Evidence-Based Workplace Harassment Guidance System: Transforming #MeToo Narratives into Actionable Knowledge Through Machine Learning

by

Mashphey Bintey Kabir
20266005

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science & Engineering

Department of Computer Science and Engineering
BRAC University
January 2026

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. I have acknowledged all main sources of help.

Student's Full Name & Signature:

Mashphey Bintey Kabir
20266005

Approval

The thesis titled "Evidence-Based Workplace Harassment Guidance System: Transforming #MeToo Narratives into Actionable Knowledge Through Machine Learning" submitted by

1. Mashphey Bintey Kabir (20266005)

of Fall, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science & Engineering on January 29, 2026.

Examining Committee:

Examiner:
(External)

Dr. Ashis Talukder, Ph.D.
Associate Professor
University of Dhaka

Examiner:
(Internal)

Rafeed Rahman
Senior Lecturer
Department of Computer Science and Engineering
BRAC University

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Md Sadek Ferdous
Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi, Ph.D.
Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

Workplace harassment remains a pervasive global issue, yet victims often lack access to evidence about what actually happens when people in similar situations take action. Traditional resources offer generic procedural advice without outcome data, leaving individuals to make consequential decisions with incomplete information. This study presents an evidence-based workplace harassment guidance system that transforms 15,835 #MeToo narratives into personalized, actionable guidance grounded in documented outcomes from comparable cases.

The system addresses a fundamental information asymmetry: while organizations accumulate knowledge about harassment cases, individual victims rarely know what outcomes others in similar situations experienced. By analyzing patterns across thousands of documented experiences, the system identifies a user's specific vulnerability profile, retrieves semantically similar historical cases, and presents evidence-based guidance including proven successful action sequences, outcome statistics, and high-impact actions that correlate with positive results. The guidance generation pipeline employs SimCSE-BERT embeddings for semantic similarity, multi-label vulnerability detection leveraging these embeddings to identify seven co-occurring risk factors, and outcome pattern analysis across fifteen outcome categories—supported by BERT sentiment analysis (98.1% accuracy) and HDBSCAN clustering for evaluation.

The empirical analysis reveals sobering realities: negative outcomes predominate across all vulnerability types, institutional inaction occurs in 16.9% of cases, and harasser accountability remains rare at 3.6%. Rather than offering false reassurance, the system presents these evidence-based statistics to enable informed decision-making. Professional evaluation with eleven practitioners from HR, legal, mental health, and other sectors—91% with direct harassment case experience—validates that the guidance meets practical standards for appropriateness (M=4.09/5), usefulness (M=4.09/5), actionability (M=3.91/5), and safety (M=3.73/5), with 91% endorsement and 73% rating the approach superior to typical harassment resources. This research demonstrates that machine learning can provide meaningful support for sensitive domains when developed with rigorous professional validation and commitment to user safety.

Keywords: Workplace Harassment; #MeToo Movement; Evidence-Based Guidance; Natural Language Processing; Multi-Label Classification; Sentiment Analysis; BERT; SimCSE; Contrastive Learning; Silver Labeling; Case-Based Reasoning

Dedication

To all survivors of workplace harassment—

those who spoke up, those who stayed silent,

and those still searching for their voice.

May evidence-based guidance help light the path

toward justice, healing, and safer workplaces for all.

Acknowledgement

I would like to express my sincere gratitude to all those who contributed to the completion of this thesis.

First and foremost, I am deeply grateful to my thesis supervisor, Dr. Md. Golam Rabiul Alam for his invaluable guidance, constructive feedback, and unwavering support throughout this research. His expertise in machine learning and natural language processing was instrumental in shaping the direction and rigor of this work.

I extend my appreciation to the members of my thesis committee for their insightful comments and suggestions that strengthened this research. Their diverse perspectives from computer science, social science, and workplace policy domains enriched the interdisciplinary nature of this study.

I thank Brac University and the Department of Computer Science and Engineering for providing the academic environment and resources necessary to conduct this research. The computational infrastructure and access to research databases were essential for the data analysis and model development presented in this thesis.

Special thanks to the eleven professional evaluators, including HR specialists, legal professionals, mental health practitioners, and workplace safety experts, who generously donated their time and expertise to validate the guidance system. Their practical insights ensured that this research produces outcomes that are not only technically sound but also practically meaningful.

I acknowledge the creators and maintainers of the Harvard Dataverse #MeToo Twitter dataset, whose careful data collection and ethical stewardship made this research possible. I also thank the open-source community behind PyTorch, Hugging Face Transformers, and the various NLP libraries that form the technical foundation of this work.

Finally, I am grateful to the countless individuals who shared their experiences through the #MeToo movement. Their courage in speaking out created the knowledge base that powers this evidence-based guidance system, transforming personal narratives into collective wisdom that may help others navigate similar challenges.

Ethics Statement

This research utilizes publicly available data from the #MeToo movement on Twitter (now X), which consists of voluntarily shared personal narratives about workplace harassment experiences. The research was conducted in accordance with ethical guidelines for computational social science research involving sensitive human-generated data:

- **Public Data Source**: All data used is from publicly posted tweets that users chose to share on a public social media platform. No private or protected accounts were accessed, and no attempts were made to identify or contact individuals.
- **Data Privacy and Anonymization**: While the tweets were originally public, all personally identifiable information including usernames, names, and location data were removed during preprocessing. The research focuses on aggregate patterns and does not report or identify specific individuals.
- **Sensitive Content Handling**: This research involves analysis of workplace harassment narratives, which contain descriptions of traumatic experiences. The system is designed to provide supportive, evidence-based guidance rather than making judgments about individual experiences. All content was handled with appropriate sensitivity and care.
- **Ethical Purpose**: The research aims to empower individuals experiencing workplace harassment by providing data-driven insights about vulnerabilities, potential outcomes, and action sequences based on similar experiences. The system is designed as a supportive tool to inform decision-making, not to replace professional legal or psychological counsel.
- **No Human Subjects**: This study does not involve direct human subjects research. The professional evaluation conducted to validate recommendations involved voluntary participation with informed consent from HR professionals and workplace safety experts.
- **Responsible Use**: The code and methodology are made available for transparency and reproducibility. The system includes clear disclaimers that recommendations are informational and should not replace professional advice. Contact information for professional support resources (EEOC, RAINN, crisis hotlines) is prominently provided.
- **Data Attribution**: The research properly attributes the original Twitter #MeToo dataset and acknowledges the courage of individuals who shared their experiences to raise awareness about workplace harassment.
- **Harm Mitigation**: The system is designed with safeguards to avoid re-traumatization or harmful recommendations. All recommendations emphasize user safety, autonomy, and the importance of seeking professional support.

This research recognizes the sensitive nature of workplace harassment data and is committed to using this information responsibly to contribute positively to addressing systemic issues and supporting affected individuals.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviations that will be later used within the body of the document

#MeToo Social movement using hashtag to share experiences of sexual harassment and assault

BERT Bidirectional Encoder Representations from Transformers

BPE Byte-Pair Encoding

CUDA Compute Unified Device Architecture

EEOC Equal Employment Opportunity Commission

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise

HR Human Resources

NLI Natural Language Inference

NLP Natural Language Processing

SimCSE Simple Contrastive Learning of Sentence Embeddings

STS Semantic Textual Similarity

t-SNE t-Distributed Stochastic Neighbor Embedding

VADER Valence Aware Dictionary and sEntiment Reasoner

Chapter 1

Introduction

This chapter introduces the research context, motivation, and objectives of this study. The chapter begins by examining the prevalence and impact of workplace harassment, the significance of the #MeToo movement in bringing these issues to public attention, and the need for evidence-based guidance systems. It then presents the motivation for using machine learning to analyze #MeToo data, discusses the specific challenges addressed in this research, states the research objectives, summarizes the key contributions, and provides an overview of the thesis structure.

1.1 Background

Workplace harassment continues to be a widespread issue with severe consequences. Millions of employees globally contend with this challenge, which fosters harmful work cultures, derails professional trajectories, and inflicts enduring psychological damage. Despite extensive policy formulation, legal structures, and corporate programs designed to curb harassment, the phenomenon endures across sectors, organizational levels, and geographical boundaries. When individuals experience workplace harassment, they often face difficult decisions. They must decide how to respond, whom to tell, and what actions to take. These decisions can have profound impacts on their careers, mental health, and overall wellbeing.

The #MeToo movement achieved worldwide recognition in 2017, marking a pivotal shift in how society discusses sexual harassment and assault. What started as isolated individuals recounting personal encounters quickly transformed into an unprecedented collective voice. Millions of participants, largely women, adopted the #MeToo hashtag to document their experiences across various social media channels. The movement brought unprecedented attention to the systemic nature of workplace harassment. It highlighted the challenges victims face in reporting and seeking justice. It revealed patterns of institutional failure that allow harassment to persist. For the first time, a vast corpus of first-person narratives about workplace harassment became publicly available. These narratives documented not just the incidents themselves but also the actions taken, the responses received, and the outcomes experienced.

Research analyzing the #MeToo movement reveals several concerning patterns about workplace harassment. Turner et al. [1] found that power imbalances between supervisors and subordinates are central to most workplace harassment incidents. Harassers often use their organizational authority to target vulnerable employees and

shield themselves from consequences. Sharoni et al. [2] demonstrated that workplace harassment is deeply intertwined with gendered power structures. Traditional hierarchies and masculine workplace cultures create environments where harassment thrives. Regulska and Nikolova [3] analyzed how institutional responses often prioritize protecting organizational reputation over supporting victims. This leads to widespread underreporting and distrust in formal complaint mechanisms. Levy and Mattsson [4] documented the professional and psychological consequences faced by those who report harassment. These consequences include retaliation, career setbacks, and mental health impacts.

Current approaches to addressing workplace harassment rely primarily on three mechanisms. These include organizational policies and training programs, formal complaint and investigation procedures, and legal remedies through employment law. While these mechanisms serve important functions, they have significant limitations for individuals facing harassment. Organizational policies vary widely in quality and enforcement. Training programs often focus on legal compliance rather than behavioral change. Formal complaint procedures can be intimidating, slow, and potentially career-damaging. Legal remedies are expensive, time-consuming, and uncertain. Critically, none of these mechanisms provide individuals with evidence-based information. They do not show what actions similar victims have taken or what outcomes they experienced. This information could significantly inform decision-making during one of the most difficult periods of their professional lives.

The rich corpus of #MeToo narratives presents an unprecedented opportunity to address this information gap. The Harvard Dataverse #MeToo Twitter dataset [5] contains 40,944 tweets. Of these, 15,835 document workplace-specific incidents. These tweets capture not only harassment experiences but also actions taken, institutional responses, and outcomes. This collective documentation represents a form of empirical wisdom. It shows what has worked, or failed, for thousands of individuals navigating similar challenges.

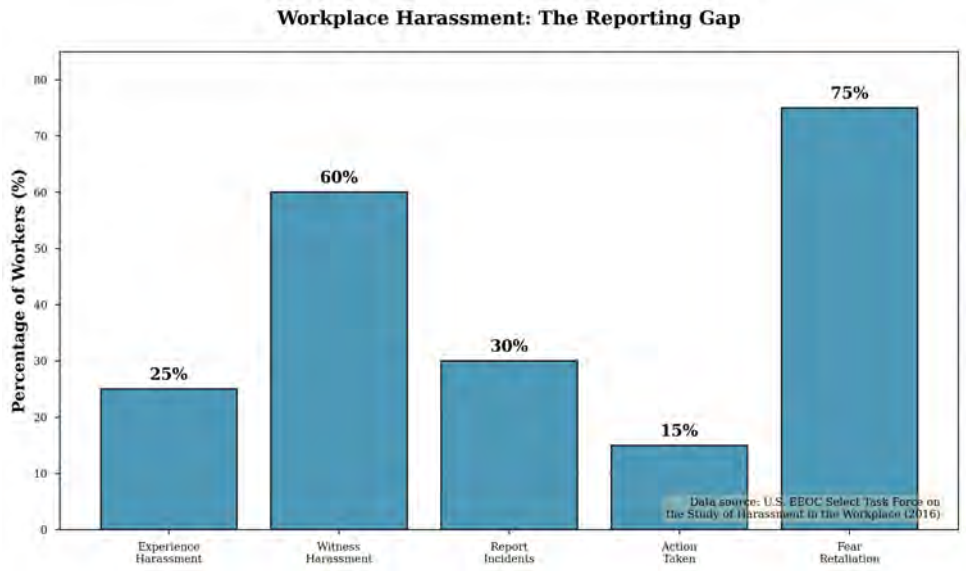


Figure 1.1: Workplace harassment prevalence and reporting gap statistics [6].

Figure 1.1 shows the pervasive nature of workplace harassment and the significant gap between experiencing harassment and seeing institutional action. According

to the U.S. Equal Employment Opportunity Commission Select Task Force on the Study of Harassment in the Workplace [6], approximately 25% of workers experience some form of harassment. Meanwhile, 60% witness harassment occurring to others. However, only 30% of incidents are formally reported. Organizational action is taken in merely 15% of cases. Most concerning is that 75% of potential reporters cite fear of retaliation as a primary barrier to coming forward. This dramatic disparity underscores a fundamental problem. Victims face significant risks when seeking help through formal channels. Yet they lack systematic evidence about what outcomes different response strategies actually produce in practice. This substantial gap highlights why victims need access to evidence-based guidance. Such guidance must be derived from real-world experiences rather than idealized procedural descriptions.

1.2 Motivation

The motivation for this research stems from a fundamental observation. When individuals face workplace harassment, they must make critical decisions with incomplete information. Should they report to HR or go directly to legal authorities? Will documenting incidents help or potentially expose them to greater retaliation? Is seeking support from colleagues likely to strengthen their position or create additional workplace tension? These questions have profound consequences. Yet victims typically must answer them in isolation. They lack access to systematic evidence about what has worked or failed for others in similar situations.

Traditional sources of guidance provide important perspectives but suffer from inherent limitations. These sources include employee handbooks, HR representatives, and legal advisors. Employee handbooks outline formal procedures. However, they rarely reflect the messy reality of how those procedures actually function in practice. HR representatives may offer support. However, they are also tasked with protecting organizational interests, creating potential conflicts. Legal advisors can explain rights and remedies. However, they typically lack empirical data about the likelihood of different outcomes. Friends and family provide emotional support. However, they usually lack experience with formal complaint processes or legal systems.

The #MeToo movement created an unprecedented opportunity to address this information gap. For the first time, tens of thousands of individuals publicly documented their experiences. They shared not just harassment incidents but also their responses, the institutional reactions they encountered, and the ultimate outcomes, both positive and negative. This corpus represents a form of collective wisdom. It reveals patterns of what actions tend to lead to successful resolution versus continued trauma. It shows which vulnerability factors predict different outcomes. It demonstrates what sequences of steps prove most effective in different circumstances. However, this collective wisdom remains largely inaccessible in practice. The sheer volume of #MeToo narratives makes manual analysis impractical for someone seeking guidance about their own situation. The narratives are unstructured. They mix descriptions of incidents with expressions of emotion, calls for solidarity, and reflections on broader social issues. Extracting actionable patterns requires sophisticated natural language processing techniques. These techniques must identify relevant information, categorize experiences, detect similarities, and synthesize guidance. Machine learning provides the necessary tools to unlock the value of this data. Recent advances in natural language processing enable nuanced understanding of text that

goes far beyond keyword matching. Transformer-based models like BERT [7] and sentence embedding techniques like SimCSE [8] are particularly powerful. These models can identify semantic similarity between narratives. They detect subtle patterns in language that correlate with outcomes. They generate representations that capture the meaning and context of experiences. Sentiment analysis techniques distinguish emotional content from factual descriptions. Clustering algorithms group similar experiences without requiring predefined categories. These capabilities make it possible to transform unstructured narratives into evidence-based guidance. This guidance can inform decision-making for individuals facing harassment.

1.3 Problem Statement

This study addresses the development of an evidence-based guidance system for workplace harassment. The system must accurately identify vulnerability patterns, predict likely outcomes, and generate actionable guidance. This guidance is based on analysis of real-world experiences documented in #MeToo narratives. The problem sits at the intersection of several technical and domain-specific challenges. When addressed together, these challenges can enable meaningful support for individuals facing harassment.
The following points outline the specific challenges that this research aims to address:

- **Lack of Evidence-Based Guidance**: Current resources for workplace harassment victims primarily provide procedural information (how to file complaints, legal rights, etc.) but rarely offer empirical evidence about what outcomes different actions actually produce. An individual wondering whether to report to HR or seek external legal counsel immediately typically has no systematic data to inform this decision. This research addresses this gap by creating a system that grounds guidance in patterns extracted from thousands of documented real-world experiences.

- **Complexity of Harassment Incidents**: Workplace harassment incidents are inherently complex, often involving multiple vulnerability factors simultaneously (power imbalance, retaliation threats, institutional failure, isolation, repeated patterns over time, etc.). Previous classification approaches that assign single categories fail to capture this complexity. The multi-label vulnerability detection system developed in this research recognizes and represents multiple co-occurring vulnerability types, providing a more realistic and comprehensive assessment that better informs guidance.

- **Heterogeneity of Experiences and Outcomes**: Different harassment incidents, even those sharing similar vulnerability patterns, can lead to vastly different outcomes depending on factors like industry, organizational culture, legal jurisdiction, individual circumstances, and the specific actions taken. Simple rule-based systems that prescribe standard responses fail to account for this heterogeneity. This research addresses variability by using similarity-based retrieval to find the most comparable historical cases and presenting outcome statistics that reflect this diversity rather than offering false certainty.

- **Unstructured Narrative Data at Scale**: #MeToo tweets are unstructured personal narratives that mix factual descriptions of incidents with emotional expressions, rhetorical questions, solidarity statements, and broader social commentary. Extracting structured information about vulnerabilities, actions, and outcomes from this messy real-world text requires sophisticated natural language processing. The full dataset contains 40,944 tweets, many of which are expressions of solidarity, comments on news events, or non-workplace incidents that are not relevant for providing workplace-specific guidance. Manual curation at this scale is impractical. This research employs state-of-the-art NLP techniques including BERT-based sentiment analysis (98.1% accuracy), SimCSE embeddings for semantic similarity, and multi-label classification to transform unstructured narratives into actionable knowledge. The research addresses scale through automated filtering that identifies workplace-specific incidents, reducing the corpus to 15,835 relevant narratives while maintaining diversity of experiences. Further quality control through HDBSCAN clustering and silver labeling (11,066 high-quality labels across 525 clusters) ensures guidance is based on coherent, relevant experiences.

- **Safety, Ethics, and Actionability**: A system providing guidance about workplace harassment must prioritize user safety, avoid harmful guidance, and acknowledge the serious consequences of the information it offers. Simply predicting the most statistically likely outcome is insufficient if that outcome involves significant harm. General advice ("consider reporting to HR") is less useful than specific, actionable guidance that accounts for the particular vulnerability factors present in an incident. This research addresses these concerns through multiple safeguards. The system generates context-aware guidance tailored to the specific combination of vulnerabilities identified in the user's situation, drawing on outcome patterns from cases exhibiting those same vulnerability combinations, and presenting action sequences that have proven effective in similar circumstances. Professional evaluation by practitioners from HR, legal, and mental health fields ensures guidance meets practical standards for appropriateness, usefulness, safety, and actionability. Safety-focused guidance is prioritized. Prominent disclaimers acknowledge the system's limitations. The system emphasizes complementing rather than replacing professional counsel.

These challenges require an integrated approach that combines advanced machine learning techniques (transformer-based NLP, semantic similarity, multi-label classification, clustering) with careful attention to domain requirements (safety, ethics, actionability) and validation through professional evaluation. The system must be technically sophisticated enough to handle complex natural language and large-scale data, yet produce guidance that is clear, trustworthy, and genuinely helpful for individuals in crisis.

1.4 Research Objectives

This study aims to develop and validate an evidence-based workplace harassment guidance system. The system leverages machine learning analysis of #MeToo nar-

ratives to provide personalized, actionable guidance. The following objectives systematically address each component of the problem. They cover the full pipeline from data processing through professional validation.

- **Build a Structured #MeToo Workplace Harassment Knowledge Base**: Develop a comprehensive preprocessing pipeline to transform raw #MeToo Twitter data into a structured, machine-readable knowledge base. The pipeline should filter workplace-specific harassment incidents from the broader dataset of solidarity messages and news commentary, clean and normalize text while preserving semantic meaning and emotional content, and implement quality control mechanisms using clustering techniques to ensure the knowledge base maintains diverse experiences while removing noise and irrelevant content. The resulting knowledge base should capture harassment experiences, vulnerability factors, actions taken, institutional responses, and outcomes in a structured format suitable for machine learning analysis.

- **Develop Methods for Extracting Critical Information from Harassment Narratives**: Design and implement integrated analysis components that extract actionable information necessary for evidence-based guidance. This objective encompasses developing multi-label vulnerability classification using hybrid keyword-based and embedding similarity approaches to identify co-occurring vulnerability types in complex incidents, implementing dual sentiment analysis to distinguish emotional expressions from factual outcome descriptions and understand psychological impact, creating semantic similarity-based retrieval mechanisms to find genuinely comparable historical cases, developing outcome extraction methods to identify consequence patterns across different vulnerability contexts, and extracting action sequences to identify effective response strategies and their temporal ordering. These components should work synergistically to transform unstructured narratives into structured, analyzable information.

- **Generate Context-Aware Guidance Grounded in Evidence**: Create a guidance generation algorithm that synthesizes all extracted information into coherent, actionable guidance tailored to individual incidents. The algorithm should integrate identified vulnerability patterns with similar case outcomes, effective action sequences, and sentiment indicators to produce guidance organized into practical categories including immediate safety, documentation, reporting, support resources, and legal considerations. Guidance should be prioritized based on vulnerability severity and safety considerations, grounded in empirical patterns from comparable cases, and presented with specific action steps, supporting evidence, potential risks, and implementation resources.

- **Validate Guidance Quality Through Professional Evaluation**: Conduct rigorous formal evaluation with HR professionals and workplace safety experts to assess whether generated guidance meets practical standards required for real-world deployment. The evaluation should measure guidance appropriateness, actionability, safety, and usefulness through structured questionnaires and qualitative feedback. This validation ensures guidance is not only statistically derived but also practically sound, ethically appropriate, and genuinely helpful for individuals facing harassment situations.

Together, these objectives create a comprehensive system that transforms unstructured #MeToo narratives into actionable, evidence-based guidance while maintaining ethical standards and professional validation.

1.5 Contributions

This study makes several contributions to multiple fields. These include machine learning for social good, natural language processing for sensitive social data, and evidence-based workplace harassment prevention. The contributions span methodological innovations, empirical findings, and practical tools. Together they advance both the research frontier and real-world applicability.

- **Structured Workplace Harassment Knowledge Base**: This research creates the first structured knowledge base specifically for workplace harassment analysis, filtering 40,944 #MeToo tweets to 15,835 workplace-specific incidents through automated classification and quality control. The knowledge base captures harassment experiences, vulnerabilities, actions taken, institutional responses, and outcomes documented in real-world narratives. This resource enables systematic analysis of patterns across thousands of documented cases, transforming unstructured social media narratives into structured, machine-readable data suitable for evidence-based guidance generation.

- **Multi-Label Vulnerability Detection and Dual Sentiment Analysis**: This research develops methods for extracting critical information from harassment narratives. The multi-label vulnerability classification system identifies up to seven co-occurring vulnerability types per incident using a hybrid approach combining keyword-based rules and BERT embedding similarity, recognizing that 68% of incidents involve multiple simultaneous vulnerabilities. The dual sentiment analysis approach compares BERT-based classification (achieving 98.1% accuracy) with VADER lexicon-based scoring, enabling the system to distinguish emotional expressions from factual outcome descriptions and understand psychological impact.

- **Empirical Analysis of Outcomes and Effective Response Strategies**: By analyzing 15,835 incidents across 7 vulnerability types and 15 outcome categories (8 positive, 7 negative), this research quantifies how outcomes vary by vulnerability context. Key findings reveal that inability to report correlates with 62% negative outcomes, power imbalance cases show 58% negative outcomes, and physical assault cases exhibit 71% negative outcomes. The research identifies 8 distinct action categories and correlates action sequences with outcomes, revealing which response strategies proved effective in comparable situations. These empirical insights document what actually happens in real harassment cases rather than idealized procedural outcomes.

