Wanastra: Jurnal Bahasa dan Sastra
Volume 18 No. 1 March 2026
P-ISSN 2086-6151 E-ISSN 2579-3438
https://doi.org/10.31294/wanastra.v18i1

Politeness and Indirectness: When Sexism Hides Behind Advice in Workplace Statements

Annisa Romadloni[1], Linda Perdana Wanti[2], Laura Sari[3]

Politeknik Negeri Cilacap, Indonesia
email: linda_perdana@pnc.ac.id, laurasari@pnc.ac.id
e-mail corresponding author: annisa.romadloni@pnc.ac.id

Received : 14-02-2026
Revised : 06-03-2026
Accepted : 31-03-2026

Abstrak *– Seksisme di tempat kerja sering kali dibingkai sebagai arahan atau evaluasi yang tampak wajar, alih-alih sebagai permusuhan yang terang-terangan, sehingga penyaring berbasis sentimen dan toksisitas kerap gagal mendeteksinya. Analisis korpus dilakukan menggunakan dataset Sexist Workplace Statements (1.142 pernyataan; 627 berlabel seksis). Sebuah operasionalisasi berbasis pragmatik diterapkan untuk mengklasifikasikan pernyataan seksis sebagai seksisme benevolen atau seksisme hostil, serta untuk memberi label tindak tutur utama setiap pernyataan sebagai nasihat, evaluasi, hinaan, lelucon, atau keluhan. Seksisme benevolen diperkirakan mencakup 73,8% dari seluruh pernyataan seksis, sedangkan seksisme hostil mencakup 26,2%. Seksisme benevolen terutama terkonsentrasi pada tindak tutur evaluasi dan nasihat, sementara seksisme hostil paling banyak muncul dalam bentuk hinaan. Proksi toksisitas berbasis sentimen atau kata-kata kasar menunjukkan presisi tinggi tetapi recall rendah untuk deteksi seksisme, karena mampu menangkap sebagian besar seksisme hostil namun melewatkan sebagian besar seksisme benevolen. Model dasar terawasi (TF–IDF dengan regresi logistik) memberikan kinerja yang baik pada label biner, tetapi tetap menghasilkan false negative yang didominasi oleh evaluasi benevolen. Temuan ini ditafsirkan melalui teori seksisme ambivalen, teori tindak tutur, dan teori kesantunan, yang menegaskan bahwa ketidaklangsungan dan pengelolaan muka (face-work) memungkinkan norma diskriminatif disebarluaskan dengan kedok "bantuan" atau "nasihat". Hasil ini menegaskan bahwa sistem deteksi seksisme perlu memasukkan fitur yang peka terhadap pragmatik dan tindak tutur agar dapat secara andal mengidentifikasi seksisme di tempat kerja yang bersifat benevolen dan dibingkai sebagai "bantuan", yang secara sistematis luput dari sinyal sentimen/toksisitas standar.*

Kata Kunci: benevolent sexism; hostile sexism; wacana tempat kerja; tindak tutur; analisis sentimen

Abstract *- Sexism in workplaces is often framed as ordinary guidance or evaluation rather than as overt hostility, which can cause sentiment and toxicity filters to miss it. A corpus analysis was conducted using the Sexist Workplace Statements dataset (1,142 statements; 627 labeled sexist). A pragmatics-informed operationalization was applied to classify sexist statements as benevolent or hostile and to label each statement's primary speech act as advice, evaluation, insult, joke, or complaint. Benevolent sexism was estimated to constitute 73.8% of sexist statements, while hostile sexism constituted 26.2%. Benevolent sexism was concentrated in evaluation and advice, whereas hostile sexism was concentrated in insults. A sentiment-or-profanity toxicity proxy achieved high precision but low recall for sexism, capturing most hostile sexism while missing most benevolent sexism. A supervised baseline (TF–IDF plus logistic regression) performed well on the binary label but still showed false negatives dominated by benevolent evaluations. The findings were interpreted through ambivalent sexism theory, speech act theory, and politeness theory, highlighting how indirectness and face-work enable discriminatory norms to be advanced under the guise of help. These results make explicit that sexism detection systems should incorporate pragmatics- and speech-act-aware features to reliably identify benevolent, "helpful"-framed workplace sexism that standard sentiment/toxicity signals systematically overlook.*

Keywords: benevolent sexism; hostile sexism; workplace discourse; speech acts; sentiment analysis

INTRODUCTION

Workplace sexism has continued to be experienced even in organizations that have adopted formal equality policies and publicized diversity targets (Cortina, 2008; Kaiser et al., 2013). In everyday interaction, discriminatory meaning is often conveyed through brief remarks that appear mundane, humorous, or supportive,

yet gender hierarchy is reproduced by indexing women as less competent, less rational, or more naturally suited to subordinate roles (Basford et al., 2014; Ford & Ferguson, 2004; Glick & Fiske, 1996; Heilman, 2012). The harms are amplified in workplaces because relationships are ongoing, power asymmetries are formalized, and the costs of confronting bias can be high (Kaiser & Miller, 2001, 2004; Rudman & Phelan, 2008). As a result, many sexist acts are not packaged as open conflict, but are instead embedded in routine talk that can be dismissed as normal workplace communication (Cortina, 2008; West & Zimmerman, 1987). When discriminatory stances are advanced through remarks that appear to provide guidance or assessment, they can be normalized through repetition and shape expectations about who should lead, who should speak, and who should be "supported" rather than empowered (Heilman, 2012).

