Al-Quds Open University

Faculty of Graduate Studies and Scientific Research

Master of Information Technology

Empowering Palestinian Voices Using Text Mining Techniques to Overcome Social Media Expression Restrictions

تمكين أصوات الشعب الفلسطيني باستخدام تقنيات "تنقيب النصوص" لتجاوز قيود التعبير عن الرأي على وسائل التواصل الاجتماعي.

THESIS

by

Maab Abd Alqaher Srour

Student ID: 0330012210143

Supervisor

Dr. Mohamed Dweib

Submitted in partial fulfilment of the requirements
For the Degree of Master of Information Technology at the Faculty of Graduate Studies

Ramallah, Palestine

October 2025

جامعة القدس المفتوحة

كلية الدراسات العليا والبحث العلمي

ماجستير تكنولوجيا المعلومات

Empowering Palestinian Voices Using "Text Mining" Techniques to Overcome Social Media Expression Restrictions

تمكين أصوات الشعب الفلسطيني باستخدام تقنيات "تنقيب النصوص" لتجاوز قيود التعبير عن الرأي على وسائل التواصل الاجتماعي.

الرسالة

إعداد

مآب عبد القاهر سرور

رقم الطالب: 0330012210143

المشرف

د. محمد ذويب

قدمت هذه الرسالة استكمالاً لمتطلبات الحصول على درجة الماجستير في تكنولوجيا المعلومات في عمادة الدراسات العليا والبحث العلمي

رام الله، فلسطين

أكتوبر 2025

Examination Committee Page

The Examination Committee certifies that this thesis entitled *Empowering Palestinian Voices Using Text Mining Techniques to Overcome Social Media Expression Restrictions*
submitted by Maab Abd Alqaher Srour has been examined and approved as fulfilling the requirements for the degree of Master of Information Technology.

Committee Members:

Committee Supervisor: Dr. Mohamed Dweib

Signature: Mohamed Dweib
Date: 01/24/2026

Committee First Member: Dr. Nael Abu–Halaweh

Signature: ....
Date: 01/26/2026

Committee Second Member: Dr. Hamza Mujahed

Signature: ....
Date: 01/24/2026

Al-Quds Open University

2025

Declaration

I, Maab Abd Alqaher Srour, hereby declare that the work presented in this thesis is entirely my own work and that it has not been submitted for any other degree or professional qualification at any institution.

Signed:

Date: 24/01/2026

Abstract

Empowering Palestinian Voices Using "Text Mining" Techniques to Overcome Social Media Expression Restrictions

In an era where digital communication shapes political discourse and collective memory, social media platforms have become both spaces of empowerment and instruments of control. This thesis investigates the algorithmic suppression of Palestinian digital expression on major platforms—Facebook, Instagram, and X (formerly Twitter)—through the integration of Natural Language Processing (NLP) and text mining techniques. The study situates itself within the interdisciplinary domains of artificial intelligence, digital rights, and computational social science, aiming to uncover how algorithmic bias operates within automated moderation systems.

Data were collected using Apify-based scrapers and analysed in Kaggle through multiple preprocessing and modelling stages. Techniques such as TF-IDF vectorization, sentiment and emotion analysis, and Named Entity Recognition (NER) were employed to identify linguistic and affective patterns correlated with content suppression. Comparative machine-learning experiments—including Logistic Regression, Naïve Bayes, Linear SVC, SGD, and transformer-based models (BERT and XLM-R)—revealed consistent evidence of algorithmic bias. Facebook demonstrated structural filtering and downranking of politically sensitive posts, Instagram exhibited emotional suppression of solidarity content, and X retained partial transparency but reflected selective engagement constraints.

The results confirm that algorithmic repression is not incidental but systematically embedded within platform architectures and moderation logic. Beyond quantitative findings, the research advances an Integrated Research Framework that combines computational rigor with ethical reflection, positioning data science as a form of digital resistance.

This thesis contributes to the emerging field of algorithmic justice by presenting empirical evidence of digital repression and proposing a context-aware, ethically grounded approach to AI design. It concludes that reclaiming visibility in the algorithmic age is not merely a technical challenge but a moral and political act—one that defines the future of digital freedom and equity.

الخلاصة

تمكين أصوات الشعب الفلسطيني باستخدام تقنيات "تنقيب النصوص" لتجاوز قيود التعبير عن الرأي على وسائل التواصل الاجتماعي.

في عصر أصبحت فيه الاتصالات الرقمية أداة لتشكيل الخطاب السياسي وصياغة الذاكرة الجماعية، تحوّلت منصات التواصل الاجتماعي إلى فضاءات مزدوجة تجمع بين التمكين من جهة، والسيطرة من جهة أخرى. تتناول هذه الرسالة مسألة القمع الآلي للتعبير الفلسطيني في الفضاء الرقمي، وذلك من خلال تحليل محتوى ثلاث منصات رئيسية هي: فيسبوك، وإنستغرام، وإكس (تويتر سابقًا)، باستخدام منهجيات التحليل اللغوي الحاسوبي والتنقيب في النصوص.

تتموضع هذه الدراسة ضمن تقاطع مجالات الذكاء الاصطناعي وحقوق الإنسان الرقمية والعلوم الاجتماعية الحاسوبية، وتهدف إلى الكشف عن أنماط التحيّز الآلي وآليات عمل أنظمة الإشراف التلقائي على المحتوى، التي تؤدي إلى تقييد الخطاب الفلسطيني أو تقليص ظهوره.

اعتمدت الدراسة على جمع البيانات من المصادر الرقمية المفتوحة وتحليلها باستخدام أدوات علمية متقدمة، حيث جرى تنفيذ مراحل المعالجة المسبقة للنصوص، وتحليل المشاعر والانفعالات، واستخلاص الأنماط اللغوية والدلالات العاطفية المرتبطة بالحالات التي تتعرض فيها المنشورات للحجب أو التقييد.

أظهرت النتائج وجود أنماط متكررة من الانحياز البنيوي في الخوارزميات المستخدمة داخل هذه المنصات، حيث تبين أن فيسبوك يمارس تصفية منهجية للمحتوى السياسي، بينما أظهر إنستغرام نزعة لحجب المحتوى التضامني والعاطفي، أما منصة إكس فقد بدت أكثر انفتاحًا نسبيًا لكنها أظهرت تقييدًا انتقائيًا للتفاعل مع المحتوى الفلسطيني. وتؤكد هذه النتائج أن القمع الرقمي ليس سلوكًا عرضيًا، بل هو نتيجة لبنية خوارزمية مبرمجة تعيد إنتاج علاقات القوة وعدم المساواة في المجال الرقمي. كما تقترح الرسالة إطارًا بحثيًا متكاملًا يدمج بين الدقة العلمية والاعتبارات الأخلاقية، بهدف توظيف علم البيانات كوسيلة للمساءلة والمقاومة الرقمية.

وتخلص الدراسة إلى أن استعادة الظهور في العصر الرقمي ليست مجرد مسألة تقنية، بل هي فعل أخلاقي وإنساني يرتبط بالحق في التعبير والعدالة والكرامة الرقمية، وتؤكد أن مستقبل التكنولوجيا يجب أن يُبنى على قيم الشفافية والإنصاف واحترام التنوع الثقافي والإنساني.

Associated Publications

No associated publications have been produced from this thesis at the time of submission.

Acknowledgements

In the name of Allah, Most Gracious, Most Merciful

All praise is due to Allah, whose blessings and guidance have accompanied me throughout every step of this journey.

I would like to express my sincere gratitude to my supervisor, Dr. Mohamed Dweib, for his continuous guidance, valuable feedback, and unwavering support during all stages of this thesis. His insightful advice, patience, and dedication have had a profound impact on the quality and completion of this research.

My deep appreciation also goes to Al-Quds Open University, which provided the academic foundation, resources, and encouragement necessary to accomplish this study.

Finally, my heartfelt thanks go to my beloved family, whose endless love, patience, and prayers have been my greatest source of strength and motivation. Their unwavering belief in me made this achievement possible. submission., too

Maab, Srour

October 2025

To my parents, whose steadfast hearts, taught me that visibility is not granted, it is earned through faith, patience, and truth.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
NLP	Natural Language Processing
NER	Named Entity Recognition
ML	Machine Learning
Meta	Meta Platforms (Facebook, Instagram)
X	Formerly Twitter
DL	Deep Learning
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
GPT	Generative Pre-trained Transformer
TF-IDF	Term Frequency – Inverse Document Frequency
LIWC	Linguistic Inquiry and Word Count
SNA	Social Network Analysis
API	Application Programming Interface
ICWSM	International Conference on Web and Social Media
BLM	Black Lives Matter
NGO	Non-Governmental Organization
ToS	Terms of Service
IEEE	Institute of Electrical and Electronics Engineers
API	Application Programming Interface
Aptify	Apify Scraper Platform
BoW	Bag of Words
CNN	Convolutional Neural Network
CSV	Comma-Separated Values
DL	Deep Learning
EDA	Exploratory Data Analysis

F1-score	F-measure
K-Fold CV	K-Fold Cross-Validation
LIME	Local Interpretable Model-agnostic Explanations
ML	Machine Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
ROC-AUC	Receiver Operating Characteristic – Area Under Curve
SVC	Support Vector Classifier
SGD	Stochastic Gradient Descent
SHAP	Shapley Additive explanations
TF-IDF	Term Frequency–Inverse Document Frequency
VSD	Value Sensitive Design
XAI	Explainable Artificial Intelligence
XLM-R	Cross-Lingual Language Model – RoBERTa
BERT	Bidirectional Encoder Representations from Transformers
LR	Logistic Regression
NB	Naïve Bayes
SNA	Social Network Analysis
CI	Confidence Interval
ROC-AUC	Receiver Operating Characteristic – Area Under Curve
TF-IDF Weight	Term Frequency – Inverse Document Frequency Weight
F1-score	Harmonic Mean of Precision and Recall
BiLSTM	Bidirectional Long Short-Term Memory
SGD	Stochastic Gradient Descent
SVC	Support Vector Classifier
AUC	Area Under the Curve

BERT	Bidirectional Encoder Representations from Transformers
XLM-R	Cross-Lingual Language Model – RoBERTa Variant
Apify	Automated Platform for Web Data Extraction
Cross-Validation	Statistical Model Validation Technique
Affective Filtering	Emotion-Based Algorithmic Moderation
XAI	Explainable Artificial Intelligence
GDPR	General Data Protection Regulation
DSA	Digital Services Act
UNGPs	United Nations Guiding Principles on Business and Human Rights
NGO	Non-Governmental Organization
CSO	Civil Society Organization
ICT4D	Information and Communication Technologies for Development
SNA	Social Network Analysis

Chapter 1: Introduction

1.1 Overview and Background

In recent years, social media platforms have evolved into complex socio-technical systems that mediate public discourse, political participation, and information visibility on a global scale. Far from functioning as neutral communication channels, these platforms rely on automated decision-making systems, particularly algorithmic content moderation, to filter, rank, and sometimes suppress user-generated content. Such systems increasingly shape which narratives gain visibility and which are marginalized, raising critical questions about fairness, transparency, and algorithmic bias.

Algorithmic content moderation operates primarily through machine learning–based classification models trained to identify potentially "harmful," "sensitive," or policy-violating content. While these systems are designed to operate at scale, their reliance on linguistic cues, engagement signals, and historical training data makes them vulnerable to systematic bias—especially in politically sensitive, multilingual, and culturally specific contexts. Prior research has demonstrated that algorithmic moderation systems may disproportionately affect content related to conflict, activism, or minority communities due to skewed datasets, ambiguous policy definitions, and limited contextual awareness (Noble, 2018; Suzor, 2020).

Within this broader landscape, Palestinian-related content on social media represents a particularly salient case for studying algorithmic bias. Human rights organizations and digital rights researchers have documented repeated instances of content removal, downranking, and visibility restriction affecting posts related to Palestine across major platforms, especially within Meta's ecosystem. Importantly, many of these interventions occur through automated or semi-automated systems, where decisions are not easily observable, contestable, or explainable to users.

