic account of politeness. It is mostly used by [154–157] to investigate politeness in Cypriot Greek, and applying this approach to other contexts is another dimension of research in the field of linguistic politeness.

While researchers often disagree about which approach is most comprehensive or applicable in all contexts, each approach contributes valuable insights into understanding politeness in communication. The choice of approach may vary depending on the specific research question, context, and cultural considerations.

2.2 Politeness Variations in Linguistic Studies

Politeness analysis is a complex phenomenon because the perception of (im)politeness varies from one individual to another, and there are multiple gradations of more- or less- (im)polite behavior [50]. Thus, politeness needs to be examined in a social context. Several factors influence the use of politeness in a particular situation. In this section, we will discuss such factors emphasizing the role of gender, age, and culture in politeness usage in interaction.

Politeness and Gender. There has been a good amount of research on language and gender traits that attempt to identify and explain the differences in speech styles of males and females. One of the most significant distinctions has been discovered in linguistic politeness. Many studies in socio-linguistics have observed the variations in the use of politeness strategies between males and females and concluded that linguistic politeness is influenced by gender [17, 79, 100, 105]. According to these studies, women are more likely to use politeness strategies and compliments than men do while communicating. The work of Lakoff [79] is a groundbreaking study in addressing the problem of gender-related differences in politeness. The author claimed that women sound much more polite than men as they tend to use tag-questions more often than men. The tag-questions do not force belief or agreement on the addressee, and hence their use by women while interacting makes them sound more polite. Moreover, the author claimed that female speech includes many super-polite forms like “*would you mind...*”, “*would it be comfortable...*” etc., and consultative

devices in utterances, for example, “*could you tell what trouble you’re facing?*”, “*will you clean the window?*”. Holmes [58] agreed that, in general, females use more tags than males, as claimed by Lakoff [79]. Besides, the author also identified that females tend to consider question-tags to indicate politeness. Males, on the other hand, use them to express uncertainty. Holmes [57] stated that compliments are considered to be positive politeness strategies as they aim to strengthen the solidarity between the interlocutors. The notable gender difference in politeness is shown in how women and men utilize compliments. According to this study, women tend to give and receive more compliments than men. Wardhaugh [168] argued that females employ polite phrases like “*please*”, “*if you don’t mind*” etc. more often than males in conversations. Montgomery [116] claimed that out of age, social status, race, and gender, gender is the single most contributing factor to the occurrence of multiple modals. This study argued that both men and women are more sensitive to the face of women with whom they are conversing than to those of men. Mills [105] provided a comprehensive examination of the assumptions underlying the relationship between gender and politeness. In this study, the author developed a dynamic conception of the association between gender and politeness and established a discourse-based analytical perspective on politeness. The author argued that politeness is a set of strategies rather than a choice of suitable utterances.

Politeness and Age. Age is one of the significant aspects that might affect one’s views, experiences, social interactions, and linguistic engagements. Age-specific views of politeness hold that people of different age groups talk and behave differently as far as politeness is concerned. Pfuderer [128] suggested that linguistic politeness is shown precisely when the listener is older or senior in the position. Lakoff [80] claimed that adults must adhere to politeness norms and maxims. Researchers have investigated the influence of age on emotional well-being and emotional language. Carstensen et al. [22] have shown that emotional well-being increases with age and that older people exhibit more positive emotions than younger ones. As politeness is an essential element of affective conversational behavior [53], recently, researchers have started investigating how age differences affect the choice of politeness strategies in interactions. Many researchers have focused on determining how and when children acquire politeness and how they use and comprehend the different polite forms while conversing. Küntay et al. [78] pointed out in their work that early studies on politeness focused chiefly on English-speaking children with an emphasis on the learning of politeness routines such as “*thank you*”, “*please*”, “*I’m sorry*”, and comparable findings have been observed by the politeness studies with the children from other cultures like the one with Italian children [8], Japanese-speaking children [119], to mention a few, demonstrating that they are also socialized into politeness norms at a younger age. Studies also suggest that the use of the word “*please*” appears early, while indirect requests and politeness devices are used with increasing age. Moreover, the works by Maratsos [99] and Shatz and Gelman [147] showed that children are sensitive to their listeners as they use polite forms according to their needs and age.

Politeness and Culture. Politeness is a crucial aspect of communication, especially in cross-cultural contexts where miscommunications or misunderstandings may adversely affect the relationship between the interlocutors. The maxim-based approach to politeness introduced by Leech [85] accommodates a broader range of cultural variations. This seminal work is likely to be the most appropriate for East-West comparisons. Lakoff [80] argued that politeness tends to avoid conflicts, and there are three rules for avoiding conflicts: camaraderie, deference, and distance. The authors mentioned that different cultures have different dominant rules; for example, Australian culture often emphasizes camaraderie, Japanese culture emphasizes deference, and British culture emphasizes distance. There are also linguistic differences in how politeness is expressed and perceived across cultures [59, 65]. For instance, there are significant differences in how politeness is conveyed and evaluated by the Chinese and Americans; what might be considered polite by the Chinese may sound utterly rude to a typical American. To illustrate, a cordial offer in Mandarin, “Come

and eat with us, or I will get mad at you!” may sound extremely impolite to an American [52]. Furthermore, there are differences between Chinese and Westerners in the ways of greeting and farewell, expressing thanks, praising, and in using addressing terms for others. For example, Westerners often use “Hi!” or “Hello!” to greet, while the Chinese prefer to ask “What are you doing here?” or “Have you eaten?”, which are perceived as interference with privacy by the Westerners and hence, considered impolite. Thus, understanding politeness across cultures is vital as it helps to enhance intercultural communication.

Some of these linguistic theories of politeness have a strong correlation with advancements in computational politeness. For instance, a seminal work by Danescu-Niculescu-Mizil et al. [33] was an initial attempt to carry forward the groundwork done in discourse pragmatics, primarily by the [17, 80], in identifying and discussing social politeness. Moreover, the evaluative nature of politeness, necessary for making judgments about people’s behavior, is essentially the different forms of contextual information that computational politeness aims to capture. In general, politeness means socially correct and appropriate speech and behavior. However, what context is required to understand and evaluate politeness forms a crucial component. This contextual knowledge can be of various forms, like cultural, gender-specific, age-specific, and/or situational contexts. To illustrate, compare the polite example, ‘*Hello! How are you?*’ with another ‘*Let’s go for a cup of tea.*’ The former is likely to be polite to all the people. The latter is likely to be polite for most people (friends, family, or long-standing colleagues). However, for the people who are new or senior in a position, let’s say, one’s employer, the statement is not polite. A direct statement like the latter one may threaten the hearer’s face in restricting their freedom of action and thus may be perceived as impolite by them. Hence, the politeness understanding and computational approaches to politeness are contingent on what information or context is known.

2.3 Relationship with Etiquette, Formality, Emotion and Offensiveness

Politeness is related to other forms of social language and behavior, such as **etiquette, formality, and emotion**. The terms “etiquette” and “politeness” are likely to conjure up images of formal curtsies and the proper dining fork to use. Although both the terms are often used interchangeably, **there is a subtle difference between the two. Etiquette is the protocol via which we communicate politeness.** It uses verbal, physical, gestural, and even more **primitive forms of communication**. For instance, people can show respect by posture, quiet speech, or by using titles and honorifics. Similar to politeness, there are differences in the use of etiquette codes across cultures [103]. Politeness and etiquette are very much at the forefront of managing social interactions and thus play a crucial role in training and predicting social interaction behaviors and perceptions across cultures.

Formality is another important dimension of stylistic variation in language [10, 61]. Stylistic distinctions can frequently have a greater influence on how a listener interprets a statement than the literal meaning does [60], and the ability to recognize and respond to differences in formality is a necessary part of complete language understanding. Heylighen and Dewaele [55] argued that formality emerges out of a communicative objective - to maximize the amount of information being conveyed to the listener while at the same time maintaining (or at least appearing to maintain) Grice’s communicative maxims of Quality, Quantity, Relevance, and Manner as much as possible [51]. Fischer [44] discussed the use of “respect vocabulary” and “formal etiquette” as the display of politeness marking a formal situation. Formality is closely linked with various stylistic dimensions, including politeness [15, 64], suggesting a positive correlation between them.

Another aspect of pragmatics that has gained importance in recent politeness studies is the emotional aspects [82, 136]. Brown et al. [17] mentioned displaying emotion or lack of control of emotions as positive politeness strategies

or potentially face-threatening acts. The authors further referred to the ‘communication of affect’ (display rules and emotional stance) as the display of affect, which is socially constructed, with cultural and situational expectations about what and how feelings should be displayed. They mentioned that communication of appropriate levels of affect is primarily concerned with linguistic politeness. Culpeper et al. [31], in his description of requests, pointed out the significance of affect and familiarity as components of B&L’s *distance* (D) variable. Kienpointner [70] argued that the variables, *power* (P) and *distance* (D), proposed by B&L, also contain the elements of emotions. Culpeper [29] asserted that “*displaying emotions such as contempt or anger has nothing in itself to do with impoliteness. However, somebody displaying great contempt for and anger at someone and doing so publicly may be judged (..) to have acted in an inappropriately and unfairly hurtful way (..), causing an emotional reaction such as embarrassment or anger.*” Numerous linguistic studies indicate that emotions are essential in appraisals and relational work. Therefore, investigating the relationship between emotions and politeness can benefit linguistic politeness research.

Politeness is often discussed in connotation with another linguistic aspect, offensiveness. Politeness and offensiveness are essential aspects of linguistic communication that influence interpersonal relationships and social dynamics. Politeness encompasses various linguistic strategies employed to convey respect, courtesy, and consideration for others, facilitating a positive and harmonious atmosphere. Offensiveness, on the other hand, pertains to the use of language that is disrespectful, rude, or hurtful to others. The notion of offensiveness lies at the core of impoliteness [29], and it can take various forms from general and often harmless profanity to derogatory and abusive language intended to cause harm and emotional distress, such as cyberbullying and hate speech [169]. Bousfield [13] and Culpeper [29] delved into the concept of impoliteness, equating the intent to offend with posing a threat or causing damage to one’s face. Culpeper [29, 30], and Haugh [54] proposed a theoretical framework to explain impoliteness, considering it as an attitudinal stance of the speaker. Offense, on the other hand, can be perceived as emotional reactions experienced by recipients or the cause of such reactions. It’s important to note that acting impolitely doesn’t always equate to offending, as the perception of offense varies among interlocutors. For example, interrupting a teammate during a presentation with disruptive comments may be seen as impolite, but using derogatory language to insult their intelligence is considered offensive. Offensiveness, therefore, can be viewed as an extreme form of impoliteness involving a direct attack on someone’s dignity or identity.

Nevertheless, it cannot be denied that instances of impoliteness/offensiveness are closely linked to the concept of politeness [28]. The degree of impoliteness or offensiveness often varies depending on the expected level of politeness in a given situation. For example, asking the Vice Chancellor of a university to be quiet might be more offensive than making the same request to a young daughter. Vogel [163] mentioned that behaviors of politeness and impoliteness can modulate offensiveness. They exist on a continuum of communication behaviors and can coexist within the same interaction. A message may have a polite tone but still convey an offensive message or criticism. Conversely, impolite language may not necessarily intend to be offensive; it could stem from a lack of awareness of social norms or an expression of frustration rather than an intentional attempt to harm. The relationship between politeness and offensiveness is multifaceted, influenced by context, intent, and individual perception [28]; what one person deems polite, another may consider impolite or offensive, highlighting the subjective nature of these concepts.

These characteristics of politeness pave the way for a few works in computational politeness research, which we will discuss in forthcoming sections.

3 PROBLEM DEFINITION

The advancements in machine learning and deep learning techniques profoundly impact the performance of various Natural Language Processing (NLP) tasks, such as Part-of-Speech tagging, Named Entity Recognition, and Semantic Parsing, to name a few. These techniques have also demonstrated remarkable success in contextual language understanding and generation, enabling breakthroughs in computational discourse pragmatics. Politeness is one of the widely studied topics in discourse pragmatics. Computational models and simulations of politeness-oriented interactions hold significant utility across various applications. These include modeling customer behavior, generating responses for conversational agents and similar intelligent systems like robots, comprehending the dynamics of online communities, and facilitating language learning, among others. In computational politeness, research predominantly falls into two overarching categories: **identification** and **generation**, which we will discuss in subsequent sub-sections. Table 2 and Table 3 provide a summary of past works in computational politeness identification and politeness generation, respectively. While several interesting conclusions may be drawn from both the tables, the three notable ones are: **(i)**. requests are the pre-dominant text form for computational politeness analysis; **(ii)**. politeness is imperative in both conversational (open-domain and task-oriented) and non-conversational settings for promoting user experience, and **(iii)**. incorporation of contextual information in the form of conversational history is a recent trend in computational politeness modeling for dialogue systems.

Table 2. Summary of computational approaches to politeness identification along different parameters.

| | Datasets | | Languages | | Approach | | | Annotation | | Features | | | | Context | |
|-------------------------------------|--------------------|----------------|-----------|-------|------------|-----------------------------|------------------------|------------|-----------|----------|---------|-----------|-------|--------------|-------|
| | Non-conversational | Conversational | English | Other | Supervised | (Semi/Weakly/Un-)Supervised | Reinforcement Learning | Manual | Automatic | N-grams | Lexical | Syntactic | Other | Conversation | Other |
| Alexandrov et al. [2] | | ✓ | | ✓ | ✓ | | | | | | ✓ | ✓ | | | |
| Danescu-Niculescu-Mizil et al. [33] | ✓ | | ✓ | | ✓ | | | ✓ | | ✓ | ✓ | ✓ | | | |
| Li et al. [88] | ✓ | | ✓ | ✓ | ✓ | | | ✓ | | | ✓ | ✓ | | | |
| Aubakirova and Bansal [4] | ✓ | | ✓ | | ✓ | | | | | | | | | | |
| Chhaya et al. [24] | ✓ | | ✓ | | ✓ | | | ✓ | | | ✓ | ✓ | | | |
| Kumar [76] | ✓ | | | ✓ | ✓ | | | ✓ | | ✓ | | | ✓ | | |
| Mishra et al. [108] | | ✓ | ✓ | | ✓ | | | ✓ | ✓ | | | | | ✓ | |
| Dasgupta et al. [34] | ✓ | | ✓ | | ✓ | | | | | | | | | | |
| Priya et al. [129] | | ✓ | ✓ | | ✓ | | | ✓ | | | | | | ✓ | |

3.1 Politeness as Natural Language Understanding Task

Recently, a few research works on computational politeness have been inclined toward identifying various linguistic features of politeness and predicting whether the given text is polite or not. Thus, given a sentence, “*Would you please help me?*” should be predicted as polite, while the sentence “*Would you please stop?*” should be predicted as impolite. It is important to note that though both sentences contain the phrase ‘Would you please’, which is generally considered a polite way to request or ask for something, both are not polite. The perception of politeness or impoliteness in these

sentences depends on the context and the nature of the request. The former sentence is typically perceived as polite because it is a polite request for assistance. In this sentence, ‘Would you please’, signifies respect and deference towards the person from whom the help is needed. In the latter sentence, though the polite marker ‘Would you please’ is used, it is perceived as impolite. It is a request for someone to cease or desist from a particular action. If used with a harsh tone, it exacerbates the perception of impoliteness. Table 2 provides a brief overview of the different datasets, languages, approaches, challenges, etc., on natural language understanding of politeness.