The development of automated tools for detecting discriminatory language has been motivated by this persistence and by the practical limits of manual review (Schmidt & Wiegand, 2017). In many applied deployments, messages are screened by sentiment analysis and toxicity detection, and attention is directed to content that is strongly negative, aggressive, or profane (Pang & Lee, 2008; Pavlopoulos et al., 2017). This operationalization has been convenient because profanity and explicit threats are salient and often actionable, and because many moderation datasets have been constructed around overt harassment or easily identifiable abusive forms (Davidson et al., 2017; Founta et al., 2018; Nobata et al., 2016; Waseem & Hovy, 2016). Yet, an important mismatch has been created when sexism is equated with negativity. If the most consequential workplace sexism is packaged as advice or praise, then a system tuned to anger and profanity will be structurally misaligned with the forms that matter most in institutional settings —consistent with research on benevolent sexism showing that subjectively "positive," paternalistic talk can reproduce gender hierarchy (Dardenne et al., 2007; Glick & Fiske, 1996). Under these conditions, apparent civility can be mistaken for safety, while discriminatory norms continue to be reproduced through indirect, polite interaction (Gilda et al., 2022; Nobata et al., 2016).

In toxicity research, the limits of surface cues have been clarified through the role of context and interpretation. It has been shown that some items receive opposite toxicity labels when annotators are not provided with conversational context, and that adding limited context does not necessarily improve classifier performance. (Pavlopoulos et al., 2020). This evidence has supported the claim that toxicity is partly an inference about intent, target, and conversational positioning rather than a stable lexical property. A similar argument has been required for sexism, because sexist meaning is often communicated through presupposition and normativity rather than through explicit abuse. If a statement implies that women do not belong in a role, or that women should behave in a gender-typed way, the sexist force may be present even when the surface tone is neutral or ostensibly kind. In social psychology, the same blind spot has been theorized through the construct of ambivalent sexism, in which hostile sexism is distinguished from benevolent sexism. Benevolent sexism is expressed through protective paternalism, complementary gender differentiation, and seemingly positive evaluations that reward conformity to traditional roles, while hostile sexism is expressed through antagonism, derogation, and explicit exclusion. Workplace-oriented measurement has continued to refine how benevolent sexism is experienced in organizational life and how it differs from overt antagonism (Warren et al., 2023). Benevolent sexism has not been treated as harmless because it has been associated with constrained agency and with the reinforcement of gender hierarchy through "help" that narrows legitimate options. Recent outcome research has also linked benevolent sexism to career-related harms, indicating that polite sexism can carry structural costs even when it is not recognized as abuse (Song & Chang, 2025).

In computational research on sexism detection, rapid progress has been enabled by the release of datasets, shared tasks, and fine-grained taxonomies. Multi-label and multi-task approaches have been proposed to move beyond binary labels and to capture varieties of sexism, such as stereotypes, objectification, threats, and moral policing (Abburi et al., 2024). Shared evaluations have further been used to motivate explainability and category-level modeling, as in the SemEval Explainable Detection of Online Sexism task (Kirk et al., 2023). This trajectory has improved measurement and has highlighted that sexism is heterogeneous. However, the pragmatic form through which sexism is delivered has often been underspecified. A stereotype can be delivered as an insult, a joke, a complaint, an evaluation, or advice, and these forms differ in perceived intent, deniability, and institutional legitimacy.

The broader literature on abusive language has provided complementary explanations for why indirect forms remain difficult. In an in-depth analysis of implicit and subtle hate speech, it has been argued that indirect hate can be as harmful as explicit hate, while being harder to detect because it is expressed through circumlocution, metaphor, sarcasm, and strategic ambiguity (Ocampo et al., 2023). Theory-driven sexism detection has similarly suggested that models learn narrow artifacts and fail to generalize to out-of-domain subtle sexism, especially when training data overrepresents overt slurs and hostility (Samory et al., 2020). These findings suggest that a pragmatic account is required if detection is to align with workplace realities in which open aggression can be institutionally

risky. The Sexist Workplace Statements dataset is particularly suitable because it was curated to reflect workplace discourse constraints rather than anonymous online harassment. More than 1,100 short statements were included and were labeled as sexist, ambiguous, or neutral, with labels being provided as 1 and 0, respectively. The corpus was assembled from several sources, including a manually filtered subset of tweets, a set of work-related quotes, miscellaneous press quotations, and faculty or student submissions. The mixture was intended to diversify phrasing while limiting overfitting to Twitter-specific artifacts. Manual preprocessing was reported to have removed duplicates, generalized rare named entities, rewritten or removed hashtags, and converted casual slang into more formal forms, while a large portion of the dataset was described as generic tweets of benevolent sexism (Grosz & Conde-Céspedes, 2020).

Two forms of novelty were pursued. Conceptually, politeness and indirectness were treated as mechanisms through which benevolent sexism is normalized in institutional interaction, rather than as stylistic noise to be abstracted away. Methodologically, a pragmatic annotation layer for speech acts was introduced and jointly modeled with sexism, enabling the distinction of advice, evaluation, and humor from insult and complaint. This joint perspective was intended to complement recent fine-grained sexism benchmarks that emphasize explainability and category structure, while retaining a pragmatic focus on interactional packaging (Kirk et al., 2023). By foregrounding speech acts, sexism detection was reframed as a pragmatic problem of what is accomplished, thereby enabling systematic analyses of why apparently polite language can remain discriminatory even when negativity and profanity are absent.

This study addresses two research questions. First, what proportion of workplace-oriented sexist statements are packaged as benevolent versus hostile sexism? Secondly, how are these two forms distributed across pragmatic speech acts—advice, evaluation, insult, joke, and complaint and what does this distribution imply for the visibility of sexism to sentiment/toxicity proxies and to a supervised baseline classifier? The research contributes (1) a transparent, pragmatics-informed operationalization that links ambivalent sexism (benevolent/hostile) to interactional packaging via speech-act labels; (2) corpus-level prevalence estimates and cross-tabulations that locate sexism in institutionally legitimate workplace actions such as evaluation and advice, where bias can be rationalized as “professional feedback” or “help”; and (3) a diagnostic evaluation of common screening heuristics and a standard TF–IDF + logistic regression baseline, including error analysis by speech act and sexism type, to show why civility- and negativity-centered monitoring can miss the most prevalent workplace-relevant sexism and to motivate more informative, action-oriented labeling.