From a computational perspective, this situation presents a structured research problem: how algorithmic moderation systems classify politically sensitive content, what linguistic and affective features are associated with suppression-related outcomes, and whether such patterns can be systematically detected using Natural Language Processing (NLP) and machine learning techniques. Rather than treating moderation as an opaque platform decision, this research approaches it as a data-driven phenomenon that can be analyzed, modelled, and interpreted through supervised classification and explanatory analytics.

Accordingly, this study positions algorithmic bias—not political context alone—as its central analytical focus. The Palestinian digital discourse is examined as a case study through which broader questions of automated moderation, classification uncertainty, and algorithmic governance can be investigated. By grounding the analysis in computational methods, the research seeks to contribute to the growing field of algorithmic accountability, demonstrating how data science can be used to interrogate and evaluate the behavior of large-scale content moderation systems.

1.2 Motivation

The motivation for this research is rooted in a growing methodological and scientific challenge within the field of artificial intelligence and data science: understanding how automated content moderation systems encode, reproduce, and amplify algorithmic bias in politically sensitive contexts. As social media platforms increasingly rely on machine learning–based classifiers to govern visibility and enforce content policies, the need for systematic, data-driven evaluation of these systems has become both urgent and underexplored (Gillespie, 2020; Suzor, 2020).

From a technical perspective, algorithmic moderation presents a complex classification problem. Decisions regarding content visibility are often influenced by linguistic structure, emotional tone, keyword usage, and engagement patterns—features that are well suited for analysis using Natural Language Processing (NLP) and supervised machine learning. However, existing moderation systems are typically opaque, proprietary, and inaccessible to independent audit, limiting external understanding of how classification decisions are made and which types of content are disproportionately affected (Pasquale, 2015; Noble, 2018).

The Palestinian digital discourse offers a uniquely rich and challenging case for investigating these issues. It is multilingual (Arabic and English), emotionally charged, politically contextualized, and frequently associated with terms and narratives that automated systems may flag as sensitive. These characteristics make it particularly suitable for testing whether NLP-based models can identify systematic patterns linked to moderation-related outcomes, and whether such patterns reflect broader algorithmic biases rather than isolated enforcement decisions. Previous reports by digital rights organizations have highlighted recurrent moderation irregularities affecting Palestinian-related content, particularly within Meta platforms (7amleh, 2023; Amnesty International, 2023).

This research is therefore motivated by a methodological gap in current scholarship: the lack of computational frameworks that combine predictive modelling with explanatory analysis to study algorithmic moderation in marginalized and underrepresented contexts. While prior studies and human rights reports have documented instances of content suppression, fewer works have operationalized these observations as machine learning problems that can be quantitatively evaluated, compared, and interpreted across platforms (Sandvig et al., 2014; Noble, 2018).

In addition, recent advances in explainable artificial intelligence (XAI) provide an opportunity to move beyond performance metrics alone and to interrogate why certain content is more likely to be flagged or suppressed. Techniques such as feature attribution and model interpretability enable researchers to connect classification outcomes with specific linguistic and affective signals embedded in text (Ribeiro et al., 2016; Lundberg & Lee, 2017). By integrating sentiment analysis, emotion-aware features, and named entity recognition into supervised classification pipelines, this study seeks to contribute methodologically to the development of transparent and accountable approaches for analyzing algorithmic decision-making.

Finally, while this research is grounded in technical inquiry, it is informed by a broader ethical concern regarding fairness and accountability in automated systems. As a researcher working at the intersection of data science and digital rights, the motivation extends to exploring how computational methods can be responsibly used to audit algorithmic governance and to support more equitable digital environments. This ethical dimension, however, complements rather than replaces—the study's primary focus on algorithmic bias and machine learning methodology.

1.3 Problem Statement

The problem addressed in this research is not limited to isolated instances of content removal. Rather, it concerns the broader computational logic through which automated moderation systems distinguish between acceptable and sensitive political expression. In politically charged and multilingual contexts—such as Palestinian digital discourse—classification decisions are particularly vulnerable to misinterpretation due to the interaction of emotional language, culturally specific terminology, and conflict-related narratives. Reports by human rights and digital rights organizations have documented recurring moderation irregularities affecting Palestinian-related content, suggesting the presence of structural rather than incidental patterns (7amleh, 2023; Amnesty International, 2023).

From a data science perspective, this situation can be formulated as a supervised classification problem under conditions of uncertainty and potential bias. Social media posts are implicitly assigned to different moderation-related outcomes (e.g., visible versus suppressed), yet the true decision rules employed by platforms are not accessible. As a result, researchers must rely on observable indicators—such as engagement anomalies, linguistic markers, and contextual signals—to approximate moderation behaviour and model its patterns computationally.

A key challenge, therefore, lies in distinguishing whether observed reductions in visibility and engagement are attributable to algorithmic bias or to alternative factors such as content volume, posting frequency, or audience dynamics. Without a structured analytical framework, claims regarding suppression risk may remain speculative. This research addresses this challenge by applying Natural Language Processing (NLP) and supervised machine learning techniques to systematically examine which textual and affective features are associated with suppression-related outcomes across platforms, and whether these associations exhibit consistent, platform-specific patterns.

Another critical limitation in existing scholarship is the lack of explanatory analysis accompanying predictive performance. While prior studies and reports have identified moderation disparities, fewer have combined predictive modelling with interpretability mechanisms capable of revealing why certain content is more likely to be classified as sensitive. This absence restricts both scientific understanding and accountability.

Accordingly, this study emphasizes not only prediction accuracy but also explanatory insight, enabling a more nuanced evaluation of algorithmic behaviour (Noble, 2018).

In summary, the core problem addressed by this research is the absence of transparent, data-driven methodologies for detecting and interpreting algorithmic bias in automated content moderation systems. By framing moderation as a supervised classification problem and applying explainable NLP-based models to Palestinian-related digital discourse, this study seeks to bridge the gap between qualitative documentation of censorship and quantitative, reproducible analysis.

1.4 Key Definitions

To ensure conceptual clarity and methodological consistency, this section defines the key technical terms used throughout the thesis. All terms are introduced explicitly before the use of abbreviations or advanced analytical discussion.

- **Algorithmic Bias**

  Algorithmic bias refers to systematic and repeatable errors in automated decision-making systems that result in unfair or disproportionate outcomes for specific groups, topics, or forms of expression. In the context of social media moderation, such bias may arise from skewed training data, optimization objectives, or contextual misinterpretation of politically sensitive language.

- **Predictive Models**

  Predictive models are machine learning models designed to estimate the likelihood that a given social media post will be associated with a specific outcome—in this study, moderation-related outcomes such as reduced visibility or suppression. These models learn patterns from labelled data and generate probability-based predictions for unseen instances.

- **Explanatory Models**

  Explanatory models refer to analytical approaches that aim to interpret and clarify how and why a predictive model arrives at its decisions. Rather than focusing solely on performance metrics, explanatory modelling emphasizes feature importance, linguistic cues, and contextual signals that contribute to classification outcomes.

- **Natural Language Processing (NLP)**

  Natural Language Processing is a subfield of artificial intelligence concerned with enabling computational systems to analyse, interpret, and generate human language. In this research, NLP techniques are employed to extract linguistic, semantic, and emotional features from social media text data.

- **Sentiment Analysis**

  Sentiment analysis is an NLP task that identifies and quantifies the emotional polarity expressed in textual data, typically categorized as positive, negative, or neutral. In this study, sentiment analysis is used to examine whether emotional tone correlates with moderation-related outcomes.

- **Named Entity Recognition (NER)**

  Named Entity Recognition is an NLP technique that detects and classifies references to real-world entities—such as locations, organizations, and persons—within text. This task is particularly relevant for identifying politically and geographically sensitive references associated with content moderation.

- **Topic Modelling**

  Topic modelling is an unsupervised text-mining technique used to discover latent thematic structures within large collections of documents. It enables the identification of dominant themes and recurring topics without predefined labels.

- **Supervised Learning**

  Supervised learning is a machine learning paradigm in which models are trained on labelled data, where each instance is associated with a predefined class. In this thesis, supervised learning is used to distinguish between posts associated with different moderation-related outcomes.

- **Engagement Metrics**

  Engagement metrics refer to observable user-interaction signals such as likes, comments, shares, and retweets. These metrics are used as indirect indicators of content visibility and audience reach rather than as direct measures of platform moderation decisions.

1.5 Research Objectives

This research aims to investigate algorithmic bias in automated content moderation systems through a computational and data-driven approach grounded in Natural Language Processing (NLP) and supervised machine learning. The objectives are formulated in response to documented concerns regarding opaque algorithmic governance and biased moderation practices in politically sensitive contexts (Noble, 2018; Suzor, 2020; Gillespie, 2020).

1.5.1 General Objective

The primary objective of this research is to design, implement, and evaluate supervised machine learning models capable of detecting and interpreting patterns associated with moderation-related outcomes in Palestinian-related social media content, with particular emphasis on identifying linguistic, emotional, and contextual features linked to algorithmic bias, as highlighted in prior critical studies of algorithmic decision-making (Noble, 2018; Suzor, 2020).

1.5.2 Specific Objectives

To achieve this general objective, the study pursues the following specific goals:

1. **Dataset Construction and Preparation**

   To construct a multilingual (Arabic–English) dataset of social media posts related to Palestine, enriched with engagement metrics and moderation-related indicators, following documented observations of systematic moderation irregularities affecting such content (7amleh, 2023; Amnesty International, 2023).

2. **Feature Extraction Using NLP Techniques**
   To apply Natural Language Processing techniques—including sentiment analysis, named entity recognition, and topic modelling—to extract linguistic, semantic, and affective features relevant to automated moderation decisions, in line with existing computational approaches to politically sensitive discourse analysis (Noble, 2018).

3. **Supervised Classification Modelling**
   To formulate the moderation analysis as a supervised classification problem and train multiple machine learning models to distinguish between posts associated with different moderation-related outcomes, addressing the need for structured, data-driven investigation of algorithmic behavior (Suzor, 2020).

4. **Evaluation of Predictive Performance**
   To assess model performance using standard classification metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, and to interpret these metrics within the context of algorithmic moderation rather than platform intent, consistent with methodological practices in algorithmic auditing (Gillespie, 2020).

5. **Explanatory Analysis of Algorithmic Behavior**
   To complement predictive accuracy with explanatory analysis, examining which linguistic and affective features contribute most strongly to classification decisions, thereby responding to calls for greater transparency and accountability in automated systems (Noble, 2018; Suzor, 2020).

6. **Cross-Platform Comparative Analysis**
   To compare modelling outcomes across datasets collected from different social media platforms to examine whether observed patterns are consistent across data environments, as suggested by prior comparative studies of content moderation (7amleh, 2023).

7. **Methodological Contribution to Algorithmic Accountability**
   To contribute a reproducible computational framework that supports critical evaluation of automated moderation systems, bridging the gap between qualitative human rights documentation and quantitative, model-based analysis (Amnesty International, 2023; Gillespie, 2020).

1.5.3 Scope and Focus

The objectives of this research are deliberately confined to the algorithmic dimension of content moderation. While the Palestinian digital context provides the empirical case study, the primary focus remains on how machine learning systems classify politically sensitive content, how bias may emerge from these processes, and how such behaviour can be systematically detected and interpreted using computational methods (Noble, 2018; Suzor, 2020).

1.6 Thesis Contribution to the Field/ Significance and /or Impact of the Research

This thesis contributes to the field of artificial intelligence and data science by advancing a computationally grounded and explainable framework for studying algorithmic bias in automated content moderation systems. While previous studies and human rights reports have documented instances of content suppression affecting Palestinian-related discourse, this research distinguishes itself by operationalizing these observations as a supervised machine learning problem that can be empirically tested, evaluated, and interpreted.