The works by Danescu-Niculescu-Mizil et al. [33] and Aubakirova and Bansal [4] consider two labels, viz. polite and impolite for the classifier. Lately, Mishra et al. [107] and Mishra et al. [108] consider fine-grained labels for the classifier as: polite, somewhat_polite, somewhat_impolite, and impolite to effectively model the politeness variations in goal-oriented interactions. Thus, instead of simply classifying the given text as polite or impolite, they perform the fine-grained classification as “What is your departure city?” - impolite, “Can you provide your departure city?” - somewhat_impolite, “Please, provide your departure city?” - somewhat_polite, “For further processing, I would like to know your departure city, please, name it?” - polite.

Chhaya et al. [24] formulate the task of politeness prediction as a regression problem and a 2-class classification problem in which the model predicts whether the text has politeness or not. Yeomans et al. [174] develop a package in R that provides functions to extract politeness markers in the English language, graphically compare these markers to covariates of interest, develop a supervised model to identify politeness in new documents and inspect high- and low-politeness documents. Aljanaideh et al. [3] model the problem of politeness detection in natural language requests as a clustering task. The authors have created a set of clusters for every word in the request. Each cluster contains contextualized representations of the word (obtained using a pre-trained BERT model) based on the context in which the word appears and the labels of items the word occurs in. The authors state that examining the context of using words helps enrich the linguistic analyses of the politeness task. For example, the hedge “Maybe” can be used as a part of a polite request (“Maybe we can meet”) or an impolite one (“Maybe it would be better if we never talked directly to each other ever again”). Voigt et al. [164] approach the task of politeness detection analogous to detecting respectfulness that police officers display to community members. The authors endeavor to predict various facets of respectful language (apologizing, giving agency, softening of commands, etc.) by analyzing linguistic features derived from theories of politeness, power, and social distance and find that particular forms of politeness are implicated in perceptions of respect.

Lately, a few works delve into the dynamics of online interactions, exploring various factors beyond but associated with the politeness that contributes to positive conversational outcomes, such as constructive comments [120, 121], supportiveness [20], receptiveness [106], or empathy [20, 146, 182]. The authors in [6, 178] define a novel task of forecasting the future behavior of a conversation from early signals. Bao et al. [6] investigate the effect of prosocial behavior, including activities like helping, sharing, comforting, rescuing, and cooperating in online discussions. Inspired by the prosocial behavior theories in social psychology [14], the authors delineate eight broad categories of behavior, viz. *information sharing, gratitude, esteem enhancement, social support, social cohesion, fundraising and donating, mentoring, and absence of antisocial behavior*. They attempt to forecast the prosocial quality of a conversation, demonstrating that these outcomes can be accurately forecasted from cues early in the conversation. Zhang et al. [178] introduce the conversation prompts used to initiate different types of conversations in online collaborative settings. These prompts include *factual check, moderation, coordination, casual remark, action statement, and opinion*. The authors explore the correlation between these conversation prompts and predefined politeness strategies to anticipate potential conversation failures. A handful of studies investigate how politeness influences interactions involving individuals with opposing or

conflicting views. Jeong et al. [66] endeavor to discern the impact of politeness and respect [33] on warm and friendly communication styles, exploring how these factors influence perceptions and interactions in distributive negotiation. Yeomans et al. [175] aim to ascertain whether employing a more receptive communication style facilitates the pursuit of cooperative objectives among individuals with differing opinions, such as a readiness to collaborate in the future, interpersonal trust, and conflict de-escalation. They address conversational receptiveness by identifying the politeness cues that individuals perceive when assessing a counterpart’s openness to dialogue. They utilize these cues to create an intervention to encourage conversation partners to communicate their willingness to engage with opposing views.

Table 3. Summary of computational approaches to politeness generation along different parameters.

| | Datasets | | Languages | | Approach | | | Annotation | | Features | | | | Context | |
|------------------------|--------------------|----------------|-----------|-------|------------|-----------------------------|------------------------|------------|-----------|----------|---------|-----------|-------|--------------|-------|
| | Non-conversational | Conversational | English | Other | Supervised | (Semi/Weakly/Un)-Supervised | Reinforcement Learning | Manual | Automatic | N-grams | Lexical | Syntactic | Other | Conversation | Other |
| Madaan et al. [98] | ✓ | | ✓ | | ✓ | | | | ✓ | ✓ | | | | | |
| Fu et al. [45] | ✓ | | ✓ | | | ✓ | | | | | | | | | |
| Niu and Bansal [122] | ✓ | ✓ | ✓ | | | ✓ | ✓ | | | | | | | | |
| Sennrich et al. [144] | ✓ | | ✓ | ✓ | ✓ | | | | ✓ | | | | | | |
| Feely et al. [38] | ✓ | | ✓ | ✓ | ✓ | | | | ✓ | | | | | | |
| Golchha et al. [49] | | ✓ | ✓ | | | | ✓ | ✓ | | | | | | ✓ | ✓ |
| Firdaus et al. [40] | | ✓ | ✓ | ✓ | | | ✓ | ✓ | | | | | | ✓ | ✓ |
| Wang et al. [167] | ✓ | | ✓ | | ✓ | | | | ✓ | | | | | | |
| Firdaus et al. [43] | | ✓ | ✓ | | ✓ | | ✓ | ✓ | | | | | | ✓ | |
| Mishra et al. [107] | | ✓ | ✓ | | | | ✓ | ✓ | ✓ | | | | | ✓ | |
| Silva et al. [148] | | ✓ | ✓ | | ✓ | | ✓ | | | | | | | | |
| Saha et al. [140] | ✓ | | ✓ | | | ✓ | | | | | | | | | |
| Firdaus et al. [41] | | ✓ | ✓ | | | | ✓ | ✓ | | | | | | ✓ | |
| Mishra et al. [113] | | ✓ | ✓ | | | | ✓ | | ✓ | | | | | ✓ | |
| Mukherjee et al. [117] | | ✓ | ✓ | | ✓ | | | | | | | | | ✓ | |
| Mishra et al. [112] | | ✓ | ✓ | | | | ✓ | | ✓ | | | | | ✓ | |
| Priya et al. [130] | | ✓ | ✓ | | | | ✓ | | ✓ | | | | | ✓ | |
| Mishra et al. [109] | | ✓ | ✓ | | | | ✓ | ✓ | | | | | | ✓ | |
| Mishra et al. [111] | | ✓ | ✓ | | | | ✓ | | ✓ | | | | | ✓ | |
| Firdaus et al. [42] | | ✓ | ✓ | | ✓ | | | ✓ | | | | | | ✓ | |
| Mishra et al. [110] | | ✓ | ✓ | | | | ✓ | | ✓ | | | | | ✓ | |

3.2 Politeness as Natural Language Generation Task

Several computational politeness research works concentrate on the generation of polite sentences. Given the non-polite text, the objective of the politeness generation task is to generate a politeness-oriented text. For instance, “How can we

help?” is a non-polite sentence that can be converted into a polite one: *“Help has arrived! We are sorry to see that you are having trouble. How can we help?”*. Table 3 provides a brief overview of the different datasets, languages, approaches, challenges, etc., on computational approaches to politeness generation.

Madaan et al. [98] focus on politeness as a style transfer task. The authors devise a tag and generate framework for transforming non-polite sentences into polite ones. Fu et al. [45] propose a computational framework for generating a paraphrased version of a given message that meets the required degree of politeness in a particular communication situation. In recent times, there has been a surge in interest in incorporating politeness besides various styles, emotions, and personalities in conversational artificial intelligent agents to make them behave human-like. The work of Walker et al. [165] represents an initial effort to infuse politeness into conversational agents that provide interesting variations of character and personality in an interactive narrative application. Gupta et al. [53] explore building more affective and socially intelligent task-oriented conversational agents by incorporating different politeness strategies in the culinary instructions. Niu and Bansal [122] attempt to induce politeness in chit-chat conversations in the absence of parallel data. Golchha et al. [49] and Firdaus et al. [40] present a method for inducing polite phrases in customer-care responses to enhance user satisfaction. Wang et al. [167] infuse the social language, *viz.* politeness and positivity in customer support agents’ responses. According to Merriam-Webster dictionary⁴, positivity is the quality or state of being positive. The authors argue that the users would prefer a conversational agent exhibiting more polite and positive behavior and would be more willing to engage with, respond to, and persist in the interaction when conversing with such agents. Their work is motivated by previous research works that have shown that positivity can engage employees and improve their performance in the workplace [153] and leads to likeability as it is the language that is upbeat and cheerful [32]. Rana et al. [135] use the state-of-the-art tag and generate approach to introduce politeness in normal chatbot responses and study the effects of polite triggers among different genders, age groups, and personalities using a cross-sectional analysis. Lately, Firdaus et al. [43] also attempt to show the variation in politeness according to the specified personas. Notably, they introduce an approach to generate polite and personalized dialogue responses according to the user’s personal information, primarily based on the age and gender of the user, while being contextually consistent with the conversational history. Saha et al. [140] propose an ensemble framework for generating polite, detoxified, and emotionally charged counterspeech to combat the rising online hate speech. Silva et al. [148] extend the state-of-the-art generative [122] and rewriting [98] approaches to overcome the domain-shift problem and enable the transfer of politeness patterns to a new fashion domain. Sennrich et al. [144] model the politeness interpretation as a machine translation task for English → German. Feely et al. [38] use a similar formulation and controls the honorifics (politeness) in NMT from English → Japanese.

Several recent studies have delved into the task of generating polite responses, taking into account a range of affective states and socio-linguistic cues [41, 42, 110–112, 130]. Firdaus et al. [41] focus on incorporating user sentiment information to produce contextually appropriate polite responses. Likewise, Firdaus et al. [42] leverage the user’s affective states, encompassing emotional and sentiment information, to generate polite responses that exhibit relevant polite behaviors such as greetings, apologies, assurances, appreciations, or expressions of empathy. Mishra et al. [111] introduce a system for generating polite and empathetic responses, while their subsequent study by Mishra et al. [112] utilizes users’ emotional information for the same purpose. Priya et al. [130] introduce a dialogue system capable of generating persuasive responses that adapt politeness and empathy strategies. Mishra et al. [110] incorporate

⁴<https://www.merriam-webster.com/dictionary/positivity>

factors such as gender, age, persona, and sentiment information of users to generate contextually relevant polite and interpersonal responses.

3.3 A note on languages

Politeness is an important facet of communication and is sometimes argued to be cultural-specific, yet existing computational linguistic studies of politeness are primarily dominated by the English language. A few research works in other languages, viz., Hindi [40, 75], German [144], Japanese [38] and Chinese [88] have also been reported. Similar to politeness research in the English language, there exists a dual focus on the research on politeness analysis in other languages, with some studies concentrating on politeness identification, while others center their efforts on politeness generation. The authors in [75, 88, 152] formulate the politeness analysis as an identification task. Kumar [75] and Li et al. [88] attempt to identify politeness in natural language text, focusing on Hindi and Chinese languages, respectively. Srinivasan and Choi [152] evaluate how well multilingual models can identify politeness in text. They conduct this study in nine typologically diverse languages, namely Hindi, Tamil, Korean, Spanish, French, Vietnamese, Russian, Afrikaans, and Hungarian.

Several studies like the works in [38, 40, 144, 162] approach politeness from the perspective of a generation task, aiming to generate or maintain politeness in response to given inputs. The authors in [38, 144, 162] devise a methodology to regulate the degree of formality or politeness in the neural machine translation (NMT) from one language to another. Sennrich et al. [144] and Feely et al. [38] perform a pilot study to control honorifics in NMT focusing on English \rightarrow German and English \rightarrow Japanese, respectively. Recently, Viswanathan et al. [162] develop a domain adaptation technique to bias a model to produce translations according to a desired formality or register. The authors experiment with formality levels for different language pairs: English \rightarrow French, English \rightarrow Spanish, and English \rightarrow Czech. Firdaus et al. [40] propose methods to induce courteous behavior in generic customer-care responses in multilingual scenarios (English and Hindi).

Some of the aforementioned studies have made significant contributions to dataset creation, while others have contributed significantly to the development of novel approaches and methodologies. The specifics of the datasets and methodologies introduced in these works are elaborated upon in the Dataset and Approaches sections of this manuscript.

4 DATASETS

In this section, we outline the datasets utilized for computational politeness research. We categorize these datasets into non-conversational data (such as request posts on online forums) and conversational data (such as user-agent interaction).

4.1 Non-conversational Dataset

Social media, websites, blogs, and emails facilitate access to a vast amount of user-generated content in the form of text. These textual data may be of various types like user queries, complaints, etc. All these texts involve making requests in one or another form; requests for seeking information, requests for help, or requests for specific action. It is generally anticipated that such requests will be couched in remarkably polite language to ingratiate the requestor with the person in a position to honor the request. However, due to text length constraints set by some online platforms, this text tends to be brief and lacks the necessary verbal and behavioral signals that help to grasp and interpret a given piece of communication as polite or impolite. Despite such barriers, datasets of requests have been popular for studying

politeness from a computational point of view. The reason is that the requests involve the use of B&L’s negative politeness strategy: minimizing the imposition on the addressee, for example, by apologizing for the imposition “*I’m really sorry but...*” or by being indirect, “*Would you mind if...*”.

The pioneering research by Danescu-Niculescu-Mizil et al. [33] proposes a benchmark corpus called Stanford Politeness Corpus (SPC) of requests. The SPC consists of requests from two large online communities: Wikipedia, where the requests involve editing and other administrative functions, and Stack Exchange (SE), where the requests center around a diverse range of topics (e.g., programming, gardening, cycling). The intuition leading to domain selection is that both communities exhibit a notable abundance of user-to-user requests characterized by substantive interactions rather than superficial social exchanges. These requests typically involve soliciting specific information or concrete actions, underscoring the expectation of a meaningful response from participants. The corpus consists of more than 10,000 requests (4,353 requests from Wikipedia and 6,604 from Stack Exchange). Each request is annotated with a discrete score ranging from -3 representing “very impolite” through 0 indicating “neutral” to +3 indicating “very polite” by five Amazon’s Mechanical Turk (AMT) workers. The authors eventually assign the politeness score of a request as the average of the five scores assigned by the annotators. These requests are categorized into 5 positive politeness strategies (*Gratitude, Deference, Greeting, Positive Lexicon, Negative Lexicon*) and 15 negative politeness strategies (*Apologizing, Please, Please start, Indirect (btw), Direct question, Direct start, Counterfactual modal, Indicative modal, 1st person pl., 1st person, 2nd person, 2nd person start, Hedges, Factuality*).

Li et al. [88] create a politeness corpus containing 5,300 posts each posted on Twitter and Weibo in 2014. Two native language annotators independently label each post with a politeness score using a similar annotation scheme as in Danescu-Niculescu-Mizil et al. [33]. Madaan et al. [98] introduce a large dataset comprising 1.39 million sentences annotated with politeness scores. These sentences are curated from email exchanges of requests in the Enron corpus [72]. The sentences are labeled with a politeness score using a politeness classifier [122], and sentences with a score of over 90% are considered as polite. Chhaya et al. [24] create a dataset consisting of 1,050 emails extracted from the Enron email dataset and annotate each email for frustration, formality, and politeness labels following the annotation protocol of the Likert Scale. Frustration is labeled on a 3-point scale, with neutral equating to not frustrated; and frustrated and very frustrated are marked with -1 and -2, respectively. Formality and politeness follow a 5-point scale from -2 to +2, where both extremes mark the higher degree of presence and absence of the respective dimension. Each email is assigned a final score as an average of 10 scores annotated by the AMT workers.