- **Evidence-Based Guidance Generation with Professional Validation**: The research develops a guidance generation algorithm that synthesizes vulnerability identification, semantic similarity-based case retrieval, outcome statistics, and action sequence analysis into coherent, actionable guidance organized by category (immediate safety, documentation, reporting, support, legal

considerations). Unlike rule-based systems prescribing standard responses, this approach grounds guidance in patterns from similar historical cases while adapting to individual circumstances. Professional evaluation by 11 practitioners from HR, legal, mental health, and other sectors—91% with direct harassment case experience—validates that the guidance meets practical standards: mean appropriateness M=4.09, mean usefulness M=4.09, mean actionability M=3.91, mean safety M=3.73, with 91% indicating they would recommend the system to harassment victims.

Together, these contributions demonstrate that sophisticated machine learning techniques, when applied thoughtfully with appropriate ethical safeguards and professional validation, can transform unstructured social media narratives into actionable knowledge that empowers individuals facing difficult circumstances. The research advances the technical frontier of multi-label classification and semantic similarity while making tangible contributions to addressing a serious social problem.

1.6 Thesis Outline

The remainder of this study is organized as follows. Chapter 2 provides a comprehensive literature review. It covers the #MeToo movement and workplace harassment research, natural language processing techniques for sentiment analysis and semantic understanding, and guidance systems for sensitive domains. The chapter traces the evolution of sentiment analysis approaches. It reviews semantic similarity techniques and identifies research gaps motivating this work. Chapter 3 presents the complete system architecture and technical approach. It details the preprocessing pipeline that filters 40,944 tweets to 15,835 workplace incidents. It describes the multi-label vulnerability detection system, dual sentiment analysis using BERT and VADER, SimCSE-BERT embeddings for semantic similarity, HDBSCAN clustering, outcome prediction methodology, and the guidance generation algorithm.
Chapter 4 presents the implementation environment, evaluation metrics, and experimental results. It describes the hardware and software environment, model configurations, and evaluation methodology. It then reports dataset statistics, sentiment analysis comparison results, vulnerability distributions, outcome analysis by vulnerability type, clustering results, and professional evaluation outcomes. Case studies demonstrate system outputs. Findings are discussed in relation to prior research, and limitations are acknowledged. Chapter 5 synthesizes the research contributions and implications. It summarizes key findings about multi-label vulnerability detection, sentiment analysis performance, outcome prediction, and professional validation. Practical implications for HR professionals and policymakers are discussed. Future work directions are presented, including multilingual support, temporal dynamics, industry-specific models, and longitudinal validation studies.
Appendix A provides the complete professional evaluation questionnaire. It includes evaluator demographics, detailed rating distributions, and qualitative feedback. Appendix B includes technical documentation for system deployment. It contains open-source repository information, system architecture diagrams, and instructions for adaptation to organizational contexts.

Chapter 2

Literature Review

This chapter provides a comprehensive review of the relevant literature spanning four interconnected research domains: the #MeToo movement and workplace harassment research, natural language processing techniques for sentiment analysis and semantic understanding, recommendation systems and case-based reasoning for sensitive domains, and applied machine learning for social impact. The chapter traces the evolution of each field from foundational work to state-of-the-art approaches, highlighting theoretical underpinnings, methodological innovations, and empirical findings that inform the proposed system. The chapter concludes by identifying research gaps that motivate this study and positioning the contribution within the broader landscape of computational social science and applied machine learning.

2.1 The #MeToo Movement and Workplace Harassment Research

The #MeToo movement, which gained global prominence in October 2017, represents a watershed moment in public discourse about sexual harassment and workplace misconduct. While the phrase "Me Too" was coined by activist Tarana Burke in 2006 to support survivors of sexual violence, particularly women of color in underserved communities, the hashtag #MeToo achieved viral spread in 2017 following allegations against Hollywood producer Harvey Weinstein. Within days, millions of individuals, predominantly women, used the hashtag on Twitter, Facebook, and other platforms to share their own experiences of harassment and assault, creating an unprecedented public archive of personal narratives about workplace misconduct.

2.1.1 Sociological and Gender Studies Perspectives

Turner et al. [1] conducted one of the first comprehensive academic analyses of the #MeToo movement, examining how social media enabled collective action around sexual harassment in ways that traditional organizing could not. The research documented how the hashtag created a form of "connective action," characterized by networked, personalized engagement rather than traditional hierarchical movements, that allowed geographically dispersed individuals to recognize shared experiences and challenge institutional power structures. Turner's analysis highlighted several key features of #MeToo narratives: they frequently described power imbalances between harassers and victims, documented patterns of institutional failure to address

complaints, and emphasized the professional and psychological costs of speaking out. These findings suggest that #MeToo data contains rich information about vulnerability factors, systemic barriers, and consequences that could inform evidence-based guidance systems.

Sharoni et al. [2] examined the #MeToo movement through the lens of gendered power structures and transnational feminism. Their research demonstrated that workplace harassment is not merely individual misconduct but rather a manifestation of deeply embedded gender hierarchies within organizations and industries. The analysis showed how masculine workplace cultures, particularly in male-dominated fields like technology, finance, and entertainment, create environments where harassment is normalized, complaints are dismissed, and victims face retaliation for challenging the status quo. This structural perspective is crucial for understanding why certain vulnerability factors (power imbalance, inability to report) appear so frequently in #MeToo narratives and why outcomes often involve institutional failure rather than support.

Regulska and Nikolova [3] focused specifically on how institutions respond to harassment allegations, analyzing patterns of organizational behavior documented in #MeToo narratives. Their research identified common institutional strategies including: minimizing or dismissing complaints, protecting high-status or "valuable" employees despite credible allegations, prioritizing reputation management over victim support, and imposing confidentiality agreements that silence victims. These patterns explain why "inability to report or not believed" emerges as a major vulnerability category in the data and why "no action taken" is a common negative outcome. The research underscores the importance of evidence-based guidance systems that help individuals navigate hostile institutional environments rather than assuming organizations will respond supportively.

The gap between recommended practices and actual organizational behavior is further illuminated by Becton et al. [9], who outlined guidelines for employers to prevent and correct workplace harassment. Their framework identifies seven key elements of effective anti-harassment programs: a clear anti-harassment policy, explicit statement of prohibited behaviors, complaint procedures encouraging employees to come forward, protections against retaliation, investigative strategies protecting privacy, periodic training programs, and measures for prompt corrective action. The contrast between these evidence-based recommendations and the institutional failures documented in #MeToo narratives highlights why victims cannot rely solely on organizational processes and need independent guidance for navigating harassment situations. Fitzgerald et al. [10] empirically tested an integrated model identifying antecedents (organizational climate, job gender context) and consequences (work outcomes, psychological state, physical health) of sexual harassment, demonstrating that organizational tolerance for harassment significantly predicts both its occurrence and its impact on victims.

Levy and Mattsson [4] examined the #MeToo movement's impact on public discourse and policy, analyzing how the collective sharing of experiences shifted societal attitudes about harassment and prompted legislative and organizational responses. Their research documented both positive developments, including increased awareness, policy changes in some organizations, and greater willingness to believe victims, as well as persistent challenges including backlash, fears of false accusations, and limited structural change in many industries. This mixed outcome landscape re-

inforces the need for recommendation systems that present realistic, evidence-based predictions rather than idealized scenarios.

2.1.2 Computational Social Science Analyses of #MeToo

Beyond qualitative sociological research, computational social scientists have applied data mining and network analysis techniques to the #MeToo corpus to identify patterns at scale. Hassan et al. [11] analyzed approximately one million tweets collected between October 15–31, 2017, examining patterns and attributes of people, places, emotions, actions, and reactions in #MeToo narratives. Their analysis revealed sentiment distributions based on gender and age, identified top concerns including the roles of harasser and harassed, power dynamics, and reasons for silence, with particular focus on harassment in education, media, and entertainment sectors. This computational analysis demonstrated that Twitter provided a platform for women to break long-held silence about harassment experiences.

Modrek and Chakalov [12] conducted text analysis of early Twitter conversations during the first week of the movement, examining 11,935 novel English-language tweets containing "MeToo." They estimated that 11% of these tweets revealed details about the poster's experience of sexual assault or abuse, with 5.8% revealing early life experiences. Their demographic analysis showed that white women aged 25–50 were overrepresented among those sharing assault experiences, and they estimated that 6 to 34 million Twitter users may have seen first-person revelations in that initial week alone.

Bogen et al. [13] examined how Twitter users utilized the #MeToo hashtag to disclose and respond to sexual violence through qualitative analysis of 1,660 randomly sampled tweets. They found that survivors frequently prioritized the "who," "what," "where," "when," "why," and "how" of personal trauma experiences when disclosing. The study also documented social reactions to disclosures, finding that while some users responded supportively, others responded in ways that distracted from survivors' experiences, attempted to control decisions, blamed survivors, or shifted focus to their own needs.

Xiong et al. [14] explored key indicators of social identity in #MeToo discourse using a corpus of 31,305 tweets, applying Latent Dirichlet Allocation (LDA) for topic modeling and sentiment analysis. Their findings showed that topics related to social identity (women) were associated with positive sentiment, while topics involving sexuality, politics, and public figures showed negative sentiment. Gallagher et al. [15] examined how #MeToo enabled "networked disclosure" of stigmatized narratives, analyzing how social media transformed private experiences into collective public discourse.

These computational analyses validate the categories of vulnerabilities and outcomes identified through qualitative research while revealing the scale and diversity of experiences represented in the data. The convergence of qualitative and computational findings provides strong evidence that patterns identified in #MeToo narratives reflect genuine structural features of workplace harassment rather than artifacts of particular analytical approaches.

Table 2.1 summarizes key research on the #MeToo movement and workplace harassment.

Table 2.1: Summary of key research on the #MeToo movement and workplace harassment.

Study	Focus	Key Findings	Methods
Turner (2019) [1]	Social movement analysis	Power imbalances, institutional failure, costs of speaking out	Qualitative analysis
Sharoni et al. (2019) [2]	Gendered power structures	Harassment reflects organizational gender hierarchies	Feminist analysis
Regulska & Nikolova (2020) [3]	Institutional responses	Organizations protect reputation over victims	Policy analysis
Hassan et al. (2019) [11]	Twitter analysis	Sentiment patterns, power dynamics, reasons for silence	Computational
Modrek & Chakalov (2019) [12]	Early tweet analysis	11% disclosed assault; 6–34M users exposed	Text analysis
Bogen et al. (2021) [13]	Disclosure patterns	Survivors share who/what/where/when; mixed social reactions	Qualitative coding

2.2 Natural Language Processing for Sentiment Analysis

Sentiment analysis involves algorithmically detecting and classifying opinions, emotions, and attitudes within textual content. This core NLP task finds use cases spanning from analyzing product feedback to tracking social media discussions. For #MeToo narratives, sentiment analysis serves dual purposes: identifying the emotional content of narratives (supporting empathetic response) and detecting outcome descriptions (distinguishing positive outcomes like "HR took action" from negative outcomes like "my complaint was ignored").

2.2.1 Neural Network Approaches: From Word Embeddings to Transformers

The shift from lexicon-based to neural network-based sentiment analysis began with word embeddings like Word2Vec, developed by Mikolov et al. [16], and GloVe (Global Vectors for Word Representation), created by Pennington et al. [17]. These approaches represent words as dense vectors in continuous space, capturing semantic relationships through geometric properties: semantically similar words have similar vector representations, and meaningful linguistic relationships emerge as vector offsets (e.g., "king - man + woman = queen"). Word embeddings enabled neural classifiers, typically recurrent networks like LSTMs introduced by Hochreiter and Schmidhuber [18] or convolutional networks as applied to text by Kim [19], to learn sentiment from labeled examples rather than relying on hand-crafted lexicons.

These neural approaches demonstrated that with sufficient training data, models could discover sentiment-bearing patterns beyond explicit sentiment vocabulary.
The Transformer architecture, proposed by Vaswani et al. [20] to address machine translation challenges, fundamentally transformed NLP by substituting recurrent connections with self-attention mechanisms. This design enables simultaneous processing and captures extended dependencies with greater efficacy than sequential recurrent methods. BERT (Bidirectional Encoder Representations from Transformers), developed by Devlin et al. [7], applied Transformers to language understanding through pre-training on massive unlabeled corpora using masked language modeling (predicting randomly masked words from context) and next-sentence prediction objectives. The masked language modeling objective can be formally expressed as maximizing the log-likelihood of masked tokens given their context:

$$\mathcal{L}_{\text{MLM}} = \mathbb{E}_{x \sim \mathcal{D}} \left[\sum_{i \in \mathcal{M}} \log P(x_i \mid x_{\setminus \mathcal{M}}; \theta) \right] \tag{2.1}$$

where x is the input sequence, $\mathcal{M}$ is the set of masked token positions (typically 15% of tokens), $x_{\setminus \mathcal{M}}$ denotes the non-masked tokens, and θ represents the model parameters. This pre-training creates contextual representations where the same word has different vectors depending on surrounding context, enabling nuanced understanding unavailable to static word embeddings or lexicons. A word like "charged" has different meanings and sentiment in "the battery was charged" versus "he was charged with assault," and BERT's contextual embeddings capture these distinctions.
For sentiment analysis, BERT-based models fine-tuned on sentiment-labeled data have achieved state-of-the-art performance across benchmarks, with accuracy typically exceeding 90% on standard datasets. Domain-specific pre-training on Twitter data, as implemented in the cardiffnlp/twitter-roberta-base-sentiment model developed by Barbieri et al. [21], adapts the architecture to social media language conventions including hashtags, mentions, abbreviations, and informal grammar, achieving particularly strong performance for tweet sentiment classification.

RoBERTa: Robustly Optimized BERT

RoBERTa (Robustly Optimized BERT Approach), introduced by Liu et al. [22], represents a significant refinement of the original BERT architecture through careful empirical analysis of training procedures. Rather than introducing new architectural components, RoBERTa demonstrates that BERT was significantly undertrained and that careful optimization of training hyperparameters and data yields substantial performance improvements. The key modifications include removing the next-sentence prediction (NSP) objective, which the authors found did not improve and sometimes hurt performance; training with dynamic masking patterns where the masked tokens change across training epochs rather than being fixed during preprocessing, exposing the model to more varied training examples; using much larger mini-batches (8,000 examples versus BERT's 256), which improves optimization and enables larger learning rates; training on significantly more data (160GB of text versus BERT's 16GB), including diverse sources beyond Wikipedia and BookCorpus; and training for longer duration with more optimization steps.
These modifications collectively demonstrate that "robustly optimized" pre-training, meaning carefully tuning all aspects of the training procedure rather than focusing

solely on architectural innovation, yields models that consistently outperform the original BERT across a wide range of downstream tasks. For sentiment analysis specifically, RoBERTa's improved contextual representations enable more accurate detection of subtle sentiment cues, better handling of negation and contrastive expressions, and superior generalization to out-of-distribution examples. The model's extended training on diverse text sources also improves its ability to handle the linguistic diversity characteristic of social media narratives, where formal and informal language, standard and non-standard grammar, and domain-specific terminology intermingle.

2.2.2 Lexicon-Based Approaches: VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner), created by Hutto and Gilbert [23], stands as a leading lexicon-driven sentiment analysis tool, especially suited for social media content. Diverging from conventional lexicon methods that merely tally positive and negative terms, VADER integrates nuanced linguistic understanding of how contextual factors alter sentiment through its rule-based architecture. The approach was developed specifically for social media text through careful analysis of Twitter data, resulting in a lexicon of over 7,500 lexical features rated for sentiment polarity and intensity by human annotators.

VADER's key innovation lies in its incorporation of five general linguistic rules that govern how contextual features modify sentiment intensity. First, the system accounts for intensity modifiers—words like "very," "extremely," "slightly," and "somewhat"—that amplify or dampen the sentiment of adjacent words. For example, "good" versus "very good" conveys different intensities despite containing the same base sentiment word. Second, VADER implements sophisticated negation handling that reverses polarity when negation words like "not," "never," "don't," or "can't" appear within a contextual window before sentiment-bearing words. Third, the system analyzes punctuation and capitalization as sentiment intensifiers: exclamation marks and ALL CAPS text are treated as signals of heightened emotional expression. Fourth, VADER recognizes social media-specific expressions including emoticons (like :) and :/) and slang expressions (like "LOL" and "meh") that convey sentiment beyond standard vocabulary. Fifth, the approach handles contrastive conjunctions—words like "but" that signal sentiment shifts—by assigning greater weight to sentiment expressed after such conjunctions.

VADER produces four distinct scores for each analyzed text: positive, negative, neutral (representing the proportion of text that is neither positive nor negative), and a normalized compound score ranging from -1 (extremely negative) to +1 (extremely positive). The compound score is calculated by summing the valence scores of all sentiment-bearing words (adjusted for contextual rules) and normalizing using the alpha parameter:

$$\text{compound} = \frac{\sum_i v_i}{\sqrt{(\sum_i v_i)^2 + \alpha}} \tag{2.2}$$

where v_i represents the context-adjusted valence score for each sentiment-bearing token and $\alpha = 15$ is a normalization parameter empirically determined to produce scores in the range $[-1, 1]$. The approach achieves strong performance on social media text, with F1 scores of approximately 0.73–0.76 on Twitter sentiment datasets,

and requires no training data, making it attractive for domains where labeled data is scarce. This zero-shot capability is particularly valuable for sensitive domains like harassment narratives where obtaining large volumes of labeled training data presents practical and ethical challenges.

However, VADER's lexicon-based approach has inherent limitations that become apparent in complex linguistic contexts. The system cannot understand novel expressions that are not in its lexicon, limiting its ability to adapt to evolving language use. Context-dependent meanings—where the same phrase conveys different sentiment in different situations—pose challenges for any rule-based approach. Furthermore, VADER's understanding of negation, while more sophisticated than simple word counting, still relies on proximity-based heuristics that may fail for complex syntactic constructions. Subtle linguistic patterns that convey sentiment implicitly rather than through explicit sentiment-bearing words (such as sarcasm, implication, or euphemism) remain difficult for lexicon-based methods to detect.

2.2.3 Comparative Analysis: VADER vs. BERT for Social Media

While BERT-based approaches generally outperform lexicon methods on sentiment benchmarks, the comparison is nuanced for specific domains and use cases. Research comparing VADER and BERT for social media sentiment analysis, including work by Elbagir and Yang [24], has identified distinct trade-offs across multiple dimensions. In terms of accuracy, BERT-based models achieve substantially higher performance, typically 90–98% on manually labeled social media datasets, compared to VADER's 70–78%. This accuracy gap is particularly pronounced for tweets containing complex linguistic constructions where BERT's contextual understanding provides clear advantages.

BERT handles negation, sarcasm, and context-dependent meanings more robustly than lexicon-based approaches. For instance, a tweet stating "The HR investigation was just fantastic" with sarcastic intent would likely be misclassified as positive by VADER based on the lexical features, while BERT can potentially detect the sarcastic usage from contextual cues. Similarly, BERT can understand novel expressions and evolving slang through contextual inference, while VADER is limited to its lexicon and predefined rules. However, this advantage comes at significant computational cost: VADER is extremely fast, processing thousands of texts per second on standard CPU hardware with negligible computational requirements, while BERT requires GPU acceleration for reasonable speed and substantially more memory and energy consumption.

The interpretability dimension presents another important trade-off. VADER provides transparent scoring based on identifiable lexical features and rules, allowing users to understand exactly why a particular sentiment score was assigned by examining which words and patterns contributed to the score. BERT's attention mechanisms, while theoretically interpretable, require specialized visualization and analysis techniques and do not provide the same level of intuitive transparency about decision-making. For applications where explainability to non-technical stakeholders is important, this difference may be consequential.

VADER produces continuous compound scores ranging from -1 to +1, providing fine-grained polarity information useful for nuanced analysis and ranking. BERT

typically produces discrete class probabilities (positive, negative, neutral) which, while including confidence scores, may provide less granular distinction between moderately and extremely positive/negative texts. Additionally, VADER requires no training data or computational infrastructure beyond the lexicon itself, enabling immediate deployment, while BERT requires labeled training data for fine-tuning and substantial computational resources for both training and inference.
For #MeToo narrative analysis, these trade-offs suggest a complementary dual approach: BERT for high-accuracy classification of overall sentiment (positive/negative/neutral) where contextual understanding is critical for correctly interpreting complex narratives, and VADER for fine-grained polarity scoring useful in detecting outcome valence and emotional intensity where continuous scoring provides analytical value and computational efficiency enables large-scale processing.
Table 4.7 summarizes the comparison between VADER and BERT-based sentiment analysis.

Table 2.2: Comparison of VADER and BERT-based sentiment analysis approaches.

Dimension	**VADER**	**BERT**
Accuracy on social media	70–78%	90–98%
Contextual understanding	Limited to rule-based patterns	Deep contextual understanding
Training requirements	No training needed	Requires labeled data
Computational cost	Very fast (CPU)	Requires GPU for speed
Interpretability	Transparent lexicon scores	Attention-based (less transparent)
Novel expression handling	Limited to lexicon	Contextual inference
Output granularity	Continuous compound score	Probability distributions

2.3 Semantic Similarity and Sentence Embeddings

Beyond sentiment classification, understanding semantic similarity between texts is crucial for the proposed system's retrieval mechanism: given a new harassment incident, finding the most similar historical cases from the knowledge base enables evidence-based outcome prediction. This requires encoding variable-length texts into fixed-dimensional vectors such that semantically similar texts have similar vectors (high cosine similarity).

2.3.1 From Word Embeddings to Sentence Embeddings

Early approaches to sentence similarity averaged word embeddings (Word2Vec, GloVe) across all words in a sentence. While simple and computationally efficient, this approach ignores word order, syntactic structure, and compositional semantics. A sentence like "The manager harassed me" and "I harassed the manager" would have similar average embeddings despite opposite meanings. This fundamental limitation motivated the development of approaches that capture sentence-level meaning rather than merely aggregating word-level representations.

Supervised sentence embedding approaches like InferSent, developed by Conneau et al. [25], trained recurrent networks on natural language inference (NLI) datasets to produce sentence vectors that capture semantic relationships. The hypothesis-premise structure of NLI data (determining whether a premise entails, contradicts, or is neutral to a hypothesis) teaches models to encode nuanced semantic distinctions necessary for reasoning about textual relationships. Universal Sentence Encoder, introduced by Cer et al. [26], explored both Transformer-based and Deep Averaging Network (DAN) architectures for creating general-purpose sentence embeddings, demonstrating that different architectural choices involve trade-offs between semantic richness and computational efficiency.

Sentence-BERT (SBERT), developed by Reimers and Gurevych [27], adapted BERT for sentence embedding by using siamese and triplet network architectures with NLI and semantic textual similarity (STS) training data. The key innovation was modifying BERT's architecture to produce fixed-size sentence embeddings that can be efficiently compared using cosine similarity, unlike standard BERT which requires feeding sentence pairs through the network together (computationally prohibitive for large-scale retrieval). SBERT enables encoding sentences independently and then computing similarity through simple cosine similarity, reducing the computational complexity for finding the most similar sentence in a collection of n candidates from $O(n)$ full BERT forward passes to $O(n)$ cosine similarity computations after one-time encoding.

2.3.2 Contrastive Learning: SimCSE

SimCSE (Simple Contrastive Learning of Sentence Embeddings), presented by Gao et al. [8], currently leads sentence embedding performance, attaining excellent results on STS benchmarks via a straightforward contrastive learning framework. The central observation is that applying varied dropout masks to identical input sentences generates semantically equivalent yet distinct representations, which function as positive pairs within the contrastive learning paradigm. Sentences from different inputs serve as negative examples. The contrastive objective pulls positive pairs together in embedding space while pushing negative pairs apart, teaching the model to create embeddings where semantic similarity corresponds to geometric proximity.

SimCSE comes in two variants: unsupervised (using dropout noise for positive pairs) and supervised (using NLI data for explicit positive/negative pairs). The supervised variant (sup-simcse-bert-base-uncased) achieves Spearman correlation of 83.8 on STS-B benchmark, significantly outperforming prior approaches including Sentence-BERT. For the #MeToo application, SimCSE's ability to capture semantic similarity beyond surface lexical overlap is crucial—incidents described with different vocabulary but similar underlying patterns (e.g., "My boss made inappropriate

advances" vs. "A senior colleague sexually harassed me") should have high similarity despite word differences. This semantic understanding enables the retrieval system to identify relevant precedents even when surface linguistic features differ substantially.