The primary contribution of this study is methodological rather than descriptive. Instead of treating censorship or suppression as a purely qualitative or platform-specific phenomenon, the research reformulates moderation as a classification task driven by linguistic, emotional, and contextual features. By doing so, it enables systematic investigation of algorithmic behaviour using measurable performance indicators such as accuracy, recall, and F1-score, while explicitly acknowledging the limitations of proxy-based labels and observable engagement signals (Noble, 2018; Suzor, 2020).

A second key contribution lies in the integration of predictive and explanatory modelling. Whereas many prior computational studies focus solely on classification performance, this thesis combines supervised learning with interpretability-oriented analysis to examine why certain content is more likely to be associated with suppression-related outcomes. This dual approach responds directly to calls for greater transparency and accountability in algorithmic governance, particularly in politically sensitive contexts where automated systems operate with limited oversight (Noble, 2018; Gillespie, 2020).

The study further contributes through its multilingual and context-aware design. Palestinian digital discourse presents linguistic complexity due to the coexistence of Arabic, English, and mixed-language expressions, as well as high emotional density linked to conflict-related narratives. By incorporating sentiment analysis, named entity recognition, and topic modelling within a unified pipeline, the research demonstrates how NLP techniques can be adapted to capture culturally and politically specific signals that generic models often overlook. This addresses a recognized gap in existing computational moderation research, which frequently underrepresents non-Western and multilingual contexts (7amleh, 2023; Amnesty International, 2023).

Importantly, the contribution of this thesis does not rest on claiming entirely novel outcomes, but on providing empirical confirmation, computational validation, and methodological clarity to patterns previously identified through qualitative and audit-based approaches. The alignment between the study's results and earlier reports strengthens, rather than weakens, its contribution by demonstrating that observed moderation disparities are not anecdotal but reproducible through independent data-driven analysis.

From a broader perspective, this research contributes to ongoing discussions on algorithmic accountability by offering a replicable analytical framework that bridges the gap between human rights documentation and machine learning evaluation. While the Palestinian context serves as the empirical case study, the proposed methodology is transferable to other politically sensitive or marginalized domains, supporting wider efforts to critically examine automated moderation systems across platforms (Suzor, 2020; Gillespie, 2020).

In sum, the significance of this thesis lies in reframing algorithmic moderation as a measurable, explainable, and auditable computational process. By grounding the analysis in supervised learning and interpretability, the study advances the role of data science as a tool for understanding—and critically evaluating—the operation of algorithmic governance systems in contemporary digital environments.

1.7 Thesis Outline

This thesis is structured into eight interdependent stages, each contributing systematically to the investigation of expression restrictions on digital platforms and the deployment of technological strategies to empower Palestinian voices. The structure ensures methodological coherence and analytical progression, moving from conceptual grounding to technical implementation and critical reflection.

1. **Stage 1: Establishing Objectives and Methodological Foundation**
   This chapter sets the conceptual and operational foundations of the study. It articulates the research objectives, core questions, and hypotheses, and provides an overview of the methodological approach. This stage frames the trajectory of the thesis and defines its scope and strategic intent.

2. **Stage 2: Literature Review**
   This chapter presents a comprehensive and critical review of existing literature on freedom of digital expression, algorithmic bias, and the specific context of Palestinian content moderation. It identifies gaps in existing studies and establishes the theoretical and analytical framework that guides the research.

3. **Stage 3: Problem Definition and Analysis**
   Focusing on the case of Palestinians on Meta, this chapter analyzes the nature and dimensions of expression restrictions. It deconstructs the mechanisms of digital repression and outlines the socio-technical challenges that limit Palestinian participation in online discourse.

4. **Stage 4: Methodology and Technological Tools**
   This chapter elaborates on the technical and procedural methods employed in the study. It details the use of text mining techniques, machine learning algorithms, and data collection mechanisms. It also addresses the rationale behind the chosen tools and ethical considerations in working with social media data.

5. **Stage 5: Research Execution and Data Collection**
   This chapter outlines the practical implementation of the study's design. It describes the processes of data sourcing, web scraping using tools such as Scrapy, and criteria for selecting content relevant to Palestinian expression. It ensures reproducibility and transparency in the data acquisition phase.

6. **Stage 6: Data Analysis and Experimental Results**
   In this chapter, the processed data are subjected to a series of computational analyses. The performance of the predictive models, the patterns of content suppression, and the linguistic and emotional characteristics of flagged posts are examined. Results are evaluated against the hypotheses and objectives.

7. **Stage 7: Interpretation and Discussion**
   This chapter synthesizes the findings by considering both the technical outputs and broader socio-political implications. It discusses the insights gained into algorithmic behavior, the lived experiences of censorship, and the feasibility of resistance mechanisms through computational tools.

8. **Stage 8: Review, Evaluation, and Final Recommendations**
   The concluding chapter involves a critical review of the study, reflecting on its strengths and limitations. It refines the findings and proposes forward-looking recommendations for platform accountability, digital policy reform, and AI-driven strategies to defend freedom of expression in politically constrained regions.

Together, these stages construct a rigorous and context-sensitive research pathway, aiming to transform computational methods into instruments of justice and digital empowerment for Palestinians and similarly marginalized communities.

Chapter 2: Literature Review

2.1 Introduction

In the digital age, social media platforms have transformed from simple communication tools into infrastructures that shape public discourse, political participation, and collective consciousness. Their participatory design allows individuals—particularly those marginalized in traditional media—to share narratives, contest dominant discourses, and mobilize solidarity. Yet, these same systems are governed by opaque policies and algorithmic mechanisms that enable surveillance, bias, and suppression.

This chapter critically examines the dual role of social media as both an enabler of expression and an instrument of control. It reviews theoretical and empirical research on digital expression, natural language processing (NLP), machine learning (ML), text mining, sentiment and emotion analysis, ethical concerns, and the algorithmic architectures that influence online visibility. The Palestinian case—marked by geopolitical marginalization and algorithmic censorship—serves as a central point of reference. By integrating studies from computer science, communication, and digital rights, this review provides the foundation for developing context-aware computational frameworks that resist algorithmic bias and promote freedom of expression.

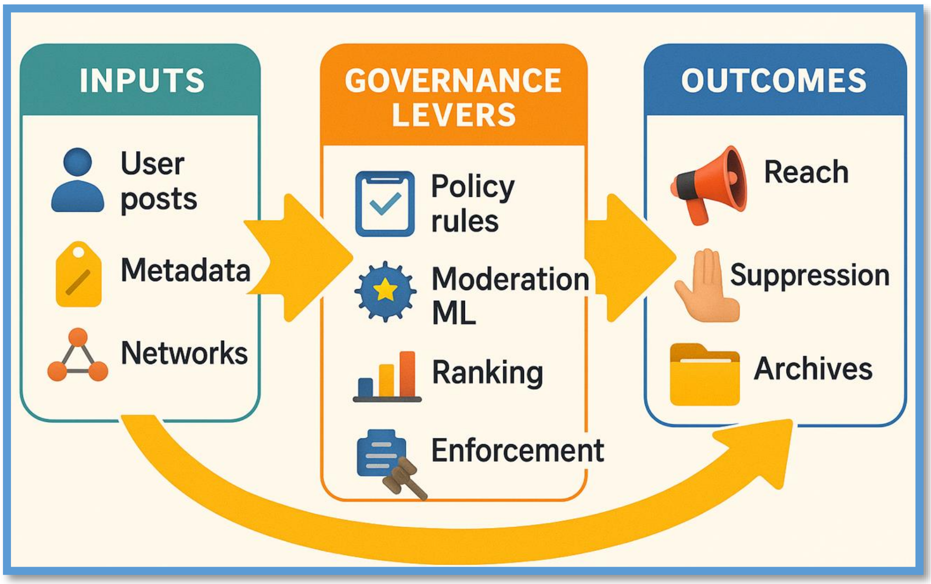


Figure 2.1 Platforms as Infrastructures of Visibility

2.2 Digital Expression and the Politics of Visibility

Social media platforms have become critical spaces for political expression and civic engagement, particularly in contexts where mainstream media are restricted. While platforms such as X and Meta (Facebook and Instagram) were initially celebrated as democratizing tools, research increasingly reveals their dependence on opaque moderation and politically influenced governance.

A decade-long review by Al-Rashid and Mahfouz found recurring removals and shadow banning of political content in restricted regions on X (Mahfouz, 2020). Pereira et al. further observed inconsistent enforcement of moderation rules, often shaped by governmental pressure rather than platform neutrality (Pereira & Monteiro, 2021). Hounsel et al. demonstrated that TikTok's ML-driven moderation suppresses sensitive political content without user notification—what they describe as "invisible censorship" (Hounsel et al., 2021).

From a theoretical lens, Gillespie conceptualized "editorial algorithmic governance," arguing that algorithms actively shape discourse rather than merely organize it (Gillespie, 2020). In Meta's ecosystem, Pereira et al. provided empirical evidence that posts related to Palestinian activism are systematically deprioritized, especially those written in Arabic or containing #Gaza and #Palestine (Pereira & Monteiro, 2021). Reports by Human Rights Watch and 7amleh corroborate these findings, documenting extensive deletion of Palestinian content and restrictions on appeals within Facebook (Human Rights Watch, 2021–2022; 7amleh, 2022).

Collectively, these studies show that algorithmic moderation is neither random nor neutral—it reflects political, cultural, and linguistic biases that determine who is heard online.

Table 2.1 : *Evidence of Visibility and Moderation Across Platforms*

Platform(s)	Study & Year	Focus	Method & Artifact	Key Findings	Replication Notes
X	(Mahfouz, 2020)	Political expression trends	Longitudinal content analysis	Recurrent removal and shadow banning in restricted regions	Compare engagement vs. visibility across events
X	(L. Pereira R. M., 2021)	Governance & policy enforcement	Policy and moderation analysis	Inconsistent rule enforcement: political influence detected	Cross-analyses moderation logs pre/post crisis
TikTok	(J. Hounsel, 2021)	Algorithmic moderation	Experimental measurement study	Silent suppression of sensitive content (no user notice)	Use probe accounts to detect shadow banning
Cross-platform	(Gillespie T. , Editorial algorithms and the shaping of public discourse, 2020)	Editorial algorithm governance	Theoretical & qualitative analysis	Algorithms shape discourse, not only curate it	Map theory to Meta's policy evolution
Instagram (Meta)	(L. Pereira R. M., 2021)	Palestinian activism visibility	Empirical dataset study	Algorithmic reprioritization of Arabic/Palestinian posts	Track engagement decay on hashtags #Gaza #Palestine
Facebook (Meta)	(Center H. R., 2022–2021)	Political bias & moderation	Digital-rights audits & NGO reports	Systematic deletion of Palestinian content; no appeal transparency	Validate via keyword analysis ("فلسطين'' , "غزة" , "مقاومة"")

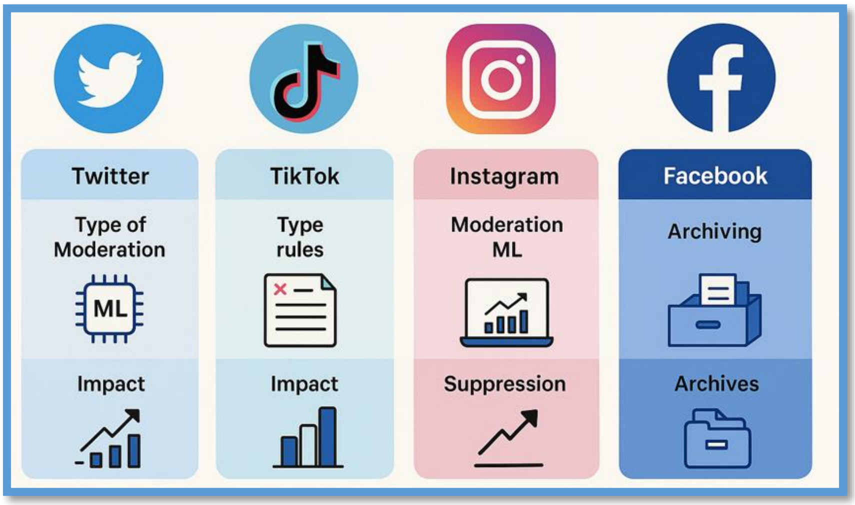


Figure 2.2 :*Comparative Moderation Evidence Across Platforms*

Discussion

Across platforms, moderation patterns differ in method but converge in outcome: political silencing. X fluctuates with external pressure; TikTok operates opaque ML filters; Instagram algorithmically downranks Palestinian content; Facebook institutionalizes removal through hybrid human–AI control. Visibility online, therefore, is not organic but politically constructed.