Sennrich et al. [144] utilize OpenSubtitles [159] and present a politeness annotated corpus for machine translation task from English → German. Each sentence in the corpus is automatically annotated with the rules based on a morphosyntactic annotation by [145]. The sentences with imperative verbs are labeled as informal, while sentences containing informal or polite German address pronouns are labeled with the corresponding class. Lately, Bharti et al. [9] introduce *PolitePEER*, a large-scale corpus of peer reviews annotated with politeness labels. It comprises review sentences from various sources such as ICLR, NeurIPS, Publons, and ShitMyReviewersSay, covering diverse topics including machine learning, natural language processing, artificial intelligence, environment and ecology, clinical medicine, psychiatry and psychology, and engineering. A total of 1348 review sentences from ICLR, 58 from NeurIPS, 888 from ShitMyReviewersSay, and 206 from Publons, are annotated with five politeness labels, namely *highly_impolite*, *impolite*, *neutral*, *polite*, and *highly_polite*. The authors argue that the *PolitePEER* dataset is one of its kind, which may enable researchers to gain deeper insights into the language used in peer reviews and lead to the development of more accurate models for predicting politeness in this domain.

Politeness is a phenomenon that is highly culture and language-specific, and what is polite in one language may be impolite in another, resulting in numerous cross-cultural misunderstandings and conflicts. Thus, it is essential to have a sound knowledge of politeness across languages and ensure that the machines have the same understanding. Politeness in low-resource Indian languages is a little-explored aspect of the language. Recently, with the rise in online content in Hindi, a few researchers focus on studying politeness in Hindi texts. Kumar [75] introduce a politeness annotated corpus consisting of over 479,000 blog posts and blog comments in Hindi. The blogs are collected using a popular aggregator of blogs in Hindi named *Chitthajagat*. The blogs in the dataset are annotated with one of the four levels of politeness, viz. neutral, appropriate, polite, and impolite in a semi-supervised manner. A total of 30,000 texts from the corpus are manually annotated by four annotators. The manually annotated text is utilized for training classifiers and annotating the remaining blogs in the corpus. Kumar [74] present a Corpus of Computer-Mediated Communication in Hindi (CO3H) for studying politeness in Hindi. The data for the corpus is collected from blogs, web portals, e-forums, emails, and public and private chats over the web. The data consists of comments, including requests, offers, compliments, criticism, etc. After a fundamental analysis of how politeness or impoliteness is expressed in the language, an annotation scheme has been developed for annotating the corpus. Approximately 26,000 comments are manually annotated with one of the four politeness categories, viz. neutral, appropriate, polite, and impolite [76]. Table 4 presents a summary of these non-conversational datasets.

Table 4. Summary of politeness-labeled non-conversational datasets. Details of the evaluation metrics can be found in Section 6.

| Reference | Dataset | Domain | Language | Task | Evaluation Metrics |
|-------------------------------------|--|--|---------------------|---|-----------------------------|
| Danescu-Niculescu-Mizil et al. [33] | Stanford Politeness Corpus of requests | Wikipedia and Stack Exchange Request Posts | English | Politeness classification | Accuracy |
| Li et al. [88] | Politeness Corpus | Twitter and Weibo Posts | English and Chinese | Politeness classification | Accuracy, ROC-AUC |
| Madaan et al. [96] | Enron Corpus | Email Exchanges | English | Style transfer (Non-polite to polite) | BLEU-s |
| Chhaya et al. [24] | Enron Dataset | Email | English | Frustration, formality and politeness classification | Accuracy, Mean Square Error |
| Sennrich et al. [144] | OpenSubtitles | Movies Subtitles | English, German | Politeness preservation in neural machine translation | BLEU |
| Bharti et al. [9] | PolitePEER | Peer Reviews | English | Politeness classification | F1-score |
| Kumar [74] | CO3H | Online Forums Requests | Hindi | Politeness classification | Accuracy |
| Kumar [75] | Politeness Corpus of Hindi Blogs | Blogs Posts and Comments | Hindi | Politeness classification | Accuracy |

4.2 Conversational Dataset

Politeness has also been explored in the conversational settings. Wang et al. [167] propose a dataset consisting of 233,571 message pairs exchanged between the ride-sharing company's on-boarding drivers and customer support representative (CSR). Each message pair is annotated automatically using the state-of-the-art pre-trained classifiers with politeness and sentiment scores. Golchha et al. [49] introduce a conversational dataset, CYCCD (Courteously Yours Customer Care Dataset), which consists of the interactions between customers and professional customer care agents of companies on their Twitter handles. The dialogues in the dataset consist of generic utterances in English and their corresponding courteous versions.

Inspired by their work, Firdaus et al. [40] follow their guidelines and build a multi-lingual conversational dataset for English and Hindi to generate polite customer care responses in both languages, thereby enhancing the performance and usability of conversational agents. Firdaus et al. [41] annotate the CYCCD [49] with sentiment (positive, neutral, negative) to generate sentiment-guided polite responses. Firdaus et al. [42] annotate the CYCCD [49] with polite behavior (greeting, apology, assurance, appreciation, and empathy) to generate polite behavior-aware responses in

customer support dialogues. Bothe and Wermter [12] explore the DailyDialog [90] dataset and annotate each utterance of this dataset with a degree of politeness on the scale of [1,5] using a politeness analyzer from the work of [6].

Mishra et al. [107] utilize the recently released multi-domain task-oriented conversational dataset, MultiDoGo [127], and annotate each utterance of the dataset with one of the four politeness labels, *viz.* polite, somewhat_polite, somewhat_impolite, and impolite. The authors perform the annotation in two stages. In the first stage, the politeness score of each utterance of the MultiDoGo dataset is obtained via the Stanford Politeness Classifier (trained on Wikipedia requests data [33]). The utterance is considered polite if the score ≥ 0.50 , else impolite. In the second phase, the authors randomly select approximately 10K utterances from each domain and manually annotate each utterance with one of the four fine-grained politeness classes to obtain the ground-truth data. Mishra et al. [108] follow similar guidelines as in [107] and create two politeness-annotated conversational datasets utilizing the DSTC1 [173] and Microsoft Dialogue Challenge (MDC) [89] datasets. Mishra et al. [109] enrich the TaskMaster [21] dataset by manually annotating it with fine-grained politeness labels with the help of crowd-workers from Amazon Mechanical Turk (AMT).

Mishra et al. [113] annotate the PersuasionForGood dataset [166] with the politeness strategy labels: bald on-record, positive politeness, negative politeness, and off-record. Priya et al. [129] curate a large-scale, novel dialogue dataset, POEM, consisting of 5,000 dialogues in English for mental health counseling and legal aid of women and children crime victims. The POEM dataset is manually annotated with politeness labels and multiple emotion labels at the utterance level. For politeness annotation, the authors consider ternary labels, *viz.* polite, neutral, and impolite. For emotion annotation, they consider a rich set of 17 emotion categories, namely anticipation, confident, hopeful, anger, sad, joy, compassion, fear, disgust, annoyed, grateful, impressed, apprehensive, surprised, guilty, trust, and neutral.

Mishra et al. [111] also curate a novel mental health and legal counseling dialogue (MHLCD) dataset in English comprising 1006 dialogues between the victim and the counselor. Further, they annotate the counselor's utterances in this dataset with politeness and empathy labels. The authors use ternary politeness labels - polite, neutral, and impolite, and binary empathy labels - empathetic and non-empathetic for politeness and empathy annotations. Mishra et al. [112] introduce two large-scale counseling conversational datasets, EPE-enEIH and EPE-HLCC, that focus on supporting individuals dealing with crime victimization and substance addiction. The dataset focused on delivering counseling support to crime victims is prepared by translating the Hindi conversational dataset, EmoInHindi [149] to English. The other dataset used in the study is prepared using High-quality and Low-quality Counseling Conversations data [126], consisting of conversations between counselors and substance addicts. They annotate the agent's utterances in these datasets with one of the 11 emotion categories (caring, hopeful, content, surprised, afraid, confident, proud, trusting, joy, sad and angry), one of three politeness labels (impolite, low_polite and highly_polite) and ternary empathy labels (non-empathetic, low_empathetic and highly_empathetic).

Priya et al. [130] curate another counseling conversational dataset, HEAL, consisting of 216 conversations between the counseling bot and the crime victim. The counseling bot's responses are annotated with novel empathy strategies (reflective listening, confidential comforting, evoke motivation, express emotional support, offer counseling, escalate assurance and no strategy) and politeness strategies (positive politeness, negative politeness, and bald-on-record) [16]. Mishra et al. [110] introduce a novel psychotherapeutic conversations dataset, PsyCON, comprising of conversations between the therapist and the user suffering from one of the seven most common psychological issues, *viz.* depression, anxiety, stress, bipolar disorder, disruptive behavior and dissocial disorders, post-traumatic stress disorder (PTSD), and schizophrenia [172]. The user's utterances are annotated with sentiment labels (positive, negative, and neutral). The therapist's utterances are annotated with politeness labels (polite, moderately_polite, and impolite) and interpersonal

behavior labels (directing, helpful, understanding, complaint, imposing, confrontational, dissatisfied, and uncertain). Table 5 provides a comprehensive summary of these conversational datasets across various dimensions.

Table 5. Summary of politeness-labeled conversational datasets. Details of the evaluation metrics can be found in Section 6.

| Reference | Dataset | Domain | Language | Task | Evaluation Metrics |
|--|--|--|----------------|--|---|
| Golchha et al. [49] | CYCCD | Customer Care | English | Polite response generation | BLEU, ROUGE-1, ROUGE-2, ROUGE-L, Perplexity, Content Preservation, Emotion Accuracy |
| Firdaus et al. [40] | Hindi Customer Care Dialogue | Customer Care | English, Hindi | Polite response generation | BLEU, ROUGE-1, ROUGE-2, ROUGE-L, Perplexity, Content Preservation, Emotion Accuracy |
| Wang et al. [167] | CSR-Driver Exchanges | Customer Care | English | Polite response generation | BLEU, word2vec-based similarity |
| Bothe and Wermter [12], Priya et al. [129] | Politeness-annotated DailyDialog | Daily-life Conversations | English | Politeness classification | Accuracy, F1-score |
| Mishra et al. [107] | Politeness-annotated MultiDoGO | Airline, Fastfood, Finance, Insurance, Media, Software | English | Polite dialogue generation | Success Rate, Politeness Adaptability, Turn Completion Rate |
| Mishra et al. [108] | Politeness-annotated DSTC-1, MDC | DSTC-1: Bus Schedule; MDC: Movie-Ticket Booking, Restaurant reservation, Taxi Ordering | English | Politeness classification | Accuracy, Geometric Mean, Index Balance Accuracy |
| Mishra et al. [109] | Politeness-annotated TaskMaster | Flight, Food Ordering, Hotel Movie, Music, Restaurant, Sports | English | Politeness-adaptive dialogue generation | Dialogue Length, METEOR, ROUGE-2-L, Success Rate, Average Politeness Score |
| Firdaus et al. [41] | Sentiment-annotated CYCCD | Customer Care | English | Sentiment-guided polite response generation | Perplexity, BLEU, Rouge-L, Politeness Adaptability |
| Mishra et al. [113] | Politeness-annotated PersuasionForGood | Charity Donation | English | Polite and empathetic persuasive dialogue generation | Perplexity, Response Length, PerStr, PolStr, Emp |
| Priya et al. [129] | POEM | Mental Health and Legal Counseling | English | Politeness and Emotion Classification | Accuracy and F1 |
| Mishra et al. [111] | MHLCD | Mental Health and Legal Counseling | English | Polite and empathetic dialogue generation | Perplexity, R-LEN, CoStr, Pol, Emp |
| Mishra et al. [112] | EPE-enEIH and EPE-HLCC | Mental Health and Legal Counseling | English | Emotion-adaptive polite and empathetic dialogue generation | Perplexity, Response Length, EPC, EEC, PC, EC |
| Priya et al. [130] | HEAL | Mental Health and Legal Counseling | English | Politeness and empathy strategies- adaptive persuasive dialogue generation | Perplexity, Response Length, CoAct, EmpStr, PolStr |
| Firdaus et al. [42] | Polite Behavior-annotated CYCCD | Customer Care | English | Polite behavior-aware response generation | Perplexity, BLEU, ROUGE-L, Politeness Adaptability |
| Mishra et al. [110] | PSYCON | Mental Health Counseling | English | Polite and interpersonal behavior-aware response generation | Perplexity, BERTScore-F1, Response Length, GA _c , P _c , CT _c , Po _c , IB _c |

5 APPROACHES

In this section, we present the approaches used for computational politeness. In general, approaches to computational politeness can be categorized into statistical and deep learning-based approaches. We will examine these methods in the following sections.

5.1 Statistical Approaches

This section describes the feature sets and the learning algorithms that have been reported for statistical computational politeness. The majority of the approaches use bag-of-words as features. Nevertheless, several other linguistic features have been reported in addition to these features. Table 6 summarizes the features used in past work.

Table 6. Features used for statistical classifiers. (# N) represents the number of respective features used in the experiment.

| Reference | Features |
|-------------------------------------|---|
| Danescu-Niculescu-Mizil et al. [33] | Lexical and syntactic features (# 20) |
| Li et al. [88] | PoliteLex (# 26), LIWC, EmoLex, Stanford Politeness API |
| Kumar [76] | Unigrams, Bigrams and Linguistic features (# 8) |
| Chhaya et al. [24] | Lexical (# 12), Syntactic (# 6), Derived (# 8), Affect-based (# 29), Word embeddings (Word2Vec) |
| Madaan et al. [98] | N-grams tf-idf |
| Fu et al. [45] | Lexical and syntactic features (# 18) |

Approaches to Politeness Identification. Alexandrov et al. [1] present numerical estimations for politeness based on the lexical-grammatical properties of the text and subjective expert opinions. They have considered three factors of politeness: the first greeting, polite words, and polite grammar forms, and used these factors to construct a series of polynomial models for automatic evaluation of the level of politeness. Alexandrov et al. [2] propose a linear regression model for politeness estimation using the same approach as in [1].

In the past, several works in statistical computational politeness have been reported that rely on Support Vector Machines (SVM). Danescu-Niculescu-Mizil et al. [33] construct a classifier for predicting politeness in natural language requests in English with various domain-independent lexical and syntactic features about key politeness theories, for instance, a first or second person start vs. plural, please start vs. please, etc. Specifically, they propose two classifiers - a bag of words classifier (BOW) and a linguistically informed classifier (Ling.). BOW classifier is an SVM with unigram feature representation. The Ling. classifier is an SVM using the linguistic features⁵ besides the unigram features. They then apply the logistic regression model on top of the SVM output to predict the politeness scores for requests between 0 and 1. However, Hoffman et al. [56] challenge the work of Danescu-Niculescu-Mizil et al. [33] and revalidates the politeness labeling tool for broadly addressing politeness in social computing data.

Li et al. [88] develop a politeness feature set, *PoliteLex*, that is compatible with both English and Mandarin to study the resemblances and distinctions between the US and China's politeness. They have trained several SVM classifiers using features from one or more feature sets, viz. *PoliteLex*, *Linguistic Inquiry and Word Count (LIWC)*⁶ [125], *EmoLex*⁷ [115] and *Stanford Politeness API* [33]. The authors have shown that *PoliteLex*, combined with *LIWC* and *EmoLex* outperform the *Stanford Politeness API* in predicting politeness on Twitter, Weibo, and Stanford Politeness Corpus. Further, the findings of this research reveal that within the Mandarin Weibo platform, discussions centered around future-oriented themes, expressions of group affiliation, and expressions of gratitude are generally regarded as more polite in comparison to English Twitter. Conversely, discussions involving topics related to death, the use of pronouns (excluding honorifics), and informal language tend to exhibit higher degrees of impoliteness on Mandarin Weibo compared to English Twitter.

Kumar [76] propose an SVM-based classifier for identifying linguistic politeness in Hindi. The authors have proposed two classifiers based on the bag-of-words (BoW) model. Of these two classifiers, one employs unigram feature representation, and the other employs unigram and bigram feature representations. They have also developed one more classifier which utilizes unigrams, bigrams, and manually identified linguistic structures as features such as formulaic expressions (*dhanyawad*, *kripya*, etc.), use of particle 'ji', 'zara', subjunctive verb form (*mujhe*, *samjhe*, etc.), suggestion markers/deontics (*cahiye*, etc.), honorific pronominals and verb form (*bataiega*, etc.).