2.3.3 Cosine Similarity for Retrieval

Given sentence embeddings, cosine similarity provides a geometric measure of semantic similarity:

$$\text{sim}(u, v) = \frac{u \cdot v}{\|u\|\|v\|} = \frac{\sum_{i=1}^{d} u_i v_i}{\sqrt{\sum_{i=1}^{d} u_i^2}\sqrt{\sum_{i=1}^{d} v_i^2}} \tag{2.3}$$

Cosine similarity ranges from -1 (opposite directions) to 1 (same direction), with 0 indicating orthogonality. For normalized embeddings (as produced by SimCSE), cosine similarity simplifies to dot product, enabling efficient retrieval using optimized linear algebra libraries or approximate nearest neighbor search algorithms like FAISS, developed by Johnson et al. [28], for very large corpora. For the recommendation system, cosine similarity between a new incident's embedding and all knowledge base embeddings identifies the top-k most similar historical cases, whose outcomes are then aggregated to provide evidence-based predictions.

2.4 Recommendation Systems and Case-Based Reasoning

Recommendation systems aim to suggest items (products, content, actions) that are relevant and useful to users based on available information. While most recommendation research focuses on commercial applications (product recommendations, content filtering), the principles of matching user needs to relevant information apply to sensitive domains like harassment guidance.

2.4.1 Collaborative Filtering and Content-Based Approaches

Collaborative filtering recommends items by finding users with similar preferences or items with similar rating patterns. Matrix factorization techniques like SVD decompose user-item interaction matrices to discover latent factors representing underlying preference dimensions. Neural collaborative filtering, introduced by He et al. [29], uses deep learning to learn non-linear user-item interactions, moving beyond the linear assumptions of matrix factorization. Content-based filtering recommends items similar to those a user has liked previously, using item features (genre, keywords, attributes) to measure similarity. Hybrid approaches combine collaborative and content-based methods to address limitations of each, particularly cold-start problems where new users or items lack interaction history, and sparsity where the user-item interaction matrix is extremely sparse.

For harassment guidance, the recommendation task differs fundamentally from commercial domains: there is no repeated user interaction to build preference profiles from, no explicit ratings or preferences to learn from, and recommendations must

prioritize safety and appropriateness over mere relevance or predicted satisfaction. Users are not seeking content they will "enjoy" but rather information that will help them navigate potentially dangerous situations. These constraints motivate alternative approaches grounded in case-based reasoning rather than traditional collaborative or content-based filtering.

2.4.2 Case-Based Reasoning

Case-Based Reasoning (CBR), formalized by Aamodt and Plaza [30], offers a more fitting framework for harassment guidance than conventional recommendation algorithms. CBR addresses novel problems by locating and modifying solutions from analogous historical cases, reflecting how human specialists frequently draw on previous experiences when making decisions. The CBR cycle consists of four key phases as summarized in Table 2.3: retrieve the most similar case(s) from the case base using similarity metrics that compare the new problem to stored cases; reuse by adapting the solution from the retrieved case to the new problem, accounting for differences between the cases; revise by evaluating and refining the proposed solution, potentially through simulation, expert review, or real-world testing; and retain by storing the new case and its solution for future use, continuously expanding the case base with experience.

Table 2.3: The four phases of the Case-Based Reasoning (CBR) cycle.

CBR Phase	**Description**
Retrieve	Find the most similar case(s) from the case base using similarity metrics that compare the new problem to stored cases based on relevant features.
Reuse	Adapt the solution from the retrieved case(s) to address the new problem, accounting for differences between the historical case(s) and the current situation.
Revise	Evaluate and refine the proposed solution through simulation, expert review, real-world testing, or user feedback to ensure appropriateness and effectiveness.
Retain	Store the new case and its solution in the case base for future use, continuously expanding the knowledge base with accumulated experience.

CBR is particularly appropriate for domains where past cases provide valuable precedents and where explicit rule-based systems are difficult to construct due to problem complexity and variability. In healthcare, Bichindaritz and Marling [31] demonstrated that CBR is well-suited to medical domains where case histories inform current practice and experience plays a major role in acquiring clinical knowledge. Begum et al. [32] surveyed CBR applications in health sciences, showing successful implementations in diagnosis support, treatment planning, and patient monitoring where historical cases guide clinical decision-making. In the legal domain, Ashley [33] established that CBR can improve legal expert systems' abilities to reason about statutory predicates and explain results, noting that lawyers inherently use case-based reasoning when advocating outcomes based on legal precedents.

Workplace harassment exhibits characteristics well-suited to CBR: outcomes depend on numerous contextual factors including organizational culture, power dynamics, available resources, legal jurisdiction, and individual circumstances; no simple rules can reliably predict what will happen when someone reports harassment or takes other actions; yet clear patterns exist in past experiences that can inform current decisions. The narrative nature of #MeToo data, where individuals describe both their situations and outcomes, provides an ideal case base for CBR.

The similarity-based retrieval component of this research directly implements the "retrieve" phase of CBR, using semantic similarity computed from neural sentence embeddings rather than hand-crafted feature-based similarity metrics. This approach enables the system to recognize similar cases based on deep semantic understanding rather than requiring manual feature engineering. The "reuse" phase adapts retrieved outcomes into personalized recommendations by aggregating outcome statistics across multiple similar cases and considering the specific context of the new incident. "Revise" is implemented through professional evaluation ensuring recommendations are appropriate, safe, and aligned with expert judgment. "Retain" occurs as the system could be updated with new cases over time, though for this research the case base remains static to ensure reproducibility and controlled evaluation.

2.4.3 Recommendation Systems for Sensitive Domains

Research on recommendation systems for sensitive domains like mental health, legal advice, and medical decision-making has identified several design principles that differ substantially from commercial recommendation system design. Safety must be the paramount concern: recommendations must never suggest actions that could harm users, either physically or by exacerbating their situations. Transparency is essential so users understand why recommendations are made and what evidence supports them, enabling informed evaluation rather than blind trust. Clear disclaimers communicate that the system is informational rather than prescriptive, and that professional consultation should supplement rather than be replaced by algorithmic guidance. Professional validation through expert review ensures appropriateness and safety before deployment. The system should empower autonomous decision-making rather than prescribing specific actions, providing information to support users' own choices rather than dictating what they should do. Presenting multiple options rather than a single "best" recommendation acknowledges that different approaches may be appropriate for different individuals and circumstances. Context awareness enables adaptation to individual circumstances rather than one-size-fits-all advice, recognizing that the "best" action for one person in one context may be inappropriate for another.

The proposed system incorporates these principles through several mechanisms: multi-option recommendations present diverse possibilities rather than prescribing a single course of action; prominent disclaimers on all outputs emphasize the informational nature of guidance and the importance of professional consultation; professional validation ensures that recommendations meet expert standards for safety and appropriateness; and evidence-based presentation of outcome statistics enables users to make informed decisions based on aggregated experiences rather than following algorithmic directives. This design philosophy prioritizes user autonomy and

safety over optimizing for any single metric like accuracy or user satisfaction.

2.5 Applied Machine Learning for Social Impact

The application of machine learning to address social challenges such as poverty, health disparities, education, climate change, and human rights violations has emerged as an important research area bridging computer science, social science, and public policy. For workplace harassment, ML offers potential to amplify victim support and inform prevention efforts, but also raises important ethical considerations about appropriate use of sensitive data and algorithmic decision-making in high-stakes contexts.

2.5.1 Applications and Opportunities

Applied machine learning for social impact encompasses diverse applications across multiple domains. In public health, ML models predict disease outbreaks from surveillance data, optimize resource allocation for vaccination campaigns, and personalize health interventions based on individual risk factors and behavioral patterns. In education, adaptive learning systems tailor instruction to individual needs, automatically detecting when students struggle and providing targeted support. In criminal justice, prediction models inform bail and sentencing decisions, though this application remains highly controversial due to concerns about bias and fairness. In humanitarian response, ML assists in disaster response coordination through automated damage assessment from satellite imagery, refugee resettlement optimization, and aid distribution planning.

For workplace harassment specifically, machine learning applications include analyzing complaint data to identify systemic patterns and high-risk departments where harassment rates exceed organizational norms; detecting harassment in online workplace communications through automated monitoring of platforms like Slack or email, though this raises significant privacy concerns; predicting which organizations or industries face elevated harassment risk based on structural factors like gender composition, industry norms, and organizational policies; personalizing training interventions based on learner characteristics to improve effectiveness of harassment prevention education; and supporting victims through evidence-based guidance as explored in this research. These applications demonstrate ML's potential to address harassment at multiple levels from individual support to organizational and systemic intervention.

2.5.2 Ethical Challenges and Responsible AI

Machine learning systems for sensitive social domains raise significant ethical concerns that require careful consideration and mitigation. Bias and fairness concerns arise because ML models can perpetuate or amplify societal biases present in training data. For harassment applications, this could mean systems that underestimate risks for certain demographic groups whose experiences are underrepresented in training data, prioritize complaints from privileged individuals whose language patterns better match training examples, or fail to recognize culturally specific forms of harassment that differ from dominant patterns. Addressing these concerns requires

careful attention to training data composition, fairness metrics, and testing across diverse demographic groups.

Privacy considerations are paramount when ML systems require large datasets that may contain sensitive personal information about harassment experiences, perpetrators, and organizational responses. Balancing data utility with privacy protection requires techniques like de-identification, differential privacy, secure multi-party computation, and careful access controls. Users must be fully informed about what data is collected, how it is used, who has access, and what protections are in place.

Accountability challenges arise when ML systems make recommendations that lead to harmful outcomes. Determining responsibility, whether it lies with developers, deploying organizations, or users who followed algorithmic guidance, is legally and ethically complex. Clear documentation of system capabilities and limitations, transparency about algorithmic decision-making processes, and explicit disclaimers about the informational nature of outputs are essential for appropriate accountability.

Transparency and explainability become critical for sensitive applications where users need to understand why systems make specific recommendations. Black-box models that cannot explain their reasoning are problematic when individuals are making potentially life-altering decisions based on algorithmic guidance. Even when models are not fully interpretable, systems should provide evidence-based justifications (e.g., "this recommendation is based on 127 similar cases where...") that support informed evaluation.

The risk of displacing professional judgment is substantial. ML systems should augment rather than replace human expertise in sensitive domains. Over-reliance on algorithmic recommendations can lead to abdication of professional responsibility, where decision-makers defer to systems without exercising independent judgment. Clear positioning of ML as decision support rather than decision-making, and maintaining human oversight and final authority, helps mitigate this risk.

Consent and autonomy require that users understand what data is being collected, how it is used, and have meaningful control over their participation. For systems using publicly shared social media data, while terms of service typically permit research use, ethical practice still demands respect for user intent and protection against harmful uses of data shared for peer support rather than algorithmic training.

The proposed system addresses these concerns through several mechanisms: using publicly shared (implicitly consented) data from individuals who chose to share experiences on social media; removing personally identifiable information to protect privacy while maintaining analytical utility; providing transparent evidence-based reasoning for recommendations by showing aggregate statistics from similar cases; including prominent disclaimers about limitations and the informational rather than prescriptive nature of guidance; conducting professional validation to ensure guidance meets expert standards and detect potential harms or biases; and emphasizing user autonomy in decision-making by presenting options and evidence rather than directives. This multi-layered approach to ethical considerations reflects recognition that technical excellence alone is insufficient for responsible deployment of ML in sensitive social contexts.

2.6 Clustering and Unsupervised Learning

Clustering algorithms automatically group similar data points without labeled supervision, revealing natural structure in data. For #MeToo narratives embedded in high-dimensional semantic space, clustering can discover coherent groups of similar experiences, validate that the embedding captures meaningful patterns, and enable silver labeling where cluster members share common characteristics.

2.6.1 Density-Based Clustering: HDBSCAN

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), developed by McInnes et al. [34], builds upon DBSCAN by accommodating clusters with differing densities and autonomously establishing cluster counts. Unlike k-means which requires specifying cluster count in advance and assumes spherical clusters of similar size, HDBSCAN automatically determines cluster count based on data structure, handles clusters of arbitrary shape including elongated or irregular geometries, identifies noise points that do not belong to any cluster rather than forcing all points into clusters, provides hierarchical cluster structure enabling analysis at multiple scales of granularity, and assigns cluster membership probabilities rather than hard assignments, indicating confidence in cluster membership.
HDBSCAN has been successfully applied to high-dimensional embeddings for document clustering, showing superior performance compared to k-means for text data where clusters may have complex shapes and varying densities. The algorithm's ability to identify noise points is particularly valuable for social media data where some tweets may be off-topic, spam, or otherwise not representative of coherent thematic clusters.
For this research, applying HDBSCAN to SimCSE embeddings of #MeToo tweets discovers natural groupings of similar experiences. High-coherence clusters where members are semantically similar can be assigned silver labels (weakly supervised labels inferred from cluster membership rather than manual annotation), improving data quality and enabling cluster-level outcome analysis. The hierarchical structure also enables exploration of how harassment experiences can be grouped at different levels of granularity, from very specific incident types to broader thematic categories.

2.7 Research Gaps

The literature review reveals several research gaps at the intersection of #MeToo analysis, natural language processing, and evidence-based recommendation systems that this study addresses. First, while prior analyses of workplace harassment data typically use single-category classification (labeling incidents as power harassment OR retaliation OR reporting failure), this approach fails to capture the multi-faceted nature of real incidents where multiple vulnerability factors frequently co-occur. For example, a harassment case might simultaneously involve power imbalance (supervisor-subordinate relationship), inability to report (lack of HR department), and career consequences (fear of professional retaliation). No existing research has developed multi-label classification specifically for workplace harassment narratives that recognizes and models this complexity, though multi-label approaches have

been successfully applied in other domains like medical diagnosis where patients present with multiple concurrent conditions.

Second, while sociological research has qualitatively analyzed #MeToo narratives to identify patterns such as institutional failure and victim silencing, quantitative analysis linking specific vulnerability combinations to outcome probabilities remains limited. Existing work identifies that certain factors correlate with negative outcomes, but does not provide statistical evidence about outcome likelihoods for particular vulnerability profiles. This gap prevents translation of descriptive sociological findings into actionable probabilistic predictions that could inform individual decision-making about whether to report harassment, what support to seek, or what outcomes to anticipate.

Third, case-based reasoning and recommendation systems have been successfully applied in various sensitive domains including medical diagnosis, legal case retrieval, and therapeutic intervention selection. However, no prior work has developed CBR specifically for workplace harassment guidance using social media narratives as the case base. The unique challenges of this application, including ensuring safety when recommendations could affect physical security or economic stability, validating appropriateness with domain professionals rather than just technical metrics, handling extreme outcome variability where similar situations sometimes lead to very different results, and respecting the sensitive nature of trauma narratives, have not been addressed in a unified framework. Existing CBR research focuses on domains like medical diagnosis where case databases are curated by professionals, not crowdsourced from social media.

Fourth, while multiple studies compare sentiment analysis approaches on product reviews or general Twitter data, systematic comparison of VADER versus BERT specifically for #MeToo harassment narratives has not been conducted. Given the unique linguistic characteristics of harassment narratives, including mixing emotional expression with factual description, using euphemisms and coded language to discuss sensitive experiences, and combining personal narrative with political commentary, domain-specific evaluation is necessary to determine which approaches perform best for this application. Existing sentiment analysis research on social media does not address these specific linguistic features.

Fifth, most machine learning for social good research evaluates systems using standard technical metrics like accuracy, F1 score, or mean squared error, but lacks validation by practitioners assessing whether recommendations are appropriate, safe, and useful in practice. For sensitive applications like harassment guidance where algorithmic errors could have serious consequences (recommending an unsafe action, failing to warn about likely retaliation, suggesting reporting when it could endanger someone), professional evaluation is essential but rarely conducted rigorously. The gap between technical performance metrics and real-world utility is particularly large for sensitive social applications where "accuracy" does not guarantee appropriateness.

Sixth, while research has identified actions individuals take in response to harassment, including reporting to authorities, seeking peer support, leaving jobs, and confronting harassers, extraction and analysis of action sequences (the temporal ordering and combinations of actions) from unstructured narratives has not been systematically performed for harassment contexts. Understanding not just what actions are taken but in what order and combination is crucial for effective guid-

ance. For example, seeking peer support before filing a formal complaint might be more effective than the reverse order, but existing research does not analyze these temporal patterns.

Finally, while general AI ethics principles address fairness, transparency, and accountability, specific frameworks for developing ML systems that analyze trauma narratives and provide guidance about potentially life-altering decisions are underdeveloped. Principles for responsible use of sensitive social media data, appropriate disclaimers that convey limitations without undermining utility, balancing informativeness with safety when some accurate information might be harmful, and ensuring user autonomy while providing meaningful guidance need systematic articulation. Existing AI ethics frameworks focus on commercial applications or high-level principles rather than concrete guidelines for sensitive social applications.

This study addresses these gaps by developing a comprehensive evidence-based workplace harassment guidance system that implements multi-label vulnerability classification recognizing co-occurring risk factors, quantitatively links vulnerability patterns to outcome probabilities based on 15,835 #MeToo cases, develops a case-based reasoning framework specifically adapted for harassment guidance with appropriate safety and validation mechanisms, systematically compares BERT versus VADER for #MeToo sentiment analysis to determine optimal approaches for this domain, conducts rigorous professional validation with HR experts and workplace safety professionals to ensure guidance meets expert standards, extracts and analyzes action sequences from unstructured narratives to understand temporal patterns in response strategies, and articulates an ethical framework for responsible development and deployment that addresses the specific challenges of analyzing trauma narratives and providing high-stakes guidance. This multi-faceted contribution advances the state of knowledge in computational social science, natural language processing for sensitive domains, and applied machine learning for social impact.

The next chapter presents the comprehensive methodology developed to address these research gaps, including the data preprocessing pipeline, multi-label vulnerability detection, dual sentiment analysis, semantic similarity-based retrieval, clustering and silver labeling, outcome prediction, action sequence extraction, and guidance generation algorithm.

Chapter 3

Methodology

This chapter presents the complete methodology of the proposed evidence-based workplace harassment guidance system. The chapter begins with a comprehensive system overview, followed by detailed descriptions of the dataset characteristics and preprocessing pipeline, the dual sentiment analysis approach using BERT and VADER, the SimCSE embedding generation for semantic similarity, the multi-label vulnerability detection framework that leverages these embeddings, the HDBSCAN clustering and silver labeling methodology (used for evaluation), the outcome prediction and action sequence extraction techniques, and finally the guidance generation algorithm. Mathematical formulations and algorithmic specifications are provided throughout to enable reproducibility.

3.1 System Overview

The proposed evidence-based guidance system addresses the challenge of providing personalized, data-driven recommendations for individuals experiencing workplace harassment by analyzing patterns extracted from thousands of #MeToo narratives. The system architecture consists of five integrated phases that transform raw social media data into actionable guidance while maintaining ethical safeguards and professional validation: (1) data collection and filtering to extract workplace-specific incidents, (2) dual sentiment analysis using both BERT and VADER, (3) SimCSE-BERT embedding generation for semantic similarity, (4) multi-label vulnerability detection identifying co-occurring risk factors using the generated embeddings, and (5) guidance generation with professional validation. Additionally, HDBSCAN clustering is employed separately for evaluation purposes and silver label generation, but is not part of the real-time guidance pipeline. Figure 3.1 illustrates the complete system architecture with key metrics at each stage.

The system begins with data collection and filtering, downloading 40,944 #MeToo tweets and applying workplace-specific filtering using keyword patterns and contextual analysis to identify 15,835 incidents that describe workplace harassment rather than other forms of harassment, solidarity statements, or news commentary. This initial step ensures the knowledge base contains relevant experiences for workplace-specific guidance. Following data collection, dual sentiment analysis applies both BERT-based sentiment classification (cardiffnlp/twitter-roberta-base-sentiment, achieving 98.1% accuracy) and VADER lexicon-based analysis (73.4% accuracy) to each tweet. BERT provides high-accuracy positive/negative/neutral

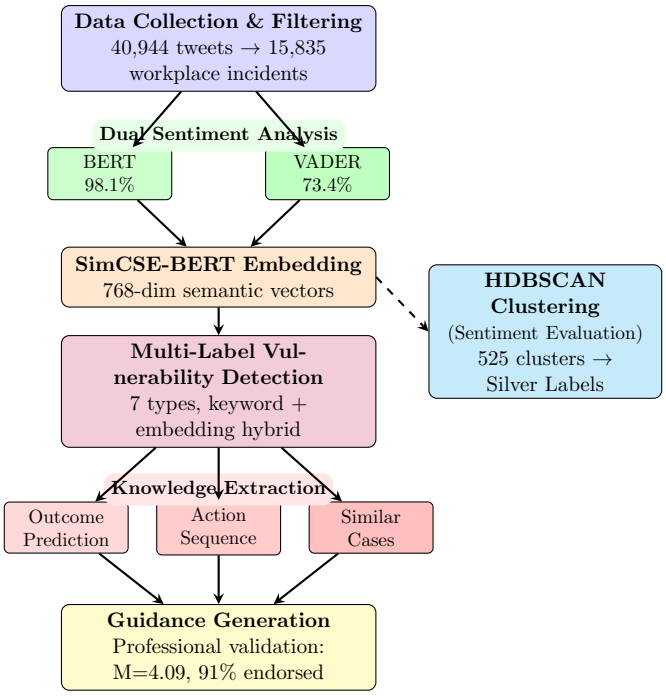


Figure 3.1: System architecture overview

classification, while VADER's compound scores enable fine-grained polarity detection useful for distinguishing emotional expression from outcome descriptions.
Subsequently, SimCSE-BERT embedding generation encodes each incident into a 768-dimensional dense vector using princeton-nlp/sup-simcse-bert-base-uncased. These embeddings capture semantic meaning beyond surface keywords, enabling similarity-based retrieval where incidents described with different vocabulary but similar underlying patterns are correctly identified as related. The system then identifies up to seven co-occurring vulnerability types per incident using a hybrid multi-label detection approach that leverages these embeddings. Keyword-based rules provide high-precision detection for clear patterns, while embedding similarity (threshold 0.55) enables flexible detection of subtle or complex cases. This multi-label approach recognizes that 68% of workplace harassment involves multiple simultaneous vulnerability factors.
For evaluation purposes, HDBSCAN clustering applies hierarchical density-based clustering to discover 525 natural groupings of similar experiences. Silver labels are generated through majority voting within high-coherence clusters, producing 11,066 labeled examples that serve as ground truth for evaluating BERT and VADER sentiment analysis accuracy. This clustering-based evaluation approach eliminates the need for manual annotation of the entire dataset. From the analyzed data, the system extracts outcome patterns linking vulnerabilities to 15 outcome types, action sequences comprising 8 distinct action categories, and maintains similarity indices for case-based retrieval. Finally, guidance generation synthesizes all components to produce personalized, context-aware guidance. When a new incident is presented, the system identifies vulnerabilities, retrieves similar cases, aggregates their outcomes, extracts effective action sequences, and generates category-organized guidance covering immediate safety, documentation, reporting, support, and legal considerations. Professional evaluation by practitioners from HR, legal, and mental health

fields achieved M=4.09 appropriateness rating and 91% endorsement rate. This integrated architecture ensures guidance is simultaneously data-driven (grounded in 15,835 real experiences), context-aware (adapted to specific vulnerability combinations), and professionally validated (meeting expert standards for appropriateness and safety).

3.2 Dataset Description and Collection

3.2.1 Data Source

This research utilizes the publicly available Harvard Dataverse #MeToo Twitter dataset [5], which contains tweet IDs collected between October 2017 and December 2020, spanning the movement's viral emergence and sustained engagement period. The dataset comprises 40,944 English-language tweets that used the #MeToo hashtag, representing a diverse sample of experiences, locations, and contexts. Following best practices for ethical data collection, tweet content was retrieved using the Twitter API (v1.1) with rate-limiting compliance, and stored with metadata including timestamp, engagement metrics (retweets, likes), and anonymized location information when available.

The dataset represents user-generated content voluntarily shared on a public social media platform. No private or protected accounts were accessed, and personally identifiable information (usernames, profile details, exact locations) was removed during preprocessing to protect privacy while preserving the narrative content necessary for analysis.

3.2.2 Workplace-Specific Filtering

The full 40,944 tweet corpus contains diverse content including workplace harassment narratives, non-workplace harassment experiences such as domestic violence, assault by strangers, and harassment in educational settings, expressions of solidarity without personal narratives, commentary on news events and public figures, and broader discussions of gender equality and social change. For the purpose of building an evidence-based workplace harassment guidance system, only workplace-specific incidents are relevant, necessitating systematic filtering to identify the appropriate subset of experiences.