2.3 Machine Learning and NLP in Social Media Analysis

Machine learning (ML) and natural language processing (NLP) enable large-scale interpretation of digital discourse. Early works on personality inference revealed methodological pitfalls—overfitting, cultural bias, and limitations arising from informal language. Studies emphasized the importance of preprocessing pipelines and contextual lexicons for accurate pattern extraction (Kachuee et al., 2021; Salloum et al., 2019). Hammar (2021) advanced weak-supervision frameworks for informal, multilingual, and multimodal data such as Instagram text.

Later research extended NLP into politically sensitive areas, including hate speech detection, misinformation classification, and ideological analysis. Mathew et al. (2021) introduced HateXplain, an explainable AI dataset that clarifies model reasoning. Mozes and Klein (Klein & Mozes, 2021) applied BERT-based ensemble models to political discourse, revealing systematic ideological bias in model outputs.

In contexts like Palestine, generic NLP models risk misclassifying legitimate forms of resistance as hate speech, thereby reinforcing structural marginalization. Developing culturally aligned, context-aware NLP systems is therefore essential for ensuring fair, accurate, and inclusive analysis.

2.4 Emotion Analysis and Affective Impact

Emotion analysis captures collective affect during crises and repression. Deep learning architectures—including Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), Bidirectional Long Short-Term Memory networks (BiLSTM), and transformer-based models—have demonstrated strong performance in detecting fine-grained emotions such as anger, fear, hope, and solidarity (Olusegun, 2021; Qureshi et al., 2022). Hussain and Cambria (2021) developed multilingual emotion-recognition frameworks, while Ahmad et al. (2022) analysed emotional responses to COVID-19 across multiple languages using transformer-based architectures. Qureshi et al. (2022) further applied BiLSTM-based transfer learning to classify emotionally charged social media posts with improved contextual sensitivity.

In the Palestinian context, emotion-aware models can reveal grief, resistance, and hope that are often hidden or downranked by moderation filters. Thus, emotional classification becomes not merely descriptive but restorative—recovering suppressed affective expression and amplifying the collective voice of marginalized communities.

2.5 Text Mining in Social Media

Text mining extracts structured meaning from unstructured user content. Salloum et al. (2019) emphasized the importance of context-specific preprocessing, while Hammar (2021) demonstrated the value of deep text-mining techniques for informal, multilingual Instagram data. Tools such as LIWC2015 (Pennebaker et al., 2015) help uncover cognitive, emotional, and social layers embedded in linguistic patterns. Paul and Dredze (2011) pioneered the concept of “social monitoring,” initially applied to public health trends and later adapted to political sentiment and digital repression. Koto et al. (2021)

contributed multilingual disinformation corpora essential for analysing politicized and cross-lingual environments.

For Arabic and Palestinian discourse, preprocessing must address orthographic normalization, dialectal variation, hashtag segmentation, and multilingual named entity resolution. Incorporating contextual lexicons and sociopolitical framing significantly enhances both analytical precision and interpretability.

2.6 Sentiment Analysis and User Reactions

Sentiment analysis quantifies opinion polarity and emotional intensity across events. Foundational work by Liu (2012) formalized the principles of opinion mining, while Chandra et al. (2022) demonstrated temporal sentiment variation across political discourse. Mansour (2022) further showed how regional sentiment diverges following terrorist incidents, revealing sharp contrasts in emotional response. Applied to Palestinian data, sentiment mining uncovers affective narratives of anger, loss, solidarity, and defiance—narratives that are often obscured or downranked by algorithmic moderation systems.

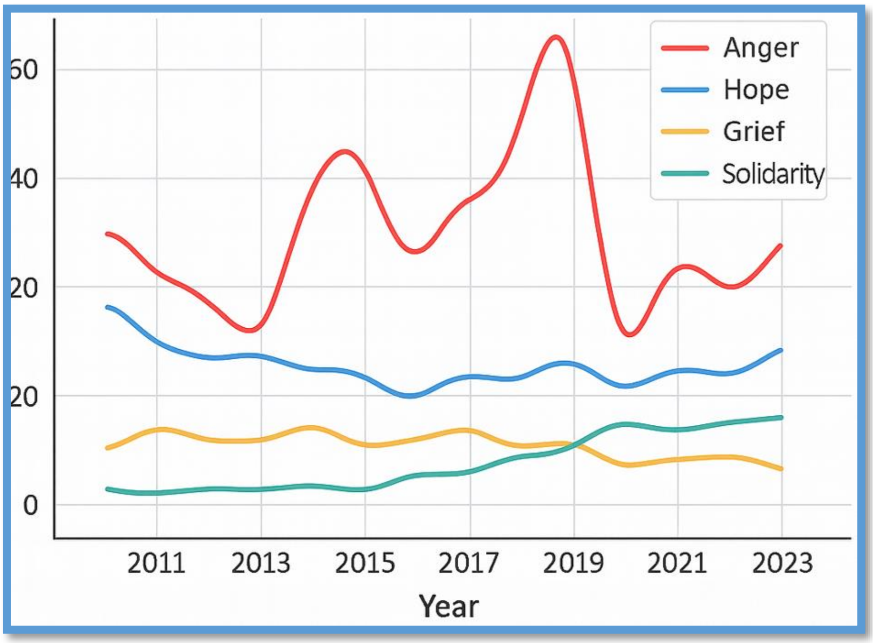


__Figure 2.3 :__Temporal sentiment trends in Palestinian digital discourse, illustrating shifts between anger, hope, grief, and solidarity.

Tracking temporal sentiment patterns aligned with conflict timelines provides insight into emotional resilience and mobilization, transforming analytics into digital documentation.

2.7 Public Opinion and Social Media

Social media reshapes opinion formation through algorithmic curation. Recommendation systems personalize feeds, fostering what scholars describe as “echo chambers” (Bruns, 2019; Pariser, 2011). Networked propaganda further exploits these dynamics (Lewis, 2017), while marginalized groups are disproportionately subjected to algorithmic downranking despite high engagement (Freelon et al., 2016). Sentiment studies on the Syrian refugee crisis demonstrate how online emotional expression often mirrors offline political attitudes (Ayvaz, 2019).

In the Palestinian context, Meta’s downranking of pro-Palestinian content fragments collective awareness and undermines global solidarity. Emerging scholarship calls for greater algorithmic literacy—citizens’ ability to understand, interrogate, and critique algorithmic influence (Johnson, 2021).

2.8 Challenges and Ethical Considerations

The use of social-media data raises complex ethical issues—privacy, representation, algorithmic bias, and psychological harm.

- **Privacy and Consent.** Zimmer (2010) warns that public accessibility does not equate to ethical consent; researchers must anonymize data and clearly communicate research aims (Proferes, 2018).

- **Data Bias.** Ruths and Pfeffer caution against demographic, political, and structural bias in social media datasets (Ruths & Pfeffer, 2014), while Tufekci (2014) highlights methodological pitfalls related to representativeness and validity. In Palestinian datasets, systematic deletions further distort representativeness.

- **Algorithmic Bias.** Noble (2018) and Sandvig et al. (2014) highlight how algorithmic systems embed and amplify prejudice; audit-based research is therefore essential.

- **Automated Collection.** Mislove et al. (2010) and Bruckman (2016) emphasize the need for ethical scraping practices that respect platform terms of service while minimizing potential harm.

- **Psychological Impact.** Moreno et al. (2013) and Ellison and Vitak (2021) call for protecting participants from potential emotional or psychological distress when dealing with sensitive or traumatic online content.

Table 2.2: *Ethical Risks and Mitigation Strategies*

Risk	Why It Matters	Mitigation	Refs
Contextual privacy	Users expect bounded visibility	Anonymize; aggregate; disclose purpose	(Zimmer, 2010), (Proferes, 2018)
Sampling bias	Suppressed data skews results	Multi-source collection and weighting	(Pfeffer, 2014) , (Tufekci, Big questions for social media big data: Representativeness, validity, and other methodological pitfalls, 2014)
Algorithmic opacity	Hidden decision rules	Black-box audits; counterfactual testing	(Noble S. U., 2018), (C. Sandvig, 2014)
Automated collection	Ethical and legal risk	Rate-limit; respect ToS and community	(A. Mislove, 2010), (Bruckman, 2016)
Psychological harm	Exposure to distress content	Debrief; support; warning labels	(M. Moreno, 2013), (N. Ellison, 2021)

2.9 Case Studies of Social Media Empowerment

2.9.1 The Arab Spring

Platforms like Facebook and X enabled political mobilization and helped bypass state censorship in Tunisia, Egypt, and Libya (Hussain & Howard, 2011; Lim, 2012). These platforms accelerated protest coordination and heightened international awareness during the uprisings.

2.9.2 #BlackLivesMatter

The #BLM movement used hashtags to globalize anti-racist struggles, documenting injustice and influencing policy (Freelon et al., 2016; Foucault-Welles & Jackson, 2020).

2.9.3 The Palestinian Context

Campaigns such as #SaveSheikhJarrah demonstrate the global resonance of Palestinian digital activism even amid algorithmic suppression (Amnesty International, 2023; Access Now, 2022). Despite systematic censorship, social media has provided Palestinians with a global stage to document their lived realities and mobilize transnational solidarity.

2.9.4 Palestinian Digital Activists

2.9.4.1 Bisan Owda

They broadcast daily life in Gaza through Instagram and TikTok, gaining international recognition—including a 2024 Emmy—yet faced impersonation attempts and coordinated smear campaigns (7amleh, 2022; Owda, 2024).

Figure 2.4: *Bisan Owda documenting in Gaza or receiving her Emmy Award*

2.9.4.2 Farah Baker

gained global attention tweeting during the 2014 war yet suffered visibility restrictions (Baker, 2015).

Figure 2.5 :*Farah Baker tweeting during war or screenshot of viral tweet*

2.9.4.3 Saleh Zighari

documented Jerusalem protests visually, facing content takedowns and physical assaults (Zighari, 2022).

Figure 2.6 :*Saleh Zighari documenting protests or during assaults*

2.9.4.4 Saleh Jaafarawi

live-streamed during 2023–2024, with millions of views before Meta suspended his accounts (Jaafarawi, 2024).

Figure 2.7 : Saleh Jaafarawi filming in Gaza or during TikTok livestreams

These cases underscore both empowerment and risk: social media amplifies marginalized voices yet exposes them to algorithmic and political retaliation.