⁵A total of 20 politeness features have been identified by the authors, namely *Gratitude*, *Deference*, *Counterfactual modal*, to name a few.

⁶*LIWC* maps tokens into psychological categories.

⁷*EmoLex* consists of 14,182 English unigrams mapped to 7 emotion categories – joy, surprise, anticipation, sadness, fear, anger, and disgust.

Wang et al. [167] utilize statistical models to estimate the role of politeness and positivity in predicting drivers engagement. They have analyzed that use of such social language leads to greater users' responsiveness and task completion. In particular, they have employed pre-trained SVM-based classifier [33] and a rule-based sentiment analyzer, VADER [62] to predict the politeness level and positivity of messages, respectively.

Miyamoto et al. [114] propose an autonomous politeness control method for conversational agents. The authors have constructed a regression model to predict politeness and comfortableness in terms of three social factors of the politeness theory: social distance, social status difference, and conversation scene. They have conducted multiple regression analyses to obtain a model that facilitates the selection of appropriate politeness of utterance according to three social factors. Chhaya et al. [24] model frustration along with its two complementary affects, formality and politeness. The authors have proposed state-of-the-art statistical models for modeling frustration, formality, and politeness in their work. Specifically, they have developed Linear, Lasso, Ridge, and SVR regression models and different classification models, viz. Logistic Regression, SVC, Linear SVC, Random Forest, and Nearest Neighbors, with different features for determining the existence or non-existence of frustration, formality, and politeness in emails.

Approaches to Politeness Generation. Madaan et al. [98] introduce a pipeline approach, *tag and generate*, which is a two-staged approach for converting non-polite sentences into polite ones while preserving the content of the source text. The tagger model first identifies the style attribute markers from the source style and then either substitutes them with a positional token named [TAG] or adds [TAG] tokens to the input without deleting any phrase from the input. This feature of the model allows generation of these tags in an input that lacks an attribute marker. The generator model generates sentences in the target style by replacing these [TAG] tokens with stylistically relevant words inferred from the target style. In order to estimate the style markers for a specific style, a simple approach based on n-gram tf-idfs has been used. Fu et al. [45] propose a pipeline approach for generating a paraphrase of a given message that can convey the desired degree of politeness in a particular communication setting. This pipeline approach consists of three steps: *Plan*, *Delete*, and *Generate*. For a given input message, this approach initially determines the politeness strategies and their respective markers. The proposed model employs Integer Linear Programming (ILP) in the plan stage to estimate an appropriate target strategy combination, followed by the deletion of markers associated with the strategies that need to be omitted to obtain the post-deletion context. The proposed model then sequentially integrates the new strategies from the ILP solution into this context to construct the final paraphrased message.

5.2 Deep Learning-based Approaches

With the gain in popularity of deep learning-based architectures in NLP applications, many such approaches have been reported for computational politeness.

Approaches to Politeness Identification. Aubakirova and Bansal [4] propose a Convolutional Neural Network (CNN)-based architecture using the Stanford Politeness Corpus for predicting politeness in natural language requests. Furthermore, the authors have presented various visualization strategies: activation clustering, embedding space transformations, and first derivative saliency to understand what and how machines learn about politeness. The authors state that this approach not only outperforms the prior work of Danescu-Niculescu-Mizil et al. [33] using manually-defined features but also extends the existing features and identifies novel politeness strategies such as indefinite pronouns (*something/somebody*) and punctuation (question marks "???", an ellipsis "..."). Mishra et al. [108] employ a hierarchical transformer network for predicting the politeness label of a given utterance in a goal-oriented dialogue. The authors state that this approach allows them to learn contextual information from previous utterances, which aids in predicting the politeness labels accurately. Dasgupta et al. [34] propose a graph-induced transformer

network (GiTN) to automatically detect formality and politeness in text. GiTN utilizes Graph Convolution Network (GCN) [71] to encode the syntactic information present in the input text and the BERT [35] model to generate the contextual embeddings of the input text. The output of GCN and BERT are then combined and passed through an attention-based BiLSTM [176] layer followed by the fully connected layer to automatically detect stylistic characteristics like formality and politeness in the input text. The authors independently fine-tune GiTN on formality-annotated and politeness-annotated corpora for formality and politeness detection tasks, respectively.

Approaches to Politeness Generation. Sennrich et al. [144] introduce an attention-based encoder-decoder NMT system [5] for English → German translation task. The authors demonstrate that using the target side T-V annotation as an extra input during training of an NMT model can control the generation of politeness (honorifics) at test time via side constraints. Feely et al. [38] propose a formality-aware NMT system for controlling honorifics in English → Japanese translation and demonstrate that this approach improves the translation quality, especially for formal and polite sentences in each test set.

Niu and Bansal [122] propose three weakly supervised models, namely the Fusion model, Label-Fine-Tuning (LFT) model, and Polite Reinforcement Learning (Polite-RL) model. These models aim to generate natural, varied, and contextually consistent polite or impolite responses for open-domain dialogues without parallel data. Each of the three models is based on a state-of-the-art politeness classifier [33], and a sequence-to-sequence dialogue model. The Fusion model combines the decoder of an encoder-attention-decoder conversation model with a language model trained on stand-alone polite utterances. The LFT model prepends a politeness-score scaled label to each source sequence during training. The model can generate polite, neutral, and rude responses at test time by simply scaling the label embedding by the corresponding score. The Polite-RL facilitates the generation of polite responses by assigning rewards proportional to the politeness classifier score of the sampled response. The authors have illustrated that the LFT and Polite-RL models can generate better polite responses without compromising the dialogue quality.

Firdaus et al. [40], Golchha et al. [49], Wang et al. [167] provide evidence that incorporation of courteousness is essential for achieving user satisfaction and promoting user retention. Thus, Golchha et al. [49] propose a reinforced pointer-generator model for generating emotionally and contextually consistent courteous responses in task-oriented customer-care dialogues in English. Motivated by their work, Firdaus et al. [40] utilizes a pointer-generator network for generating courteous responses in both Hindi and English by modeling the user’s emotional state by learning language invariant representation using adversarial training. Wang et al. [167] propose a sequence-to-sequence learning framework to infuse politeness and positivity in goal-oriented customer support agents’ responses while preserving content. Mishra et al. [107] introduce a novel end-to-end reinforcement learning (RL) based Politeness Adaptive Dialogue System (PADS). This approach first extracts the politeness semantics of an utterance using a politeness classifier based on DistilBERT [141]. Then, the RL-based dialogue agent uses this politeness information to design its reward feedback. The authors state that polite rewards in PADS consider the user’s satisfaction information, enabling the agent to accomplish the task quickly.

Lately, Silva et al. [148] examine two different strategies for generating politeness, viz. politeness response generation [122] and rewriting [98]. They further analyze how each method handles domain changes and adapts each of them to transfer politeness patterns to the novel fashion domain. Their work is one of the earliest studies on politeness for task-oriented dialog agents that facilitated transfer learning for particular domains. Saha et al. [140] propose *CounterGeDi*, an ensemble of generative discriminators (*GeDi*) to guide the generation of a DialoGPT model toward more polite, detoxified, and emotionally charged counterspeech in order to combat the increasing online hate speech. Firdaus et al. [43] propose a Polite Personalized Dialog Generation (PoPe-DG) framework for inducing politeness in the

responses of dialogue systems in accordance with the user’s persona information such as age and gender. The authors have manually created polite templates according to the user profiles utilizing the bAbI dataset [69]. Their proposed PoPe-DG framework employed a reinforced deliberation network for encoding the age and gender information along with polite templates to generate polite personalized responses.

Mukherjee et al. [117] introduce a polite chatbot that can generate polite and contextually coherent responses. The authors formulate a three-step method for building a polite chatbot. In the first step, the authors train a BART [86] based politeness transfer model (PTM) in a supervised fashion that generates polite synthetic dialogue pairs of contexts and polite utterances. PTM takes neutral sentences as input and outputs a sentence that preserves the content while increasing politeness. The synthetic input-output pairs are obtained following [98]. Then, they apply this politeness transfer model to generate synthetic polite chat data, which they finally utilize to train a dialogue model capable of generating polite responses. For the training of the dialogue model, multiple pre-trained language models are employed, namely GPT-2 [133], DialoGPT [179], and BlenderBot [137].

Firdaus et al. [41] introduce an end-to-end Transformer Encoder-Decoder [161] based architecture to predict the sentiment of the utterances and then utilize the predicted sentiment information to generate contextually correct polite responses in customer support dialogues. In particular, they utilize a hierarchical network consisting of two transformer-based encoders, namely *sentence encoder* (SE) to encode the utterances and *context encoder* (CE) to encode the output of the sentence encoder to capture the dialog context. The authors apply softmax activation on the output of the SE to predict the sentiment information, which is then combined with the contextual information obtained from CE to generate polite responses aligned with the user’s sentiment and dialogue history. To ensure that the users’ sentiment and politeness are induced appropriately in the generated responses, the authors design task-specific rewards and jointly employ reinforcement learning (RL) and machine learning to train the proposed model. Firdaus et al. [42] present a transformer-based encoder-decoder network with an affective tracker to capture the contextual information along with the affective information, behavior-aware generators to attend the context information to compute behavior-aware polite representations, and a polite generator to generate the final polite response considering the representations from different generators. This framework aims to capture the emotional aspect from the context, generate response representations based on different polite behaviors (greeting, apology, assurance, empathy, and appreciation), and ultimately produce appropriate polite responses utilizing a weighted-sum approach.

Mishra et al. [109] propose GenPADS, a novel end-to-end dialogue system that reinforces politeness in task-oriented conversations. GenPADS incorporates a politeness classifier and a generation model in a reinforcement learning setting to adapt and generate polite responses that are informative, empathetic, and interactive. To build the politeness classifier and generation model, they employ DistilBERT [141] and BART [86], respectively, as the backbone architectures. The development of GenPADS consists of three steps: GenPADS first utilizes a transformer-based politeness classifier to extract the politeness semantics from the utterances. Subsequently, a dialogue agent, operating on a reinforcement learning framework, utilizes this extracted politeness information to formulate its reward feedback mechanism, thereby adapting its behavior to prioritize polite actions. Finally, a transformer-based generator model is employed to generate interactive and diverse responses.

5.3 Miscellaneous

In addition to the two approaches mentioned above, a few other approaches for politeness study in NLP have been reported. In this section, we will describe a few such approaches.

Johnson and Rizzo [67] attempt to teach polite behaviors to an animated pedagogical agent to build a socially intelligent system that can monitor learner performance and provide socially sensitive feedback at appropriate times. They have utilized the natural language generator [68] to build a computational framework for producing appropriate interaction tactics. The authors claim that their generation framework could present the same tutorial comment with different degrees of politeness. For example, a suggestion to save the current factory description can be stated either as bald on record (e.g., “*Save the factory now*”), as a hint (“*Do you want to save the factory now?*”), or as a suggestion of what the tutor would do (“*I would save the factory now*”) or as a suggestion of joint action (“*Why don’t we save our factory now?*”). Gupta et al. [53] present POLLy (Politeness for Language Learning), which consists of two components: an AI Planner based on GraphPlan [11] and a Spoken Language Generator (SLG). AI Planner generates a sequence of actions (plan) for completing the cooking task, and the SLG communicates the plan needed to accomplish that task collaboratively, with the ultimate objective of learning English as a second language.

Miller et al. [102, 102, 104] describe a computational model of etiquette and politeness perception across different cultures. They have designed and implemented a computational algorithm called “Etiquette Engine (EE)” based on B&L’s politeness theory to serve as a predictive model of an observer’s perceived “degree of imbalance” of an interaction, whether among humans or humans and non-player characters (NPCs). They have demonstrated EE’s ability to produce culture-specific, politeness-appropriate utterances and perceptions of utterances in controlled tests involving project team members, open surveys, and game settings.

Roman et al. [138, 139] present empirical pilot studies on the role of politeness in dialogue summarization. The authors have created four dyadic dialogues between a customer and vendor using the NECA⁸ (*Net Environment for Embodied Emotional Conversational Agents*) system. They further asked thirty participants of different qualifications, genders, and courses involved in the study to summarize these dialogues. These studies demonstrate how people deal with politeness/behavior in dialogues and how their point of view influences their summaries to facilitate the construction of an automatic dialogue summarization system. In Kumar and Jha [77], the authors present a corpus-based study of politeness across two languages. The authors have studied politeness in a translated parallel corpus of Hindi and English and showed how politeness in a Hindi text is translated into English. The authors assert that their study can be helpful for machine translation tasks from Hindi → English while maintaining the degree of politeness across languages.

In Song et al. [150], based on the B&L politeness theory [16], the authors examine the effect of e-commerce chatbot politeness strategies (apology versus appreciation) on customers’ post-recovery satisfaction, taking into consideration the variable of face concern and further investigated the moderating influence of time pressure on it. Prior investigations within the domain of politeness theory in AI have predominantly centered on customer preferences for polite chatbots and how impolite chatbots can trigger face-threatening acts among customers [36]. Their study extends this line of research by demonstrating that chatbots can employ politeness strategies in response to customer’s face-threatening acts resulting from service failures. Their findings suggest that, in post-recovery consumer satisfaction, fostering a positive human-chatbot relationship (through appreciation) proves more effective than acknowledging the chatbot’s limited competence (via apology). This insight serves as a valuable reference point for e-commerce platforms seeking to devise suitable service recovery strategies for their chatbots.

⁸NECA is a conversational agent platform that allows to control several dialogue parameters such as interlocutors’ role, interests, and politeness.

5.4 Politeness and Large Language Models

Large Language Models (LLMs), (a.k.a generative AI), are language models with tens or hundreds of billions (or more) of parameters trained on massive textual data [181]⁹. Examples of LLMs include GPT-3 (175B parameters) [19], LLaMA2 (13B, 70B parameters) [160], ChatGPT¹⁰, etc. The rapid advancements and impressive achievements of LLMs have triggered a revolutionary shift in the NLP tasks like emotion recognition [180], summarization [131], dialogue [93], etc. With the ability to generate creative and human-like text that involves knowledge utilization and complex reasoning, whether the LLMs' responses align with politeness norms or values remains a critical question. Since politeness norms substantially affect human-human interactions, several researchers are actively investigating whether the LLMs possess social competence encompassing politeness recognition, interpretation, and understanding, which is necessary for effective communication and social interactions.

Recently, Li et al. [87] evaluate the performance of state-of-the-art language models on politeness prediction using the *Stanford Politeness Corpus* (SPC) [33]. They have evaluated the ChatGPT competency in predicting politeness. This study has shown that zero-shot prediction with ChatGPT can provide reasonable results and explanations for its prediction. Ziems et al. [183] prompt the LLMs, namely Flan-T5 [26], OpenAI's GPT-3 (text-001, text-002, and text-003) [19] and ChatGPT in a zero-shot fashion for classifying the utterances in SPC into one of three politeness labels, viz. polite, neutral, and impolite. They have mentioned that these LLMs performed fairly well on the politeness classification task, which indicates the LLMs' broader ability to recognize conversational social norms.

The widespread application of LLMs in various fields, such as healthcare, education, etc. underscores the importance of assessing their adherence to social norms, particularly politeness. This raises the question of whether LLMs should exhibit politeness towards humans and vice versa. Liévin et al. [91] highlight the significant impact of politeness on human-machine interactions, indicating that adopting polite strategies by LLMs fosters trust, especially in contexts like healthcare and education. Since LLMs like ChatGPT learn from human interactions, behaving politely with them facilitates the generation of respectful and collaborative outputs. While human politeness may not affect LLM behavior, these systems typically refrain from responding to derogatory or offensive language. For example, Microsoft's GPT-powered Bing AI responds to such language by expressing disapproval, while ChatGPT advises against using offensive language or clarifies its lack of emotions.