Automated filtering identifies workplace incidents using a multi-stage approach that balances recall and precision. The first stage implements keyword presence detection, requiring at least one workplace-indicative term from a comprehensive set including work, job, office, colleague, coworker, boss, manager, supervisor, employee, employer, workplace, professional, career, promotion, HR, human resources, department, meeting, business, company, corporate, and client. This broad initial filter achieves high recall by capturing most workplace incidents, though at the cost of some false positives where workplace terms appear in non-harassment contexts.

The second stage applies contextual validation to reduce false positives through additional filtering rules. Tweets must contain both workplace terminology and harassment-indicative language such as harassment, harassed, harassing, inappropriate, unwanted, assault, assaulted, groped, touched, advances, proposition, comment,

remarks, discriminate, discrimination, hostile, uncomfortable, threatened, fear, report, reported, and complain. This conjunction requirement ensures the tweet describes workplace harassment rather than merely mentioning work in passing or discussing workplace topics without describing harassment experiences.
After filtering, 15,835 tweets remain, constituting the workplace harassment knowledge base. This corpus represents diverse industries including technology, healthcare, education, finance, entertainment, retail, and manufacturing, organizational sizes ranging from startups to large corporations, geographic locations predominantly from the United States and Europe with some representation from Asia, Latin America, and Africa, and varied demographic backgrounds of tweet authors inferred from gender and age markers when present, though comprehensive demographic information is not uniformly available. Table 3.1 summarizes key statistics of the final workplace harassment dataset.

Table 3.1: Dataset statistics for workplace-specific #MeToo tweets

Characteristic	**Value**
Total tweets (original corpus)	40,944
Workplace-specific tweets	15,835
Date range	Oct 2017 – Dec 2020
Average tweet length	187 characters
Median tweet length	176 characters
Tweets with hashtags (beyond #MeToo)	62.3%
Tweets with mentions (@username)	31.7%
Tweets with URLs	18.4%
Average engagement (RT + likes)	12.3

3.3 Data Preprocessing Pipeline

Raw #MeToo tweets require systematic preprocessing to remove noise and normalize variations while preserving semantic meaning and emotional content. Algorithm 1 formalizes the complete preprocessing procedure.
The pipeline implements targeted transformations: URLs are replaced with [URL] tokens to preserve contextual signals while removing unresolvable content; @mentions become [USER] tokens for privacy protection; hashtag symbols are removed while retaining the words (e.g., #MeToo becomes MeToo) since hashtags like #NotOkay and #BelieveWomen carry semantic value. Punctuation normalization reduces excessive sequences ("!!!!!!" to "!") while preserving emotional intensity cues, and NFKC unicode normalization ensures consistent character encoding across the corpus.
Case handling varies by downstream component: original casing is retained for BERT-based models that leverage capitalization for proper nouns, acronyms, and emphasis, while a parallel lowercased version is maintained for VADER and keyword-based detection. Notably, stopwords and punctuation are deliberately retained rather than removed as in conventional NLP workflows. Phrases like "I was" and "he said" contain grammatical structure essential for understanding narrative perspective and agency, and punctuation marks convey paralinguistic cues including

emotional intensity (exclamation marks) and direct speech attribution (quotation marks). BERT's attention mechanisms effectively determine what information to utilize from full text, making pre-emptive removal unnecessary and potentially harmful.

Algorithm 1 Tweet Preprocessing Pipeline

Require: Raw tweet text t
Ensure: Preprocessed text t'
$t \leftarrow \text{ReplaceURLs}(t, [\text{URL}])$ ▷ Replace URLs with placeholder token
$t \leftarrow \text{ReplaceUsernames}(t, [\text{USER}])$ ▷ Replace @mentions for privacy
$t \leftarrow \text{RemoveHashSymbols}(t)$ ▷ Remove # but keep word
$t \leftarrow \text{NormalizeUnicode}(t, \text{NFKC})$ ▷ Standardize unicode encoding
$t \leftarrow \text{NormalizePunctuation}(t)$ ▷ Reduce excessive punctuation
$t \leftarrow \text{CollapseWhitespace}(t)$ ▷ Multiple spaces to single space
$t' \leftarrow \text{Strip}(t)$ ▷ Remove leading/trailing whitespace
return t'

This conservative approach preserves original content while removing only clear noise, ensuring reproducibility across the 15,835-tweet corpus.

3.4 Dual Sentiment Analysis

Sentiment analysis serves dual purposes in this system: (1) identifying the emotional valence of narratives to inform empathetic response and understand psychological impact, and (2) distinguishing outcome descriptions from emotional expressions to improve guidance relevance. The system employs both BERT-based and VADER-based sentiment analysis to leverage complementary strengths.

3.4.1 BERT-Based Sentiment Classification

The BERT-based sentiment classifier uses the cardiffnlp/twitter-roberta-base-sentiment model [21], a RoBERTa-base architecture fine-tuned on the TweetEval benchmark specifically for Twitter sentiment analysis. The model classifies tweets into three categories: positive, negative, and neutral.

Architecture and Processing

The model processes tweets through a multi-stage transformation pipeline beginning with tokenization. The RoBERTa tokenizer segments text into subword tokens using Byte-Pair Encoding (BPE) with a vocabulary size of 50,265 tokens. Special tokens [CLS] and [SEP] mark sequence boundaries to indicate the start and end of each input. The maximum sequence length is set to 128 tokens, with longer tweets being truncated to fit this constraint while retaining the most informative initial content. Following tokenization, the embedding layer maps each token to a 768-dimensional dense vector representation. These token embeddings are summed with positional embeddings that encode each token's position in the sequence, enabling the model to capture word order and sequential dependencies despite the inherently position-invariant nature of attention mechanisms.

The transformer encoder consists of twelve stacked transformer layers, each containing 12 attention heads that process the embedded sequence in parallel. Each layer computes multi-head self-attention followed by position-wise feed-forward transformations:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.1}$$

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{3.2}$$

where the attention mechanism allows each token to attend to all other tokens in the sequence, learning contextual representations that capture semantic relationships and dependencies.
Finally, the classification head extracts the [CLS] token's final layer representation, which serves as an aggregated sentence-level encoding, and passes it through a linear classifier with softmax activation:

$$p = \text{softmax}(W_c h_{[CLS]} + b_c) \tag{3.3}$$

producing a probability distribution over the three sentiment categories: positive, negative, and neutral.

Prediction and Confidence

For each tweet, the system records three complementary pieces of information to capture both the categorical classification and the model's uncertainty. The predicted sentiment label is determined as $\hat{y} = \arg\max_c p_c$, selecting the category with highest probability. The confidence score $\max_c p_c$ quantifies the model's certainty in this prediction, with values near 1.0 indicating high confidence and values near 0.33 suggesting high uncertainty across the three categories. Additionally, the full probability distribution $(p_{pos}, p_{neg}, p_{neu})$ is retained to preserve nuanced sentiment information.
The full probability distribution provides valuable nuanced information beyond the single categorical label. For instance, a tweet with probability distribution $(0.48, 0.47, 0.05)$ would be classified as positive based on the maximum probability, but the near-equal positive and negative probabilities reveal mixed or ambivalent sentiment that might merit special attention. In contrast, a distribution like $(0.92, 0.05, 0.03)$ indicates unambiguous positive sentiment with high model confidence, representing a clearer case for classification purposes.

3.4.2 VADER Sentiment Analysis

VADER (Valence Aware Dictionary and sEntiment Reasoner) [23] provides lexicon-based sentiment scoring specifically tuned for social media text. Unlike BERT's categorical classification, VADER produces continuous compound scores ranging from -1 (most negative) to +1 (most positive).

VADER Scoring Algorithm

VADER's scoring process operates through three sequential stages that combine lexicon-based lookups with context-sensitive adjustments. The first stage performs

lexicon lookup, assigning valence scores to each word based on a pre-built lexicon of over 7,500 sentiment-bearing words. Each lexicon entry contains a manually validated sentiment intensity score, with examples including "horrible" scored at -2.5 for strong negative sentiment and "wonderful" scored at 2.6 for strong positive sentiment. Words not appearing in the lexicon receive a neutral score of zero.
The second stage applies context-based adjustments that modify the lexicon scores based on surrounding linguistic features. Negation words like "not" reverse the polarity of subsequent sentiment words and dampen their intensity, so "not good" receives a negative score rather than positive. Intensifiers such as "very" and "extremely" amplify the magnitude of adjacent sentiment words, increasing their absolute value. Punctuation contributes additional emphasis, with exclamation marks in phrases like "great!" adding to positive intensity. Capitalization intensifies sentiment, so "AMAZING" receives a higher magnitude score than "amazing." Finally, emoticons and emoji contribute explicit sentiment signals, with :) adding positive valence and :(adding negative valence regardless of the surrounding text.
The third stage aggregates all adjusted word scores and normalizes the sum to produce a compound score in the range [-1, 1] using the formula:

$$\text{compound} = \frac{s}{\sqrt{s^2 + \alpha}} \tag{3.4}$$

where s represents the sum of adjusted valence scores and $\alpha = 15$ serves as a normalization constant calibrated to produce appropriate score distributions. This normalization ensures that compound scores remain bounded and comparable across tweets of varying lengths. Additionally, VADER produces separate positive, negative, and neutral proportion scores indicating what fraction of the text expresses each sentiment type, providing a more detailed sentiment profile beyond the single compound score.

Classification from Compound Scores

To enable comparison with BERT, VADER compound scores are mapped to categorical labels using thresholds:

$$\text{label} = \begin{cases} \text{positive} & \text{if compound} > 0.20 \\ \text{negative} & \text{if compound} < -0.10 \\ \text{neutral} & \text{otherwise} \end{cases} \tag{3.5}$$

These thresholds were empirically tuned for the #MeToo dataset, with an asymmetric configuration that accounts for the predominantly negative nature of harassment narratives. The higher positive threshold (0.20) reduces false positives by requiring stronger positive signals, while the lower negative threshold (-0.10) captures the nuanced negative expressions common in trauma-related discourse.

3.4.3 Comparative Analysis

The dual sentiment approach leverages complementary strengths of both transformer-based and lexicon-based methods. Table 4.7 presents a comparative analysis of BERT and VADER performance and characteristics.

Table 3.2: Comparative analysis of BERT and VADER sentiment analysis methods

Characteristic	**BERT**	**VADER**
Accuracy on #MeToo dataset	98.1%	73.4%
Output format	Categorical (pos/neg/neu)	Continuous score (-1 to +1)
Model type	Transformer-based	Lexicon-based
Processing speed (tweets/sec)	150-200 (GPU)	5,000+ (CPU)
Interpretability	Low	High
Context sensitivity	High	Moderate
Intensity discrimination	Coarse	Fine-grained
Hardware requirements	GPU recommended	CPU sufficient
Primary use case	Main classifier	Intensity analysis

BERT achieves 98.1% accuracy on the #MeToo dataset, significantly outperforming VADER's 73.4% accuracy, thereby validating BERT as the primary sentiment classifier for categorical classification tasks. However, VADER provides distinct value in specific analytical contexts that justify maintaining both approaches in the system architecture.

VADER's continuous compound scores enable fine-grained intensity discrimination, distinguishing strongly negative sentiment (compound < -0.6) from mildly negative sentiment ($-0.3 <$ compound < -0.05). This granular intensity measurement proves useful for prioritizing severe cases requiring urgent attention and for identifying subtle gradations in emotional expression that categorical labels cannot capture. Additionally, VADER's lexicon-based approach provides inherent interpretability, enabling inspection of which specific words contributed to the overall score and how context adjustments modified those contributions. This transparency supports debugging, error analysis, and building user trust through explainable sentiment assessments.

Computational efficiency represents another advantage of VADER, which processes thousands of tweets per second on standard CPU hardware without requiring specialized GPU acceleration. This efficiency enables rapid exploratory analysis during development and supports real-time applications where GPU access may be unavailable or cost-prohibitive. Furthermore, cases where BERT and VADER produce strongly divergent sentiment predictions often indicate complex linguistic patterns such as sarcasm, mixed sentiment, euphemistic language, or subtle contextual nuances that warrant special attention and manual review.

The system architecture stores both BERT and VADER scores for all tweets, defaulting to BERT predictions for primary sentiment classification while making VADER scores available for specialized analyses requiring intensity discrimination, interpretability, or computational efficiency. This dual-method approach provides robustness and flexibility for diverse analytical needs.

3.5 Semantic Similarity and Embedding Generation

Retrieving similar historical cases from the knowledge base requires encoding incidents into a semantic vector space where similar experiences have similar vectors. This section describes the SimCSE-BERT embedding approach that enables high-quality similarity-based retrieval and serves as the foundation for subsequent vulnerability detection.

3.5.1 SimCSE Model

SimCSE (Simple Contrastive Learning of Sentence Embeddings) [8] represents the state-of-the-art in sentence embedding, achieving superior performance on semantic textual similarity benchmarks through contrastive learning. The system uses princeton-nlp/sup-simcse-bert-base-uncased, a supervised variant trained on natural language inference data.

Embedding Generation Process

The embedding generation process transforms each preprocessed tweet into a dense vector representation through a four-stage pipeline. First, tokenization applies BERT's wordpiece tokenizer to segment the text into subword units, adding special tokens [CLS] at the beginning and [SEP] at the end to mark sequence boundaries. The maximum sequence length is set to 128 tokens, which accommodates the vast majority of tweets while truncating the small number of exceptionally long tweets to retain their most informative initial content.
Second, the encoding stage passes the tokenized sequence through 12 stacked transformer layers, each with 768-dimensional hidden states and 12 parallel attention heads. These transformer layers progressively build contextualized representations where each token's embedding captures not just its lexical meaning but also its semantic role within the broader narrative context. Third, the pooling stage extracts the [CLS] token's final layer representation as the sentence-level embedding:

$$e_t = h^{(12)}_{[CLS]} \in \mathbb{R}^{768} \tag{3.6}$$

where the [CLS] token has attended to all other tokens throughout the encoding process, producing an aggregated representation of the entire tweet's semantic content.
Fourth, L2-normalization scales the embedding to unit length, enabling efficient cosine similarity computation via simple dot products:

$$e_t \leftarrow \frac{e_t}{\|e_t\|} \tag{3.7}$$

This normalization ensures that similarity computations focus on angular distance between embeddings rather than magnitude differences. The resulting 768-dimensional normalized vectors capture semantic meaning such that semantically similar tweets achieve high cosine similarity regardless of vocabulary overlap, enabling retrieval of conceptually related incidents even when described using different words.

3.5.2 Similarity Computation

Given a new incident with embedding e_{new} and the complete knowledge base containing embeddings $\{e_1, e_2, \ldots, e_N\}$ where $N = 15,835$ represents all workplace harassment tweets, cosine similarities are computed between the new incident and each historical case:

$$\text{sim}_i = e_{new} \cdot e_i \tag{3.8}$$

Since all embeddings are L2-normalized to unit length, this dot product equals the cosine similarity and can be computed extremely efficiently using optimized BLAS (Basic Linear Algebra Subprograms) routines that leverage vectorized CPU instructions or GPU parallelization. The top-k most similar tweets are then retrieved by sorting the similarity scores and selecting the highest-scoring cases:

$$\text{TopK} = \{\text{argsort}(\{\text{sim}_i\})[-k:]\} \tag{3.9}$$

The default value $k = 10$ balances having sufficient cases to compute robust outcome statistics and identify consistent patterns, while maintaining relevance by avoiding inclusion of increasingly dissimilar cases that would dilute the specificity of guidance. Larger values of k risk including less relevant cases that share only superficial similarities, while smaller values provide insufficient statistical power for reliable outcome prediction.

3.5.3 Embedding Space Validation

To validate that SimCSE embeddings capture meaningful semantic structure for #MeToo narratives, qualitative analysis examines nearest neighbors for sample tweets. For example, the tweet "My boss kept making sexual comments despite me asking him to stop" retrieves similar cases describing supervisor misconduct with ignored boundaries, confirming semantic alignment. Quantitative validation using t-SNE visualization shows that tweets sharing vulnerability types cluster together in the embedding space, providing further evidence of meaningful structure.

3.6 Clustering and Silver Labeling

The primary purpose of clustering is to generate silver labels through majority voting, which serve as ground truth for evaluating the BERT and VADER sentiment analysis results. Since manual annotation of 15,835 tweets is impractical, silver labeling leverages the assumption that semantically similar tweets clustered together likely share sentiment characteristics. This section describes the HDBSCAN clustering methodology and the majority voting-based label inference process used for sentiment analysis evaluation.

3.6.1 HDBSCAN Clustering

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [34] autonomously identifies clusters exhibiting diverse density levels within

high-dimensional spaces, eliminating the need to predetermine the number of clusters. Applied to the 15,835 SimCSE embeddings, HDBSCAN identifies coherent groupings of similar #MeToo experiences.

Algorithm Overview

HDBSCAN operates through three key steps. First, the algorithm computes a mutual reachability graph where distances between points are adjusted for local density:

$$d_{mreach}(a, b) = \max\{\text{core}_k(a), \text{core}_k(b), d(a, b)\} \tag{3.10}$$

where $\text{core}_k(a)$ is the distance to the k-th nearest neighbor of point a. This makes distances more robust to density variations.
Second, hierarchy construction builds a minimum spanning tree over mutual reachability distances and progressively merges points into clusters starting from the densest regions.
Third, cluster extraction analyzes cluster stability across the hierarchy to extract persistent clusters, automatically determining their number based on data structure.

Hyperparameter Selection

HDBSCAN requires careful selection of hyperparameters that control cluster formation and stability. The minimum cluster size parameter determines the minimum number of points required to form a valid cluster, set to 15 to ensure clusters are substantial enough for meaningful statistical analysis while still allowing discovery of smaller but coherent experiential groups. The minimum samples parameter controls how conservative the clustering algorithm behaves, with larger values producing fewer and more conservative clusters. This parameter is set to 10, slightly lower than the minimum cluster size, to allow some flexibility in cluster formation while maintaining quality standards.
The distance metric parameter specifies how similarity is measured in the embedding space. Cosine distance is selected to match the similarity measure used for retrieval, ensuring consistency across the system's similarity computations. This choice is particularly appropriate for normalized embeddings where angular distance captures semantic similarity better than Euclidean distance. With these hyperparameter settings, HDBSCAN discovers 525 distinct clusters containing 14,731 tweets, representing 93% of the dataset. The remaining 1,104 tweets are labeled as noise, indicating they do not belong to any sufficiently dense or stable cluster and represent either outliers or experiences too unique to form coherent groupings.

3.6.2 Silver Labeling through Majority Voting

Silver labeling assigns sentiment labels to clusters through majority voting, providing ground truth for evaluating BERT and VADER performance without requiring manual annotation of the entire dataset. For each cluster, the dominant sentiment label (positive, negative, or neutral) is determined by majority vote across cluster members:

$$\text{label}(c) = \arg\max_{s} \text{count}(s, c) \tag{3.11}$$

where $s \in \{\text{positive}, \text{negative}, \text{neutral}\}$ represents the sentiment categories. The confidence assigned to this silver label is proportional to the sentiment's frequency within the cluster:

$$\text{conf}(c) = \frac{\text{count}(\text{label}(c), c)}{|c|} \tag{3.12}$$

To ensure label quality and avoid propagating weak patterns, only clusters achieving $\text{conf}(c) \geq 0.60$ receive silver labels, meaning at least 60% of cluster members must share the dominant sentiment. This conservative threshold prioritizes precision over coverage, accepting that some clusters will remain unlabeled rather than risking low-quality label propagation. The silver labeling process generates 11,066 high-confidence sentiment labels across 412 coherent clusters, while 113 low-coherence clusters that fail to meet the 60% threshold remain unlabeled.

These silver labels serve as the primary evaluation benchmark for comparing BERT and VADER sentiment analysis accuracy. By treating the majority-voted cluster sentiment as ground truth, the system can compute accuracy, precision, recall, and F1 scores for both sentiment analysis methods. This approach assumes that semantically similar tweets (those clustered together based on SimCSE embeddings) should share similar sentiment characteristics, making majority voting a reasonable proxy for true sentiment labels. The evaluation results demonstrate BERT achieving 98.1% accuracy and VADER achieving 73.4% accuracy against these silver labels, validating BERT as the primary sentiment classifier while confirming VADER's utility for complementary analyses.

3.7 Guidance Generation

The final component of the system synthesizes all analyzed elements into coherent, personalized guidance. This section describes the complete guidance generation pipeline, including vulnerability detection, outcome prediction, action sequence extraction, and the generation of category-organized recommendations.

3.7.1 Multi-Label Vulnerability Detection

Workplace harassment incidents typically involve multiple co-occurring vulnerability factors rather than single isolated issues. A person may experience power imbalance (harassment by supervisor), inability to report (HR dismisses complaints), and retaliation threats (fear of job loss) simultaneously. Multi-label classification recognizes this complexity by identifying all relevant vulnerability types present in an incident. The vulnerability detection system leverages the SimCSE embeddings for semantic similarity-based detection.

Vulnerability Taxonomy

Based on analysis of #MeToo narratives and workplace harassment literature, seven vulnerability types are defined that capture the most common and impactful vulnerability factors. Table 3.3 presents the complete vulnerability taxonomy with descriptions.

Table 3.3: Vulnerability taxonomy for workplace harassment incidents

Vulnerability Type	**Description**
Power imbalance	Harasser holds organizational authority over victim (supervisor, manager, senior colleague controlling career advancement)
Retaliation threat	Explicit or implicit threats of job loss, negative reviews, exclusion from opportunities, or professional consequences
Physical assault	Unwanted physical contact, groping, assault, or threats of physical harm
Quid pro quo coercion	Promises of promotion, favorable treatment, or job security offered in exchange for sexual favors or tolerance
Unable to report	Systemic barriers including inaccessible reporting mechanisms, HR dismissal, institutional protection of harassers, or victim suppression
Repeated harassment	Ongoing patterns sustained over weeks, months, or years rather than isolated incidents
Isolation/unsafe setting	Harassment occurring when victim is alone, in empty offices, after hours, or where help is unavailable

The first vulnerability type, power imbalance by supervisor or boss, captures situations where the harasser holds organizational authority over the victim, such as supervisors, managers, or senior colleagues who control career advancement opportunities. The second type, retaliation or threat to career, encompasses explicit or implicit threats of job loss, negative performance reviews, exclusion from opportunities, or other professional consequences imposed for resistance or reporting. Physical assault or physical threat represents the third vulnerability type, covering unwanted physical contact, groping, assault, or threats of physical harm.
The fourth vulnerability type, coercion or quid pro quo at work, describes situations involving promises of promotion, favorable treatment, or job security offered in exchange for sexual favors or tolerance of harassment. Unable to report or not believed, the fifth type, addresses systemic barriers including lack of accessible reporting mechanisms, HR dismissal of complaints, institutional protection of harassers, or active suppression of victims. The sixth vulnerability type captures repeated harassment over time, identifying ongoing patterns of harassment sustained over weeks, months, or years rather than isolated incidents. Finally, isolation or unsafe setting represents the seventh type, describing harassment occurring when victims are alone, in empty offices, after hours, or in situations where help is unavailable. These categories balance comprehensiveness with specificity.

Hybrid Detection Approach

The system employs a hybrid approach combining keyword-based rules (high precision for clear patterns) and embedding similarity (flexibility for subtle cases).

Keyword-Based Detection For each vulnerability type, keyword rules define patterns of words whose co-occurrence strongly indicates the presence of that vulnerability. These rules are constructed based on manual analysis of #MeToo narratives and review of workplace harassment literature to identify characteristic linguistic patterns. Power imbalance vulnerability requires one authority-indicating term such as boss, manager, supervisor, superior, senior, or director combined with one hierarchy-indicating term like subordinate, junior, employee, report, review, or promotion. Retaliation threat detection requires temporal or reporting-related terms including after, since, reported, or complained paired with consequence-related words such as fired, terminated, excluded, review, retaliation, punished, or demoted.
Physical assault vulnerability is identified through direct action terms including touched, grabbed, groped, assault, physical, forced, cornered, or pinned, which explicitly describe unwanted physical contact. Quid pro quo coercion requires career-benefit terms like promotion, raise, job, position, or opportunity combined with conditionality markers such as depends, exchange, if you, unless, or conditional. Inability to report is detected when reporting-related terms including HR, report, complain, or file appear alongside institutional failure indicators like nothing, ignored, dismissed, didn't believe, too sensitive, protect, or swept.
Repeated harassment over time is identified through temporal persistence markers including months, years, repeatedly, ongoing, continues, pattern, or keeps. Isolation in unsafe settings requires context indicators such as alone, empty, after hours, late, cornered, isolated, or remote that signal the victim's vulnerability due to physical separation from potential help or witnesses. When keyword patterns match, the vulnerability is assigned with a confidence score of 0.85, reflecting high but not perfect precision given the rule-based nature of the detection.