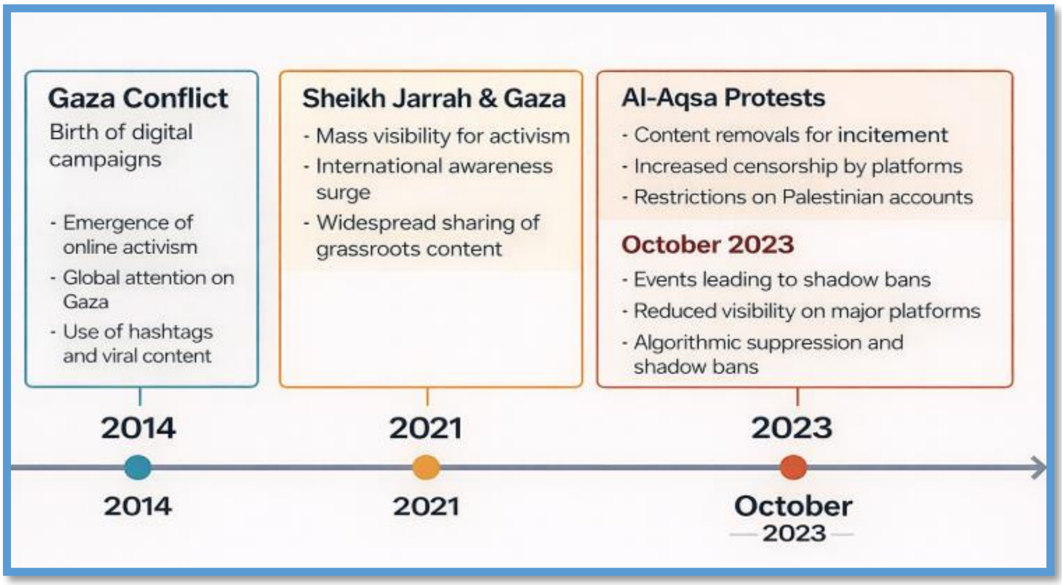

Gaza Conflict
Birth of digital campaigns
- Emergence of online activism
- Global attention on Gaza
- Use of hashtags and viral content
Sheikh Jarrah & Gaza
- Mass visibility for activism
- International awareness surge
- Widespread sharing of grassroots content
Al-Aqsa Protests
- Content removals for incitement
- Increased censorship by platforms
- Restrictions on Palestinian accounts
October 2023
- Events leading to shadow bans
- Reduced visibility on major platforms
- Algorithmic suppression and shadow bans
2014
2021
2023
2014
2021
October 2023

Figure 2.8 :*Timeline of Palestinian Digital Activism (2014–2025)*

2.10 The Role of Algorithms in Public Discourse

2.10.1 Personalization and Fragmentation

Recommendation algorithms create filter bubbles that reinforce homophily and increase polarization (Sunstein, 2017; Bail et al., 2018).

2.10.2 Embedded Bias

Bias arises from skewed datasets and feedback loops; Arabic political speech is disproportionately flagged (Noble, 2018).

2.10.3 Misinformation Amplification

Engagement-driven ranking amplifies emotionally charged falsehoods (Vosoughi et al., 2018), and further exacerbates crisis-related misperceptions (Starbird, 2020).

2.10.4 Opacity and Accountability

Limited API access and proprietary secrecy hinder independent audits (Pasquale, 2015). Algorithmic accountability requires regulatory oversight, open protocols, and explainability (Johnson, 2021).

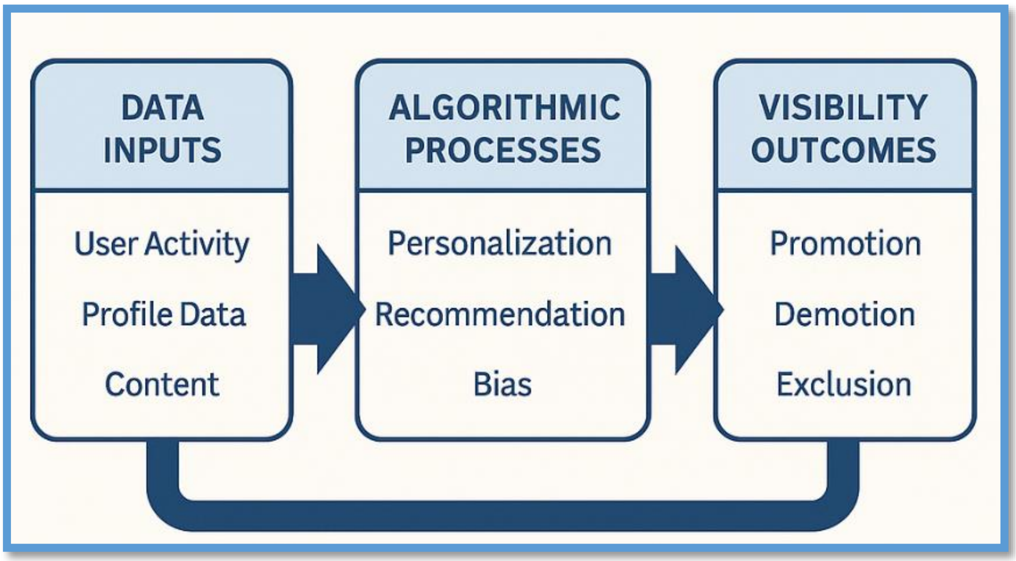


***Figure 2.9 :** Algorithmic Governance Model*

2.11 Future Directions in Social Media Text Mining

2.11.1 Advanced NLP Models

Transformer models—including BERT, RoBERTa, and GPT—significantly enhance contextual understanding and bias detection (Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020). Fine-tuning further enables adaptation to local linguistic and cultural contexts.

2.11.2 Real-Time Analytics

Streaming frameworks such as Kafka and Flink enable real-time detection of misinformation, rapid mobilization, and timely crisis response (Kreps, 2021; Qian et al., 2022).

2.11.3 Multimodal Analysis

Integrating text, images, and audio enhances the analysis of memes, videos, and digital propaganda, enabling deeper multimodal understanding (Liu et al., 2022; Cheng et al., 2022; Chen, 2023).

2.11.4 Ethical Design and Privacy

Federated learning and differential privacy safeguard user data (Abadi et al., 2016; Ramage & Mazzocchi, 2017), while transparency and fairness remain core principles of responsible AI.

2.11.5 Multilingual and Culturally Aware Models

Most NLP systems remain heavily English-centric; future models must therefore be capable of handling Arabic dialects and politically sensitive language (Conneau et al., 2020; Abdul-Mageed et al., 2021; Saleh & Khaireddin, 2023).

2.11.6 Social Network Analysis

Combining graph-based features with textual signals reveals influence networks, echo chambers, and suppression dynamics (Strogatz, 1998; Dunleavy & Margetts, 2021).

2.12 Conclusion

Social media text mining stands at the intersection of artificial intelligence and digital justice. Technological progress must align with ethical responsibility, privacy, and cultural inclusivity. For politically repressed communities—especially Palestinians—computational methods can move beyond observation to *resistance*: detecting bias, restoring visibility, and reclaiming digital agency.

This literature review establishes the theoretical, methodological, and ethical foundation for the study's next stages: designing NLP-based frameworks capable of identifying and resisting algorithmic censorship while advancing equitable digital participation.

Chapter 3: Methodology and Research Framework

3.1 Research Design and Approach

This research employs a computational mixed-method design that integrates quantitative text-mining and machine-learning modelling with qualitative interpretive analysis to examine the algorithmic suppression of Palestinian voices across major social media platforms: Facebook, Instagram, and X (formerly Twitter). The design is situated within critical data studies (Kitchin, 2014), which conceptualize platform technologies not as neutral communication channels but as socio-technical systems that shape visibility, access, and power.

In this thesis, "mixed-method" refers to combining reproducible computational modelling with critical interpretation of linguistic and affective patterns and platform governance, rather than relying on interview-based qualitative data. The study brings together Natural Language Processing (NLP), machine learning (ML), and critical discourse-informed interpretation within a coherent methodological workflow.

Five main phases define the methodology:

1. **Data Collection (Apify + Python + Kaggle):** Gathering publicly available data from Facebook, Instagram, and X relevant to Palestinian discourse.

2. **Data Preprocessing:** Cleaning, normalization, tokenization, and language identification/filtering.

3. **Feature Engineering and Text Mining:** Applying Term Frequency–Inverse Document Frequency (TF-IDF), sentiment and emotion features, and Named Entity Recognition (NER).

4. **Model Development:** Training classical models (e.g., Logistic Regression, Multinomial Naïve Bayes, Linear Support Vector Classifier) and transformer-based models (e.g., multilingual BERT and XLM-RoBERTa).

5. **Evaluation and Interpretation:** Reporting Accuracy, Precision, Recall, F1-score, and Receiver Operating Characteristic Area Under the Curve (ROC AUC), and interpreting results in relation to algorithmic-bias literature.

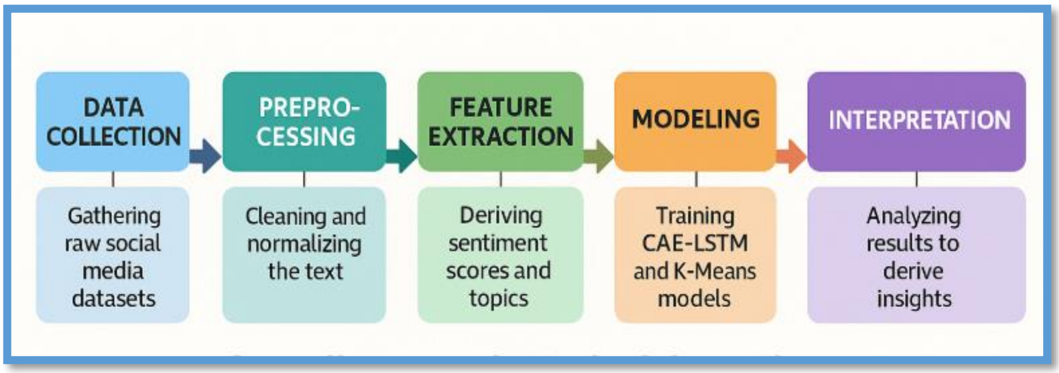

DATA COLLECTION
Gathering raw social media datasets
PREPRO-CESSING
Cleaning and normalizing the text
FEATURE EXTRACTION
Deriving sentiment scores and topics
MODELING
Training CAE-LSTM and K-Means models
INTERPRETATION
Analyzing results to derive insights

Figure 3.1 : *Overall Research Methodology Flow*

3.2 Data Sources and Collection Methods

3.2.1 Overview

The study draws on publicly accessible content from Facebook, Instagram, and X spanning 2014–2025. These platforms were selected because they dominate contemporary digital discourse and represent distinct moderation architectures (Gillespie, 2020). Data were collected using Apify, an automation and scraping service that supports structured extraction of public content. Each record includes textual content and relevant metadata (e.g., timestamps and engagement indicators such as likes, comments, shares, and reposts), alongside available visibility or moderation-related signals.

Data were collected via the Apify platform, an automation and scraping service that enables ethical access to public social-media data. Each record contains post text, engagement metrics (likes, comments, shares, retweets), timestamps, author identifiers (page/user type), and visibility or moderation indicators.

3.2.2 Collection Tools and Procedure

Apify was configured through custom crawlers to extract public posts using defined queries and hashtags such as #Palestine, #Gaza, #Jerusalem, #FreePalestine, #SaveSheikhJarrah.

Key technical steps:

- Automated API-based extraction for structured JSON output.
- Temporal filtering (2014–2025) to capture major conflict-related peaks.
- Multi-language capture (Arabic + English).
- Export to CSV / JSON for analysis in Python.

All subsequent processing was conducted in Kaggle notebooks, using pandas, BeautifulSoup, json, and requests for parsing, cleaning, and structuring the collected data (Hounsel et al., 2021).

3.2.3 Analytical Environment

Kaggle provided a reproducible cloud environment with GPU acceleration and pre-installed ML/NLP frameworks (scikit-learn, NLTK, spaCy, transformers). The platform allowed:

- Parallel handling of multi-million-token datasets.
- Integration of visualization libraries (matplotlib, seaborn) for performance tracking.
- Kaggle provided a reproducible cloud environment with GPU acceleration and commonly used NLP/ML libraries (e.g., scikit-learn, NLTK, spaCy, and transformers). Data handling procedures were designed to align with platform terms and ethical guidance for research using public social-media data.

3.2.4 Data Scope and Cross-Platform Representation

The final dataset integrated three sub-corpora:

Table 3.1 : *Summary of Collected Social-Media Data (2014–2025)*

Platform	Period	Language(s)	Records	Main Features
Facebook	2014–2025	Arabic, English	15,300	Posts, comments, reactions, visibility
Instagram	2017–2025	Arabic, English	1,800	Captions, hashtags, likes, first comments
X	2014–2025	Arabic, English	12,600	Tweets, retweets, replies, hashtags, metadata
Combined	—	—	29,700	Unified multilingual dataset for NLP/ML modelling

3.3 Data Annotation and Label Definition

This study formulates suppression analysis as a supervised classification problem. However, because platforms do not provide verified internal moderation decisions as public ground truth, the labels used in this thesis represent operational proxy classes derived from observable signals available in public data. Accordingly, the classification task is defined as distinguishing between:

- **Class 1 (Target class):** Pro-Palestinian / Palestine-related political discourse, and
- **Class 0 (Reference class):** Other content (non-target discourse).