6 REPORTED PERFORMANCE

Table 7 and Table 8 illustrate the reported values from previous research works for politeness identification and politeness generation tasks, respectively. It is important to note that the values cannot be directly compared since they are deduced from distinct datasets, experiment settings, methodologies, or metrics. Besides, in a few earlier studies, simply an analysis of datasets is conducted. These works do not report any evaluation metric such as Accuracy or F1-score, etc. Consequently, we have not included such works in the tables. Nonetheless, the table provides an approximation of the present performance of computational approaches to politeness identification and generation.

Reported Performance on Politeness Identification. In Table 7, we have highlighted the performance evaluation for the politeness identification task reported in the literature on the different datasets in varied domains. Danescu-Niculescu-Mizil et al. [33] show that linguistically-informed features outperform the bag-of-words model in both

⁹In the existing literature, there exists no established agreement regarding the minimal parameter scale deemed necessary for LLMs, primarily due to the interplay between model capacity, data magnitude, and computational resources. In this survey, we adopt a more flexible interpretation of LLMs and prioritize discussions revolving around politeness in LLMs.

¹⁰<https://chat.openai.com/>

in-domain and cross-domain settings. The authors compare the accuracy of the politeness classifier for both in-domain and cross-domain settings with the human ability to detect politeness and find that their classifier achieves near-human performance in detecting politeness. Li et al. [88] report lexical features as their top discriminating features. The authors show that the lexical approach to predicting politeness is generalizable across domains. The CNN-based classifier in Aubakirova and Bansal [4] results in the accuracy of 85.8% and 66.4% on Wikipedia and Stack Exchange (SE) corpora, respectively.

Mishra et al. [108] report metrics other than accuracy for evaluating their proposed hierarchical transformer network for politeness prediction. Specifically, they have used F1-score, geometric mean(GM) [7], and index-balance accuracy (IBA) [46] as the automatic evaluation metrics. The graph-induced transformer-based architecture in Dasgupta et al. [34] reports the F1-scores of 75%, 75%, and 81% for politeness identification on Wikipedia, Stack Exchange, and Enron Corpus. Likewise, Priya et al. [129] evaluate the performance of the politeness detection task in dialogues using accuracy and F1-scores. The authors have reported the accuracy and F1 scores of 90.30% and 87.19%, and 86.78% and 75.27% for politeness identification on the POEM and DailyDialog datasets, respectively. For the politeness identification task, it has been observed that accuracy and F1 scores stand out as the predominant evaluation metrics.

Table 7. Reported performance values (in %) for computational approaches to politeness identification.

| Reference | Details | | | Reported Performance |
|-------------------------------------|---|---------------------|---|--|
| | Dataset | Language | Domain | |
| Danescu-Niculescu-Mizil et al. [33] | Stanford Politeness Corpus of requests | English | Online Communities - Wikipedia and Stack Exchange (SE) | In Domain - Acc: 83.79 (Wiki) and 78.19 (SE) Cross Domain - Acc: 67.53 (Wiki) and 75.43 (SE) |
| Li et al. [88] | Twitter and Weibo posts, Stanford Politeness Corpus of requests | Chinese and English | Online platforms (Twitter, Weibo, Wikipedia and Stack Exchange) | Acc: 77.1, F1: 77.2, ROC-AUC: 77.5 (Twitter) Acc: 80.2, F1: 80.4, ROC-AUC: 81.5 (Weibo) Acc: 68.6, F1: 68.5, ROC-AUC: 68.4 (SPC) |
| Kumar [76] | CO3H | Hindi | Blog comments | Acc: 77.55 |
| Chhaya et al. [24] | Enron Dataset | English | Emails | MSE: 1.556, Acc: 72 |
| Aubakirova and Bansal [4] | Stanford Politeness Corpus of requests | English | Online Communities - Wikipedia and Stack Exchange | Acc: 85.8 (Wikipedia) and 66.4 (SE) |
| Mishra et al. [108] | Politeness-annotated DSTC1 Politeness-annotated Microsoft Dialog Challenge - A Dialogue State Tracking (MDC) | English | MDC - Movie, Restaurant, Taxi DSTC1-Bus schedule | In Domain - Acc: 98.1, F1:98.5, GM: 98.9, IBA:95.0 Cross Domain - Acc: 65.8, F1: 64.0, GM: 72.5, IBA: 53.1 |
| Priya et al. [129] | POEM Politeness-annotated DailyDialog | English | Mental health and legal counseling | Acc: 90.30, F1: 87.19 (POEM) Acc: 86.78, F1: 75.27 (DailyDialog) |
| Bharti et al. [9] | PolitePEER | English | Peer reviews | Class-wise F1 scores- highly polite: 98, polite: 89, neutral:81, impolite:78, highly impolite:92 |
| Dasgupta et al. [34] | Stanford Politeness Corpus of requests Enron Corpus | English | Online Communities - Wikipedia and Stack Exchange Emails | F1 Scores - 75 (Wikipedia), 75 (SE), 81 (Enron) |

Reported Performance on Politeness Generation. There have been several works that attempt to incorporate politeness in natural language generation systems. Table 8 depicts the performance evaluation for such systems as documented in the existing literature. This evaluation encompasses various datasets across diverse domains. For instance, Madaan et al. [98] uses the n-grams tf-idf-based approach to estimate the politeness markers and generate the politeness-oriented text. They have achieved a BLEU score of 0.704. Likewise, Fu et al. [45] achieves a BLEU [124] score of 0.688 for generating the paraphrased version of a given message that can convey the desired level of politeness. Niu and Bansal [122] incorporate politeness in chit-chat dialogues and achieve a BLEU score of 0.24.

Table 8. Reported performance values (in % except BLEU and PPL) for computational approaches to politeness generation.

| Reference | Details | | | | Reported Performance |
|-----------------------|-------------------------------|--|------------------|---|---|
| | Task | Dataset | Language | Domain | |
| Fu et al. [45] | Style Transfer (Paraphrasing) | Stanford Politeness Corpus of requests | English | Online community (Wikipedia) | BLEU-s: 0.688 |
| Madaan et al. [98] | Style Transfer | Enron Corpus | English | Emails | BLEU-s: 0.704 |
| Sennrich et al. [144] | NMT | OpenSubtitles | English→German | Movie subtitles | BLEU: 0.24 |
| Feely et al. [38] | NMT | ASPEC, JESC, KFTT | English→Japanese | ASPEC (Scientific paper abstracts corpus), JESC (movie and television subtitles), KFTT (Wikipedia data) | BLEU: 0.433 (ASPEC), 0.203 (JESC), 0.255 (KFTT) |
| Niu and Bansal [122] | Style Transfer | MovieTriples | English | IMSDb movie scripts | BLEU-4: 0.94, Politeness Score: 61.0 |
| Golchha et al. [49] | Dialogue Generation | CYCCD | English | Customer care | BLEU: 0.692, ROUGE-1: 73.5, ROUGE-2: 69.9 ROUGE-L: 72.3, PPL: 0.437, CP: 77.5, EA: 86.8 |
| Firdaus et al. [40] | Dialogue Generation | CYCCD Hindi Customer Care Dialogue | English Hindi | Customer care | English: BLEU: 0.724, ROUGE-L: 75.2, PPL: 0.418, CP: 79.2, EA: 87.9 Hindi: BLEU: 0.596, ROUGE-L: 61.4, PPL: 0.441, CP: 68.5, EA: 77.9 |
| Wang et al. [167] | Dialogue Generation | CSR-Driver Exchanges | English | Customer care | BLEU: 9.89, word2vec-based similarity: 9.38 |
| Firdaus et al. [43] | Dialogue Generation | bAbi | English | Personalized conversations with polite templates | BLEU-4: 0.275, ROUGE-L: 35.2, PPL: 1.004, PA: 73.0 |
| Mishra et al. [107] | Dialogue Generation | Politeness-annotated MultiDoGo | English | Task-oriented conversations covering different domains | Success rate: 65.8 (airline), 78.2 (fastfood), 71.4 (finance), 72.4 (insurance), 73.2 (media), 63.2 (software) PA: 47.0, TCR: 53.0 |
| Silva et al. [148] | Polite Response Generation | Enron and fashion domain data | English | Emails and fashion | BLEU: 0.703, ROUGE: 86.7, METEOR: 51.5 (Enron) BLEU: 0.850, ROUGE: 83.7, METEOR: 58.99 (Fashion) |
| Saha et al. [140] | Counterspeech suggestions | CONAN Reddit Gab | English | Hatespeech instances | BLEU-2: 0.136 (CONAN), 0.083 (Reddit), 0.088 (Gab) COLA: 79 (CONAN), 81 (Reddit), 85 (Gab) |
| Mishra et al. [111] | Dialogue Generation | MHLCD | English | Mental health and legal counseling | CoStr: 80.30, Pol: 92.54, Emp: 46.4, PPL: 1.91, R-LEN: 18.71 |
| Mishra et al. [112] | Dialogue Generation | EPE-HLCC, EPE-enEIH | English | Mental health and/or legal counseling | EPC: 61.4, EEC: 60.7, PC: 65.8, EC: 63.1, PPL: 12.19, R-LEN: 25.81 (EPE-HLCC) EPC: 73.7, EEC: 75.9, PC: 71.6, EC: 75.2, PPL: 2.03, R-LEN: 21.09 (EPE-enEIH) |
| Priya et al. [130] | Dialogue Generation | HEAL | English | Mental health and legal counseling | CoAct: 56.5, EmpStr: 61.8, PolStr: 69.9, PPL: 2.55, R-LEN: 16.06 |
| Mishra et al. [109] | Dialogue Generation | Politeness-annotated Taskmaster | English | Restaurants, food ordering, movie, flight, hotels, music sports | DL:10.7, POL:85.1, MET:72.1, R-2-F1:54.3, SR:79 (Flights) DL:11.5, POL:93.6, MET:64.2, R-2-F1:44.4, SR:86 (Food) DL:9.9 11.5, POL:89.3, MET:70.9, R-2-F1:57.3, SR:82 (Hotels) DL:9.5, POL:88.8, MET:70.0, R-2-F1:40.7, SR:84 (Movies) DL:9.4, POL:95.9, MET:44.0, R-2-F1:23.7, SR:86 (Music) DL:8.5, POL:92.0, MET:70.9, R-2-F1:46.7, SR:82 (Restaurant) DL:10.9, POL:94.8, MET:65.7, R-2-F1:41.8, SR:81 (Sports) |
| Firdaus et al. [41] | Dialogue Generation | Sentiment-annotated CYCCD | English | Customer care | PPL: 1.004, BLEU-4: 0.275, ROUGE-L: 0.352, PA:0.77 |
| Mishra et al. [113] | Dialogue Generation | Politeness-annotated PersuasionForGood | English | Charity | PerStr: 59.98, PolSt: 41.11, Emp: 78.1, PPL: 11.06, LEN: 16.87 |
| Firdaus et al. [42] | Dialogue Generation | Polite Behavior-annotated CYCCD | English | Customer Care | PPL: 0.987, BLEU-4: 0.310, ROUGE-L: 38.2, PA: 81 |
| Mishra et al. [110] | Dialogue Generation | PSYCON | English | Mental health counseling | GA _c : 90.1, P _c : 84.1, CT _c : 92.6, Po _c : 92.5, IB _c : 83.4 PPL: 2.52, BERTScore-F1: 0.89, R-LEN: 23.89 |

Lately, Firdaus et al. [40], Golchha et al. [49] attempt to induce politeness in task-oriented dialogues. They have used both automatic evaluation (Perplexity (PPL) [18], Emotion Accuracy (EA), Content Preservation (CP), BLEU, ROUGE [92]) as well as human evaluation (Fluency, Content Adequacy, and Politeness Consistency) metrics to evaluate the performance of their proposed framework. Mishra et al. [107] utilize task success rate and human evaluation metrics, namely Task Completion Rate (TCR) and Politeness Adaptability (PA), to show the effectiveness of their proposed framework. Saha et al. [140] generate polite, detoxified, and emotionally charged counter-speech using three hate-speech datasets, namely CONAN [27], Reddit [132], and Gab [132], and observe significant improvement across

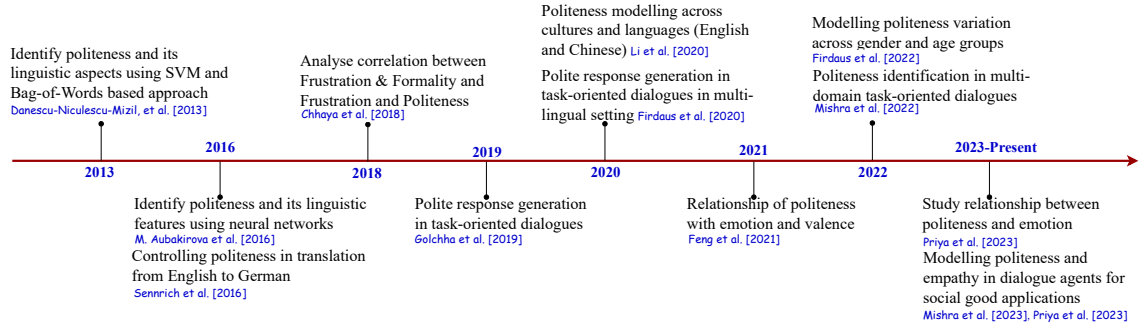


Fig. 3. Trends in computational politeness research

different attribute scores. The politeness and detoxification scores increased by around 15% and 6%, respectively, while the emotion in the counter-speech increased by at least 10 points across all the datasets.

In recent times, there have been works that employ task-specific metrics to measure the level of politeness and other related aspects like emotion and empathy in the generated responses. For instance, in Firdaus et al. [41, 43], the authors utilize Politeness Accuracy (PA) to measure the degree of politeness in the generated responses. Mishra et al. [111], Mishra et al. [112], Priya et al. [130] and Mishra et al. [113] employ several task-specific metrics, viz. <CoStr (Counseling Strategy Consistency), Pol (percentage of polite utterances generated), Emp (percentage of empathetic utterances generated)>, <PC (politeness correctness), EC (Emotion Correctness), EPC (Emotion-Politeness Consistency), EEC (Emotion-Empathy Consistency)>, <PolStr (Politeness Strategy Consistency), EmpStr (Empathy Strategy Consistency), CoAct (Counseling Act Correctness)>, and <PolSt (Politeness Strategy Consistency), Emp (Empathy), PerStr (Persuasion Strategy Consistency)>, respectively to evaluate the efficacy of their approach in terms of task success. It is to be noted that these works employ generic metrics like PPL (Perplexity) and R-LEN/DL (Response Length/Dialogue Length) as well for evaluating the generated responses.

Most of the past works do not provide an analysis of systems based on the variation of politeness across different domains, demographics, etc., mainly owing to the absence of corpora with varying degrees of politeness specified. However, an analysis of politeness transfer across domains is seen in Silva et al. [148]. They look at the examples that their model correctly generates polite responses through rewriting and conclude that politeness through rewriting provides a balance between the delivery of polite expressions and the preservation of contents. Recently, Firdaus et al. [43] analyze politeness variation across different age groups and genders, reporting a Politeness Appropriateness (PA) of 73% as a measure of the politeness quotient of response. Subsequently, Mishra et al. [110] explore variations in politeness and interpersonal behavior among individuals of different age groups, genders, and personas. Their analysis involved assessing Gender-Age Consistency (GA_c), Persona Consistency (P_c), Psychotherapeutic Approach Correctness (CT_c), Politeness Correctness (PO_c), and Interpersonal Behavior Correctness (IB_c), yielding values of 90.1%, 84.1%, 92.6%, 92.5%, and 83.4%, respectively.