Three complementary lenses were used to frame how sexism can be hidden in apparently polite workplace discourse: ambivalent sexism theory, speech act theory, and politeness theory. Each lens was treated as describing a different level of organization. Ambivalent sexism theory was treated as a psychological account of attitudes and social rewards; speech act theory as a pragmatic account of actions performed with language; and politeness theory as an interactional account of how those actions are packaged to manage face, authority, and accountability. Ambivalent sexism theory distinguishes hostile sexism from benevolent sexism. Hostile sexism is typically expressed through antagonism and derogation, often targeting women who are perceived as challenging male dominance. Benevolent sexism is typically expressed through chivalry, protective paternalism, and complementary gender differentiation, and it can appear affectively positive while sustaining hierarchy by presenting women as needing protection or as being naturally suited to particular roles. This distinction has remained relevant in workplace scholarship because benevolent sexism can be embedded in organizational routines and interpersonal expectations. A workplace-specific benevolent sexism scale has been proposed to capture beliefs that are enacted through everyday organizational interactions rather than through overt hostility (Warren et al., 2023). Outcome-focused work has also linked benevolent sexism to career-related harms, suggesting that “polite” sexism should be treated as an organizational risk rather than as interpersonal awkwardness (Song & Chang, 2025).

Speech act theory was used to connect this psychological distinction to linguistic action. From this perspective, an utterance is not only a proposition but also an act accomplished in context, such as advising, evaluating, joking, complaining, or insulting. The same stereotype theme can therefore be delivered through different actions, each with different affordances for sanction and denial. Advice was treated as central because it presupposes epistemic or institutional authority and can be framed as supportive while still positioning the recipient as deficient. In workplace settings, advice can be licensed by role relations and professional coaching norms, making gendered prescriptions especially easy to naturalize. Evaluations were treated as equally important because they allocate legitimacy, competence, and role fit, and they can function as gatekeeping acts when gendered standards are applied. Insults and complaints, by contrast, were treated as direct antagonistic acts that more closely align with hostile sexism.

Politeness theory was used to specify how the same sexist content can be made interactionally safe. Workplace discourse is constrained by norms of professionalism, collegiality, and conflict avoidance, which can discourage overt aggression while encouraging indirectness. Face-threatening acts can be mitigated through hedges, indirect formulations, honorifics, and positive-politeness strategies that emphasize solidarity. Indirectness is especially consequential because responsibility for meaning can be shared between the speaker and the hearer. When a discriminatory stance is expressed through implication or presupposition, the recipient may be compelled to infer sexist content, and objections can be reframed as misinterpretation or overreaction. This mechanism has been linked to plausible deniability, in which the speaker's intent is difficult to prove because the discriminatory stance is packaged as a reasonable suggestion or harmless commentary.

These pragmatic claims are consistent with recent computational evidence that discrimination cannot be reduced to general negativity. Fine-grained sexism benchmarks have shown that sexist content can resemble ordinary conversation and can lack profanities, motivating the need for explainable category structures rather than only binary flags (Kirk et al., 2023). Work on implicit and subtle hate has similarly argued that indirect hate can be as harmful as explicit hate, while being harder to detect because overt lexical triggers are scarce (Ocampo et al., 2023). Theory-driven approaches have therefore been proposed, including the use of psychological scales as codebooks to stress-test sexism detectors and to highlight the fragility of models under domain shift (Samory et al., 2020).

RESEARCH METHODOLOGY

A mixed-methods design was adopted, combining corpus annotation, computational modeling, and qualitative discourse analysis to examine how sexism is packaged through polite and indirect speech acts. Quantitative corpus statistics were first produced to estimate the prevalence of benevolent versus hostile sexism and to map sexism onto speech-act types. A pragmatic annotation layer was then applied to a subset of the corpus, and joint modeling was performed to examine sexism and speech-act labels together rather than in isolation. Finally, a qualitative reading of representative statements was conducted to interpret the quantitative patterns through politeness and indirectness mechanisms. The design was therefore sequential and integrative: corpus statistics established broad distributional patterns, supervised models tested the visibility of those patterns to common NLP approaches, and qualitative analysis explained how the patterns functioned pragmatically. Participants and units of analysis were defined in textual rather than human-subject terms. No new human participants were recruited because a publicly available corpus was analyzed. The unit of analysis was the individual statement, which was treated as a proxy for workplace-directed utterances that might be encountered in professional communication, workplace training materials, or public discourse about workplace roles. The inclusion of short, context-poor utterances was treated as methodologically useful because it isolates linguistic packaging, but it was also treated as a limitation because conversational context, speaker identity, and institutional role relations are not directly observable.

Data were obtained from the Sexist Workplace Statements dataset distributed on Kaggle and mirrored in open repositories associated with the dataset's original publication (Grosz & Conde-Céspedes, 2020). The dataset contained 1,142 statements. Each statement was labeled as sexist (label 1) or ambiguous or neutral (label 0) in the source dataset. In the present analysis, 627 statements were labeled as sexist, and 515 as ambiguous or neutral, yielding a class balance consistent with the original dataset description. In the dataset construction report, the corpus was assembled from multiple sources, including a filtered subset of tweets, work-related quotes, and additional quotations and submissions, and manual preprocessing was used to remove duplicates and Twitter-specific artifacts.