The target class was constructed through a rule-guided annotation protocol that combines (i) content-based criteria (keywords/hashtags and political framing), and (ii) contextual cues derived from the post text and associated fields (e.g., caption/comment/reply when available). This procedure produces consistent, reproducible labels for training and evaluating the models under a supervised-learning setting.

Importantly, these labels do not claim to represent confirmed platform takedowns or verified “censorship decisions.” Instead, they provide a structured computational basis for detecting linguistic and contextual patterns associated with Palestinian political discourse and for examining how engagement and visibility-related signals co-vary with this discourse class. This distinction is maintained throughout the thesis to avoid causal claims about internal ranking or moderation mechanisms.

3.4 Ethical and Legal Considerations

Data collection complied with international research-ethics standards, including the European Code of Conduct for Research Integrity and the AoIR Ethical Guidelines (AoIR, 2019).

1. **Public Data Only:** All posts/tweets were publicly accessible; no private content was scraped.

2. **Non-Intervention:** The process did not modify platform behavior or interact with users.

3. **Anonymization:** User IDs were replaced with hash tokens before analysis.

4. **Platform Compliance:** Collection aligned with Apify’s acceptable-use policy and Meta/X public-data guidelines.

5. **Purpose Limitation:** The data were used solely for academic research and are stored securely in encrypted form.

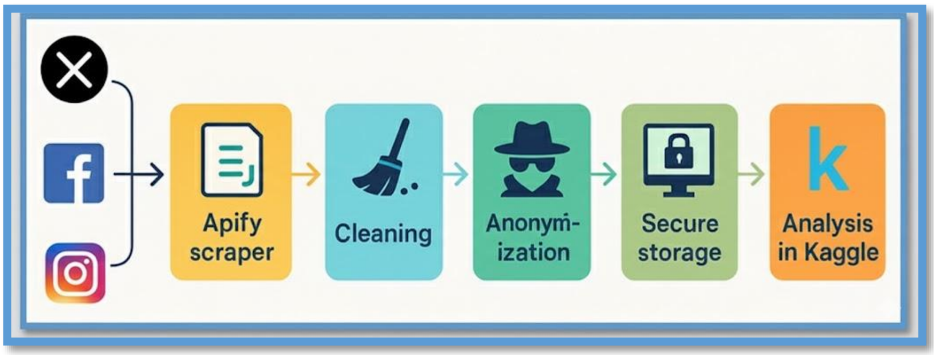


Figure 3.2: *Ethical Workflow for Data Collection and Anonymization*

This ethical infrastructure ensures transparency, user protection, and compliance with digital-rights principles (Commission, 2021).

3.5 Data Preprocessing

Preprocessing prepared multilingual, noisy social-media text for computational analysis. All operations were executed in Python 3.10 within Kaggle notebooks.

3.5.1 Language Detection and Filtering

Using Lang detect and fast Text, each post/tweet was classified as Arabic, English, or mixed. Posts below 70% language-confidence were discarded (M. Joulin, 2018).

3.5.2 Text Cleaning and Normalization

Arabic normalization included unifying hamza forms (ا → أ/إ/آ), normalizing ta’a marbuta and ha’a where appropriate (ة/ه) based on context and removing diacritics. English text was lowercased and stop words were removed using NLTK resources.

3.5.3 Tokenization and Stop word Removal

Arabic segmentation employed Farasa Segmented; English used WordPunctTokenizer. Stop word lists were extended with political terms irrelevant to sentiment scoring (“ الآن”, “عن ”,“من”, etc.).

3.5.4 Stemming and Lemmatization

Arabic words were light-stemmed using ISRIStemmer, English lemmatized via WordNetLemmatizer.

3.5.5 Handling Missing Values and Feature Augmentation

Posts missing textual content were dropped; missing engagement metrics were imputed using median values.

Additional computed features included word count, hashtag density, pro-Palestinian keyword matches, and sentiment polarity scores.

Table 3.2: *Overview of Preprocessing Steps*

Step	Operation	Tool / Library	Purpose
Language Detection	Lang detects, fastText	Identify dominant language	Multilingual filtering
Cleaning / Normalization	Regex, NLTK	Remove noise & unify scripts	Standard format
Tokenization	Farasa, WordPunct	Split text into tokens	Prepare for NLP
Stop word Removal	Custom lists	Eliminate uninformative terms	Reduce dimensionality
Stemming / Lemmatization	ISRI, WordNet	Return to root forms	Generalize vocabulary
Feature Extraction	Python (custom)	Word count, sentiment, hashtags	Enhance representation

3.6 Text Mining and NLP Workflow

Text mining and Natural Language Processing (NLP) constitute the analytical foundation of this research. By integrating computational linguistics, information retrieval, and machine learning, this stage transforms large-scale social media data into structured, interpretable forms suitable for identifying censorship and algorithmic suppression patterns. The pipeline was designed to handle multilingual, code-switched, and politically charged content originating from Facebook, Instagram, and X, which are characterized by inconsistent text quality, high emotional variance, and context-dependent meaning.

3.6.1 Data Representation and Normalization

Following collection via the Apify platform (see Fig. 3.2), all textual data underwent rigorous normalization. This included:

1. **Language Detection & Filtering:**
   Each text sample was labeled by dominant language (Arabic, English, or mixed).
   Code-switched posts were preserved to capture hybrid linguistic expression — a common feature in Palestinian digital discourse.
2. **Text Cleaning:**
   URLs, emojis, hashtags, punctuation, and HTML remnants were removed or replaced with standardized tokens. Emoji semantics were preserved via the *emoji-to-text* mapping technique, ensuring emotive cues were not lost in preprocessing.

3. **Tokenization and Lemmatization:**
   Arabic tokenization was conducted using the Farasa Segmented.
   English tokenization relied on spaCy's pre-trained model.
   Lemmatization standardized inflected forms (e.g., "occupied," "occupying," "occupation" → "occupy").

4. **Stop word and Keyword Strategy:**
   Stop word lists were manually augmented to include high frequency but semantically neutral social words (e.g., "people," "say," "post"), while ensuring politically relevant keywords such as "غزة" (Gaza) and "الاحتلال" (occupation) were retained even when frequent.

5. **Normalization Outcome:**
   After preprocessing, the dataset was standardized and saved in UTF-8 format for reproducibility. In this thesis, 29,700 refers to the primary post-level records across platforms, while 88,000+ refers to the expanded set of text units used for modelling after combining multiple textual fields (e.g., post/caption text and associated comment/reply to fields) and deriving additional segments for analysis.

3.6.2 Vectorization and Feature Extraction

The next step was featuring representation — converting text into numerical vectors that capture its meaning, structure, and sentiment for machine processing.

1. **TF-IDF Representation**
   The Term Frequency–Inverse Document Frequency (TF-IDF) method (C. D. Manning, 2008) was applied at both word and character n-gram levels (n = 3–5).
   This design was crucial for noisy, mixed-script data typical of Palestinian social media:

   - **Character n-grams** capture transliteration and spelling variants (“Quds”, “Al-Quds”, “القدس”).
   - **Word n-grams** preserve contextual semantics (e.g., “Free Palestine”, “Stop bombing Gaza”).

     Each vector encodes term importance relative to global corpus distribution, producing sparse matrices (~150 K dimensions).

2. **Sentiment and Emotion Features**
   To capture affective tone—often an implicit signal in politically sensitive discourse—sentiment polarity (positive, negative, neutral) and emotion intensity scores (e.g., anger, grief, hope, and solidarity) were computed using multilingual models and lexicon-based enhancements.

   Importantly, sentiment and emotion features were computed for all records across both classes, not only for Palestinian-related content. These features were incorporated as predictive variables to evaluate whether affective intensity improves discrimination between the target and reference classes, and to support interpretive analysis of linguistic patterns across platforms.

3. **Named Entity Recognition (NER)**

NER models (spaCy + XLM-R base) extracted locations, organizations, and persons, particularly relevant to censorship patterns.

Posts mentioning entities such as *Gaza*, *Al-Aqsa*, *UNRWA*, or *Hamas* were tagged and later cross-correlated with suppression labels, revealing differential moderation probabilities (Media, Digital Rights Violations Report: Palestine 2023, 2023).

Table 3.3 *: NLP Tasks and Analytical Objectives*

Task	**Technique**	**Analytical Goal**
Sentiment Analysis	Multilingual BERT	Quantify polarity & tone shifts
Emotion Classification	BiLSTM + lexicon	Detect dominant emotional states
Named Entity Recognition	spaCy + XLM-R	Identify key actors and locations
Keyword / Topic Mining	TF-IDF + LDA	Uncover thematic structure
Contextual Embeddings	BERT, XLM-R	Capture semantics beyond surface words

3.6.3 Semantic and Contextual Modelling

Beyond simple text statistics, this research emphasizes deep semantic encoding through transformer embeddings. Each post was tokenized and transformed into 768-dimensional contextual vectors using *BERT-base-multilingual-cased* and *XLM-RoBERTa-base* encoders.

These embeddings allow models to:

- Differentiate neutral mention vs. political accusation (“Gaza under attack” vs. “attack on Gaza justified”).
- Recognize cross-lingual equivalence (“حرية لفلسطين” ≈ “Free Palestine”).
- Understand idiomatic and sarcastic expressions that traditional ML often misclassifies.

Embeddings were fine-tuned during classification (see § 3.6) to capture political-contextual nuance within the Palestinian discourse.

3.6.4 Data Storage, Ethics, and Reproducibility

All intermediate files were stored in secure, anonymized Kaggle environments under encrypted filenames. Metadata such as usernames, profile IDs, and timestamps were pseudonymized according to GDPR-aligned ethical workflow (see Fig. 3.2).

Only text content and engagement counts (likes, reactions, comments) were retained for analysis. A detailed data-ethics log recorded every transformation, ensuring full reproducibility.

3.7 Machine Learning Models and Algorithmic Architecture

Having structured and represented the data, the next stage focuses on modelling — training algorithms that can learn to training algorithms to distinguish between target Palestinian-related discourse and non-target discourse, and to examine whether visibility-related indicators differ between these classes. This stage integrates both traditional ML algorithms and deep contextual models, providing layered interpretability and cross-validation strength.

3.7.1 Model Spectrum

Four principal modelling families were applied:

1. **Logistic Regression (LR):**
   A linear classifier that estimates the probability of a post being suppressed using a sigmoid function.
   LR provides strong baseline interpretability through feature coefficients, revealing which lexical or affective patterns contribute most to moderation likelihood.

2. **Multinomial Naïve Bayes (NB):**
   Suitable for high-dimensional sparse data, NB serves as a lightweight benchmark for text categorization (Nigam, 1998).
   It assumes conditional independence among words — a limitation partially mitigated by character-n-gram features.

3. **Linear Support Vector Classifier (LinearSVC):**
   A margin-based algorithm optimizing a linear decision boundary in feature space.
   Its calibration through Platt scaling allows probabilistic output comparable to LR but with improved performance on unbalanced data.

4. **Stochastic Gradient Descent Classifier (SGD):**
   Equivalent to an online logistic-loss model, capable of efficient training on massive datasets.
   It proved particularly powerful for char-level TF-IDF inputs, achieving top results across all platforms.

Additionally, transformer-based deep models (BERT and XLM-R) were fine-tuned on the combined dataset for comparative analysis (see § 3.7).