7 TRENDS IN COMPUTATIONAL POLITENESS

In the preceding sections of this paper, we examined the datasets, approaches, and performance values of past works in computational politeness studies in Natural Language Processing. In this section, we delve into trends observed in computational politeness research. These trends are depicted in Figure 3. In earlier research works, the supervised approaches for politeness detection were investigated. These methods concentrated on utilizing specific features for

identifying politeness in natural language text. Following fundamental studies, recently, there has been a growing trend to employ context beyond the target text for comprehending and generating polite text. Furthermore, there is a growing interest in the computational approach for modeling the link between politeness and other socio-linguistic cues; the variations in politeness across different cultures, age groups, genders, and social power; and the study of politeness for various social good applications.

Use of Context in Computational Politeness Research. Using context beyond the text to be predicted is becoming a popular trend in computational politeness research. The term context here refers to any information outside the text to be predicted and outside the common knowledge. For example, the sentence *“Let’s have a drink”* may not be polite for a new person in the group but may be polite to many others. This type of situation needs knowledge beyond the text to be categorized. In the following part of this section, we refer to the textual unit to be classified as the “target text”.

The potential of context for politeness study has sparked the interest of researchers to look at the approaches for incorporating it. The contextual information may be of various types depending upon the problem definition; for instance, in automatic sarcasm detection, the context may come from the authors’ information, topical information, or conversations of which the sarcastic utterance is a part. In computational politeness research, incorporating the conversational context is prevalent for identifying and generating politeness. By conversational context, we mean the conversational text to which the target text belongs. Consider a simple exclamation, *“Of course!”*. This exclamation may or may not be polite. To understand the politeness in the given utterance, it is necessary to look at the conversation to which the target text belongs. If the utterance before the exclamation is *“Do you want me to help you with your problem?”*, then the impoliteness in the exclamation can be inferred. On the contrary, if the utterance preceding the exclamation is *“Can I have your booking preference?”*, then the utterance, *“Of course!”* here implies *“Yes! obviously, I’m happy to share the preferences with you.”*, and is thus interpreted as polite.

Mishra et al. [108] develop a computational framework for identifying politeness in goal-oriented dialogue systems considering the context from previous utterances. They have utilized the hierarchical transformer network to obtain the context-sensitive representation of the utterances. Golchha et al. [49] utilize the contextual information from previous utterances for transforming neutral customer care responses into polite ones. Firdaus et al. [40] extend the work of Golchha et al. [49] and generate polite responses in both English and Hindi while being contextually coherent with the previous conversation turns. Mishra et al. [107] use numerous politeness-related para-linguistic features [16] and build a conversational agent with a polite-to-impolite spectrum based on a politeness classifier without relying on parallel data. The authors consider conversational history as the context for politeness classification.

Modeling Relationship between Politeness and Social Factors. Due to the importance of politeness in cross-cultural scenarios, there is a recent trend to model politeness across cultures and languages. Li et al. [88] study the differences and similarities in politeness expressions between American English and Mandarin Chinese. The authors have developed machine learning algorithms for predicting politeness in both languages. Sperlich et al. [151] examine the system of politeness covering French *tu* and *vous* from the perspective of a Korean learner of French. They state that the acquisition of *tu* is comparatively easier due to positive transfer than *vous*’s acquisition because it involves negative transfer. The explicit/implicit state of the learners’ pragmalinguistic and sociopragmatic knowledge is another factor affecting the use and interpretation of *tu/vous*.

Natural language is the most powerful tool for performing speech acts in human communication (such as making requests, giving commands, apologizing, expressing gratitude, etc.). These speech acts are executed in accordance with specific norms and principles. One of these principles is politeness. Munkova et al. [118] propose a novel approach to computational modeling of specific features of politeness in the speech act of requests in written forms. It examines the

similarities and differences in using specific social and expressive factors in the primary language (L1) and the foreign language (L2, English). In particular, the authors formulate the Transaction/Sequence model for text representation, apply the association rules for analysis, and indicate that the requests written in the primary language are less direct than those written in a foreign language. A few studies explore the link between politeness and social variables, focusing on social power dynamics. The work of Danescu-Niculescu-Mizil et al. [33] is the earliest attempt in this direction. The authors have used their computational framework to study the relationship between politeness and power. They have shown that polite editors are likely to gain higher status through elections; however, once they are at the top of their reputation, their politeness decreases. They have also demonstrated the utility of their framework in exploring the differences in politeness across other factors of interest, such as communities, geographical regions, and gender. The authors reveal that (i). different programming communities vary significantly by politeness on Stack Exchange; (ii). Wikipedia editors from the U.S. Midwest are most polite, and (iii). female editors on Wikipedia are generally more polite compared to their male counterparts.

Some studies evaluate the effect of people's age and gender on human-machine interaction and discuss politeness as an essential social skill to develop trust in such interactions and to ensure optimal interactions between machines and humans. Inbar and Meyer [63] conduct tests to examine how people behave when a more or less courteous robot responsible for access control interacts with individuals of varying ages and genders. Their study reveals that politeness is the only factor that influences the participant's perception of the peacekeeping robot's behavior. Polite peacekeeping robots are perceived as friendlier, fairer, less intimidating, and more appropriate. However, they further demonstrate that the age and gender of the people interacting with the robots had no significant effect on participants' impressions of its attributes. The authors claim that their findings are robust as the studies involved male and female participants of a wide range of ages. They have explained in support of these findings that when the gap between a polite and an impolite robot's behavior is so pronounced, it overshadows other distinctions of age and gender. Latterly, Firdaus et al. [43] study politeness variations across different age groups and genders. They utilize the age and gender of the user to decide the degree of politeness in a response. The authors state that system responses for an *elderly* person should be more polite as compared to the responses for a *young* person of the same gender. Likewise, the response for the *elderly female* user has a slightly higher politeness quotient than the *elderly male* user.

Modeling Relationship between Politeness and Other Socio-linguistic Cues. Lately, research has been inclined towards exploring the relationship between politeness and other socio-linguistic cues in interaction. Besides politeness, many socio-linguistic cues are used in the conversational analysis, such as emotion, sentiment, and dialogue acts. Detecting politeness in communication offers information about the interlocutor's social demeanor. Likewise, perceiving emotions provides information about the participants' emotional states. Sentiment or valence is another dimension in conversational analysis that helps detect the positivity and negativity of emotion, and identifying dialogue acts gives information about the performative function of an utterance. Bothe and Wermter [12] attempt to correlate the relational bonds between politeness and emotions, and politeness and dialogue acts. In particular, the authors analyze the utterances in the DailyDialog dataset [90] against the emotion and dialogue act classes in the dataset. The authors have discovered that most of the utterances in *Sadness* and *Happiness* emotion classes are polite, and the utterances with emotion classes, *Anger* and *Disgust*, are more likely to tend to be impolite. A similar phenomenon occurs with dialogue acts; *Inform* and *Commissive* contain more polite utterances than *Question* and *Directive* dialogue acts.

Feng et al. [39] use the Ortony, Clore, and Collins (OCC) model [123] and consider three aspects, viz. Elicitor or Cause, Valence, and Conduct for annotating the utterances in the dataset with the appropriate emotion classes. The authors mention that in conversational analysis, it is essential to distinguish between user and system utterances and

identify valence as it is highly related to task success or failure, thus expressing the positivity or negativity of emotion. Successfully achieving a goal is likely to evoke positive valence, while a misunderstanding with an agent or failure to achieve the objective will likely cause negative valence. Moreover, conduct that describes the politeness of users can indicate the degree of valence. For instance, the user can express extreme dissatisfaction through rudeness. It also helps to differentiate between emotions, such as those linked with apology or anger, both of which are intrinsically negative. The apologetic emotion is more likely to be associated with the polite demeanor of the interlocutor, and anger, by contrast, is more likely to be associated with impolite behavior. The authors have eventually created a set of seven emotions that differ in terms of elicitor, valence, and conduct to reflect the cognitive context of emotions while preserving the simplicity of labeling.

Following this line of the work, Priya et al. [129] investigate the inter-connectedness between politeness and emotion and devise a novel multi-task learning (MTL) framework, Caps-DGCN for simultaneous Politeness and Emotion Detection (PED) in conversations. The authors employ the Capsule Network (Caps) [37] to capture the local ordering of words in the utterance and corresponding semantic representations, and the Directional Graph Convolutional Network (DGCN) [23] to encode the syntactic information by incorporating the dependency among the words. The authors mention that the identification of politeness and emotion in conversations requires the extraction of semantic as well as syntactic information from an utterance. The authors established that their proposed Caps-DGCN framework strengthens the utterance representation by utilizing semantic and syntactic information and effectively captures the relatedness between politeness and emotion in utterances.

Modeling Politeness in Dialogue Systems for Social Good Applications. Recently, there have been studies that investigate the role of politeness and its complementary cues (emotion, empathy, etc.) in dialogue systems for various social good applications like persuasion [113], mental health support [111, 112, 130], to mention a few. Mishra et al. [113] devise a polite and empathetic dialogue system for persuasion tasks in the charity domain. The authors hypothesize that incorporating polite tone and emotional quotient into persuasive dialogue systems may make persuasive messages more appealing to the user, resulting in better persuasive outcomes. To develop a polite and empathetic persuasive dialogue system (PEPDS), the authors have fine-tuned a Maximum Likelihood Estimation loss-based dialogue model in an RL setting by designing a novel reward function. This reward function comprises five sub-rewards, viz. Politeness-strategy consistency, Persuasion, Emotion, Dialogue-Coherence, and Non-repetitiveness to ensure politeness-strategy consistency, persuasiveness, emotion acknowledgment, dialogue-coherence, and non-repetitiveness, respectively, in the generated responses. Afterward, the authors build an Empathetic transformer model on top of the RL fine-tuned model to transform non-empathetic utterances into empathetic ones.

The authors in [110–112, 130] propose counseling dialogue systems capable of conversing with the users politely. Mishra et al. [111] propose a polite and empathetic mental health and legal counseling dialogue system (**Po-Em-MHLCDS**) for crime victims. To develop **Po-Em-MHLCDS**, the authors have alternatively trained two different GPT2-medium [133] based language models to learn the distribution of the victim’s and counselor’s utterances to achieve natural language interaction between both the models. This MLE loss-based agent’s trained model is then fine-tuned in an RL framework. In order to optimize the agent’s behavior, the authors employ an RL loss considering an efficiently designed reward function to ensure appropriate counseling strategy, politeness, empathy, contextual coherence, and non-repetitiveness while counseling the victim.

Mishra et al. [112] introduce an emotion-adaptive polite and empathetic counseling conversational agent **PAL**. To develop **PAL** that can generate polite and empathetic responses while reflecting the client’s emotional state naturally and without unnecessary repetition, the authors have trained a BART-large [86] based client’s emotion-informed

dialogue model (*EIDM*) in a supervised learning framework. The *EIDM* is then fine-tuned with a Proximal Policy Optimization (PPO) [142] loss utilizing novel designed reward function consisting of preference rewards, task-specific rewards, and generic rewards. Priya et al. [130] present a polite and empathy strategies-adaptive mental health and legal counseling dialogue system, named **PARTNER** that can adopt appropriate politeness and empathy strategy together with the pertinent counseling act, including but not limited to persuasion during the counseling sessions. The authors mention that **PARTNER** learns from user interactions and improves itself depending on user feedback through rewards. In particular, the authors have initially trained a Cross-Entropy Loss based Dialogue Model (*CELDM*) - to foster natural language interaction between the *counseling bot* and the *crime victim*, and then fine-tuned the trained *CELDM* using Proximal Policy Optimization (PPO) [142] Loss based Dialogue Model (*PLDM*) - to generate politeness and empathy strategies-adaptive counseling responses. The GPT-small [133] is used as the backbone architecture for *CELDM* model.

Mishra et al. [110] propose e-Therapist, a polite and interpersonal dialogue system for psychotherapy that generates responses according to the user's age, gender, persona, and sentiment. e-Therapist is developed by first fine-tuning a language model using a supervised learning approach, which is referred to as SLLM. The SLLM then generates a set of candidate responses which are ranked using a novel reward model employing gender-age awareness, persona awareness, counseling, politeness, and interpersonal behavior rewards to reinforce the counseling agent to ensure the preferences of gender, age, persona, and sentiment-aware polite interpersonal psychotherapeutic responses. This reward model is finally optimized using the NLPO loss [134] to yield the counseling dialogue system, e-Therapist.

8 ISSUES IN COMPUTATIONAL POLITENESS

In this section, we discuss the critical issues that appear in different computational politeness works. The most important issue that is faced in most computational politeness research is the lack of annotated data.

With the success of machine learning and deep learning-based architectures, there has been a surge in interest in utilizing such architectures for computational politeness studies. It must, however, be noted that these approaches are data-intensive, and publicly available data are scarce for computational politeness study. Danescu-Niculescu-Mizil et al. [33] propose a dataset of requests annotated for politeness, which may be considered ground-breaking work in this direction. Following this work, many research works such as Aubakirova and Bansal [4], Li et al. [88] use this corpus to train classifiers for identifying politeness. Mishra et al. [107, 108] also utilize the classifier used in Danescu-Niculescu-Mizil et al. [33] for creating the conversational dataset for their work. Golchha et al. [49], Madaan et al. [98], Niu and Bansal [122], Wang et al. [167] are the few other works that introduce corpus for computational modeling of politeness. A few corpora have been available for modeling politeness in languages other than English. For instance, Firdaus et al. [40], Kumar [76] introduce dataset in Hindi, and Li et al. [88] introduce Chinese corpora annotated for politeness. Many of these datasets are created manually, either through crowdsourcing or by the researchers themselves. Recently, within computational politeness research, there has also been a growing interest in using linguistic variation related to the identity of speakers, such as age, gender, demographics, etc., for improving the identification and generation of politeness in natural language texts. Firdaus et al. [43] train gender and age-specific models for generating personalized and polite dialogue responses by manually creating the gender and age-specific polite templates. The manual transcription and annotation of data are time-intensive and costly.

Moreover, the quality of annotations is a significant concern in the case of manually annotated datasets. Because of the subjective nature of politeness, the inter-annotator agreement values mentioned in the previous research works are diverse. Li et al. [88] report Krippendorff's alpha [73] inter-rater reliability of 0.528 for Twitter and 0.661 for Weibo. Chhaya et al. [24] indicate agreements for 3 class and 5 class politeness annotations as 0.64 ± 0.03 . In Golchha et al.

[49], Mishra et al. [107, 108], the authors report a multi-rater Kappa [101] agreement ratio of approximately 80%. The value in the case of Firdaus et al. [40] is 90%. This observation highlights the need to frame appropriate guidelines for annotators. Danescu-Niculescu-Mizil et al. [33] present a corpus of requests annotated for politeness. They state that for politeness annotation, they asked the annotators to read a batch of requests, consider those requests as originating from a co-worker by email, and then annotate the requests with perceived politeness. Li et al. [88] provide the annotators with a list of politeness strategies along with their definitions and suitable examples from social media to annotate the Twitter and Weibo posts for perceived politeness values. Golchha et al. [49], Mishra et al. [107, 108] employ annotators who are highly proficient in English and have good exposure to the related task and explain them with suitable examples for politeness annotation.

9 CONCLUSION AND FUTURE DIRECTIONS

Over the past several years, research in computational approaches to politeness has received immense attention, prompting a retrospective evaluation of the overall picture that these individual works have produced. This article examines various approaches to computational politeness. We observed four milestones in the history of computational politeness research: supervised and weakly supervised feature extraction to **identify and induce politeness in a given text**, incorporation of context beyond the target text, a few studies of politeness across different social factors such as culture, age, power, gender, etc., and exploring the link between socio-linguistic indicators such as emotion, dialogue acts, etc.