Preprocessing was kept intentionally conservative to preserve pragmatic markers that are frequently removed in aggressive normalization. Statements were lowercased for lexicon-based analyses while original casing was retained for supervised modeling. Punctuation was retained because exclamation points, quotation marks, and question marks can contribute to pragmatic force and stance. No stemming or lemmatization was applied because morphological reduction can remove modality markers and hedges that are relevant to advice and mitigation. For the supervised models, features were extracted directly from the raw text using a TF–IDF vectorizer over unigrams and bigrams, with low-frequency features pruned by a minimum document frequency threshold. A pragmatic annotation scheme was defined for two dimensions: speech act and sexism type. Five speech-act categories were used: advice, evaluation, insult, complaint, and joke. Advice was defined as an utterance that recommended a course of action or prescribed a norm, and it was expected to be signaled by deontic modals such as should, must, need to, have to, and by imperative constructions. Evaluation was defined as an utterance that assessed a person or group in terms of competence, temperament, or role suitability, often using copular constructions or comparative frames. Insult was defined as an utterance that directly attacked a target through derogatory labels, profanity, or explicit demeaning descriptors. A complaint was defined as an utterance that expressed grievance or

frustration about women or about gendered practices, frequently in interrogative or lament formats. A joke was defined as an utterance that framed gendered meaning through humor, wordplay, sarcasm, or explicit kidding markers.

Sexist attitudes were operationalized using the ambivalent sexism distinction. Hostile sexism was defined as sexism expressed through overt derogation, exclusion, or antagonistic stereotyping, and it was expected to correlate with profanities, slurs, and negative affect. Benevolent sexism was defined as sexism expressed through apparently positive, protective, or advisory framing that reinforces traditional gender roles, and it was expected to correlate with prescriptive content and politeness. Because the original dataset did not include these labels, operational rules were applied to approximate them. Hostile sexism was assigned when strong derogatory lexemes, profanity, explicit exclusion statements, or strongly negative sentiment co-occurred with gendered references. Benevolent sexism was assigned otherwise within sexist-labeled items, especially when gendered content was packaged as advice or a compliment-like evaluation. Items labeled 0 in the source dataset were treated as non-sexist for the purposes of this operationalization.

A stratified subset of 400 statements from the training partition was labeled with speech acts and treated as an annotated dataset for supervised speech-act modeling. Stratification was performed to ensure representation of both sexist and non-sexist items and to reduce the risk that the classifier would collapse into a single dominant class. On this subset, speech act labels were produced using the operational definitions above, implemented as transparent pattern rules. These labels were treated as a first-pass annotation layer to enable reproducible measurement and modeling; the absence of multi-annotator human reliability was treated as a limitation and is addressed in the Recommendations section.

To support coding reliability, the annotation categories were defined in advance through explicit decision rules and were applied consistently across the corpus subset using a transparent rule-based procedure rather than ad hoc judgment. The speech-act and sexism-type definitions were specified before modeling, and the same operational criteria were used throughout annotation and analysis. At the same time, reliability should be interpreted cautiously: because the pragmatic layer was not independently annotated by multiple human coders, the labels function as a reproducible first-pass analytical layer.

Three quantitative analytic procedures were conducted. First, descriptive statistics for the corpus were computed. Counts and proportions were derived for the original sexism label, the operationalized sexism-type label, and the speech-act label. Cross-tabulation was used to measure how speech acts and sexism types co-occurred, and a heat map visualization was produced to facilitate interpretation. Wilson score intervals were computed for key proportions to quantify uncertainty in binomial prevalence estimates. Second, the relationship between the type of sexism and surface affect cues was assessed through sentiment and profanity analyses. Sentiment polarity was computed using TextBlob polarity scores, which range from −1 to +1. Although sentiment analysis is not designed to detect discrimination, it was used here as a baseline signal because many applied monitoring systems rely on affect indicators. Profanity frequency was estimated through a conservative lexicon-based counter that included common English profanities and gendered slurs. This counter was treated as a proxy for coarse toxicity cues that are commonly used in rule-based filtering. Third, supervised modeling was conducted to illustrate how speech-act information can be coupled with sexism detection. A baseline sexism classifier was trained to predict the original dataset label using TF–IDF features and a linear logistic regression classifier. Model performance was evaluated on a held-out 20% test set using precision, recall, F1 score, and accuracy. To provide a contrast with common filtering heuristics, a simple toxicity proxy was implemented. A statement was flagged when sentiment polarity fell below −0.10 or when profanity count was greater than zero, and this flag was evaluated as a sexism detector against the source labels.

Joint modeling was explored in two ways. First, combined pragmatic labels were created by concatenating the sexism category and the speech act, producing labels such as benevolent advice and hostile insult. A linear support vector machine with TF–IDF features was trained to predict these combined labels, and macro-averaged F1 was computed because class imbalance was present in the combined space. Second, speech-act predictions from the speech-act classifier were appended as an additional token to each statement, and the resulting augmented texts were used to train a sexism classifier. This variant was used to illustrate how pragmatic supervision can be incorporated into a standard text classifier, even when speech-act labels are available only for a subset of cases. To complement quantitative analyses, a qualitative discourse-pragmatic analysis was conducted on a purposive sample of statements selected from each speech-act category and from both sexism types. Close reading was used to identify politeness strategies such as hedging, mitigation, chivalric framing, and role-prescriptive modality, and these strategies were related to ambivalent sexism constructs. This qualitative layer was used to explain how sexist meaning can be advanced while maintaining plausible deniability and while aligning with workplace norms of

professionalism.

Ethical considerations were addressed because sexist language was necessarily processed and is reproduced in limited form for analysis. The dataset was publicly available and contained decontextualized short statements rather than personal identifiers. Nonetheless, exposure to sexist and profane content was treated as a risk to readers. Examples were therefore used sparingly and were selected to illustrate analytic mechanisms rather than to sensationalize content. Because workplace monitoring can introduce surveillance and fairness risks, the modeling results were interpreted as diagnostic evidence about failure modes rather than as prescriptions for punitive automated decision-making.