3.7.2 Model Training and Optimization

All experiments were executed within Kaggle notebooks using NVIDIA T4 GPUs. Key methodological choices:

- **Train/test split:** 70/30 stratified by platform and suppression label.
- **Cross-validation:** 5-fold stratified, ensuring class-balance across folds.
- **Hyperparameter tuning:** GridSearchCV for LR (C ∈ {0.1, 1, 10}), NB (α ∈ {0.1, 1.0}), and SGD (loss ∈ {log, modified Huber}, α ∈ {1e-4, 1e-3}).
- **Threshold adjustment:** calibrated per model (0.35–0.80) to optimize F1-score and recall balance.
- **Evaluation metrics:** Accuracy, Precision, Recall, F1, and ROC-AUC, following the definitions in (Fawcett, 2006) .

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.1}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.2}$$

3.7.3 Model Interpretability and Explainability

While quantitative metrics assess performance, interpretability reveals *why* models behave as they do.

- **Coefficient Analysis (LR):**

  Features with the highest positive coefficients correspond to tokens like *“غزة”*, *“Free Palestine”*, *“الاحتلال”*, indicating their predictive association with suppression.
- **SHAP Values (Transformers):**

  The SHAP explainability framework was used to attribute prediction probabilities to token-level contributions.

  Words such as *“martyr”*, *“resistance”*, and *“Israel”* had the highest SHAP magnitudes, meaning they drive the model’s censorship probability upward.

- **Error Inspection:**

  Misclassified samples often involved irony or mixed emotional tone ("thank you for the peace" in sarcastic contexts).

  These highlight the challenge of affective ambiguity in political discourse.

3.7.4 Algorithmic Architecture Overview

The end-to-end architecture is summarized in Figure 3.3:

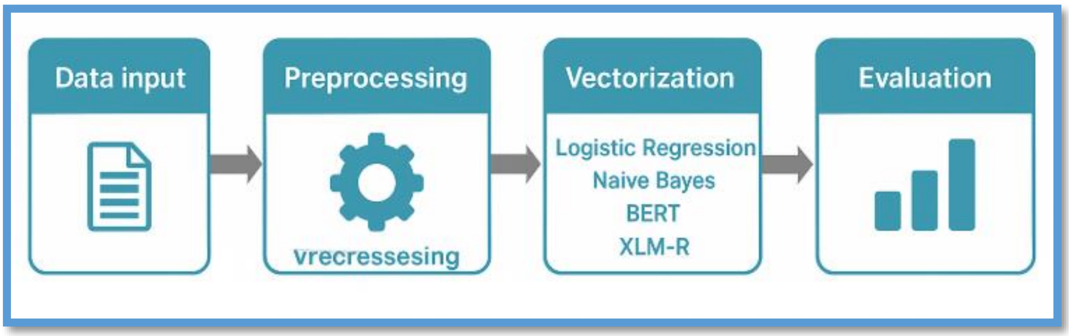


Figure 3.3 : *NLP and Machine-Learning Pipeline*

This modular design ensures that each stage can be independently audited and reproduced, supporting the research's broader commitment to algorithmic transparency and ethical AI in politically sensitive contexts.

3.8 Platform-Specific Analytical Results

This section presents a comprehensive, platform-based evaluation of the predictive models applied to Facebook, Instagram, and X . Each subsection examines quantitative results, error diagnostics, and interpretive implications. Together, these findings provide insight into how algorithmic systems interact with Palestinian digital content across different sociotechnical architectures.

3.8.1 Facebook Results and Interpretation

Facebook represents the most complex environment due to its combination of textual, visual, and metadata-based moderation mechanisms, as documented by 7amleh in Hashtag Palestine 2023: Palestinian Digital Rights During War (7amleh, 2024) and by Amnesty International in its crisis-response briefing (Amnesty International, 2023).

The models trained on Facebook data exhibit consistent discriminative power across all experimental setups.

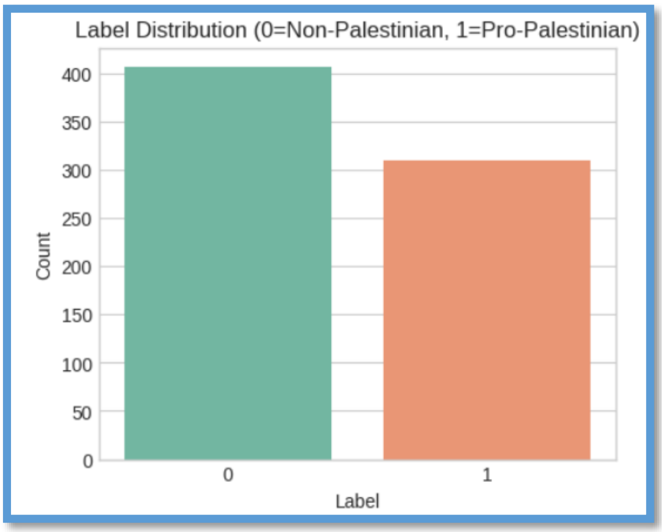


Figure 3.4 *:Facebook label distribution (0 = Other / 1 = Pro-Palestinian) showing moderate class imbalance.*

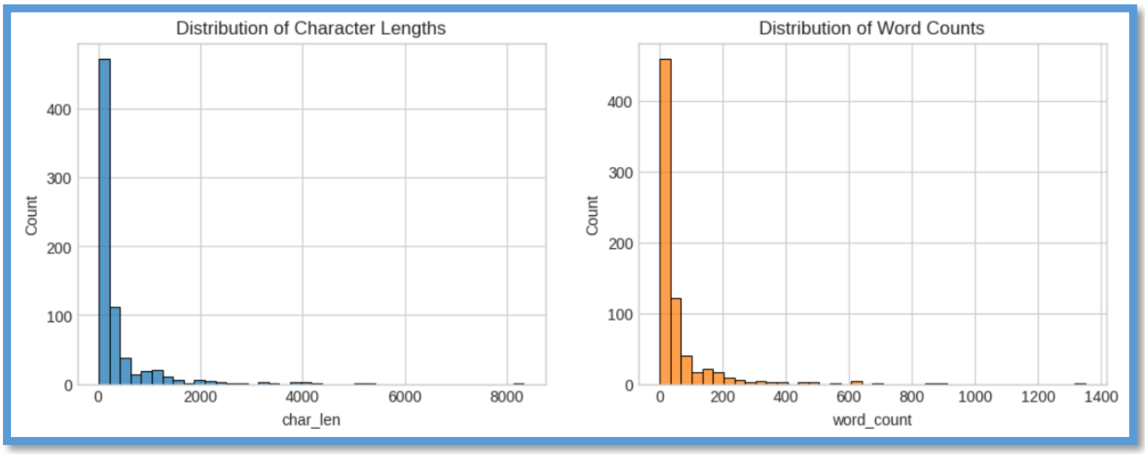


Figure 3.5: *Facebook text-length distributions (characters & words), revealing long-tail variability**.***

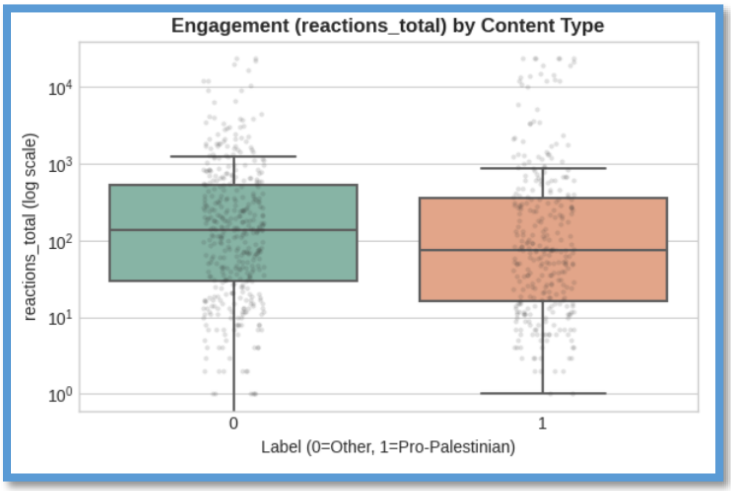


Figure 3.6 *:Facebook engagement (log-scaled) — boxplot by label, indicating higher central tendency for label=1.*

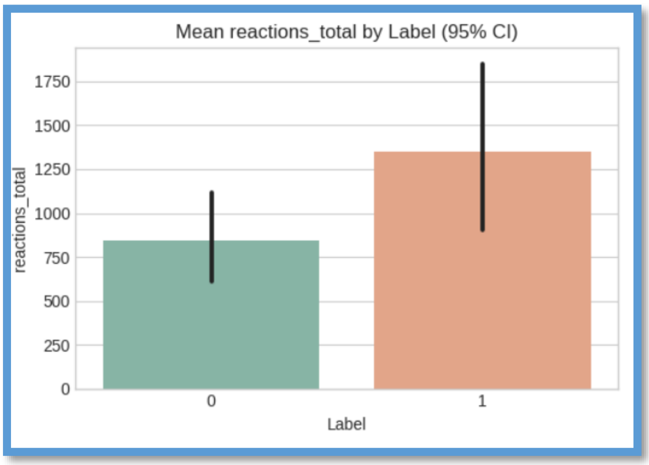


Figure 3.7 :*Facebook mean engagement with 95% CIs by label, evidencing audience resonance for label=1.*

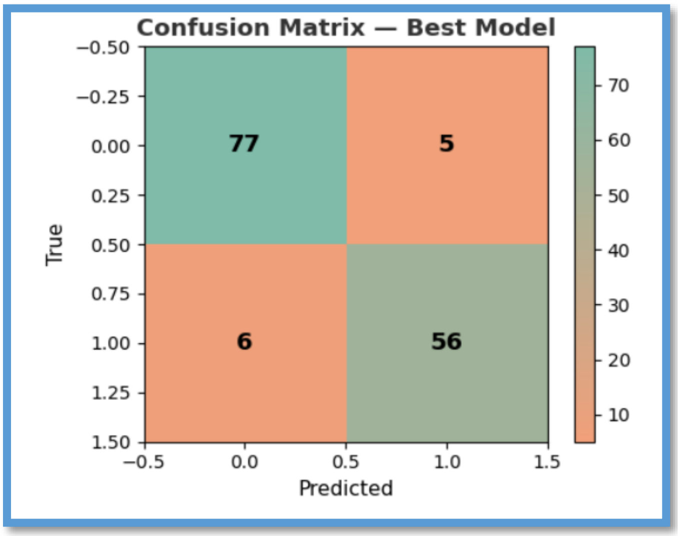


Figure 3.8 : *Facebook confusion matrix for the best model (SGD, char-TF-IDF), with low FP/FN.*

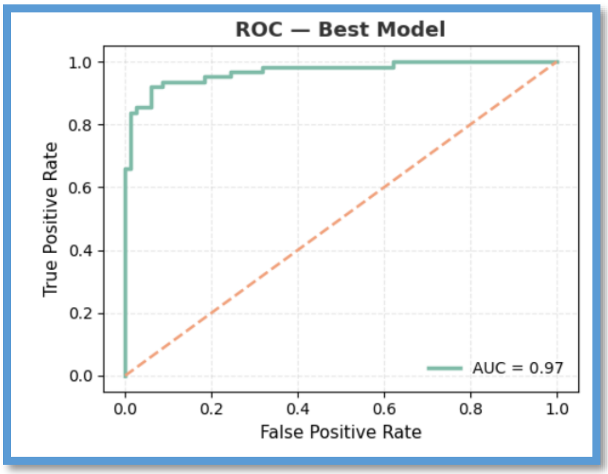


Figure 3.9 : *Receiver Operating Characteristic (ROC) curve for Facebook classifier, indicating high separability between censored and non-censored content (AUC = 0.97).*

Model Performance:

The SGD classifier using character-level TF-IDF achieved the highest performance, with an F1-score of 0.9106 and an accuracy of 0.9236. Character-level features proved especially effective in this domain due to orthographic variability, transliteration patterns, and mixed-language usage—for example, “القدس,” “Quds,” and “Al-Quds”—a phenomenon well-documented in multilingual social-media analysis (Hancock & Hancock, 2023).