This survey paper furnished illustrations containing the datasets, approaches, and performance values as stated in previous works. The two popular techniques for constructing the politeness-annotated dataset include manual labeling for politeness and automatic labeling for politeness through supervision. We observed that statistical approaches utilize features like lexical, syntactic, affect-based, and so on. Various deep learning-based approaches to automatic politeness identification and generation have also been presented in the literature. Recently, we also witnessed a growing trend in evaluating the competency of LLMs in identifying, interpreting, and generating politeness. Previously, additional conversation-specific features have been investigated to incorporate context. An underlying theme of several earlier methods is attempting to identify formality, positivity, and rapport building that is at the source of politeness. We further highlight the issues in computational politeness: **lack of politeness-annotated corpora and quality of manual annotation**. A comparative table is also provided, outlining key aspects of previous articles, including methodologies, annotation techniques, and features. We believe that the researchers who intend to comprehend the present state-of-the-art in computational politeness research will find this table informative and helpful.

Recent trends in computational politeness research emphasize contextual inclusion and modeling politeness in relation to various social and linguistic factors. While there are manually labeled gold-standard datasets, supervision is the prominent method for obtaining labels. Novel deep learning-based approaches have also been employed to infuse the context of different forms. Based on our examination of these works, we find the following emerging directions:

Modeling politeness variation across different social factors. A future area of research can be the automatic identification of variation in politeness across people of **different age groups, genders, and cultures**. A few approaches, such as [33, 88], have explored such variation to a limited extent. Lately, Firdaus et al. [43] and Mishra et al. [110] generate polite responses based on the age and gender of the user. A thorough investigation and modeling of such variation would be helpful for the development of end-to-end human-like artificial intelligent systems. It facilitates understanding of the social attitude of the interlocutors, thereby fostering the interaction among them.

Modeling politeness variation across different socio-linguistic cues. Bothe and Wermter [12] recently provided a corpus-based study and attempted to correlate the relational bonds between politeness and its **complementary cues like emotion, dialog acts, etc.** Following this, Priya et al. [129] employed a deep learning-based multi-task framework for exploring the association between politeness and emotion. This work could be extended to learn social behaviors using analyzed linguistic features with the help of more advanced deep-learning techniques and LLMs. The relatedness of socio-linguistic variables also adds a new layer to the behavioral analysis of interlocutors and their use in the virtual assistant and human-robot interaction.

Modeling of politeness for different domains and low-resource languages. Politeness is perceived on the basis of shared knowledge. As indicated in past works such as Munkova et al. [118], politeness is closely related to language-specific traits. The majority of the past works on identifying and incorporating politeness focus on the English language and a few domains, **such as customer care and fashion.** Future approaches to computational politeness will benefit from **expanding politeness understanding for new languages like Hindi, Spanish, etc., and for new domains.**

Retrieval of contextual information using deep learning-based architectures. Until now, only a handful of approaches have examined architectures based on deep learning. **As mentioned in [40, 49, 107, 108], extracting contextual information may be useful to acquire the additional shared knowledge (for example, conversation-specific or user-specific knowledge) that is required to understand politeness.**

Investigating politeness for social good applications. In recent years, there has been a notable emergence in the exploration of politeness for various social good applications. **From mental health counseling to persuasion techniques, researchers are delving deeper into understanding how politeness can positively influence various aspects of human interaction and behavior.** A handful of studies in mental health counseling [110–112, 129, 130] attempted to induce politeness in dialogue agents to foster trust, rapport, and overall therapeutic effectiveness between agents and users. Likewise, in the domain of persuasion, [113] incorporated politeness strategies to enhance the persuasiveness of messages and facilitate a more constructive and respectful atmosphere. Future approaches to computational politeness could explore politeness for other social good applications, like negotiation, education, environmental monitoring, etc., to build more empathetic, engaging, and effective communication channels that positively impact users' well-being and promote positive social change.

ACKNOWLEDGMENTS

Priyanshu Priya acknowledges the Innovation in Science Pursuit for Inspired Research (INSPIRE) Fellowship implemented by the Department of Science and Technology, Ministry of Science and Technology, Government of India for financial support. Asif Ekbal acknowledges the Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

REFERENCES

- [1] Mikhail Alexandrov, Xavier Blanco, Natalia Ponomareva, and Paolo Rosso. 2007. Constructing empirical models for automatic dialog parameterization. In *International Conference on Text, Speech and Dialogue*. Springer, 455–463.
- [2] Mikhail Alexandrov, Natalia Ponomareva, and Xavier Blanco. 2008. Regression Model for Politeness Estimation Trained on Examples. In *Proceedings of the NooJ'07 conference*. 206–13.
- [3] Ahmad Aljanaideh, Eric Fosler-Lussier, and Marie-Catherine de Marneffe. 2020. Contextualized Embeddings for Enriching Linguistic Analyses on Politeness. In *Proceedings of the 28th International Conference on Computational Linguistics*. 2181–2190.
- [4] Malika Aubakirova and Mohit Bansal. 2016. Interpreting Neural Networks to Improve Politeness Comprehension. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2035–2041.

- [5] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.
- [6] Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *Proceedings of the Web Conference 2021*. 1134–1145.
- [7] Ricardo Barandela, José Salvador Sánchez, Vicente Garcia, and Edgar Rangel. 2003. Strategies for learning in class imbalance problems. *Pattern Recognition* 36, 3 (2003), 849–851.
- [8] Elizabeth Bates and Louise Silvern. 1977. Social Adjustment and Politeness in Preschoolers. *Journal of Communication* 27, 2 (1977), 104–111.
- [9] Prabhat Kumar Bharti, Meith Navlakha, Mayank Agarwal, and Asif Ekbal. 2023. PolitePEER: does peer review hurt? A dataset to gauge politeness intensity in the peer reviews. *Language Resources and Evaluation* (2023), 1–23.
- [10] Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press.
- [11] Avrim L Blum and Merrick L Furst. 1997. Fast planning through planning graph analysis. *Artificial intelligence* 90, 1-2 (1997), 281–300.
- [12] Chandrakant Bothe and Stefan Wermter. 2022. Conversational Analysis of Daily Dialog Data using Polite Emotional Dialogue Acts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2395–2400.
- [13] Derek Bousfield. 2008. Impoliteness in interaction. *Impoliteness in interaction* (2008), 1–295.
- [14] Jonathon D Brown and S Smart. 1991. The self and social conduct: Linking self-representations to prosocial behavior. *Journal of Personality and Social psychology* 60, 3 (1991), 368.
- [15] Penelope Brown and Colin Fraser. 1979. Speech as a marker of situation. In *Social markers in speech*. Cambridge University Press, 33–62.
- [16] Penelope Brown and Stephen C Levinson. 1978. Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction*. Cambridge University Press, 56–311.
- [17] Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*. Vol. 4. Cambridge university press.
- [18] Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Jennifer C Lai, and Robert L Mercer. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics* 18, 1 (1992), 31–40.
- [19] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [20] Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling Empathy and Distress in Reaction to News Stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4758–4765.
- [21] Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4516–4525.
- [22] Laura L Carstensen, Bulent Turan, Susanne Scheibe, Nilam Ram, Hal Ersner-Hershfield, Gregory R Samanez-Larkin, Kathryn P Brooks, and John R Nesselroade. 2011. Emotional experience improves with age: evidence based on over 10 years of experience sampling. *Psychology and aging* 26, 1 (2011), 21.
- [23] Guimin Chen, Yuanhe Tian, and Yan Song. 2020. Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In *Proceedings of the 28th international conference on computational linguistics*. 272–279.
- [24] Niyati Chhaya, Kushal Chawla, Tanya Goyal, Projjal Chanda, and Jaya Singh. 2018. Frustrated, polite, or formal: Quantifying feelings and tone in email. In *Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media*. 76–86.
- [25] Christine Christie. 2007. Relevance theory and politeness. *Journal of Politeness Research* 3, 2 (2007), 269–294. <https://doi.org/doi:10.1515/PR.2007.012>
- [26] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research* 25, 70 (2024), 1–53. <http://jmlr.org/papers/v25/23-0870.html>
- [27] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. 2019. CONAN-COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2819–2829.
- [28] Jonathan Culpeper. 2011. 13. Politeness and impoliteness. *Pragmatics of society* 5 (2011), 393.
- [29] Jonathan Culpeper. 2011. *Impoliteness: Using language to cause offence*. Vol. 28. Cambridge University Press.
- [30] Jonathan Culpeper. 2016. Impoliteness strategies. *Interdisciplinary studies in pragmatics, culture and society* (2016), 421–445.
- [31] Jonathan Culpeper, Michael Haugh, and Dániel Z Kádár. 2017. *The Palgrave handbook of linguistic (im) politeness*. Springer.
- [32] Marianne Dainton, Laura Stafford, and Daniel J Canary. 1994. Maintenance strategies and physical affection as predictors of love, liking, and satisfaction in marriage. *Communication Reports* 7, 2 (1994), 88–98.
- [33] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 250–259.

- [34] Tirthankar Dasgupta, Manjira Sinha, and Chundru Geetha Praveen. 2023. Graph Induced Transformer Network for Detection of Politeness and Formality in Text. In *Companion Proceedings of the ACM Web Conference 2023*. 221–224.
- [35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [36] Doris Dippold, Jenny Lynden, Rob Shrubsall, and Rich Ingram. 2020. A turn to language: How interactional sociolinguistics informs the redesign of prompt: response chatbot turns. *Discourse, Context & Media* 37 (2020), 100432.
- [37] Chunling Du, Haifeng Sun, Jingyu Wang, Qi Qi, Jianxin Liao, Chun Wang, and Bing Ma. 2019. Investigating capsule network and semantic feature on hyperplanes for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 456–465.
- [38] Weston Feely, Eva Hasler, and Adrià de Gispert. 2019. Controlling Japanese honorifics in English-to-Japanese neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*. 45–53.
- [39] Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 4096–4113.
- [40] Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Incorporating Politeness across Languages in Customer Care Responses: Towards building a Multi-lingual Empathetic Dialogue Agent. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 4172–4182.
- [41] Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. PoliSe: Reinforcing Politeness using User Sentiment for Customer Care Response Generation. In *Proceedings of the 29th International Conference on Computational Linguistics*. 6165–6175.
- [42] Mauajama Firdaus, Priyanshu Priya, and Asif Ekbal. 2023. Mixing It Up: Inducing Empathy and Politeness using Multiple Behaviour-aware Generators for Conversational Systems. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*. 336–347.
- [43] Mauajama Firdaus, Arunav Shandilya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Being Polite: Modeling Politeness Variation in a Personalized Dialog Agent. *IEEE Transactions on Computational Social Systems* (2022).
- [44] John L Fischer. 1965. The stylistic significance of consonantal sandhi in Trukese and Ponapean. *American Anthropologist* 67, 6 (1965), 1495–1502.
- [45] Liye Fu, Susan Fussell, and Cristian Danescu-Niculescu-Mizil. 2020. Facilitating the Communication of Politeness through Fine-Grained Paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5127–5140.
- [46] Vicente Garcia, José Salvador Sánchez, and Ramón Alberto Mollineda. 2012. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems* 25, 1 (2012), 13–21.
- [47] Eelen Gino. 2001. A critique of Politeness Theories. Manchester: St.
- [48] Erving Goffman. 1967. *Interaction ritual: Essays on face-to-face interaction (1st ed.)*. Routledge, New York, NY, USA.
- [49] Hitesh Golchha, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Courteously Yours: Inducing courteous behavior in Customer Care responses using Reinforced Pointer Generator Network. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 851–860.
- [50] Sage Lambert Graham. 2007. Disagreeing to agree: Conflict(im) politeness and identity in a computer-mediated community. *Journal of pragmatics* 39, 4 (2007), 742–759.
- [51] H Paul Grice. 1975. *Logic and Conversation, Syntax and Semantics*, vol. 3 edited by P. Cole and J. Morgan.
- [52] Yueguo Gu. 1990. Politeness phenomena in modern Chinese. *Journal of pragmatics* 14, 2 (1990), 237–257.
- [53] Swati Gupta, Marilyn A Walker, and Daniela M Romano. 2007. How rude are you?: Evaluating politeness and affect in interaction. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 203–217.
- [54] Michael Haugh. 2010. When is an email really offensive?: Argumentativity and variability in evaluations of impoliteness. (2010).
- [55] Francis Heylighen and Jean-Marc Dewaele. 1999. *Formality of language: definition, measurement and behavioral determinants*. Technical Report. Center "Leo Apostel", Free University of Brussels, Pleinlaan 2, B-1050 Brussels, Belgium.
- [56] Erin R Hoffman, David W McDonald, and Mark Zachry. 2017. Evaluating a computational approach to labeling politeness: Challenges for the application of machine classification to social computing data. *Proceedings of the ACM on Human-computer Interaction* 1, CSCW (2017), 1–14.
- [57] Janet Holmes. 1988. Paying compliments: A sex-preferential politeness strategy. *Journal of pragmatics* 12, 4 (1988), 445–465.
- [58] Janet Holmes. 2013. *Women, men and politeness*. Routledge.
- [59] Thomas Holtgraves and Yang Joong-Nam. 1990. Politeness as universal: Cross-cultural perceptions of request strategies and inferences based on their use. *Journal of personality and social psychology* 59, 4 (1990), 719.
- [60] Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics* 11, 6 (1987), 689–719.
- [61] Richard Hudson. 1994. About 37% of word-tokens are nouns. *Language* 70, 2 (1994), 331–339.
- [62] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 216–225.
- [63] Ohad Inbar and Joachim Meyer. 2019. Politeness counts: Perceptions of peacekeeping robots. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 232–240.
- [64] Judith T Irvine. 1979. Formality and informality in communicative events. *American anthropologist* 81, 4 (1979), 773–790.
- [65] Richard W Janney and Horst Arndt. 1993. Universality and relativity in cross-cultural politeness research: a historical perspective. *Multilingua* 12, 1 (1993), 13–50.