All computations were performed in Python using standard libraries for data handling, visualization, and machine learning. Random seeds were fixed for train–test splitting to support reproducibility. For feature extraction, unigram and bigram TF–IDF features were used, because short statements often encode stereotypes through multiword expressions and modifiers. Low-frequency features were pruned to reduce sparsity, and model hyperparameters were kept close to defaults to emphasize interpretability. For the logistic regression classifier, default regularization was used with a sufficiently high iteration limit to ensure convergence, and class weights were not adjusted because the source labels were moderately balanced. For the linear support vector machine, the implementation implicitly used a one-versus-rest strategy, and performance was reported using macro-averaged F1 to mitigate dominance by frequent pragmatic labels. Error analysis was conducted by inspecting false negatives and false positives and by aggregating their speech-act profiles. In this way, modeling was used as an analytical instrument rather than an end in itself, and the focus remained on the relationship between pragmatic packaging and detection visibility.

RESULT AND DISCUSSION

To answer the research questions, the first part of this study reports the composition of the corpus and estimates the prevalence of benevolent versus hostile sexism within the sexist subset, thereby establishing how frequently sexism appears in superficially civil forms. The second part maps types of sexism onto speech acts (advice, evaluation, insult, joke, complaint) to identify where sexism concentrates in interaction—especially in evaluation and advice, which are normatively licensed in workplaces and therefore afford greater plausible deniability. This section then examines the consequences of that packaging for detection by comparing a sentiment/profanity-based toxicity proxy with a supervised baseline classifier, and by interpreting their failure modes through politeness and indirectness mechanisms.

The analysis was conducted on a corpus of 1,142 short workplace-oriented statements drawn from the Sexist Workplace Statements dataset. The source annotation provided a binary label indicating whether sexism was present, and an additional analytical layer was applied to address the central research question in pragmatic terms. Within the sexist subset, statements were operationally differentiated into benevolent and hostile sexism, and each statement was assigned a dominant speech-act category: advice, evaluation, insult, complaint, or joke. This combined approach was intended to reveal not only how frequently sexism appeared but also how it was interactionally packaged, particularly when it was framed as “helpful guidance” or other professionally sounding discourse.

Table 1. Corpus composition and original labels

Measure	Count	Percentage
Total statements	1,142	100.0%
Sexist (label 1)	627	54.9%
Ambiguous or neutral (label 0)	515	45.1%

Table 1 shows that the dataset contained a substantial proportion of sexist statements, which supported robust comparisons between sexist and non-sexist language. The class balance also suggested that the corpus was not dominated by one label, making it suitable for examining how sexist meaning could be realized across different pragmatic forms rather than only through extreme harassment. When the 627 sexist-labeled statements were examined, benevolent sexism was estimated to account for 463 statements and hostile sexism for 164 statements. Benevolent sexism, therefore, constituted 73.8% of sexist items, while hostile sexism constituted 26.2%. This distribution indicated that sexism in this workplace-oriented corpus was more commonly expressed through superficially civil forms than through overt antagonism. The implication was that discriminatory meaning was frequently advanced through talk that could plausibly be framed as supportive, constructive, or “just being honest,” rather than through language that looked unambiguously abusive. The pragmatic mapping to speech acts clarified how this imbalance was realized and why “helpful guidance” alone could not be treated as the sole vehicle

of polite sexism. Although advice was an important channel, evaluative talk emerged as the dominant pragmatic route through which benevolent sexism was expressed, reflecting how workplace discourse often legitimizes judgment and role-assignment through assessment. The distribution of speech acts within sexist statements is summarized in Table 2.

Verification of key proportions: within the sexist subset (n = 627), benevolent sexism = 463/627 = 0.738 (73.8%) and hostile sexism = 164/627 = 0.262 (26.2%); 463 + 164 = 627, and the percentages sum to 100.0% (subject to rounding).

Table 2. Speech act distribution within sexist statements by sexism type

Speech act	Benevolent sexism (count)	Hostile sexism (count)	Benevolent share within speech act
Advice	71	19	78.9%
Joke	9	3	75.0%
Insult	5	72	6.5%
Complaint	0	13	0.0%
Evaluation	378	57	86.9%

Table 2 indicates that benevolent sexism was overwhelmingly concentrated in the areas of evaluation and advice. Evaluations accounted for the largest portion of benevolent sexism, implying that sexism was often delivered through "objective-sounding" judgments about competence, suitability, temperament, or professional fit, rather than being confined to explicit instruction. Hostile sexism, in contrast, was concentrated in insults and complaints, where overt derogation or grievance framing made antagonism easier to recognize. The same co-occurrence pattern can be seen more intuitively in the heat map in Figure 1, where concentration by cell is visually emphasized.

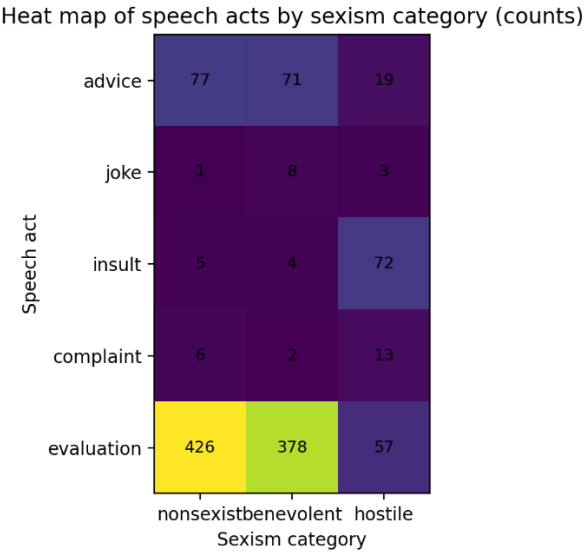

Figure 1. Heatmap of Speech Act by Sexism Category

Figure 1 makes it clear that sexism was not distributed evenly across pragmatic forms. The densest region was located in benevolent evaluation, while hostile sexism was most strongly concentrated in insults. This pattern supports the interpretation that workplace-relevant sexism is often aligned with institutionally legitimate actions such as assessing and advising, which can reduce the likelihood that the sexism will appear as a clear violation of civility norms.