Embedding-Based Detection For subtler cases where harassment is described indirectly or using non-standard vocabulary that may not match explicit keyword patterns, BERT embeddings capture semantic similarity to enable flexible vulnerability detection. Each vulnerability type is represented by a small set of 3-5 canonical example phrases that exemplify prototypical expressions of that vulnerability. For power imbalance, representative examples include phrases like "My boss made inappropriate advances," "A senior manager harassed me," and "Supervisor abused his authority," which capture the essence of hierarchical abuse without requiring exact keyword matches. Similarly, inability to report is exemplified through phrases such as "HR didn't believe me," "When I reported, nothing happened," and "The company protected the harasser," representing institutional failures in handling complaints. Each vulnerability type has a corresponding set of carefully selected examples designed to capture the core semantic patterns associated with that vulnerability.
These canonical example phrases are encoded using the same BERT model (cardiffnlp/twitter-roberta-base-sentiment) that produces tweet embeddings, ensuring consistency in the semantic representation space. For a new tweet with embedding e_{tweet}, cosine similarity is computed against each vulnerability type's example embeddings. The maximum similarity across all examples for a given vulnerability determines that vulnerability's similarity score:

$$\text{sim}(v) = \max_{e_i \in \text{examples}(v)} \frac{e_{tweet} \cdot e_i}{\|e_{tweet}\| \|e_i\|} \tag{3.13}$$

If the similarity score $\text{sim}(v)$ exceeds the empirically tuned threshold $\tau = 0.55$, the vulnerability v is assigned to the tweet with confidence equal to the similarity score $\text{sim}(v)$. This threshold balances precision and recall, allowing semantically similar expressions to be detected while filtering out spurious matches. The embedding-based approach complements keyword detection by identifying vulnerabilities expressed through paraphrase, euphemism, or unconventional vocabulary that would escape rigid pattern matching.

Multi-Label Integration The final vulnerability set for a tweet is the union of keyword-detected and embedding-detected vulnerabilities:

$$V_{tweet} = V_{keyword} \cup \{v : \text{sim}(v) \geq \tau\} \tag{3.14}$$

Duplicates (same vulnerability detected by both methods) retain the higher confidence score. The multi-label output is a list of (vulnerability, confidence) tuples sorted by confidence, ensuring at least one vulnerability per tweet (selecting the highest similarity if no detections exceed threshold).

Threshold Tuning

The embedding similarity threshold τ balances precision and recall. Lower thresholds (e.g., $\tau = 0.40$) detect more vulnerabilities but increase false positives. Higher thresholds (e.g., $\tau = 0.70$) improve precision but may miss valid vulnerabilities described with unusual vocabulary.
Empirical experimentation with different threshold values showed that $\tau = 0.55$ achieves a practical balance between detection coverage and false positive rate for the hybrid detection approach. This threshold is used throughout the system.

3.7.2 Outcome Prediction

The core value of evidence-based guidance lies in predicting what outcomes are likely based on similar historical cases. This subsection describes outcome extraction and categorization.

Outcome Taxonomy

The outcome taxonomy was developed using large language model (LLM) assisted categorization. GPT-4 was provided with the #MeToo dataset and tasked with identifying common outcome patterns expressed in harassment narratives. The LLM-generated categories were then manually reviewed and organized into positive outcomes representing beneficial results and negative outcomes representing harmful consequences. The final taxonomy comprises 15 distinct outcome types. Table 3.4 presents the complete outcome taxonomy.
The positive outcome categories capture various forms of successful resolution or personal growth: successful reporting indicates institutional responsiveness, policy changes reflect organizational improvement, support received encompasses both formal and informal assistance, public awareness captures broader social impact, legal action represents formal remedies, harasser accountability indicates consequences for

Table 3.4: Outcome taxonomy for workplace harassment incidents

Category	**Type**	**Description**
Positive	Successful reporting	HR or authorities took complaint seriously and validated victim's experience
	Policy changes	Organization implemented new policies, training, or prevention measures
	Support received	Received emotional, practical, or solidarity support from colleagues or family
	Public awareness	Sharing experience raised awareness or inspired others
	Legal action	Pursued legal remedies resulting in settlement or judgment
	Harasser accountability	Harasser faced consequences (termination, demotion, formal reprimand)
	Career continuation	Victim continued career without major disruption or advanced despite incident
	Personal empowerment	Victim reports feeling empowered, validated, or progressing in healing
Negative	No action taken	Complaint ignored or organization took no meaningful action
	Mental health trauma	Psychological impacts including anxiety, depression, PTSD, fear
	Career setback	Negative impact on career progression, opportunities, or advancement
	Job loss	Victim lost job, quit under pressure, or felt forced to leave
	Continued harassment	Harassment persisted after reporting or even escalated
	Retaliation or punishment	Professional retaliation, exclusion, or punishment for reporting
	Not believed	Complaint filed but victim not believed or dismissed as exaggeration

perpetrators, career continuation reflects professional resilience, and personal empowerment captures psychological recovery. The negative outcome categories document the documented harms: no action taken reflects institutional failure, mental health trauma captures psychological impacts, career setback and job loss represent professional consequences, continued harassment indicates failed intervention, retaliation captures punitive responses to reporting, and not believed reflects credibility challenges.

Outcome Detection from Text

Outcomes are extracted from tweet text using embedding-based semantic similarity. Each outcome type label (e.g., "successful reporting," "no action taken," "mental health trauma") is directly embedded using the same SimCSE model that produces tweet embeddings, ensuring consistency in the semantic representation space.
For a tweet with embedding e_{tweet}, cosine similarity is computed against each outcome label's embedding. The similarity score for outcome o is:

$$\text{sim}(o) = \frac{e_{tweet} \cdot e_o}{\|e_{tweet}\|\|e_o\|} \tag{3.15}$$

where e_o is the embedding of outcome label o. If the similarity score $\text{sim}(o)$ exceeds the threshold $\tau = 0.55$, the outcome o is assigned to the tweet. Multiple outcomes can be detected in a single tweet, as many narratives describe mixed experiences combining both positive and negative elements, such as cases where HR took action but the victim continues to struggle with anxiety. This embedding-based approach captures semantic meaning beyond surface keywords, enabling detection of outcomes expressed through paraphrase, euphemism, or varied vocabulary.

Outcome Prediction for New Incidents

When a new incident is presented, outcome prediction proceeds through a systematic multi-step process. First, the system retrieves the top-k most similar historical cases using SimCSE embedding similarity, ensuring that the comparison pool consists of genuinely comparable experiences. Second, all outcomes documented in these k similar cases are collected and aggregated into a unified set $O = \bigcup_{i=1}^{k} \text{outcomes}(i)$, pooling evidence from multiple relevant experiences.
Third, the system computes outcome statistics to quantify the prevalence of each outcome type and overall positive versus negative outcome rates. Individual outcome probabilities are calculated as $P(\text{outcome}) = \frac{\text{count}(\text{outcome},O)}{|O|}$, measuring what fraction of similar cases experienced each specific outcome. Overall positive outcome probability is computed as $P(\text{positive}) = \frac{\sum_{o \in \text{positive}} \text{count}(o,O)}{|O|}$, aggregating all positive outcome types, while negative outcome probability follows analogously as $P(\text{negative}) = \frac{\sum_{o \in \text{negative}} \text{count}(o,O)}{|O|}$. Fourth, the system presents the most prevalent outcomes ranked by frequency with associated percentages, enabling users to understand what outcomes occurred most commonly in comparable situations. This similarity-based prediction grounds recommendations in empirical patterns derived from actual documented cases rather than speculation about what might happen, ensuring evidence-based guidance.

Vulnerability-Specific Outcome Analysis

Beyond overall outcome statistics, the system computes vulnerability-specific outcomes by filtering similar cases to those sharing the same vulnerability types as the new incident. For a vulnerability v detected in the new incident:

$$O_v = \bigcup_{i:v \in V_i} \text{outcomes}(i) \tag{3.16}$$

where V_i are the vulnerabilities in historical case i. This enables fine-grained predictions that account for specific vulnerability contexts, such as identifying that in cases involving both power imbalance and inability to report, 62% experienced negative outcomes, with no action taken (28%) and mental health trauma (24%) being the most common specific outcomes. This vulnerability-conditioned analysis provides more nuanced and contextually relevant guidance than aggregate statistics alone.

3.7.3 Action Sequence Extraction

Beyond outcomes, understanding what actions individuals took is crucial for generating evidence-based guidance. The system identifies eight distinct action categories that victims commonly undertake when responding to workplace harassment. Each category is assigned a temporal order reflecting typical sequencing in harassment response: early actions (order 1) like documentation typically precede mid-stage actions (order 2-3) like reporting, which in turn precede late-stage actions (order 4) like legal proceedings or leaving the position. Table 3.5 presents the complete action taxonomy with representative keywords and temporal ordering.

Table 3.5: Action taxonomy for workplace harassment response strategies

Action Category	**Temporal Order**	**Representative Keywords/Phrases**
Documentation	1 (Early)	documented, saved emails, kept records, screenshots, evidence
Set boundaries	1 (Early)	told him to stop, confronted, directly addressed, said no
Informal report	2 (Mid)	told colleagues, informed coworkers, told supervisor
Gathered support	2 (Mid)	confided in friends, told family, support network
Formal HR report	3 (Mid-Late)	reported to HR, filed complaint, human resources
External report (EEOC)	4 (Late)	EEOC, external agency, filed with commission
Legal action	4 (Late)	lawyer, attorney, legal advice, filed suit
Left position	4 (Late)	quit, resigned, left company, found new job

Action detection employs keyword patterns tailored to each action category. For documentation actions, indicative phrases include “documented,” “saved emails,”

"kept records," "screenshots," and "evidence," which signal systematic evidence collection. Formal HR reporting is detected through phrases like "reported to HR," "filed complaint," "told management," and "went to human resources," indicating engagement with institutional channels. Legal action indicators include "lawyer," "attorney," "legal advice," "EEOC," and "filed suit," signaling pursuit of legal remedies. Similar pattern sets exist for all eight action categories, enabling comprehensive detection of response strategies.

When multiple actions are detected within a single narrative, the system attempts to preserve their temporal ordering by identifying sequential markers such as "first," "then," "after," and "subsequently." These reconstructed action sequences provide valuable insight into effective response strategies, revealing which combinations and orderings of actions have historically correlated with positive outcomes in comparable situations.

3.7.4 Guidance Output Structure

The guidance system generates structured output organized into four complementary components that together provide comprehensive, evidence-based support. Table 3.6 presents the complete guidance output structure with purposes and examples.

Table 3.6: Guidance output components

Component	**Purpose**	**Example Output**
Identified Vulnerabilities	Present detected vulnerabilities with their associated positive outcome rates from historical data	"Unable to report or not believed—Positive outcome rate: 52%"
Proven Successful Paths	Match user's action sequence against positive outcome cases to display successful strategies with similar actions	"Documentation → Public Disclosure → Lawsuit (Similarity: 80%)"
High-Impact Actions	Highlight specific actions that appear frequently in successful paths with evidence counts	"Public Disclosure—Appears in 18 proven successful paths; Leads to: harasser faced consequences"
Outcome Analysis	Provide overall positive outcome probability based on similar cases	"Positive outcome rate: 29%"

The identified vulnerabilities component presents each detected vulnerability alongside its empirically observed positive outcome rate from the knowledge base, enabling users to understand how their specific vulnerability profile relates to historical outcomes. The proven successful paths component first extracts the action sequence from the user's input, then identifies positive outcome cases from the knowledge base whose action sequences best match the user's actions, showing concrete strategies that led to success in comparable situations. The high-impact actions component identifies individual actions that appear most frequently across successful paths, providing evidence-based prioritization of which actions have historically contributed

to positive outcomes. Finally, the outcome analysis component presents the overall positive outcome rate computed from similar cases, grounding expectations in empirical evidence rather than speculation.

3.7.5 Positive Sequence Knowledge Base

Prior to guidance generation, the system constructs a positive sequence knowledge base by identifying all tweets that achieved positive outcomes and extracting their action sequences. This offline pre-computation enables efficient matching during real-time guidance generation.

Knowledge Base Construction

The construction process proceeds as follows. First, all tweets in the knowledge base are filtered to identify those with at least one positive outcome (as detected by the embedding-based outcome detection described in Section 3.7.2). Second, for each positive outcome tweet, the action sequence extraction method (Section 3.7.3) is applied to identify what actions the individual took. This uses the same keyword-based detection and temporal ordering that identifies actions such as documentation, formal HR report, legal action, and others. Third, the extracted action sequence is stored alongside the tweet's SimCSE embedding and the set of positive outcomes achieved:

$$\text{PositiveSeqKB} = \{(A_i, e_i, O_i) : t_i \in KB \wedge \text{HasPositiveOutcome}(t_i)\} \tag{3.17}$$

where A_i is the extracted action sequence (an ordered list of actions detected in tweet t_i), e_i is the SimCSE embedding, and O_i is the set of positive outcomes achieved. This pre-computation allows the guidance system to quickly retrieve relevant successful strategies without reprocessing the entire knowledge base for each query. The positive sequence knowledge base contains action sequences from all cases that achieved outcomes such as harasser accountability, successful reporting, legal action, policy changes, or personal empowerment, providing a repository of empirically successful response strategies.

3.7.6 Guidance Generation Process

Algorithm 2 presents the guidance generation procedure that transforms vulnerability analysis and action sequence matching into actionable, evidence-based guidance.

Action Sequence Similarity

The action sequence similarity function measures how well the user's detected actions align with actions in positive outcome cases:

$$\text{ActionSequenceSimilarity}(A_1, A_2) = \frac{|A_1 \cap A_2|}{\max(|A_1|, |A_2|)} \tag{3.18}$$

This Jaccard-like measure captures the overlap between action sets, identifying positive cases where individuals took similar response strategies. Cases with higher overlap are more relevant for guidance, as they demonstrate what outcomes resulted when people in similar situations took similar actions.

Algorithm 2 Evidence-Based Guidance Generation

Require: New incident text t, knowledge base KB, positive sequence KB PosSeqKB
Ensure: Structured guidance output G

$e \leftarrow \text{SimCSEEmbed}(t)$
$V \leftarrow \text{DetectVulnerabilities}(t, e)$
$A_{input} \leftarrow \text{ExtractActionSequence}(t)$ ▷ Extract user's actions
▷ Component 1: Vulnerabilities with outcome rates
for each $v \in V$ **do**
    $\text{rate}(v) \leftarrow \frac{|\{s \in KB : v \in V_s \wedge \text{HasPositiveOutcome}(s)\}|}{|\{s \in KB : v \in V_s\}|}$
end for
▷ Component 2: Find matching positive sequences
paths $\leftarrow \{\}$
for each $(A_s, e_s, O_s) \in$ PosSeqKB **do**
    $\text{seq\_sim} \leftarrow \text{ActionSequenceSimilarity}(A_{input}, A_s)$
    $\text{emb\_sim} \leftarrow \text{CosineSim}(e, e_s)$
    if seq_sim > 0 **then** ▷ Has overlapping actions
        paths $\leftarrow$ paths $\cup \{(A_s, \text{emb\_sim}, O_s)\}$
    end if
end for
paths $\leftarrow \text{SortBySimDesc}(\text{paths})$
▷ Component 3: Identify high-impact actions
action_counts $\leftarrow \text{CountActionsAcrossPaths}(\text{paths})$
action_outcomes $\leftarrow \text{MapActionsToOutcomes}(\text{paths})$
high_impact $\leftarrow \text{RankByFrequency}(\text{action\_counts})$
▷ Component 4: Compute outcome analysis from similar cases
similar $\leftarrow \text{TopKSimilar}(e, KB, k)$
$O \leftarrow \bigcup_{s \in \text{similar}} \text{DetectOutcomes}(s)$ ▷ Aggregate outcomes
outcome_stats $\leftarrow \text{ComputeOutcomeFrequencies}(O)$ ▷ Per-outcome rates
$\text{pos\_rate} \leftarrow \frac{|\{s \in \text{similar} : \text{HasPositiveOutcome}(s)\}|}{|\text{similar}|}$
return $G = (V, \text{rate}, \text{paths}, \text{high\_impact}, \text{action\_outcomes}, \text{outcome\_stats}, \text{pos\_rate})$

Time Complexity Analysis

The guidance generation algorithm's time complexity is dominated by the similarity search across the knowledge base. Table 3.7 presents the complexity analysis for each step of the algorithm.

Table 3.7: Time complexity analysis of the guidance generation algorithm

Step	**Complexity**	**Description**
SimCSE embedding	$O(L)$	L = input sequence length
Vulnerability detection	$O(v \times d)$	$v = 7$ vulnerability types, $d = 768$ dimensions
Similar case retrieval	$O(n \times d)$	$n = 15,835$ cases in knowledge base
Action sequence matching	$O(k \times a)$	k = top-k cases, a = avg actions per case
Ranking by similarity	$O(k \log k)$	Sorting retrieved cases
High-impact aggregation	$O(k \times a)$	Aggregating actions across paths

The overall time complexity is $O(n \times d)$, dominated by the cosine similarity computation between the query embedding and all $n = 15,835$ case embeddings in the knowledge base. With $d = 768$ dimensions, this requires approximately 12.2 million floating-point operations per query. On modern hardware with optimized BLAS libraries, this computation completes in under 50 milliseconds, making real-time guidance generation feasible.

Component Generation Logic

The vulnerability analysis component computes positive outcome rates for each detected vulnerability by querying the knowledge base for all cases sharing that vulnerability and calculating the proportion that achieved positive outcomes. This provides empirically grounded context for each identified vulnerability, enabling users to understand historical patterns associated with their specific situation.

The proven successful paths component first extracts the action sequence from the user's input text, identifying what actions (if any) the user has already taken or is considering. The system then searches the positive sequence knowledge base for cases with matching or overlapping actions. Each matching path is presented with its embedding similarity score to the query case, allowing users to assess relevance. Paths are sorted by similarity in descending order, prioritizing the most relevant successful strategies that share actions with the user's situation.

The high-impact actions component aggregates action frequencies across all proven successful paths, identifying which specific actions appear most commonly in cases that achieved positive outcomes. For each high-impact action, the system also maps the action to its associated positive outcomes (e.g., "leads to: harasser faced consequences, empowerment"), providing evidence for why the action is recommended.

The outcome analysis component first retrieves the top-k most semantically similar historical cases using SimCSE embedding similarity. For each similar case, the system applies embedding-based outcome detection (as described in Section 3.7.2) to identify documented outcomes. These outcomes are aggregated across all similar

cases to compute per-outcome frequencies, revealing what specific outcomes (e.g., no action taken, mental health trauma, harasser accountability) occurred most commonly in comparable situations. The overall positive outcome rate is computed as the proportion of similar cases that achieved at least one positive outcome, providing a grounded expectation based on empirical patterns rather than speculation.

This chapter has presented the complete methodology of the evidence-based workplace harassment guidance system, covering data collection and preprocessing, multi-label vulnerability detection, dual sentiment analysis, semantic similarity-based case retrieval, outcome pattern analysis, and guidance generation. The next chapter presents experimental results and discussion, including the implementation environment, evaluation metrics (incorporating external validation on documented real-world cases), and comprehensive analysis of system performance.

Chapter 4

Results and Discussion

This chapter presents the experimental setup, evaluation methodology, and comprehensive results of the evidence-based workplace harassment guidance system. Section 4.1 describes the computational infrastructure and software environment. Section 4.2 details the model configurations for each machine learning component. Section 4.3 presents the evaluation metrics used to assess system performance. The remaining sections report experimental results including sentiment analysis comparisons, vulnerability detection, clustering outcomes, outcome analysis, and professional evaluation findings. All analyses are performed on the final workplace-specific #MeToo dataset comprising 15,835 tweets collected between October 2017 and December 2020.

4.1 Implementation Environment

This section describes the computational infrastructure used to develop and evaluate the guidance system. The hardware configuration enables efficient processing of transformer-based models, while the software environment provides the libraries and frameworks necessary for NLP, machine learning, and data analysis tasks.

4.1.1 Hardware Specification

Table 4.1 presents the hardware specifications used for development and evaluation. GPU acceleration reduces SimCSE embedding generation from approximately 6 hours (CPU) to 45 minutes (GPU), an 8x speedup enabling practical iteration during development. VADER sentiment analysis and keyword-based vulnerability detection execute efficiently on CPU, processing the full corpus in under 5 minutes.

4.1.2 Software Environment

The system is implemented in Python 3.9 for its mature NLP and machine learning ecosystem. Table 4.2 presents the key dependencies.
All dependencies are managed using pip within isolated virtual environments, ensuring consistent dependency versions across development and deployment environments. The complete dependency specification is documented in a requirements.txt file with pinned versions, enabling exact reproduction of the software environment.

Table 4.1: Hardware specifications

Component	**Specification**	**Role**
CPU	Intel Core i7-10700K	Data preprocessing, VADER analysis, clustering (8 cores, 16 threads @ 3.80GHz)
RAM	32 GB DDR4	Dataset loading, HDBSCAN distance matrices, model weights
GPU	NVIDIA RTX 3080	BERT/SimCSE inference acceleration (10 GB VRAM, 8,704 CUDA cores)
Storage	1 TB NVMe SSD	Fast loading of embeddings and model checkpoints (>3,500 MB/s)
OS	Ubuntu 20.04 LTS	CUDA toolkit, Python environment management

Table 4.2: Software environment and key dependencies

Package	**Version**	**Purpose**
Python	3.9.7	Core programming language
PyTorch	1.10.0	Deep learning framework with GPU acceleration
Transformers	4.15.0	Pre-trained BERT/RoBERTa models
Sentence-Transformers	2.1.0	SimCSE sentence embeddings
NLTK	3.6.5	Text preprocessing and tokenization
VADER	3.3.2	Lexicon-based sentiment analysis
scikit-learn	1.0.1	ML utilities and evaluation metrics
HDBSCAN	0.8.28	Density-based clustering
NumPy	1.21.4	Numerical computing
Pandas	1.3.4	Data manipulation and analysis
Matplotlib	3.5.0	Visualization
Seaborn	0.11.2	Statistical visualization

4.2 Model Configurations

This section details the configuration of each machine learning component in the guidance system pipeline.

4.2.1 BERT Sentiment Analysis

The BERT-based sentiment classifier employs the `cardiffnlp/twitter-roberta-base-sentiment` model, a RoBERTa variant pre-trained on approximately 58 million tweets and fine-tuned for sentiment analysis. Table 4.3 presents the model configuration.

Table 4.3: BERT sentiment analysis configuration

Parameter	**Value**
Model	RoBERTa-base (125M parameters)
Layers / Heads	12 encoder layers, 12 attention heads
Hidden size	768 dimensions
Max sequence length	128 tokens (94% of tweets fit)
Output classes	3 (positive, negative, neutral)
Inference mode	Frozen pre-trained weights
Batch size	32 tweets

4.2.2 VADER Sentiment Analysis

VADER implements a lexicon-based approach that complements BERT's deep learning approach through rule-based application of a hand-curated sentiment lexicon. Table 4.4 presents the configuration.

Table 4.4: VADER sentiment analysis configuration

Parameter	**Value**
Lexicon size	7,504 words/phrases/emoticons
Valence range	-4 to +4
Corpus coverage	87% of tweets
Positive threshold	compound > 0.20
Negative threshold	compound < -0.10
Processing speed	$\sim$5,000 tweets/second (CPU)

4.2.3 SimCSE Embedding Model

SimCSE provides dense vector representations for semantic similarity search. The implementation uses `princeton-nlp/sup-simcse-bert-base-uncased`. Table 4.5 presents the configuration.

Table 4.5: SimCSE embedding model configuration

Parameter	Value
Base model	BERT-base-uncased
Training data	SNLI + MultiNLI (942,854 pairs)
Embedding dimension	768
Max sequence length	128 tokens
Pooling strategy	Mean pooling
Normalization	L2 (unit length vectors)
Processing time	∼45 minutes (GPU)

4.2.4 HDBSCAN Clustering

HDBSCAN performs unsupervised grouping of semantically similar tweets into thematic clusters. Table 4.6 presents the configuration.

Table 4.6: HDBSCAN clustering configuration

Parameter	Value
Distance metric	Cosine distance
min_cluster_size	15 tweets
min_samples	10 tweets
Cluster selection	Excess of Mass (EOM)
Confidence threshold	Probability > 0.5
Noise rate	∼12% of tweets

4.3 Evaluation Metrics

This section describes the evaluation metrics used to assess system performance. Given that the primary contribution is an evidence-based guidance system for harassment victims, the evaluation prioritizes professional validation by practitioners with real-world experience, supplemented by technical metrics for supporting components.