- [66] Martha Jeong, Julia Minson, Michael Yeomans, and Francesca Gino. 2019. Communicating with warmth in distributive negotiations is surprisingly counterproductive. *Management Science* 65, 12 (2019), 5813–5837.
- [67] W Lewis Johnson and Paola Rizzo. 2004. Politeness in tutoring dialogs: “run the factory, that’s what I’d do”. In *International Conference on Intelligent Tutoring Systems*. Springer, 67–76.
- [68] W Lewis Johnson, Paola Rizzo, Wauter Bosma, Sander Kole, Mattijs Ghijsen, and Herwin van Welbergen. 2004. Generating socially appropriate tutorial dialog. In *Tutorial and Research Workshop on Affective Dialogue Systems*. Springer, 254–264.
- [69] Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503* (2017).
- [70] Manfred Kienpointner. 2008. Impoliteness and emotional arguments. (2008).
- [71] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations* (Palais des Congrès Neptune, Toulon, France) (ICLR ’17). <https://openreview.net/forum?id=SJU4ayYgl>
- [72] Bryan Klimt and Yiming Yang. 2004. Introducing the Enron corpus.. In *CEAS*.
- [73] Klaus Krippendorff. 2007. Computing Krippendorff’s Alpha-Reliability. <http://www.asc.upenn.edu/Krippendorff/> (2007).
- [74] Ritesh Kumar. 2012. Challenges in the development of annotated corpora of computer-mediated communication in Indian languages: A case of Hindi. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*. 299–302.
- [75] Ritesh Kumar. 2014. Developing Politeness Annotated Corpus of Hindi Blogs.. In *LREC*. 1275–1280.
- [76] Ritesh Kumar. 2021. Towards automatic identification of linguistic politeness in Hindi texts. *arXiv preprint arXiv:2111.15268* (2021).
- [77] Ritesh Kumar and Girish Nath Jha. 2010. Translating politeness across cultures: Case of Hindi and English. In *Proceedings of the 3rd international conference on Intercultural collaboration*. 175–178.
- [78] Aylin C Küntay, Keiko Nakamura, and B Ateş Şen. 2014. Crosslinguistic and crosscultural approaches to pragmatic development. *Pragmatic development in first language acquisition* (2014), 317–342.
- [79] Robin Lakoff. 1973. Language and woman’s place. *Language in society* 2, 1 (1973), 45–79.
- [80] Robin Lakoff. 1973. The logic of politeness: Or, minding your p’s and q’s. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, Vol. 9. Chicago Linguistic Society, 292–305.
- [81] Robin Tolmach Lakoff. 1989. The Limits of Politeness: Therapeutic and Courtroom Discourse in Linguistic Politeness II. *Multilingua* 8, 2-3 (1989), 101–129.
- [82] Andreas Langlotz and Miriam A Locher. 2017. (Im) politeness and emotion. In *The Palgrave Handbook of Linguistic (Im) politeness*. Springer, 287–322.
- [83] GN Leech. 1983. Principles of Pragmatics London: Longman Group Ltd. (1983).
- [84] Geoffrey Leech. 1992. Pragmatic principles in Shaw’s you never can tell. *Language, text and context: essays in stylistics* (1992), 259–78.
- [85] Geoffrey Leech. 2005. Politeness: is there an East-West divide. *Journal of foreign languages* 6, 3 (2005), 1–30.
- [86] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
- [87] Can Li, Bin Pang, Wenbo Wang, Lingshu Hu, Matthew Gordon, Detelina Marinova, Bitty Balducci, and Yi Shang. 2023. How Well Can Language Models Understand Politeness?. In *2023 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 230–231.
- [88] Mingyang Li, Louis Hickman, Louis Tay, Lyle Ungar, and Sharath Chandra Guntuku. 2020. Studying Politeness across Cultures Using English Twitter and Mandarin Weibo. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–15.
- [89] Xiujun Li, Yu Wang, Siqi Sun, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *arXiv preprint arXiv:1807.11125* (2018).
- [90] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 986–995.
- [91] Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143* (2022).
- [92] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
- [93] Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, Yun-Nung Chen and Abhinav Rastogi (Eds.). Association for Computational Linguistics, Toronto, Canada, 47–58. <https://doi.org/10.18653/v1/2023.nlp4convai-1.5>
- [94] Miriam A Locher. 2006. Polite behavior within relational work: The discursive approach to politeness. *Multilingua* 25, 3 (2006), 249–267.
- [95] Miriam A Locher and DJ Glick. 2006. Power and politeness in action: Disagreements in oral communication. *LANGUAGE IN SOCIETY-LONDON*-35, 5 (2006), 729–732.
- [96] Miriam A Locher and Richard J Watts. 2005. Politeness theory and relational work. (2005).
- [97] Miriam A Locher and Richard J Watts. 2008. Relational work and impoliteness: Negotiating norms of linguistic behaviour. *Language Power and Social Process* 21 (2008), 77.
- [98] Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabás Póczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness Transfer: A Tag and Generate Approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1869–1881.

- [99] Michael P Maratsos. 1973. Nonegocentric communication abilities in preschool children. *Child development* (1973), 697–700.
- [100] Bonnie McElhinny. 1996. Language and Gender. *Sociolinguistics and Language Teaching* (1996), 218.
- [101] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.
- [102] Christopher Miller, Peggy Wu, H Funk, Lewis Johnson, and H Viljalmsson. 2007. A computational approach to etiquette and politeness: An “Etiquette Engine™” for cultural interaction training. *Proceedings of BRIMS 2007* (2007).
- [103] Christopher Miller, Peggy Wu, H Funk, Peggy Wilson, and Lewis Johnson. 2006. A computational approach to etiquette and politeness: Initial test cases. In *proceedings of 2006 BRIMS Conference*. 15–18.
- [104] Christopher A Miller, Peggy Wu, and Harry B Funk. 2008. A computational approach to etiquette: Operationalizing Brown and Levinson’s politeness model. *IEEE Intelligent Systems* 23, 4 (2008), 28–35.
- [105] Sara Mills. 2003. *Gender and politeness*. Number 17. Cambridge University Press.
- [106] Julia A Minson, Frances S Chen, and Catherine H Tinsley. 2020. Why won’t you listen to me? Measuring receptiveness to opposing views. *Management Science* 66, 7 (2020), 3069–3094.
- [107] Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Please be polite: Towards building a politeness adaptive dialogue system for goal-oriented conversations. *Neurocomputing* 494 (2022), 242–254.
- [108] Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Predicting Politeness Variations in Goal-Oriented Conversations. *IEEE Transactions on Computational Social Systems* (2022).
- [109] Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2023. GenPADS: Reinforcing politeness in an end-to-end dialogue system. *Plos one* 18, 1 (2023), e0278323.
- [110] Kshitij Mishra, Priyanshu Priya, Manisha Burja, and Asif Ekbal. 2023. e-THERAPIST: I suggest you to cultivate a mindset of positivity and nurture uplifting thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 13952–13967.
- [111] Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023. Help Me Heal: A Reinforced Polite and Empathetic Mental Health and Legal Counseling Dialogue System for Crime Victims. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14408–14416.
- [112] Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023. PAL to Lend a Helping Hand: Towards Building an Emotion Adaptive Polite and Empathetic Counseling Conversational Agent. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 12254–12271.
- [113] Kshitij Mishra, Azlaan Mustafa Samad, Palak Totala, and Asif Ekbal. 2022. PEPDS: A polite and empathetic persuasive dialogue system for charity donation. In *Proceedings of the 29th International Conference on Computational Linguistics*. 424–440.
- [114] Tomoki Miyamoto, Daisuke Katagami, and Mayumi Usami. 2020. A politeness control method for conversational agents considering social relationships with users. In *Annual Conference of the Japanese Society for Artificial Intelligence*. Springer, 224–231.
- [115] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence* 29, 3 (2013), 436–465.
- [116] Michael B Montgomery. 1998. Multiple Modals in LAGS. *From the gulf states and beyond: The legacy of Lee Pederson and LAGS* (1998), 90.
- [117] Sourabrata Mukherjee, Vojtěch Hudeček, and Ondřej Dušek. 2023. Polite Chatbot: A Text Style Transfer Application. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*. 87–93.
- [118] Dasa Munkova, Michal Munk, and Zuzana Fráterová. 2013. Identifying social and expressive factors in request texts using transaction/sequence model. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. 496–503.
- [119] Keiko Nakamura, Y Shirai, S Kobayashi, and S Miyata. 2002. Polite language usage in mother–infant interactions: A look at language socialization. *Studies in language sciences* 2 (2002), 175–191.
- [120] Courtney Napoles, Aasish Pappu, and Joel Tetreault. 2017. Automatically identifying good conversations online (yes, they do exist!). In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11. 628–631.
- [121] Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenza. 2017. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the 11th linguistic annotation workshop*. 13–23.
- [122] Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics* 6 (2018), 373–389.
- [123] Andrew Ortony, Gerald L Clore, and Allan Collins. 1988. *The cognitive structure of emotions* Cambridge. UK: Cambridge University Press9 (1988).
- [124] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
- [125] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.
- [126] Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 926–935.
- [127] Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (multidogo): Strategies toward curating and annotating large scale dialogue data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4526–4536.
- [128] C Pfuderer. 1968. A scale of politeness of request forms in English. Term paper for Speech 164A. *University of California, Berkeley. Cited by Ervin-Tripp, S.(1973). Language Acquisition and Communicative Choice, Stanford University Press, Stanford, Calif* (1968).

- [129] Priyanshu Priya, Mauajama Firdaus, and Asif Ekbal. 2023. A multi-task learning framework for politeness and emotion detection in dialogues for mental health counselling and legal aid. *Expert Systems with Applications* 224 (2023), 120025.
- [130] Priyanshu Priya, Kshitij Mishra, Palak Totala, and Asif Ekbal. 2023. PARTNER: A Persuasive Mental Health and Legal Counselling Dialogue System for Women and Children Crime Victims. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, Edith Elkind (Ed.). International Joint Conferences on Artificial Intelligence Organization, 6183–6191. <https://doi.org/10.24963/ijcai.2023/686> AI for Good.
- [131] Dongqi Pu and Vera Demberg. 2023. ChatGPT vs Human-authored Text: Insights into Controllable Text Summarization and Sentence Style Transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, Vishakh Padmakumar, Gisela Vallejo, and Yao Fu (Eds.). Association for Computational Linguistics, Toronto, Canada, 1–18. <https://doi.org/10.18653/v1/2023.acl-srw.1>
- [132] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4755–4764.
- [133] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [134] Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2022. Is Reinforcement Learning (Not) for Natural Language Processing?: Benchmarks, Baselines, and Building Blocks for Natural Language Policy Optimization. *arXiv preprint arXiv:2210.01241*. <https://arxiv.org/abs/2210.01241>
- [135] Kanishk Rana, Rahul Madaan, and Jainendra Shukla. 2021. Effect of Polite Triggers in Chatbot Conversations on User Experience across Gender, Age, and Personality. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 813–819.
- [136] Bernd Renner. 2020. *(In) directness as an (im) politeness strategy in the contact between german and Brazilian portuguese as additional languages*. Ph. D. Dissertation. Universidade de Brasília, Brasília.
- [137] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 300–325.
- [138] Norton Trevisan Roman, Paul Piwek, and AMBR Carvalho. 2004. Politeness and summarization: an exploratory study. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications (AAAI-EAAT 2004)*.
- [139] Norton Trevisan Roman, Paul Piwek, and Ariadne Maria Brito Rizzoni Carvalho. 2006. Politeness and bias in dialogue summarization: Two exploratory studies. In *Computing attitude and affect in text: theory and applications*. Springer, 171–185.
- [140] Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. CounterGeDi: A Controllable Approach to Generate Polite, Detoxified and Emotional Counterspeech. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 5157–5163. <https://doi.org/10.24963/ijcai.2022/716> AI for Good.
- [141] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *NeurIPS EMC² Workshop*.
- [142] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- [143] John R Searle and John Rogers Searle. 1969. *Speech acts: An essay in the philosophy of language*. Vol. 626. Cambridge university press.
- [144] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 35–40.
- [145] Rico Sennrich, Martin Volk, and Gerold Schneider. 2013. Exploiting synergies between open resources for german dependency parsing, pos-tagging, and morphological analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. 601–609.
- [146] Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5263–5276.
- [147] Marilyn Shatz and Rochel Gelman. 1973. The development of communication skills: Modifications in the speech of young children as a function of listener. *Monographs of the society for research in child development* (1973), 1–38.
- [148] Diogo Silva, David Semedo, and João Magalhães. 2022. Polite Task-oriented Dialog Agents: To Generate or to Rewrite?. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*. 304–314.
- [149] Gopendra Vikram Singh, Priyanshu Priya, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. EmoInHindi: A Multi-label Emotion and Intensity Annotated Dataset in Hindi for Emotion Recognition in Dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 5829–5837.
- [150] Mengmeng Song, Huixian Zhang, Xinyu Xing, and Yucong Duan. 2023. Appreciation vs. apology: Research on the influence mechanism of chatbot service recovery based on politeness theory. *Journal of Retailing and Consumer Services* 73 (2023), 103323.
- [151] Darcy Sperlich, Jaiho Leem, and Eui-Jeen Ahn. 2016. The interaction of politeness systems in Korean learners of French. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*. 163–171.
- [152] Anirudh Srinivasan and Eunsol Choi. 2022. TyDiP: A Dataset for Politeness Classification in Nine Typologically Diverse Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 5723–5738.

- [153] David Sweetman and Fred Luthans. 2010. The power of positive psychology: Psychological capital and work engagement. *Work engagement: A handbook of essential theory and research* 54 (2010), 68.
- [154] Mariana Terkourafi. 2001. *Politeness in Cypriot Greek: A frame-based approach*. Ph. D. Dissertation. Citeseer.
- [155] Marina Terkourafi. 2002. Politeness and formulaicity: evidence from Cypriot Greek. *Journal of Greek Linguistics* 3, 1 (2002), 179–201.
- [156] Marina Terkourafi. 2005. Beyond the micro-level in politeness research. *Journal of Politeness Research* 1 (2005), 237262.
- [157] Marina Terkourafi. 2005. Pragmatic correlates of frequency of use: The case for a notion of “minimal context”. *Trends in Linguistics Studies and Monographs* 161 (2005), 209.
- [158] Jenny A Thomas. 2014. *Meaning in interaction: An introduction to pragmatics*. Routledge.
- [159] Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Lrec*, Vol. 2012. Citeseer, 2214–2218.
- [160] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [161] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
- [162] Aditi Viswanathan, Varden Wang, and Antonina Kononova. 2019. Controlling Formality and Style of Machine Translation Output Using AutoML. In *Annual International Symposium on Information Management and Big Data*. Springer, 306–313.
- [163] Carl Vogel. 2015. Some puzzles of politeness and impoliteness within a formal semantics of offensive language. *Conflict and Multimodal Communication: Social Research and Machine Intelligence* (2015), 223–241.
- [164] Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences* 114, 25 (2017), 6521–6526.
- [165] Marilyn A Walker, Janet E Cahn, and Stephen J Whittaker. 1997. Improvising linguistic style: Social and affective bases for agent personality. In *Proceedings of the first international conference on Autonomous agents*. 96–105.
- [166] Xuewei Wang, Weiyang Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5635–5649.
- [167] Yi-Chia Wang, Alexandros Papangelis, Runze Wang, Zhaleh Feizollahi, Gokhan Tur, and Robert Kraut. 2020. Can You be More Social? Injecting Politeness and Positivity into Task-Oriented Conversational Agents. *arXiv preprint arXiv:2012.14653* (2020).
- [168] Ronald Wardhaugh. 2006. An Introduction to Sociolinguistics.
- [169] Zeerak Waseem, Thomas Davidson, Dana Warnsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*. 78–84.
- [170] Richard J Watts. 2003. *Politeness*. Cambridge University Press.
- [171] Richard J Watts. 2005. 2. Linguistic politeness and politic verbal behaviour: Reconsidering claims for universality. *Politeness in Language Studies in its History, Theory and Practice* (2005), 43–69.
- [172] World Health Organization WHO. 2022. *Mental disorders*. <https://www.who.int/news-room/fact-sheets/detail/mental-disorders>
- [173] Jason D Williams, Matthew Henderson, Antoine Raux, Blaise Thomson, Alan Black, and Deepak Ramachandran. 2014. The dialog state tracking challenge series. *AI Magazine* 35, 4 (2014), 121–124.
- [174] Michael Yeomans, Alejandro Kantor, and Dustin Tingley. 2018. The politeness Package: Detecting Politeness in Natural Language. *R Journal* 10, 2 (2018).
- [175] Michael Yeomans, Julia Minson, Hanne Collins, Frances Chen, and Francesca Gino. 2020. Conversational receptiveness: Improving engagement with opposing views. *Organizational Behavior and Human Decision Processes* 160 (2020), 131–148.
- [176] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation* 31, 7 (2019), 1235–1270.
- [177] George Yule. 2020. *The study of language*. Cambridge university press.
- [178] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1350–1361.
- [179] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 270–278.
- [180] Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is ChatGPT Equipped with Emotional Dialogue Capabilities? *arXiv preprint arXiv:2304.09582* (2023).
- [181] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
- [182] Naitian Zhou and David Jurgens. 2020. Condolence and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 609–626.
- [183] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can Large Language Models Transform Computational Social Science? *arXiv preprint arXiv:2305.03514* (2023).

Received 4 August 2022; revised 27 February 2024; accepted 10 March 2024