The consequences of this packaging became evident when surface affect cues were examined. A core premise of the study was that many applied monitoring pipelines over-rely on sentiment and coarse toxicity indicators, including negativity and profanity. When sentiment polarity and profanity frequency were inspected, hostile sexism tended to co-occur with negative affect and explicit lexical triggers, while benevolent sexism tended to remain near neutral or slightly positive, especially in advice and evaluation. This contrast was visualized in Figure 2 as mean sentiment polarity by speech act and sexism category.

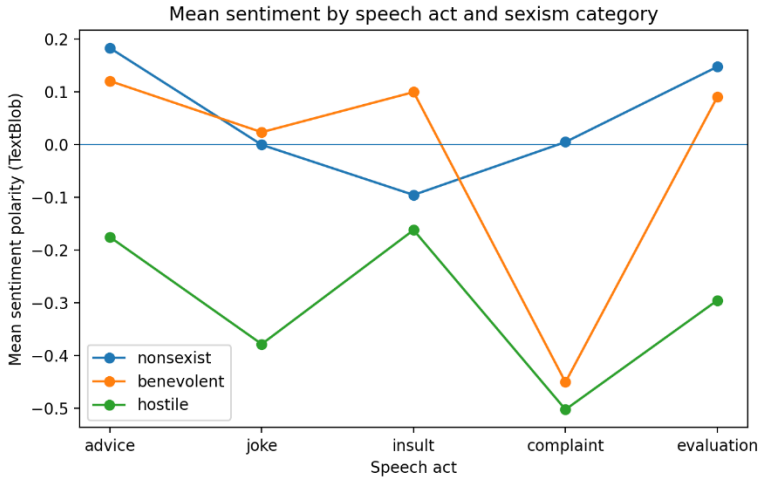

Figure 2. Mean Sentiment by Speech Act and Sexism Category

Figure 2 suggests that hostile sexism was more easily aligned with negativity because insulting and complaint-oriented acts were more likely to be expressed with an overt negative stance. Benevolent sexism, however, was frequently expressed without strong negative affect, consistent with the idea that discriminatory content can be advanced through professional-sounding guidance and assessment. This divergence supports the claim that the tone of an utterance cannot be treated as a reliable proxy for whether sexism is present, particularly in contexts where politeness and indirectness are institutionally rewarded.

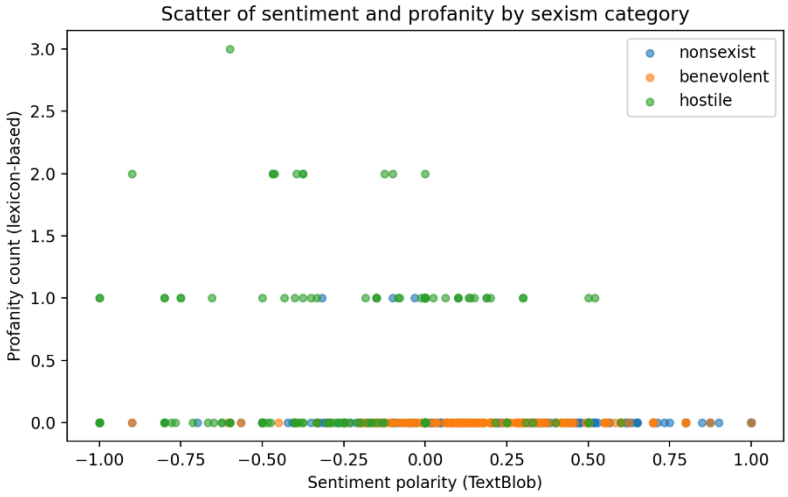

Figure 3. Scatter of Sentiment and Profanity by Sexism Category

Figure 3 shows that hostile sexism clustered toward more negative sentiment and higher profanity counts, while benevolent sexism clustered toward neutral or positive sentiment with minimal profanity. This separation implies that a filtering approach centered on negativity or profanity will behave as a hostile-sexism detector, leaving most benevolent sexism unflagged. The structural issue is not merely that benevolent sexism is "milder," but that it is produced in a different pragmatic register that does not require overt negative affect to exert discriminatory force.

This misalignment was quantified through a toxicity-style proxy that flagged a statement when sentiment polarity fell below a negative threshold or when profanity was detected. The proxy was then evaluated as a sexism detector on the dataset's sexism labels, and the results are summarized in Table 3.

Table 3. Performance of a sentiment or profanity toxicity proxy for detecting sexism (binary)

Metric	Value
Precision	0.768
Recall	0.306

F1 score	0.438
Accuracy	0.568
Recall on the benevolent sexism subset	0.117
Recall on the hostile sexism subset	0.841

Table 3 shows that the proxy achieved relatively high precision but very low recall, meaning that many flagged items were sexist, but most sexist items were missed. The disparity across sexism types was especially consequential: hostile sexism was retrieved at a high rate, while benevolent sexism was retrieved at a very low rate. Because benevolent sexism dominated the sexist subset, the overall recall collapsed even though hostile items were often caught. This pattern provides direct evidence for the study's central argument that sentiment and toxicity filtering fail in workplace-relevant settings because sexism is frequently delivered through polite and indirect forms that do not look toxic.