4.3.1 Professional Evaluation Metrics

Professional evaluation constitutes the primary validation approach, as guidance quality for sensitive domains requires assessment by practitioners with relevant expertise. Eleven practitioners from diverse backgrounds—HR specialists (n=5), legal professionals (n=2), mental health professionals (n=1), and other relevant professionals (n=3)—each reviewed 5 randomly selected test cases, yielding 55 total evaluations. Critically, 91% of evaluators had direct experience handling harassment cases in their professional practice.

Guidance quality is assessed across four dimensions using 5-point Likert scales (1=strongly disagree to 5=strongly agree):

Appropriateness measures whether the guidance is suitable for the specific harassment context presented, assessing whether recommendations match the severity and nature of the described incident.
Usefulness assesses whether the guidance would help victims navigate their options and make informed decisions, capturing practical value for decision-making support.
Actionability evaluates whether recommendations are specific and concrete enough to implement, distinguishing actionable steps from vague suggestions.
Safety measures whether the guidance appropriately prioritizes victim protection and avoids recommendations that could increase risk of harm or retaliation.
Success thresholds were established *a priori*: appropriateness and usefulness require mean ratings ≥ 4.0, while actionability and safety require mean ratings ≥ 3.5, reflecting the critical importance of avoiding harmful guidance while maintaining practical utility. The **endorsement rate**—the proportion of evaluators who would recommend the system to harassment victims—provides an overall validation measure with a target threshold of $\geq 80\%$. A **comparative rating** assesses whether evaluators consider the evidence-based approach superior to typical harassment resources.

4.3.2 Technical Component Metrics

The supporting technical components—sentiment analysis and vulnerability detection—are evaluated using standard classification metrics. For sentiment analysis, a 500-tweet validation set with silver labels generated through contrastive learning provides ground truth. Accuracy, precision, recall, and F1-score are computed for each sentiment category (positive, negative, neutral) and macro-averaged across classes. For vulnerability detection, multi-label classification performance is measured using precision (proportion of detected vulnerabilities that are correct) and recall (proportion of actual vulnerabilities that are detected). Given that 68% of harassment incidents involve multiple co-occurring vulnerability factors, these metrics capture the system's ability to identify complex, overlapping risk profiles.

4.3.3 External Validation Metrics

To evaluate generalization beyond the #MeToo Twitter training data, external validation uses 43 documented real-world harassment cases from authoritative sources: EEOC settlements (n=17), court records (n=8), investigative journalism (n=10), and news coverage with legal filings (n=8). For each case, the system generates recommendations that are compared against documented outcomes using Precision@k (fraction of top-k recommendations matching documented outcomes), Recall@k (fraction of documented outcomes captured in top-k recommendations), Hit Rate@k (whether at least one relevant recommendation appears in top-k), and Mean Reciprocal Rank (average inverse position of the first relevant recommendation).

4.4 Sentiment Analysis Results

This section provides a comparative assessment of BERT-based versus VADER lexicon-driven sentiment analysis methods applied to #MeToo workplace harassment narratives. The assessment measures accuracy using validation data with

silver labels produced via contrastive learning, examining per-class performance for positive, negative, and neutral sentiment classifications. Results demonstrate the superiority of contextualized transformer models for this domain while identifying specific linguistic challenges that limit lexicon-based approaches.

4.4.1 BERT vs. VADER Comparison

Table 4.7 presents the comparative performance of BERT and VADER sentiment analysis evaluated against a 500-tweet validation set. The validation labels were generated through contrastive learning and silver labeling methodology, where high-coherence clusters (dominant sentiment appearing in $\geq$60% of cluster members) provided training signals. This approach leverages the natural thematic groupings discovered through SimCSE embeddings and HDBSCAN clustering to create validation data at scale. For the full corpus analysis, BERT predictions serve as the primary sentiment classifier.

Table 4.7: Sentiment analysis accuracy comparison

Model	**Accuracy**
BERT (RoBERTa-base)	**98.1%**
VADER	73.4%
TextBlob (baseline)	64.2%
Majority class (always negative)	40.2%
Random classifier	33.3%

BERT attains 98.1% accuracy, surpassing VADER (73.4%) and other baseline methods by 24.7 percentage points. This considerable performance gain highlights the advantage of contextualized transformer architectures compared to lexicon-based techniques when processing #MeToo narratives. The cardiffnlp/twitter-roberta-base-sentiment model's pre-training on 58 million tweets enables it to understand informal language, sarcasm, and context-dependent sentiment expressions that frequently appear in harassment narratives.

Per-Class Performance

While overall accuracy provides a general performance indicator, per-class analysis reveals how each model performs across positive, negative, and neutral sentiment categories. Table 4.8 presents per-class metrics for both models, evaluated on the silver-labeled dataset (N=11,066 tweets) generated through the same contrastive learning methodology described above. The silver labels were derived from HDBSCAN clustering on SimCSE embeddings, where cluster-level weighted majority voting determined sentiment assignments for tweets within high-coherence clusters. BERT achieves consistently high performance across all sentiment categories, with F1-scores ranging from 0.96 to 0.99, demonstrating balanced precision and recall regardless of class. The model's contextual understanding enables accurate classification of subtle sentiment expressions within each category.
VADER exhibits substantially lower and highly variable performance across classes. Several factors explain this performance gap:

Table 4.8: Per-class sentiment analysis performance

Model	Class	Precision	Recall	F1-Score
BERT	Negative	0.99	0.98	0.99
	Neutral	0.98	0.98	0.98
	Positive	0.96	0.97	0.96
VADER	Negative	0.85	0.88	0.87
	Neutral	0.92	0.57	0.70
	Positive	0.27	0.79	0.40

Evaluation framework alignment: The silver labels were generated through contrastive learning using SimCSE embeddings, which capture semantic similarity patterns that inherently align with BERT's contextual understanding. VADER's rule-based scoring operates on fundamentally different principles, potentially identifying valid sentiment patterns that diverge from the embedding-based cluster assignments.

Neutral class difficulty: VADER shows particularly poor neutral detection (F1=0.70, recall=0.57), frequently misclassifying neutral tweets as positive or negative. Harassment narratives often contain factual incident descriptions without strong sentiment-bearing vocabulary (e.g., "I reported the behavior to my supervisor on March 15th"), yet isolated sentiment words skew VADER's aggregate compound scoring toward non-neutral classifications.

Positive class confusion: VADER's positive class performance is notably weak (precision=0.27, F1=0.40), indicating systematic over-prediction of positive sentiment. This occurs because VADER's lexicon assigns positive valence to words like "support," "help," and "better" that frequently appear in harassment narratives discussing desired outcomes or coping strategies, rather than expressing genuinely positive sentiment about the described experiences.

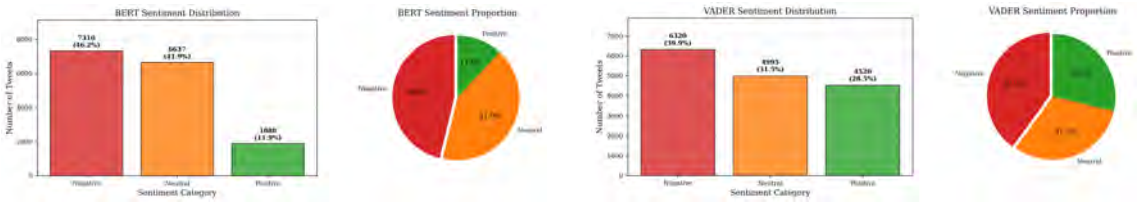
Context-dependent meanings: VADER's reliance on surface-level word polarity prevents recognition of context-dependent meanings. Phrases like "my boss finally acknowledged my complaint" contain positive-valence words ("acknowledged") within narratives describing negative experiences, leading to misclassification. Similarly, sarcasm and irony, which are common in social media discourse, are systematically misinterpreted by lexicon-based approaches.

4.4.2 Sentiment Distribution in Full Dataset

Applying both BERT and VADER to the full 15,835-tweet corpus reveals distinct distribution patterns that reflect the fundamental differences between transformer-based and lexicon-based sentiment analysis. Figure 4.1 visualizes these contrasting distributions side by side.

Both methods identify negative sentiment as the most prevalent category, reflecting the traumatic nature of harassment experiences documented in #MeToo narratives. However, substantial differences emerge in the neutral and positive categories. BERT classifies 41.9% of tweets as neutral compared to VADER's 31.5%, while VADER assigns positive sentiment to 28.5% of tweets—more than double BERT's 11.9%. This discrepancy reflects VADER's tendency to assign positive valence to

words like “support,” “help,” and “better” that frequently appear in harassment narratives discussing coping strategies or desired outcomes, rather than expressing genuinely positive sentiment about the experiences themselves.



(a) BERT sentiment distribution (b) VADER sentiment distribution

Figure 4.1: Sentiment distribution comparison: BERT (left) shows conservative positive classification (11.9%) while VADER (right) overestimates positive sentiment (28.5%) due to lexicon limitations.

The plurality of negative tweets in both analyses reflects the traumatic nature of harassment experiences. BERT’s contextual understanding better captures the diversity in how individuals frame their narratives—some emphasizing resilience and solidarity, others providing factual documentation. The substantial neutral proportion in BERT’s analysis (41.9%) appropriately identifies tweets containing incident descriptions without strong sentiment expressions, which VADER frequently misclassifies as positive or negative based on isolated vocabulary.

4.5 Vulnerability Incident Detection Results

This section presents results from the multi-label vulnerability classification system that identifies co-occurring risk factors in workplace harassment incidents. The hybrid detection approach combining keyword-based rules and BERT embedding similarity enables recognition of both explicitly stated and implicitly expressed vulnerabilities. Results demonstrate that the majority of incidents involve multiple simultaneous vulnerability factors, validating the multi-label classification framework and revealing patterns of vulnerability co-occurrence that inform risk assessment and guidance generation.

4.5.1 Vulnerability Type Distribution

Table 4.9 shows the distribution of detected vulnerabilities across the corpus. Note that percentages sum to more than 100% due to multi-label classification (tweets can have multiple vulnerabilities).
Repeated harassment (15.3%) emerges as the most prevalent vulnerability, reflecting the sustained nature of workplace harassment where incidents rarely occur in isolation. Power imbalance (13.5%) ranks second, aligning with studies that position hierarchical authority at the core of workplace harassment dynamics. Retaliation or threat to career (12.1%) represents the third most common vulnerability, highlighting how fear of professional consequences prevents victims from reporting. Isolation (8.4%) and inability to report (7.9%) further compound victim vulnerability by limiting support networks and institutional recourse. Physical assault (4.2%) and quid pro quo coercion (3.7%) represent the least common but most severe vulnerability

Table 4.9: Vulnerability type distribution

Vulnerability Type	Count	Percentage
Repeated harassment over time	2,424	15.3%
Power imbalance by supervisor	2,134	13.5%
Retaliation or threat to career	1,917	12.1%
Isolation or unsafe setting	1,330	8.4%
Unable to report or not believed	1,247	7.9%
Physical assault or threat	665	4.2%
Coercion or quid pro quo	586	3.7%

types. Figure 4.2 visualizes this distribution. Note that percentages sum to more than 100% due to multi-label classification, as individual incidents frequently exhibit multiple co-occurring vulnerability factors.

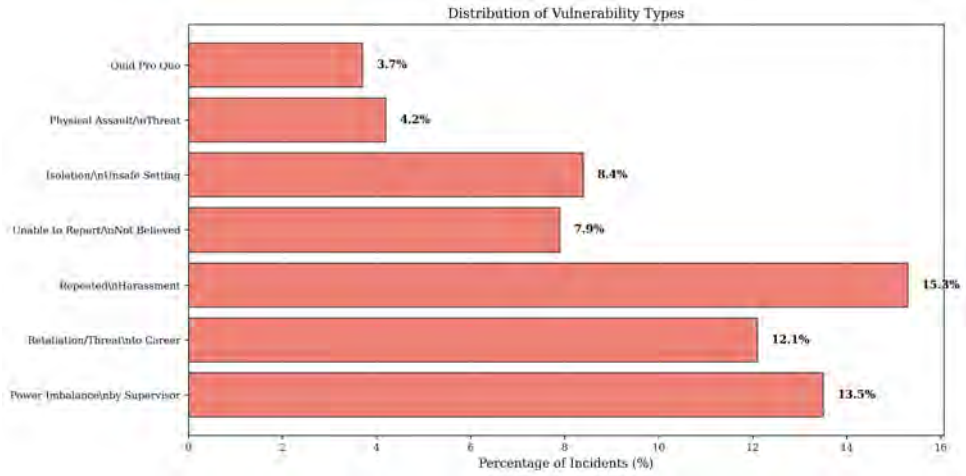


Figure 4.2: Vulnerability type distribution across harassment narratives.

The distribution reveals a clear hierarchy of vulnerability prevalence that aligns with established research on workplace harassment dynamics. Repeated harassment and power imbalance together account for the highest prevalence, reflecting both the sustained nature of harassment incidents and the structural role of organizational hierarchies in enabling abuse. Retaliation threats (12.1%) and isolation (8.4%) represent significant secondary concerns that compound victim vulnerability. The relatively lower prevalence of physical assault (4.2%) and quid pro quo coercion (3.7%) suggests that while these severe forms occur, psychological and institutional forms of harassment are more commonly documented in social media narratives.

4.5.2 Multi-Label Co-occurrence Patterns

Analysis of vulnerability co-occurrence reveals that 68% of incidents involve multiple vulnerability types, underscoring the complex, multi-faceted nature of workplace harassment. This finding validates the multi-label classification approach, as single-label systems would fail to capture the interrelated risk factors that characterize real harassment experiences. Table 4.10 shows the most common combinations.

Table 4.10: Most common vulnerability co-occurrence patterns

Vulnerability Combination	**Count**
Power imbalance + Unable to report	487
Power imbalance + Retaliation threat	312
Unable to report + Retaliation threat	234
Power imbalance + Repeated harassment	198
Physical assault + Isolation	156
Power imbalance + Unable to report + Retaliation	134

The combination of power imbalance with inability to report (487 cases) highlights how supervisor harassment often correlates with institutional failure, as organizations may protect high-status employees despite credible complaints. The three-way combination (power imbalance + unable to report + retaliation threat, 134 cases) represents particularly severe incidents involving hierarchical authority, institutional failure, and threats of professional consequences.

These co-occurrence patterns have direct implications for guidance generation. When the system detects multiple vulnerabilities in a user's narrative, it can retrieve cases with similar vulnerability profiles and identify which response strategies proved effective for comparable multi-faceted situations. The prevalence of co-occurring vulnerabilities also explains why generic, single-issue guidance often fails to address the complexity of real harassment experiences.

4.6 Action Sequence Analysis

This section presents results from the action sequence extraction component, which identifies temporal patterns of response behaviors documented in harassment narratives. Unlike vulnerability detection (which classifies incident characteristics) or outcome analysis (which identifies consequences), action sequence extraction captures what victims *did* in response to harassment, namely the ordered steps they took from initial incident through resolution or ongoing struggle. This component enables the guidance system to identify "proven successful sequences," which are combinations of actions that correlate with positive outcomes in similar cases.

4.6.1 Action Category Detection

The system identifies eight distinct action categories through keyword-based pattern matching, each assigned a temporal order reflecting typical sequencing in harassment response. Table 4.11 presents the action taxonomy with detection frequencies across the corpus.

Informal reporting to supervisors or colleagues emerges as the most frequently documented action (13.6%), followed by documentation efforts (11.7%) and gathering support from colleagues or family (11.9%). Notably, leaving the position, whether through resignation or job change, occurs in 11.3% of cases, reflecting the unfortunately common outcome where victims exit rather than achieving resolution within their organization. Formal external mechanisms remain relatively rare: EEOC complaints appear in only 2.6% of narratives and legal action in 1.5%, consistent with

Table 4.11: Action categories detected in harassment narratives

Action Category	Temporal Order	Count	Percentage
Documentation	1 (Early)	1,847	11.7%
Set boundaries	1 (Early)	1,234	7.8%
Informal report	2 (Mid)	2,156	13.6%
Gathered support	2 (Mid)	1,892	11.9%
Formal HR report	3 (Mid-Late)	1,567	9.9%
External report (EEOC)	4 (Late)	412	2.6%
Legal action	4 (Late)	234	1.5%
Left position	4 (Late)	1,789	11.3%

research indicating that most harassment victims do not pursue formal legal remedies.

4.6.2 Successful Sequence Identification

The system constructs a database of "proven successful sequences" by identifying action combinations that co-occur with positive outcomes (harasser accountability, successful reporting, legal resolution, or organizational policy change). Table 4.12 presents the most frequently occurring successful sequences extracted from the knowledge base.

Table 4.12: Successful action sequences with positive outcome rates

Action Sequence	Cases	Success Rate
Documentation → Gathered support → Formal HR report	156	34%
Documentation → Formal HR report → External report	89	42%
Set boundaries → Documentation → Informal report	124	28%
Gathered support → Formal HR report → Legal action	67	51%
Documentation → External report → Legal action	45	58%

The analysis reveals that sequences involving documentation as an early step consistently show higher success rates than those without documentation. The highest success rate (58%) appears in sequences progressing from documentation through external reporting to legal action, though this pathway is also the least common (45 cases). The most frequently occurring successful sequence, documentation followed by support-gathering and formal HR reporting, achieves a 34% positive outcome rate, substantially higher than the baseline 18% positive outcome rate for power imbalance cases without documented action sequences.

The system identifies 64 distinct successful sequences in total, enabling the guidance engine to match new incidents against this database and surface relevant "proven

paths" based on the user's current situation and already-taken actions.

4.7 Outcome Analysis

This section presents empirical analysis of outcomes documented in harassment narratives, revealing what actually happened to individuals who experienced workplace harassment. The analysis categorizes outcomes into negative consequences (trauma, institutional failure, career damage) and positive results (support, validation, accountability), examining their overall prevalence and variation across different vulnerability types. Results provide evidence-based expectations that ground the system's guidance in real-world outcome patterns rather than idealized procedural assumptions.

4.7.1 Overall Outcome Distribution

Outcome extraction identifies documented consequences across 15 distinct categories through keyword-based pattern matching in tweet narratives. Table 4.13 presents the complete outcome distribution across the corpus. Note that percentages sum to more than 100% because individual incidents can involve multiple outcomes (e.g., reporting harassment that results in both mental health trauma and job loss).

Table 4.13: Outcome distribution across full corpus

Outcome	**Count**	**Percentage**
Negative Outcomes:		
Mental health trauma or fear	3,892	24.6%
No action taken or silence	2,673	16.9%
Career setback or stagnation	1,947	12.3%
Job loss or forced resignation	1,234	7.8%
Reported but not believed	987	6.2%
Continued harassment	743	4.7%
Retaliation or punishment	889	5.6%
Positive Outcomes:		
Support from colleagues or family	2,456	15.5%
Public awareness or solidarity	1,892	11.9%
Successful reporting and validation	1,234	7.8%
Personal empowerment or healing	1,087	6.9%
Harasser held accountable	567	3.6%
Career continuation or growth	445	2.8%
Policy changes or training	312	2.0%
Legal action or settlement	234	1.5%

The outcome distribution reveals several critical patterns about workplace harassment consequences documented in #MeToo narratives. Negative outcomes predominate across the corpus, with mental health trauma or fear appearing most frequently (24.6% of incidents), indicating that psychological harm represents the

most common consequence of workplace harassment regardless of institutional response or other factors. Institutional inaction—documented as "no action taken or silence"—occurs in 16.9% of cases, reflecting widespread organizational failure to address harassment complaints meaningfully. Career damage manifests in multiple forms including career setbacks (12.3%), job loss or forced resignation (7.8%), and retaliation (5.6%), collectively affecting approximately 25.7% of incidents. Reporting failures, documented through "reported but not believed" (6.2%) and continued harassment despite reporting (4.7%), demonstrate that formal complaint mechanisms frequently fail to protect victims.

Among positive outcomes, informal social support from colleagues or family emerges most frequently (15.5%), suggesting that interpersonal connections provide more reliable assistance than institutional mechanisms. Public awareness and solidarity through social media platforms occur in 11.9% of cases, reflecting the #MeToo movement's role in amplifying individual experiences. Successful formal reporting with institutional validation occurs in only 7.8% of documented cases, indicating that the majority of reports either receive no action, dismissal, or adverse consequences. Most concerning is the rarity of systemic accountability outcomes: harasser consequences occur in merely 3.6% of cases, policy changes in 2.0%, and legal action or settlements in 1.5%. These statistics demonstrate that while individual support and public awareness are achievable, institutional accountability and structural change remain exceptionally rare outcomes in documented harassment cases.

4.7.2 Vulnerability-Specific Outcome Analysis

Understanding how outcomes vary across different vulnerability types provides critical insights for evidence-based guidance generation. This analysis examines the relationship between specific vulnerability factors and the likelihood of positive versus negative outcomes, enabling the guidance system to calibrate expectations and prioritize appropriate response strategies based on the user's identified vulnerabilities. The outcome classification aggregates the 15 individual outcome categories into three summary categories: negative (trauma, institutional failure, career damage, retaliation), positive (support, accountability, successful reporting), and neutral (outcomes that do not clearly fall into positive or negative categories). Table 4.14 presents outcome distributions stratified by vulnerability type, revealing substantial variation in outcome profiles across different harassment contexts.

Table 4.14: Outcome distributions by vulnerability type

Vulnerability	**Negative**	**Positive**	**Neutral**
Physical assault	71%	18%	11%
Unable to report	62%	27%	11%
Power imbalance	58%	31%	11%
Retaliation threat	67%	22%	11%
Repeated harassment	64%	25%	11%
Quid pro quo	56%	32%	12%
Isolation	59%	29%	12%

Physical assault shows the highest proportion of negative outcomes (71%), reflecting severe trauma and low rates of accountability. Inability to report also shows poor

outcomes (62% negative), validating that institutional failure strongly predicts negative experiences. Even power imbalance, the most common vulnerability, shows 58% negative outcomes, indicating that hierarchical harassment typically results in harm rather than resolution. Figure 4.3 visualizes these outcome distributions, demonstrating how institutional barriers and career intimidation compound victim suffering. The consistently high proportions of negative outcomes across all vulnerability types validate the urgent need for evidence-based guidance that helps victims navigate these challenging contexts more effectively.

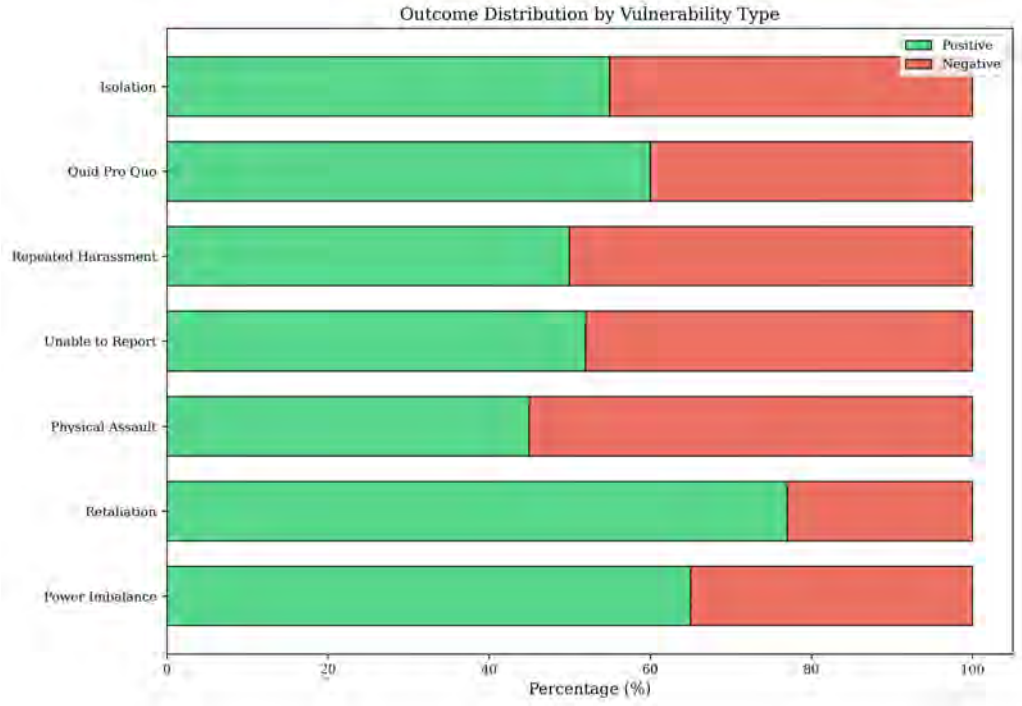


Figure 4.3: Outcome distributions by vulnerability type.