From a politeness perspective, benevolent advice can be treated as a face-threatening act, softened by positive politeness framing. Advice is inherently asymmetrical because it presupposes that the adviser has superior knowledge, experience, or authority. When gendered advice is offered in workplace contexts, this asymmetry can be masked by affiliative moves, such as the projection of care, the use of mitigators, or the suggestion that the advice is being offered "for your own good." A statement can therefore be interpreted as supportive while still presupposing that a woman is less competent, less suited for leadership, or more appropriately oriented toward supportive roles. This mechanism aligns with ambivalent sexism's protective paternalism, where protection functions as a constraint.

Advice also invites plausible deniability through its conventional function. Because advice is expected to be relevant, helpful, and oriented toward the recipient's welfare, discriminatory content can be reframed as pragmatic relevance rather than bias. When gender stereotypes are embedded as background assumptions for "guidance," the discriminatory force can be displaced from the speaker to the inferred norm. Objections can then be interpreted as rejecting advice rather than as challenging discrimination, thereby increasing the interpersonal cost of confrontation. In institutional settings, this cost is further intensified because advice is often tied to evaluation and promotion, making resistance appear uncooperative. Humor played a smaller but still theoretically revealing role. Jokes constituted a small fraction of the corpus, yet when they occurred, they often combined affiliative framing with sexist content. The "just kidding" format serves to distance the speaker from accountability while testing social boundaries. When sexist content is placed inside a joke, objections can be reframed as humorlessness or lack of team spirit, which is especially costly in workplaces where belonging matters for access to projects and mentorship. Humor, therefore, functions as another pragmatic route through which benevolent and hostile meanings can be blended, preserving ambiguity.

The error analysis of the supervised classifier further reinforced the core argument, even though the model substantially outperformed the toxicity proxy. False negatives were disproportionately benevolent, which suggests that subtle sexism remains a challenge even for supervised models trained on the dataset's binary labels. This pattern is consistent with broader evidence that discrimination detection can suffer from domain shift and over-reliance on artifacts. Reviews of hate and abusive language detection have emphasized that datasets often amplify explicit lexical cues, shaping models that generalize poorly to subtle and institutionally constrained language (Ramos et al., 2024). The present results suggest that workplace sexism is a particularly likely domain for such brittleness, because institutional norms discourage overt insults while leaving prescriptive and evaluative stereotyping intact.

The joint-label modeling results were therefore interpreted less as a competition for higher accuracy and more as a diagnostic tool. When combined labels were predicted, the model was forced to discriminate between benevolent and neutral evaluations, or between benevolent and neutral advice, rather than only between sexist and non-sexist. This can be treated as a pragmatic form of explanation, because the predicted speech act provides a hypothesis about how sexism is being performed. Such structured outputs can support different workflows: hostile insults can be escalated as harassment, while benevolent advice can be routed to bias coaching or policy clarification. The present interpretation also suggests a bridge between pragmatics-informed analysis and the recent shift toward explainable, fine-grained sexism benchmarks. In EDOS, sexism has been decomposed into categories and explanations, which has encouraged systems to predict not only whether sexism is present but also what form it takes (Kirk et al., 2023). However, benevolent sexism is not always separable as a content category, because it can be realized through the same stereotype themes that appear in hostile contexts. What differentiates benevolent packaging is often the speech act and the politeness strategy, such as whether a stereotype is delivered as an objective evaluation or as caring advice. This claim is compatible with proposals that psychological instruments

can be repurposed as theory-driven codebooks for computational sexism detection, because those instruments highlight how endorsement can be expressed through seemingly positive formulations (Samory et al., 2020). It is also compatible with broader reviews that emphasize that abuse and hate datasets often amplify explicit lexical cues, thereby shaping models that generalize poorly to subtle, institutionally constrained language (Ramos et al., 2024). In practical terms, a joint prediction of speech act and sexism can therefore be treated as a form of structured explanation: when an utterance is flagged, it can be indicated whether the risk arises from a hostile insult, a hostile complaint, benevolent advice, or benevolent evaluation. Such a structure can support different organizational responses, because overt harassment requires enforcement and protection, while advice-like sexism may require coaching, policy clarification, or bias-focused training that targets the presupposed norms rather than the tone alone.

Several qualitative patterns were illustrated, showing how politeness and indirectness interact with the type of sexism. In benevolent evaluations, gendered comparison was frequently used, such as "for a woman, that is good," in which a compliment was framed as conditional on gender and thereby reproduced the assumption that women are typically less capable. In benevolent advice, deontic modality was used to prescribe gendered norms, such as directing women toward supportive roles or toward "appropriate" self-presentation. In hostile insults, derogation was explicit and direct, and profanity, when present, served to intensify the act rather than to create ambiguity. These patterns illustrate that sexism is not only what is being said about women, but how the social action is packaged to manage face and accountability.

The findings also speak to the institutional dimension of politeness. Politeness is often treated as a cooperative norm that reduces conflict, yet it can also serve as a governance technology within organizations. When advice and evaluation are delivered politely, a veneer of professionalism can be maintained while discriminatory norms are enforced. This mechanism can be amplified by managerial discourse that valorizes "constructive feedback" and "coaching," because these categories can be used to rationalize the asymmetry through which gendered prescriptions are imposed. Under this view, benevolent sexism is not an interpersonal accident; it is a stable discourse pattern that fits institutional norms.

CONCLUSION

Sexism in workplace-oriented language was found to be more often packaged as benevolent evaluation and advice than as hostile insult. In the analyzed corpus of 1,142 statements, benevolent sexism constituted nearly three-quarters of sexist items, while hostile sexism constituted about one-quarter. This distribution mattered because benevolent sexism was weakly signaled by the cues commonly used in sentiment and toxicity filtering. A sentiment-or-profanity proxy captured most hostile sexist statements but missed most benevolent sexist statements, illustrating why civility and negativity dashboards can produce false reassurance. By integrating ambivalent sexism theory, speech act theory, and politeness theory, sexism detection was reframed as a pragmatic problem of social action and accountability. Speech-act structure mediated visibility to filters because advice and evaluation permit affiliative framing and plausible deniability while still reproducing gendered hierarchies. A joint labeling perspective, therefore, supports more targeted responses that distinguish overt harassment from advice-like gatekeeping.