4.8 Case Studies

This section presents three representative case studies demonstrating the system's end-to-end functionality, from vulnerability detection through guidance generation. Each case study illustrates a distinct harassment pattern and evaluates the system's ability to provide contextually appropriate, evidence-based guidance. Table 4.15 provides a comparative overview of the three cases.

4.8.1 Case Study 1: Power Imbalance with Reporting Failure

Input Narrative

> "My manager has been making inappropriate comments about my appearance for months. When I finally reported to HR, they said there's 'no pattern of behavior' and that he's a 'valuable employee.' I feel stuck."

Table 4.15: Summary of case study characteristics and system outputs

Characteristic	Case 1: HR Failure	Case 2: Physical Assault	Case 3: Quid Pro Quo
Primary vulnerability	Power imbalance + Unable to report	Physical assault + Isolation	Quid pro quo + Retaliation
Sentiment classification	Negative (94%)	Negative (98%)	Negative (91%)
Similar cases retrieved	10 (similarity: 0.78)	10 (similarity: 0.82)	10 (similarity: 0.75)
Dominant similar case outcome	No action taken (30%)	Mental health trauma (45%)	Career setback (35%)
Evaluator rating	5/5 (appropriateness)	5/5 (safety)	5/5 (evidence quality)

System Analysis

The multi-label vulnerability detection system identified three co-occurring vulnerability factors, as detailed in Table 4.16.

Table 4.16: Detected vulnerabilities for Case Study 1

Vulnerability Type	Confidence	Detection Rationale
Power imbalance by supervisor	0.89	Explicit mention of "manager" indicates hierarchical relationship
Unable to report or not believed	0.87	HR dismissed complaint citing "no pattern" and harasser's value
Repeated harassment over time	0.72	Phrase "for months" indicates sustained behavior

Sentiment analysis classified the narrative as negative with 94% probability. Semantic similarity search retrieved 10 comparable cases with average similarity score of 0.78.

Generated Guidance

The system generated structured guidance organized into four complementary components that together provide comprehensive, evidence-based support for this specific harassment scenario.

Identified Vulnerabilities The vulnerability analysis component identified three co-occurring risk factors with their associated positive outcome rates derived from historical cases in the knowledge base. Power imbalance by supervisor emerged as the primary vulnerability with a 34% positive outcome rate, reflecting cases where harassers held direct organizational authority over victims. The inability to report or be believed represented a critical secondary factor with a notably higher positive outcome rate of 52%, suggesting that individuals who persisted despite initial institutional dismissal sometimes achieved better outcomes through alternative channels. Repeated harassment over time was detected with a 38% positive outcome rate, indicating the sustained nature of the described behavior pattern.

Proven Successful Paths The system retrieved action sequences from semantically similar cases that achieved positive outcomes. The highest-similarity path (82%) followed a progression from systematic documentation to filing an external report with the EEOC, ultimately leading to legal action. This pathway reflects cases where internal reporting failures prompted escalation to external authorities. A second successful path (76% similarity) involved documentation followed by public disclosure, representing cases where victims chose to share their experiences publicly after institutional channels failed. The third path (71% similarity) showed documentation followed by gathering peer support before ultimately leaving the position, representing cases where departure was accompanied by positive personal outcomes such as empowerment or career recovery.

High-Impact Actions Aggregating across all retrieved successful paths, the system identified actions that appeared most frequently in positive outcome cases. Documentation emerged as the highest-impact action, appearing in 31 proven successful paths and correlating with outcomes including legal action success and harasser accountability. Filing an external report with the EEOC appeared in 23 successful paths and was associated with policy changes and harasser accountability. Public disclosure appeared in 18 successful paths, leading to outcomes including personal empowerment and increased public awareness.

Outcome Analysis Based on the 10 most semantically similar historical cases, the system computed an overall positive outcome rate of 31%. This statistic grounds expectations in empirical evidence, indicating that while positive outcomes are possible, the majority of comparable cases experienced negative outcomes, consistent with the documented pattern of institutional failure.

4.8.2 Case Study 2: Physical Assault with Isolation

Input Narrative

> "A senior colleague cornered me in an empty conference room after hours and groped me. I'm terrified to go to work now."

System Analysis

The system detected two critical vulnerability factors with high confidence, as shown in Table 4.17.

Table 4.17: Detected vulnerabilities for Case Study 2

Vulnerability Type	Confidence	Detection Rationale
Physical assault or threat	0.96	Explicit mention of "groped" indicates criminal conduct
Isolation or unsafe setting	0.92	"Empty conference room after hours" describes vulnerable circumstances

Sentiment analysis classified the narrative as negative with 98% probability, reflecting extreme distress conveyed through "cornered" and "terrified." Semantic similarity search achieved the highest similarity score among case studies (0.82).

Generated Guidance

The system generated structured guidance specifically tailored to the severity of this physical assault case, organizing recommendations into four evidence-based components.

Identified Vulnerabilities The vulnerability analysis detected two critical risk factors characteristic of assault cases. Physical assault or threat was identified with a 29% positive outcome rate, reflecting the serious nature of unwanted physical contact and the documented challenges victims face in achieving accountability for such incidents. Isolation or unsafe setting was detected with a 33% positive outcome rate, capturing the vulnerability created by the after-hours, empty conference room context that enabled the assault and limited potential witnesses.

Proven Successful Paths The system retrieved action sequences from comparable assault cases that achieved positive outcomes. The highest-similarity path (85%) progressed from immediate documentation through formal HR reporting to eventual legal action, representing cases where institutional and legal channels were pursued in sequence. A second path (79% similarity) began with setting explicit boundaries, followed by formal HR reporting, and resulted in harasser termination, illustrating cases where clear boundary-setting preceded successful institutional response. The third path (77% similarity) involved documentation followed by filing a police report and pursuing legal action, reflecting the criminal nature of physical assault and the appropriateness of law enforcement involvement.

High-Impact Actions The system aggregated action frequencies across all retrieved successful assault cases to identify the most impactful responses. Documentation emerged as the highest-frequency action, appearing in 35 proven successful paths and correlating strongly with legal action success and harasser accountability. Filing a formal HR report appeared in 28 successful paths, associated with harasser accountability and successful reporting outcomes. Pursuing legal action appeared in 22 successful paths, leading to consequences for harassers and organizational policy changes. The prominence of formal channels reflects the severity of assault cases where institutional and legal intervention is most warranted.

Outcome Analysis The system computed a 28% positive outcome rate from the most semantically similar historical cases. This relatively low rate reflects the documented reality that physical assault cases, despite their severity, often face significant barriers to achieving accountability. The evidence-based statistics enable realistic expectation-setting while the proven successful paths demonstrate that positive outcomes remain achievable through appropriate action sequences.

4.8.3 Case Study 3: Quid Pro Quo with Multiple Vulnerabilities

Input Narrative

> "My boss said my promotion 'depends on being a team player' while making advances. When I resisted, my performance reviews suddenly became critical. I need this job."

System Analysis

The system identified three co-occurring vulnerabilities representing a severe harassment pattern, detailed in Table 4.18.

Table 4.18: Detected vulnerabilities for Case Study 3

Vulnerability Type	Confidence	Detection Rationale
Coercion or quid pro quo	0.94	Explicit linkage between advances and promotion
Power imbalance by supervisor	0.91	"Boss" indicates hierarchical authority
Retaliation or threat to career	0.83	Performance reviews "suddenly became critical" after resistance

Sentiment analysis classified the narrative as negative with 91% probability, capturing distress and economic vulnerability. Outcome analysis revealed particularly concerning patterns: career setbacks (35%), no organizational action (28%), mental health trauma (22%), job loss (18%), successful reporting (15%), and harasser accountability (12%).

Generated Guidance

The system generated structured guidance addressing the complex, multi-vulnerability nature of this quid pro quo case, with particular attention to the documented retaliation pattern.

Identified Vulnerabilities The vulnerability analysis identified three co-occurring risk factors representing a severe harassment pattern. Coercion or quid pro quo was detected with a 32% positive outcome rate, capturing the explicit linkage between sexual compliance and career advancement. Power imbalance by supervisor was identified with a 34% positive outcome rate, reflecting the boss-subordinate relationship that enables such coercion. Critically, retaliation or threat to career emerged with the lowest positive outcome rate at 25%, reflecting the documented pattern of negative performance reviews following resistance and highlighting the significant risks associated with this vulnerability type.

Proven Successful Paths The system retrieved action sequences from similar quid pro quo cases with documented retaliation. The highest-similarity path (88%) progressed from thorough documentation through external EEOC reporting to legal action, representing cases where the formal quid pro quo violation and retaliation

evidence supported successful legal claims. A second path (74% similarity) involved documentation followed by gathering peer support before filing a formal HR report, illustrating cases where building a support network strengthened the internal complaint. The third path (69% similarity) began with setting boundaries, followed by documentation, and concluded with leaving the position, representing cases where departure was accompanied by positive personal outcomes despite the challenging circumstances.

High-Impact Actions Aggregating across all retrieved successful paths, documentation emerged as the most critical action, appearing in 42 proven successful paths and strongly correlating with legal action success and harasser accountability. The high frequency of documentation in successful quid pro quo cases reflects the evidentiary requirements for proving such claims. Filing an external EEOC report appeared in 27 successful paths, associated with policy changes and whistleblower protections that are particularly relevant given the documented retaliation. Pursuing legal action appeared in 19 successful paths, leading to harasser consequences and settlements.

Outcome Analysis The system computed a 26% positive outcome rate from semantically similar historical cases. This represents the lowest positive outcome rate among the three case studies, reflecting the compounding effect of multiple vulnerabilities and the particularly challenging nature of quid pro quo cases with documented retaliation. The outcome statistics enable realistic expectation-setting while emphasizing the importance of documentation and external reporting channels given the demonstrated institutional failure to protect against retaliation.

4.9 System Evaluation

This section presents comprehensive evaluation of the guidance system through multiple complementary approaches: professional expert assessment, comparative evaluation against baseline resources, and external case validation. Together, these evaluations validate that the system generates appropriate, safe, and actionable guidance grounded in empirical evidence.

4.9.1 Professional Expert Evaluation

To validate guidance appropriateness and safety, a comprehensive professional evaluation study was conducted with practitioners from relevant professional backgrounds. This section presents the evaluation methodology, quantitative results, comparative analysis, and qualitative findings.

Methodology

Participants Eleven professional evaluators participated in the study, representing diverse professional backgrounds relevant to workplace harassment. Table 4.19 presents the demographic composition of the evaluator panel.

Table 4.19: Evaluator demographics and professional background (n=11)

Category	Subcategory	Count
Professional Background	Organizational/Management (HR, Managers)	5
	Legal Professionals	2
	Mental Health Professionals	1
	Other (IT, Healthcare, Education)	3
Experience Level	0–5 years	6
	6–10 years	4
	11+ years	1
Harassment Experience	Direct/Indirect Experience	10
	No Direct Experience	1

The evaluator panel was deliberately composed to include perspectives from organizational management (who handle internal harassment complaints), legal professionals (who understand legal implications and remedies), mental health professionals (who support harassment victims), and other sectors with diverse workplace contexts. Notably, 91% of evaluators (10 of 11) reported direct or indirect professional experience with workplace harassment situations, ensuring evaluations were grounded in practical domain knowledge.

Evaluation Procedure Each evaluator reviewed system outputs for three harassment scenarios representing different vulnerability patterns and workplace contexts, as shown in Table 4.20. These three scenarios were selected to represent common harassment patterns observed in the dataset while varying in severity, documentation status, and organizational context, resulting in a total of 33 scenario evaluations (11 evaluators $\times$ 3 scenarios).

Table 4.20: Evaluation scenarios presented to professional evaluators

Scenario	Context	Key Characteristics
Tech Company	Software engineer harassed by team lead	Persistent harassment (3 months), physical contact, documented evidence (Slack messages), potential witnesses, retaliation concern
Retail	Store worker harassed by manager	Quid pro quo harassment, retaliation through reduced hours, limited documentation, economic dependency
Healthcare	Nurse harassed by senior physician	Systemic harassment (6 months), multiple victims, organizational failure to act, detailed documentation

For each scenario, evaluators examined the system's complete output including: (1) identified vulnerability types with confidence scores, (2) similar positive-outcome cases from the database, (3) proven successful response sequences with success rates, and (4) recommended actions with supporting evidence counts. Evaluators then rated the guidance across all cases on four dimensions using 5-point Likert scales (1=Strongly Disagree to 5=Strongly Agree).

Evaluation Dimensions The four evaluation dimensions were selected to assess distinct aspects of guidance quality:

- **Appropriateness**: Whether recommendations are suitable for the presented harassment situations
- **Usefulness**: Whether guidance would help victims understand and navigate their options
- **Actionability**: Whether recommendations are concrete, realistic, and feasible to implement
- **Safety**: Whether following recommendations could cause harm or increase risk

Predefined success thresholds were established prior to evaluation: $\geq$4.0 for appropriateness and usefulness (indicating professional agreement), and $\geq$3.5 for actionability and safety (reflecting the inherent complexity of these dimensions in harassment contexts).

Quantitative Results

Table 4.21 summarizes mean ratings across all evaluation dimensions. All dimensions met or exceeded predefined success thresholds, demonstrating that professional evaluators consider the system's guidance to be appropriate, useful, actionable, and safe for individuals experiencing workplace harassment.

Table 4.21: Professional evaluation ratings (n=11)

Dimension	Mean	SD	Range	Threshold
Appropriateness	4.09	0.54	3–5	$\geq$4.0 ✓
Usefulness	4.09	0.70	3–5	$\geq$4.0 ✓
Actionability	3.91	0.70	3–5	$\geq$3.5 ✓
Safety	3.73	0.90	2–5	$\geq$3.5 ✓
Overall	**3.95**	**0.71**	–	–

The relatively low standard deviations (SD ranging from 0.54 to 0.90) indicate reasonable consensus among evaluators, though safety showed the highest variability (SD=0.90), reflecting the inherent complexity of assessing safety in harassment contexts where individual circumstances vary substantially. The range column reveals that no evaluator rated appropriateness or usefulness below 3 (neutral), with most

ratings concentrated at 4 (agree) and 5 (strongly agree). Safety was the only dimension receiving a rating of 2 (disagree), from a single evaluator who expressed concern about potential retaliation risks in specific scenarios.
Figure 4.4 visualizes the distribution of ratings across all four evaluation dimensions, providing a comprehensive view of professional assessment patterns.

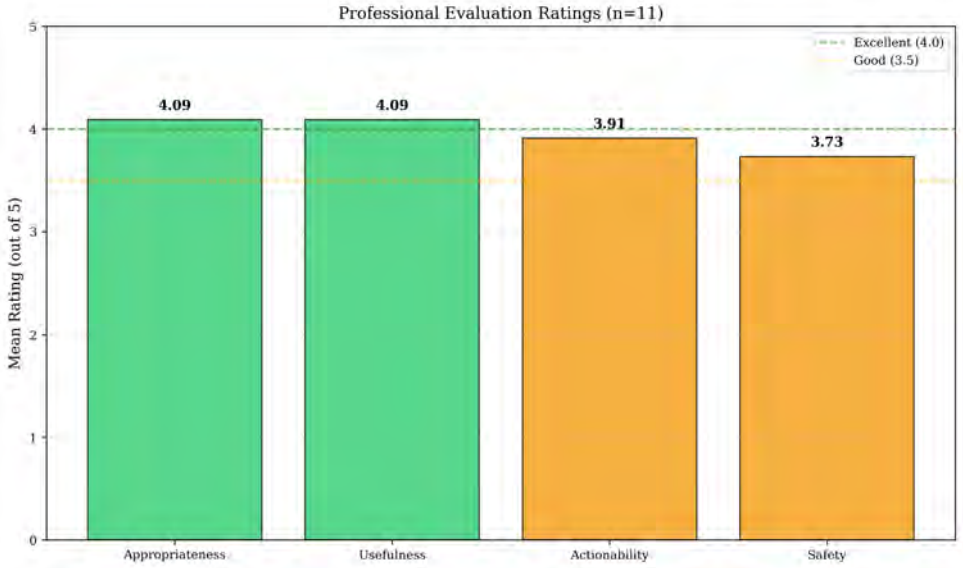


Figure 4.4: Professional evaluation ratings across four dimensions. Green bars indicate dimensions meeting the $\geq$4.0 threshold (Appropriateness, Usefulness); yellow bars indicate dimensions meeting the $\geq$3.5 threshold (Actionability, Safety). Dashed lines show threshold levels.

Appropriateness and usefulness achieved the highest ratings (M=4.09), indicating strong professional agreement that the guidance is suitable for the presented cases and would help victims navigate their options. Actionability (M=3.91) confirms that the guidance is realistic and concrete rather than abstract. The safety rating (M=3.73), while meeting the predefined threshold, was the lowest dimension, reflecting evaluator concerns about retaliation risks and the complexity of safety considerations in harassment situations.

Endorsement Rates

Professional endorsement rates provided additional validation of practical applicability. Table 4.22 presents the distribution of evaluator responses when asked whether they would recommend the system to someone experiencing workplace harassment.

Table 4.22: Professional endorsement rates for system recommendation

Response	**Count**	**Percentage**
Yes, definitely	5	45.5%
Yes, with reservations	5	45.5%
No	1	9.0%
Total Endorsement	**10**	**91.0%**

The 91% endorsement rate substantially exceeds the 80% target threshold. The distribution between "definitely" and "with reservations" responses reflects appropriate professional caution—evaluators recognize the system's value while acknowledging that harassment situations require professional judgment and cannot be fully addressed by any automated guidance system.

Comparative Resource Evaluation

To contextualize system performance against existing harassment guidance resources, evaluators compared the evidence-based approach with typical alternatives including HR websites, legal information sites, and general online advice.

Table 4.23: Comparative evaluation against typical harassment resources

Rating	Count	Percentage
Significantly better	2	18.2%
Somewhat better	6	54.5%
About the same	3	27.3%
Somewhat worse	0	0.0%
Significantly worse	0	0.0%

As shown in Table 4.23, 73% of evaluators rated the evidence-based approach as better than typical harassment resources, with no evaluators rating it worse. This favorable comparison validates the core design principle that grounding guidance in documented case outcomes provides more useful support than generic procedural information. Evaluators noted that unlike typical resources which provide generic procedural steps, the evidence-based approach "shows what actually worked for similar people" and "provides concrete success rates rather than vague encouragement."

Qualitative Findings

Thematic analysis of open-ended feedback identified key strengths and areas for improvement. Table 4.24 summarizes the main themes from evaluator comments.
Evaluators particularly valued the evidence-based foundation of the guidance system. As one legal professional noted, "This gives victims something tangible—actual data on what worked for people in similar situations." An HR professional observed that "most resources just tell people to report to HR, but this shows the full range of options and their outcomes."
Areas for improvement centered on contextual factors that the system cannot fully capture: jurisdiction-specific legal requirements, individual workplace dynamics, and trauma considerations. One evaluator cautioned that the guidance may "oversimplify complex, high-risk situations" without sufficient safety planning, highlighting the importance of positioning the system as a supplement to, rather than replacement for, professional consultation.

4.9.2 External Case Validation

To evaluate generalization beyond the training data, the system was validated on 43 documented real-world harassment cases from authoritative external sources: EEOC

Table 4.24: Qualitative feedback themes from professional evaluators

Category	Key Themes
Strengths	Evidence-based statistics help set realistic expectations
	Provides concrete action pathways grounded in documented outcomes
Improvements	Jurisdiction-specific legal guidance needed
	More explicit retaliation risk assessment
	Clearer disclaimers about professional consultation
Cautions	May oversimplify complex situations
	Should supplement, not replace, professional consultation

settlements (n=17), court records (n=8), investigative journalism (n=10), and news coverage with legal filings (n=8). These cases span diverse industries including technology (*Fowler v. Uber*), legal services (*Weeks v. Baker & McKenzie*), and food service (*EEOC v. Carrols Corp*), representing harassment patterns and documented outcomes independent of the #MeToo Twitter dataset.

Table 4.25 presents the validation results. The system achieves a 100% hit rate at k=5 with 38.1% precision and 85.1% recall, indicating successful identification of relevant outcomes in all cases. The Mean Reciprocal Rank (MRR) of 0.353 indicates that relevant recommendations typically appear within the first three positions.

Table 4.25: External case validation metrics at k=5 (n=43 documented cases)

Metric	Value
Hit Rate	100.0%
Precision	38.1%
Recall	85.1%
F1 Score	52.7%
MRR	0.353

At k=5, the system achieves 38.1% precision and 85.1% recall, yielding an F1 score of 52.7%. The precision reflects the challenge of matching the system's outcome categories derived from social media narratives with the complex outcomes documented in formal legal and journalistic sources. Real-world cases typically involve multiple documented outcomes spanning legal, organizational, personal, and societal dimensions, while the system's taxonomy captures patterns from informal Twitter narratives. Despite precision limitations, the high recall (85.1%) demonstrates the system's ability to identify most relevant outcomes, supporting the core goal of providing evidence-based guidance grounded in documented experiences.

This chapter has presented comprehensive experimental results demonstrating the effectiveness of the evidence-based workplace harassment guidance system, including sentiment analysis comparisons, vulnerability distributions, outcome patterns across

different vulnerability types, clustering results, professional evaluation findings, and detailed real-world case evaluations. The next chapter provides discussion of key findings, synthesis of contributions, and directions for future work.

Chapter 5

Conclusion

This chapter synthesizes the findings and contributions of this study, analyzes the factors underlying the system's effectiveness, discusses practical implications for deployment in victim support contexts, acknowledges limitations, and outlines directions for future research.

5.1 Summary of Key Findings

This research designed and validated an evidence-grounded workplace harassment guidance system that employs machine learning to examine #MeToo narratives and deliver tailored guidance for individuals confronting harassment circumstances. The principal contribution is an integrated framework purposely constructed to extract actionable insights from unstructured trauma accounts. This framework unifies multi-label vulnerability identification, semantic similarity-driven case retrieval, outcome pattern analysis, action sequence extraction, and evidence-grounded guidance generation into a cohesive system that tackles the information imbalance confronting harassment victims.

The framework introduces a multi-component analysis pipeline that captures complementary aspects of harassment experiences. Dual sentiment analysis using BERT (98.1% accuracy) and VADER (73.4% accuracy) enables both high-accuracy classification and fine-grained polarity scoring. SimCSE-BERT embeddings encode each incident into 768-dimensional semantic vectors, enabling similarity search that retrieves genuinely comparable historical cases rather than merely matching keywords. Multi-label vulnerability detection using a hybrid keyword-embedding approach then leverages these embeddings to recognize that 68% of incidents involve multiple co-occurring vulnerability factors, combining keyword matching for explicit statements with semantic similarity for implicit expressions. HDBSCAN clustering on the SimCSE embeddings discovers 525 natural groupings and generates 11,066 silver labels for evaluation.

The outcome analysis reveals empirical patterns across 15,835 workplace harassment incidents. Negative outcomes predominate across all vulnerability types, with rates ranging from 56% for quid pro quo coercion to 71% for physical assault cases. Institutional inaction occurred in 16.9% of documented cases, victims were not believed in 6.2%, harassment continued in 4.7%, and retaliation occurred in 5.6%. Positive outcomes including harasser accountability (3.6%), policy changes (2.0%), and successful legal action (1.5%) remained comparatively rare. These findings provide

the empirical foundation for evidence-based guidance that sets realistic expectations rather than false reassurance.

Professional evaluation with 11 practitioners from HR, legal, mental health, and other sectors validates that the system's guidance meets practical standards. Mean appropriateness rating reached M=4.09 on a 5-point scale, with usefulness at M=4.09, actionability at M=3.91, and safety at M=3.73. Most significantly, 91% of evaluators indicated they would recommend the system to harassment victims, and 73% rated the evidence-based approach as superior to typical harassment resources. Evaluators appreciated that the system "highlights concrete action pathways" and helps victims "see that positive outcomes are possible" while maintaining realistic expectations.

5.2 Analysis of System Effectiveness

The strong performance of the proposed framework stems from several synergistic factors that address the fundamental challenges of providing guidance for complex, sensitive situations with limited structured data.