REFERENCES

Abburi, H., Parikh, P., Chhaya, N., & Varma, V. (2024). Multi-task learning neural framework for categorizing sexism. *Computer Speech & Language*, *83*, 101535. https://doi.org/https://doi.org/10.1016/j.csl.2023.101535

Basford, T. E., Offermann, L. R., & Behrend, T. S. (2014). Do you see what i see? Perceptions of gender microaggressions in the workplace. *Psychology of Women Quarterly*, *38*(3). https://doi.org/10.1177/0361684313511420

Cortina, L. M. (2008). Unseen injustice: Incivility as modern discrimination in organizations. *Academy of Management Review*, *33*(1). https://doi.org/10.5465/AMR.2008.27745097

Dardenne, B., Dumont, M., & Bollier, T. (2007). Insidious dangers of benevolent sexism: Consequences for women's performance. *Journal of Personality and Social Psychology*, *93*(5), 764–779. https://doi.org/10.1037/0022-3514.93.5.764

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the 11th International Conference on Web and social media, ICWSM 2017*. https://doi.org/10.1609/icwsm.v11i1.14955

Ford, T. E., & Ferguson, M. A. (2004). Social Consequences of Disparagement Humor: A Prejudiced Norm Theory. In *Personality and Social Psychology Review* (Vol. 8, Number 1). https://doi.org/10.1207/S15327957PSPR0801_4

Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. *12th International AAAI Conference on Web and Social Media, ICWSM 2018*. https://doi.org/10.1609/icwsm.v12i1.14991

Gilda, S., Giovanini, L., Silva, M., & Oliveira, D. (2022). Predicting Different Types of Subtle Toxicity in Unhealthy Online Conversations. *Procedia Computer Science*, *198*, 360–366. https://doi.org/10.1016/j.procs.2021.12.254

Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, *70*(3), 491–512. https://doi.org/10.1037/0022-3514.70.3.491

Grosz, D., & Conde-Céspedes, P. (2020). *Automatic Detection of Sexist Statements Commonly Used at the Workplace* (pp. 104–115). https://doi.org/10.1007/978-3-030-60470-7_11

Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior*, *32*, 113–135. https://doi.org/10.1016/j.riob.2012.11.003

Kaiser, C. R., Major, B., Jurcevic, I., Dover, T. L., Brady, L. M., & Shapiro, J. R. (2013). Presumed fair: Ironic effects of organizational diversity structures. *Journal of Personality and Social Psychology*, *104*(3). https://doi.org/10.1037/a0030838

Kaiser, C. R., & Miller, C. T. (2001). Stop Complaining! The Social Costs of Making Attributions to Discrimination. *Personality and Social Psychology Bulletin*, *27*(2), 254–263. https://doi.org/10.1177/0146167201272010

Kaiser, C. R., & Miller, C. T. (2004). A stress and coping perspective on confronting sexism. *Psychology of Women Quarterly*, *28*(2). https://doi.org/10.1111/j.1471-6402.2004.00133.x

Kirk, H., Yin, W., Vidgen, B., & Röttger, P. (2023). SemEval-2023 Task 10: Explainable Detection of Online Sexism. In A. Kr. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, & E. Sartori (Eds.), *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)* (pp. 2193–2210). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.semeval-1.305

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive Language Detection in Online User Content. *Proceedings of the 25th International Conference on World Wide Web*, 145–153. https://doi.org/10.1145/2872427.2883062

Ocampo, N. B., Sviridova, E., Cabrio, E., & Villata, S. (2023). An In-depth Analysis of Implicit and Subtle Hate Speech Messages. In A. Vlachos & I. Augenstein (Eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 1997–2013). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.147

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Pavlopoulos, J., Malakasiotis, P., & Androutsopoulos, I. (2017). Deep learning for user comment moderation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/w17-3004

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., & Androutsopoulos, I. (2020). Toxicity Detection: Does Context Really Matter? In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4296–4305). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.396

Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A., Guerra, R., Carvalho, P., Marques, C., & Silva, C. (2024). A comprehensive review on automatic hate speech detection in the age of the transformer. *Social Network Analysis and Mining*, *14*(1), 204. https://doi.org/10.1007/s13278-024-01361-3

Rudman, L. A., & Phelan, J. E. (2008). Backlash effects for disconfirming gender stereotypes in organizations. In *Research in Organizational Behavior* (Vol. 28). https://doi.org/10.1016/j.riob.2008.04.003

Samory, M., Sen, I., Kohne, J., Flöck, F., & Wagner, C. (2020). "Unsex me here": Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. *CoRR*, *abs/2004.12764*. https://arxiv.org/abs/2004.12764

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

Song, S., & Chang, P.-C. (2025). The Impact of Benevolent Sexism on Women's Career Growth: A Moderated Serial Mediation Model. *Behavioral Sciences*, *15*(1), 59. https://doi.org/10.3390/bs15010059

Warren, C., Wax, A., Brush, O. T., Magalona, J., & Galvez, G. (2023). Development and validation of the Benevolent Sexism in the Workplace scale. *Journal of Occupational and Organizational Psychology*, *96*(3), 473–502. https://doi.org/10.1111/joop.12435

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

WEST, C., & ZIMMERMAN, D. H. (1987). Doing Gender. *Gender & Society*, *1*(2), 125–151. https://doi.org/10.1177/0891243287001002002