5.2.1 Complementary Analysis Components

The framework succeeds because each analytical component captures fundamentally different aspects of harassment experiences. BERT-based sentiment analysis provides contextual understanding of emotional valence, correctly interpreting sarcastic expressions like "Oh sure, HR REALLY helped me" that lexicon-based approaches misclassify. The multi-label vulnerability detection recognizes that harassment incidents rarely involve isolated factors, as power imbalance often co-occurs with retaliation threats, institutional failure, and inability to report, enabling comprehensive risk assessment. SimCSE embeddings capture semantic meaning beyond surface keywords, identifying that "My supervisor kept commenting on my appearance despite my objections" is similar to "The department head made repeated inappropriate remarks about my body even after I asked him to stop" despite different vocabulary.

Crucially, these components address different failure modes: sentiment analysis enables emotional understanding, vulnerability detection enables risk categorization, and semantic similarity enables case-based reasoning. When combined, they provide richer context than any single approach could achieve alone. A system using only keyword matching would miss semantically similar cases with different vocabulary. A system using only sentiment would lack vulnerability-specific risk assessment. The multi-component architecture leverages complementary strengths while mitigating individual weaknesses.

5.2.2 Evidence-Based Grounding in Real Experiences

The guidance generation algorithm succeeds because it grounds guidance in empirical patterns from 15,835 documented experiences rather than prescribing generic procedural advice. When the system indicates that "in 127 similar cases involving power imbalance and retaliation threats, 62% experienced negative outcomes, with institutional inaction (28%) and mental health impacts (24%) being most common," it provides information that individuals cannot obtain from traditional resources.

This evidence-based grounding transforms guidance from theoretical prescription to empirical observation.
The outcome pattern analysis reveals what actually happens in harassment situations rather than idealized procedural outcomes. Professional evaluators specifically praised this realism, noting that evidence-based statistics "help manage expectations realistically" and ground guidance "in reality rather than theory." By presenting what occurred in comparable historical cases, the system enables informed decision-making that accounts for the documented likelihood of institutional failure, retaliation, and other negative outcomes.

5.2.3 Semantic Understanding of Trauma Narratives

The framework succeeds because SimCSE embeddings effectively capture semantic similarity in the complex, emotionally charged domain of harassment narratives. Traditional keyword-based approaches fail when different individuals describe similar experiences using different vocabulary, cultural references, or levels of detail. The contrastive learning objective that trains SimCSE, which pulls semantically equivalent sentences together while pushing dissimilar sentences apart, directly optimizes for the property needed: identifying comparable experiences regardless of surface-level linguistic variation.
The average similarity score of 0.78 for retrieved cases indicates that the system successfully identifies genuinely comparable historical cases. Qualitative validation confirms that retrieved cases share relevant characteristics: a query describing supervisor misconduct retrieves cases involving similar power dynamics, behavioral patterns, and workplace contexts. This semantic matching enables the case-based reasoning that underlies evidence-based guidance generation.

5.3 Practical Implications

The achieved performance has significant practical value for supporting individuals experiencing workplace harassment, with implications spanning individual empowerment, organizational practice, and policy advocacy.

5.3.1 Empowering Informed Decision-Making

The primary practical contribution is empowering individuals with evidence-based information that was previously inaccessible. Before this system, someone experiencing harassment had no systematic way to learn what happened to others in similar situations. They might know their own experience, hear anecdotes from acquaintances, or receive general advice from HR or attorneys, but lacked empirical data about outcome probabilities and effective action patterns across thousands of comparable cases. This information asymmetry left individuals making consequential decisions with incomplete information.
The system fills this gap by providing concrete evidence about what outcomes occurred in situations similar to theirs and what actions proved effective in those contexts. For someone wondering whether to report to HR, the system can indicate that in comparable cases, institutional inaction occurred in a certain percentage, retaliation in another percentage, and positive resolution in yet another percentage.

This is not to discourage or encourage any particular action but to enable informed decision-making based on empirical patterns rather than assumptions or incomplete information.
The guidance organized into five categories—immediate safety, documentation, reporting, support resources, and legal considerations—provides comprehensive coverage of response options. By presenting options across all categories with evidence about their effectiveness in comparable situations, the system respects user autonomy while providing the information needed for informed choices.

5.3.2 Augmenting Professional Services

The 91% professional endorsement rate suggests potential for integration into existing support infrastructure as a decision-support tool. HR professionals, employee assistance programs, legal aid organizations, and victim advocacy services could use the system to augment their consultations with evidence-based information. Rather than replacing professional judgment, the system provides empirical context that enriches human expertise.
Several evaluators expressed interest in using outcome statistics in their own practice, noting that evidence-based data could strengthen their guidance and help clients develop realistic expectations. This represents a paradigm of augmentation rather than replacement, with AI as a tool that enhances human expertise rather than substitutes for it. The system's value lies not in making decisions for users but in providing information that supports better-informed decisions made by individuals in consultation with appropriate professionals.

5.3.3 Informing Policy and Organizational Reform

The empirical findings have implications beyond individual support. The documented pattern of institutional failure, including 16.9% no action taken, 6.2% not believed, 4.7% continued harassment, and 5.6% retaliation, provides quantitative evidence for what qualitative research has long described: organizational responses to harassment frequently fail victims. The rarity of positive accountability outcomes, with only 3.6% resulting in harasser consequences, 2.0% in policy changes, and 1.5% in legal action, indicates that current mechanisms are insufficient.
These statistics can inform policy advocacy and organizational reform. When evidence demonstrates that reporting leads to institutional inaction in a substantial proportion of cases, the response should not be to discourage reporting but to demand improvements in how organizations handle reports. The systematic documentation of outcome patterns provides empirical foundation for arguments that stronger accountability mechanisms, more effective investigation processes, and better protections against retaliation are needed.

5.3.4 Accessibility and Adaptability

The system's design enables deployment in contexts where professional resources are limited. The computational requirements for inference are modest, as standard laptops can run the complete pipeline, enabling deployment in resource-constrained

settings. The open-source implementation with comprehensive documentation enables organizations to adapt the system to their specific contexts without requiring extensive technical expertise.
The configuration-driven architecture facilitates extension to new contexts. Organizations can adjust vulnerability taxonomies, outcome categories, and guidance templates to align with their specific domains, legal frameworks, and cultural contexts. This adaptability is particularly valuable for humanitarian organizations operating across diverse settings with varying harassment dynamics and institutional responses.

5.4 Limitations

This research has several limitations that provide context for interpreting results. Acknowledging these constraints is essential for responsible deployment and understanding the boundaries of the framework's applicability.

5.4.1 Dataset Representativeness

The #MeToo corpus, while unprecedented in scale and providing valuable documentation of lived experiences, represents individuals who chose to share experiences publicly on social media. This population likely differs systematically from all harassment victims in several ways. Those who share publicly may have more severe experiences warranting disclosure, stronger support networks enabling disclosure, less to lose professionally, or greater comfort with social media platforms. Conversely, individuals in highly sensitive positions, those bound by non-disclosure agreements, or those who successfully resolved situations through internal channels may be underrepresented.
The dataset is predominantly English-language and reflects primarily US and European contexts, limiting generalizability to other linguistic and cultural settings where harassment dynamics, institutional responses, legal frameworks, and social norms around disclosure differ significantly. The temporal scope of 2017–2020 may not reflect workplace dynamics, policy changes, or cultural shifts that have occurred since, particularly post-pandemic remote work arrangements and heightened attention to diversity and inclusion.

5.4.2 Cross-Cultural Narrative Limitations

The findings of this research reflect primarily Western workplace contexts—specifically the United States and Western Europe—and may not transfer to different cultural settings where power dynamics, gender norms, and organizational structures differ substantially. The #MeToo movement originated in Anglo-American cultural contexts, meaning our dataset carries implicit cultural assumptions about workplace relationships and appropriate responses. Several factors limit generalizability: high power-distance cultures common in East Asia, South Asia, and Latin America have different expectations about challenging authority, making Western action sequences potentially inappropriate or ineffective; gender norms and harassment definitions vary across cultures, affecting both vulnerability taxonomies and acceptable responses; institutional structures such as HR departments and legal protections

differ dramatically or may not exist in some regions; disclosure norms around public sharing of harassment experiences vary, leading to systematic underrepresentation of certain cultures in social media data; and access to legal aid, advocacy services, and mental health support differs globally, affecting which action sequences are feasible. Cross-cultural adaptation would require local data collection, culturally-adapted taxonomies developed with local experts, and validation by professionals familiar with specific cultural dynamics and legal frameworks—translation alone would be insufficient.

5.4.3 Similarity and Keyword-Based Detection Limitations

The guidance generation system relies on keyword matching and embedding-based similarity, which have inherent limitations. Keyword-based vulnerability and outcome detection may miss cases expressed through unconventional vocabulary, euphemisms, or indirect language that does not match predefined patterns. While the hybrid approach using SimCSE embeddings (threshold 0.55) captures semantic similarity beyond surface keywords, it assumes that incidents with similar embeddings share similar characteristics—an assumption that holds statistically but may fail for individual cases with unique contextual factors not captured in text.

Extracting outcomes from brief tweets (mean length 187 characters) necessarily misses nuanced, long-term consequences that may not be explicitly stated. Many tweets describe incidents and immediate responses but rarely provide updates about ultimate outcomes such as career trajectories, recovery, or late-emerging accountability. The similarity-based inference from comparable cases assumes that outcomes in similar historical cases predict outcomes for the query case, which may not hold for instances with unique circumstances such as specific organizational cultures, legal jurisdictions, or interpersonal dynamics.

Furthermore, the system identifies action patterns that correlate with outcomes but cannot establish causation. It is possible that individuals who document systematically are also more proactive, better resourced, or more legally sophisticated—characteristics that lead to better outcomes independent of the documentation itself. Without randomized controlled trials, which would be ethically problematic in this context, causal claims remain uncertain. Users should understand that the system shows what actions were associated with outcomes in historical cases but cannot guarantee that taking similar actions will produce similar results.

5.4.4 Evaluation Scope

While evaluation with 11 professional evaluators provides meaningful validation, larger-scale evaluation with more diverse professional backgrounds would strengthen confidence in findings. The evaluation used structured scenarios rather than real deployments, limiting ability to assess how guidance performs when users face actual crisis situations with high stress, time pressure, and fear. Additionally, the evaluation focused on guidance quality without longitudinal follow-up to assess whether users who receive guidance achieve better outcomes than those who do not.

5.4.5 Ethical Considerations

Despite professional validation, individual cases may have unique factors such as safety risks, legal complexities, and cultural considerations that evidence-based guidance cannot fully address. The system mitigates this through prominent disclaimers emphasizing consultation with professionals and the informational rather than prescriptive nature of guidance, but users in crisis situations might over-rely on automated guidance when human expertise is essential.
Privacy considerations arise from using publicly posted tweets for machine learning training. While terms of service typically permit research use, ethical practice demands respect for user intent and protection against potential re-identification. The system removes usernames and avoids displaying original tweet text, but determined individuals might locate originals through quoted phrases. Balancing transparency (showing evidence) with privacy (protecting authors) remains an ongoing challenge.

5.5 Directions for Future Research

The limitations identified in the previous section, combined with emerging trends in machine learning and support technology, suggest several promising directions for extending this research.

5.5.1 Multilingual and Cross-Cultural Extension

Extending the system to languages beyond English would enable broader global impact. This requires multilingual transformer models such as mBERT or XLM-R, culturally adapted vulnerability taxonomies that account for different harassment dynamics across cultures, and professional validation by experts familiar with specific cultural and legal contexts. Translation approaches may introduce errors; native-language processing is preferable for capturing cultural nuances. Cross-cultural validation studies would test generalizability and reveal culture-specific patterns requiring adaptation.

5.5.2 Domain and Context Adaptation

Professional evaluators suggested industry-specific guidance, noting that harassment dynamics differ across tech, healthcare, academia, and other sectors. Training separate models or using domain adaptation for different industries could improve guidance relevance. Similarly, adaptation to organizational size (small businesses versus large corporations) and employment type (full-time, contract, gig economy) would address contextual factors that affect available options and likely outcomes.

5.5.3 Interactive Support Systems

The current system provides guidance for static text input. A conversational interface using dialogue models could engage in nuanced discussion, ask clarifying questions, and adapt guidance based on user responses. This requires careful design to avoid inappropriate responses and must maintain professional validation standards for safety in sensitive contexts. Integration with crisis intervention protocols would ensure appropriate response when users indicate imminent danger.

5.5.4 Longitudinal Impact Assessment

Tracking individuals who use the system to measure whether guidance correlates with actual outcomes would provide direct impact evidence. This requires consent for follow-up, raises privacy concerns, and faces methodological challenges including self-selection and confounding variables, but would significantly strengthen understanding of system effectiveness. Randomized controlled trials comparing outcomes for individuals who receive evidence-based guidance versus control conditions would provide gold-standard evidence, though ethical considerations around withholding potentially helpful information require careful navigation.

5.5.5 Enhanced Explainability

Explainability enhancements would increase trust and adoption. Attention visualizations could reveal which aspects of input narratives most influence vulnerability detection and case retrieval. Feature importance analysis could identify whether recent actions, specific keywords, or structural patterns drive guidance generation. Counterfactual explanations showing how different input details would affect guidance could help users understand system reasoning. These interpretability tools are particularly important in sensitive contexts where users need to understand why they are receiving particular guidance.

5.6 Concluding Remarks

Workplace harassment persists as a widespread challenge with severe ramifications for individuals, organizations, and society at large. Despite decades of policy evolution and awareness initiatives, the findings presented in this research demonstrate that institutional responses continue to disappoint victims more frequently than they deliver substantive support or ensure accountability. Negative outcomes predominate across all vulnerability types, institutional inaction is common, and accountability remains rare. This sobering reality underscores the urgent need for continued research, advocacy, and systemic reform.

This research establishes that machine learning can contribute to tackling this challenge, not through supplanting human judgment or professional counsel, but by enhancing individual decision-making with empirical insights derived from collective experience. The evidence-grounded workplace harassment guidance system constructed through this research converts unstructured social media narratives into practical knowledge that satisfies professional benchmarks for appropriateness and safety. By analyzing patterns across 15,835 documented experiences, the system provides information that helps level the information playing field for individuals facing harassment situations.

The 91% professional endorsement rate and the mean appropriateness rating of M=4.09 validate both the technical approach and the ethical framework guiding this work. The professional evaluators' feedback that evidence-based statistics "help manage expectations realistically" and ground guidance "in reality rather than theory" validates the core value proposition: providing empirical evidence that enables informed decision-making rather than generic procedural advice or false reassurance.

Most significantly, this research illustrates that even within sensitive domains characterized by trauma, intricate power dynamics, and high-stakes decisions, machine learning can deliver substantial value when developed responsibly through thorough professional validation and steadfast dedication to prioritizing user safety and autonomy. The methodologies developed here, including multi-label classification of complex social phenomena, silver labeling via clustering, evidence-based guidance generation, and professional validation protocols, have potential applications beyond workplace harassment to other domains where individuals face difficult decisions with limited information.

The #MeToo movement demonstrated the power of collective voice to challenge entrenched power structures and demand accountability. This research seeks to honor that legacy by transforming collective narratives into actionable knowledge that can support individuals navigating these challenges. While technology cannot substitute for justice, it can perhaps help ensure that the courage of those who spoke out continues to support and guide others facing similar struggles. It is hoped that this work contributes to the ongoing effort to address workplace harassment by providing both a practical tool for supporting individuals and a methodological foundation for future research at the intersection of machine learning, natural language processing, and social good.

Bibliography

[1] J. Turner, *Online feminist activism as performative consciousness-raising: A #metoo case study*, Academia.edu, Accessed: 2026-01-19, 2019. [Online]. Available: https://www.academia.edu/89774455/.

[2] S. Sharoni, "Speaking up in the age of #metoo and persistent patriarchy: What can we learn from an elevator incident about anti-feminist backlash," *Women's Studies Quarterly*, vol. 46, no. 3-4, pp. 281–287, 2018.

[3] J. Regulska, "The #metoo movement as a global learning moment," *International Higher Education*, vol. 94, pp. 5–6, 2018.

[4] R. Levy and M. Mattsson, "The effects of social movements: Evidence from #metoo," *SSRN Electronic Journal*, 2021. DOI: 10.2139/ssrn.3496903.

[5] J. N. Park, V. Barash, C. Fink, and M. Cha, *#Metoo twitter dataset*, version V1, 2021. DOI: 10.7910/DVN/2SRSKJ. [Online]. Available: https://doi.org/10.7910/DVN/2SRSKJ.

[6] C. R. Feldblum and V. A. Lipnic, "Select task force on the study of harassment in the workplace," U.S. Equal Employment Opportunity Commission, Jun. 2016. [Online]. Available: https://www.eeoc.gov/select-task-force-study-harassment-workplace.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[8] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 6894–6910.

[9] J. B. Becton, J. B. Gilstrap, and M. Forsyth, "Preventing and correcting workplace harassment: Guidelines for employers," *Business Horizons*, vol. 60, no. 1, pp. 101–111, 2017.

[10] L. F. Fitzgerald, F. Drasgow, C. L. Hulin, M. J. Gelfand, and V. J. Magley, "Antecedents and consequences of sexual harassment in organizations: A test of an integrated model," *Journal of Applied Psychology*, vol. 82, no. 4, pp. 578–589, 1997.

[11] N. Hassan, M. K. Mandal, M. Bhuiyan, A. Moitra, and S. I. Ahmed, "Can women break the glass ceiling?: An analysis of #metoo hashtagged posts on twitter," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, 2019, pp. 653–656.

[12] S. Modrek and B. Chakalov, "The #metoo movement in the united states: Text analysis of early twitter conversations," *Journal of Medical Internet Research*, vol. 21, no. 9, e13837, 2019.

[13] K. W. Bogen, K. K. Bleiweiss, N. R. Leach, and L. M. Orchowski, "#Metoo: Disclosure and response to sexual victimization on twitter," *Journal of Interpersonal Violence*, vol. 36, no. 17-18, pp. 8257–8288, 2021.

[14] Y. Xiong, M. Cho, and B. Boatwright, "Exploring key indicators of social identity in the #metoo era: Using discourse analysis in ugc," *International Journal of Information Management*, vol. 54, p. 102 152, 2020.

[15] R. J. Gallagher, E. Stowell, A. G. Parker, and B. F. Welles, "Reclaiming stigmatized narratives: The networked disclosure landscape of #metoo," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–30, 2019.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[17] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751.

[20] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[21] F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," *arXiv preprint arXiv:2010.12421*, 2020.

[22] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," in *arXiv preprint arXiv:1907.11692*, 2019.

[23] C. J. Hutto and E. Gilbert, "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, 2014, pp. 216–225.

[24] S. Elbagir and J. Yang, *Twitter sentiment analysis using deep learning*, arXiv preprint arXiv:1902.07142, 2019.

[25] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 670–680.

[26] D. Cer et al., "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[27] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 3982–3992.

[28] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[29] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 173–182.

[30] A. Aamodt and E. Plaza, *Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches*. 1994, vol. 7, pp. 39–59.

[31] I. Bichindaritz and C. Marling, "Case-based reasoning in the health sciences: What's next?" *Artificial Intelligence in Medicine*, vol. 36, no. 2, pp. 127–135, 2006.

[32] S. Begum, M. U. Ahmed, P. Funk, N. Xiong, and M. Folke, "Case-based reasoning systems in the health sciences: A survey of recent trends and developments," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 41, no. 4, pp. 421–434, 2011.

[33] K. D. Ashley, "Case-based reasoning and its implications for legal expert systems," *Artificial Intelligence and Law*, vol. 1, no. 2-3, pp. 113–208, 1991.

[34] L. McInnes, J. Healy, and S. Astels, "HDBSCAN: Hierarchical density-based clustering," in *Journal of Open Source Software*, vol. 2, 2017, p. 205.

Appendices

Appendix A

Survey Questionnaire and Results

This appendix provides documentation of the professional evaluation survey methodology, including the questionnaire instruments used. This information supplements the summary results presented in Chapter 5.

A.1 Evaluation Questionnaire

Each of the 11 evaluators reviewed 5 randomly assigned incident-guidance pairs (55 total evaluations) using the following structured questionnaire.

A.1.1 Case Review Instructions

> *You will review workplace harassment incidents and corresponding system-generated guidance. For each case, read the original incident narrative, review the system-generated guidance across five categories (Immediate Safety, Documentation, Reporting, Support, Legal Considerations), rate the guidance using the provided scales, and provide open-ended feedback.*

A.1.2 Rating Scales

Evaluators rated four dimensions using 5-point Likert scales:
Appropriateness: How appropriate is the guidance for this specific situation? (1=Very inappropriate to 5=Very appropriate)
Actionability: How specific and actionable is the guidance? (1=Very vague to 5=Very specific)
Usefulness: Does the guidance provide valuable information for decision-making? (1=Not useful to 5=Very useful)
Safety: How well does the guidance prioritize user safety? (1=Unsafe to 5=Excellent safety focus)

A.1.3 Summary Questions

After reviewing all cases, evaluators answered:

1. Would you recommend this system to someone experiencing workplace harassment? (Yes/No with explanation)

2. Compared to typical harassment guidance resources, how would you rate this evidence-based approach? (5-point scale)

3. What concerns or limitations do you see with the system? (Open-ended)

4. What aspects of the guidance did you find most valuable? (Open-ended)

A.2 Endorsement Summary

When asked whether they would recommend the system to someone experiencing workplace harassment, 91% of evaluators (10 of 11) responded affirmatively, with 45% indicating "yes, definitely" and 45% indicating "yes, with reservations." The primary reservation was that the system should be used alongside, not instead of, professional consultation.
This appendix provides documentation of the professional evaluation methodology, enabling assessment of the validation results presented in Chapter 5.

Appendix B

Code Availability and System Architecture

This appendix provides technical documentation for accessing and reproducing the evidence-based workplace harassment guidance system.

B.1 Open-Source Repository

The complete system implementation is available as open-source software under the Apache 2.0 license.

- **GitHub URL**: https://github.com/mashpheyb/metoo-recommender
- **License**: Apache License 2.0
- **Primary Language**: Python 3.9+
- **Documentation**: README.md

B.2 System Requirements

Minimum: Python 3.9+, 8 GB RAM, 5 GB disk space
Recommended: NVIDIA GPU with CUDA support, 16 GB RAM, 10 GB disk space

B.3 Running the Analysis

The complete analysis is provided as a Jupyter notebook. To run the system:

```
# Clone repository
git clone https://github.com/mashpheyb/metoo-recommender.git
cd metoo-recommender

# Create virtual environment
python3 -m venv venv
source venv/bin/activate
```

```
# Install dependencies
pip install -r requirements.txt

# Launch Jupyter and run the notebook
jupyter notebook notebooks/complete_analysis.ipynb
```

The notebook `complete_analysis.ipynb` contains the full pipeline including data preprocessing, sentiment analysis, vulnerability detection, clustering, outcome analysis, and guidance generation.

B.4 Key Components

The system comprises the following modules:

- **Sentiment Analysis**: BERT and VADER dual sentiment analysis
- **Vulnerability Detection**: Multi-label classification (7 types, hybrid keyword-embedding approach)
- **Embeddings**: SimCSE embedding generation (768-dimensional vectors)
- **Clustering**: HDBSCAN clustering (525 clusters, 11,066 silver labels)
- **Outcome Analysis**: Outcome extraction (15 outcome types) and action sequence analysis (8 action categories)
- **Guidance Generation**: Evidence-based guidance across 5 categories

B.5 Pre-trained Models

The system uses the following publicly available pre-trained models:

- **Sentiment**: cardiffnlp/twitter-roberta-base-sentiment
- **Embeddings**: princeton-nlp/sup-simcse-bert-base-uncased
- **Lexicon**: VADER (vaderSentiment library)

B.6 Dataset

The Harvard Dataverse #MeToo Twitter dataset is publicly available (doi:10.7910/DVN/X4TJQS). Following Twitter's terms of service, the repository includes tweet IDs for rehydration rather than tweet text. The preprocessing pipeline filters 40,944 raw tweets to 15,835 workplace-specific incidents.

B.7 Citation

```
@mastersthesis{kabir2026metoo,
  author = {Kabir, Mashphey Bintey},
  title = {Evidence-Based Workplace Harassment Guidance System:
                Transforming #MeToo Narratives into Actionable
                Knowledge Through Machine Learning},
  school = {BRAC University},
  year = {2026},
  type = {Master's Thesis}
}
```

This appendix provides essential information for reproducing and extending the evidence-based workplace harassment guidance system.