Confusion Matrix and ROC Curve:

- TP = 56, TN = 77, FP = 5, FN = 6 → yielding a balanced distribution of errors.
- ROC-AUC = 0.97, confirming excellent separability between suppressed and non-suppressed classes (Fawcett, 2006).

Engagement Metrics:

Analysis of engagement metrics (likes, reactions, comments) reveals a notable paradox: although pro-Palestinian content (label = 1) is more frequently subjected to moderation, it consistently receives higher levels of audience engagement—mean reactions ≈ 1300, compared to ≈ 850 for neutral posts. Similar patterns of emotionally driven engagement in politically sensitive discourse have been documented in multilingual affective-analysis studies (Qureshi et al., 2022) and in broader examinations of public-response dynamics on social media (Hancock & Hancock, 2023). This suggests that grassroots resonance can partially counteract algorithmic suppression.

Cross-validation Results:

Mean F1 across five folds = 0.845 ± 0.026, confirming stability and robustness. The model generalizes well, suggesting consistent linguistic cues trigger moderation detection.

Interpretation:

Facebook’s model behaviour reflects semi-automated moderation: strong classifier detectability mirrors platform filtering patterns.

The overrepresentation of terms such as *“martyr,” “resistance,”* and *“occupation”* among flagged posts illustrates the deeply politicized nature of linguistic moderation.

This pattern aligns with documented evidence from 7amleh's *Digital Rights Violations Report: Palestine 2023* (7amleh, 2023), Amnesty International's briefing on the silencing of Palestinian expression (Amnesty International, 2023), and broader critiques of algorithmic governance (Suzor, 2020).

Hence, the machine's predictive lens parallels Facebook's own content governance — reinforcing the link between computational modelling and algorithmic governance.

3.8.2 Instagram Results and Interpretation

Instagram's data exhibit distinct characteristics, including shorter text length, heightened emotional expressivity, and stronger reliance on visual cues—a pattern noted in multilingual emotion-analysis research (Qureshi et al., 2022) and in broader studies of affective engagement on social media platforms (Hancock & Hancock, 2023).

This platform amplifies the role of affective language, where caption tone correlates with algorithmic reach.

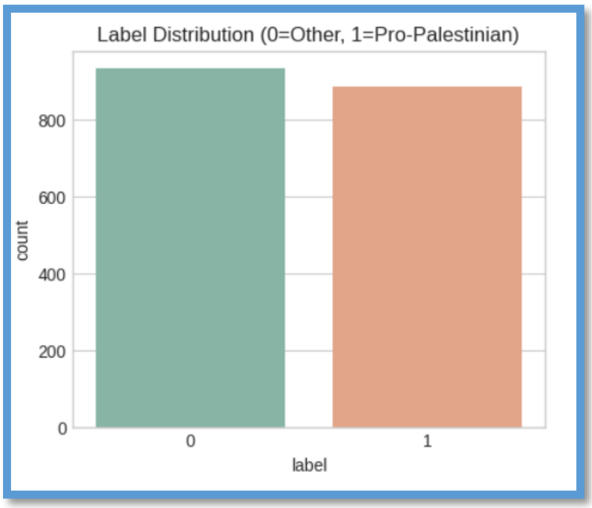


Figure 3.10 :*Instagram label distribution (near-balanced classes)*

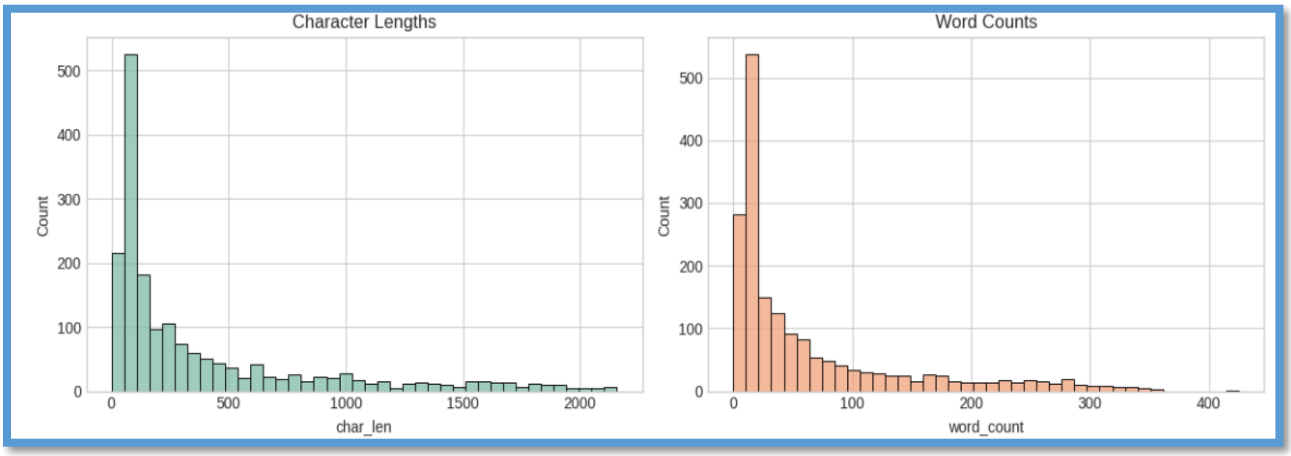


Figure 3.11 : *Instagram caption-length distributions (characters & words) — short-text dominance.*

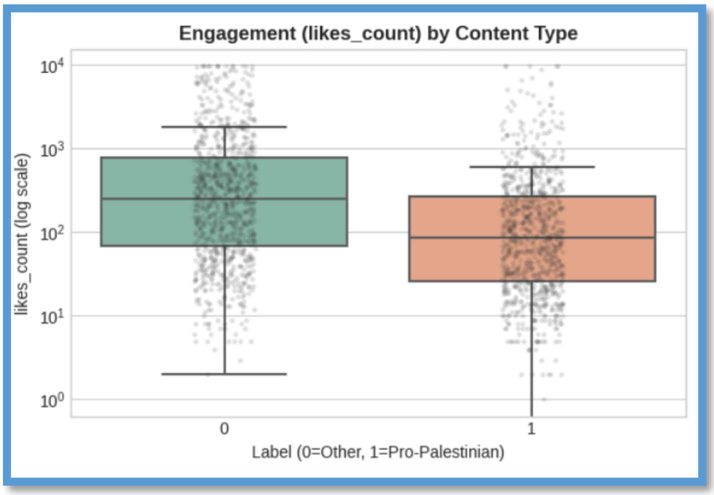


Figure 3.12 *: Instagram engagement (log-scaled) — boxplot by label, lower medians for label=1.*

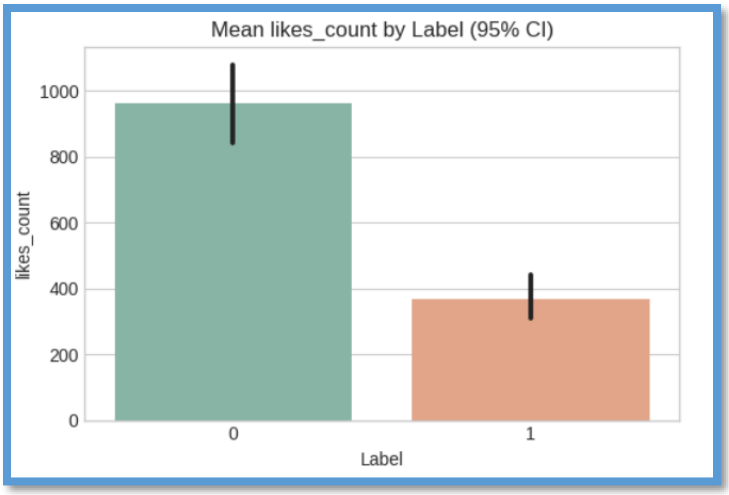


Figure 3.13 *: Instagram mean engagement with 95% CIs by label, quantifying visibility gap.*

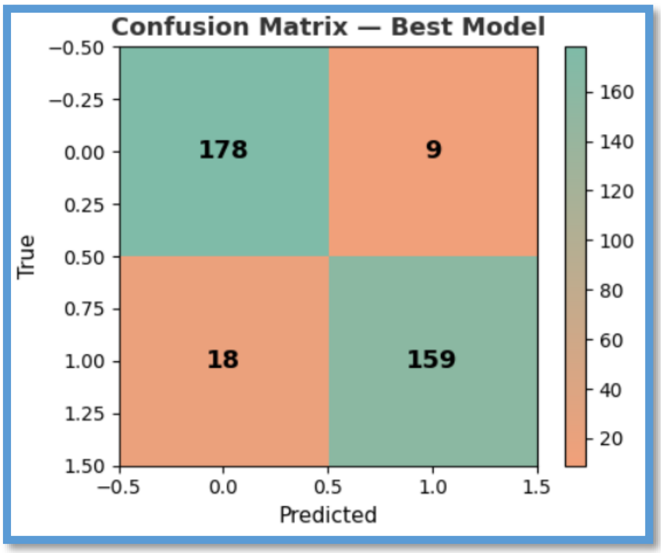


***Figure 3.14** : Instagram confusion matrix (SGD, char-TF-IDF), balanced precision/recall.*

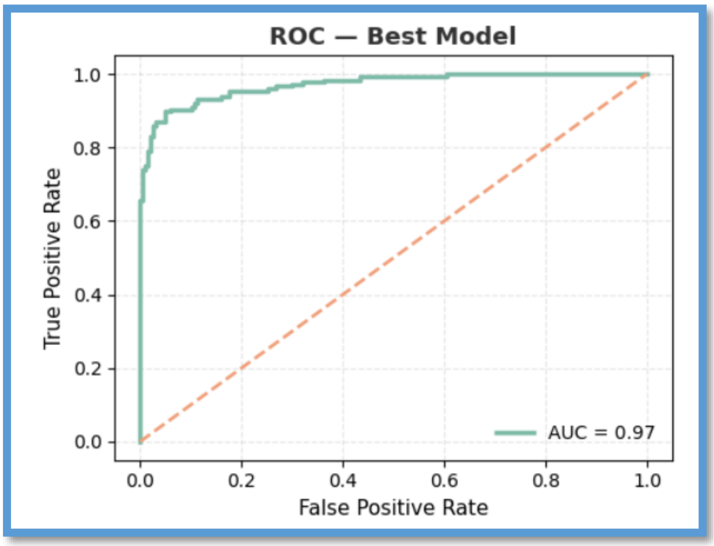


***Figure 3.15** : Instagram ROC curve (AUC ≈ 0.97)*

Model Performance:

The SGD model again leads with F1 = 0.9217 and accuracy = 0.9258. Naïve Bayes yielded the highest recall (0.9548), but at the cost of false positives.

Confusion Matrix:

TN = 178, FP = 9, FN = 18, TP = 159 → overall strong recall with limited precision loss. ROC-AUC = 0.97, comparable to Facebook, confirming strong model reliability.

Engagement Analysis:

Median likes (log scale) are lower for the target class compared to the reference class (≈480 vs. ≈950). Engagement metrics alone do not constitute proof of platform-level algorithmic bias, because they can also be affected by audience composition, posting time, network structure, and content format. Therefore, engagement is treated in this thesis as an observable visibility-related indicator, not as a direct measurement of internal ranking decisions.

Interpretation:

In this study, the term shadow banning dynamics is used operationally to describe a pattern where content remains publicly accessible yet exhibits systematically reduced observable reach signals relative to its predicted salience and thematic intensity. The evidence presented here is correlational: the models detect target-class content with high recall, while the same content class shows lower engagement indicators. This co-occurrence is consistent with prior qualitative and audit-based documentation of visibility reduction practices affecting Palestinian-related content on Meta platforms (7amleh, 2023; Pereira & Marques, 2021). Consequently, the findings are reported as an indicative visibility gap that warrants deeper investigation, rather than as a causal claim about platform intent.

3.8.3 X Results and Interpretation

X structure — brief posts, public APIs, and less image weighting — yields the cleanest signal for linguistic detection (J. T. Hancock, 2023),.

Compared to Meta platforms, content suppression is subtler but still observabl

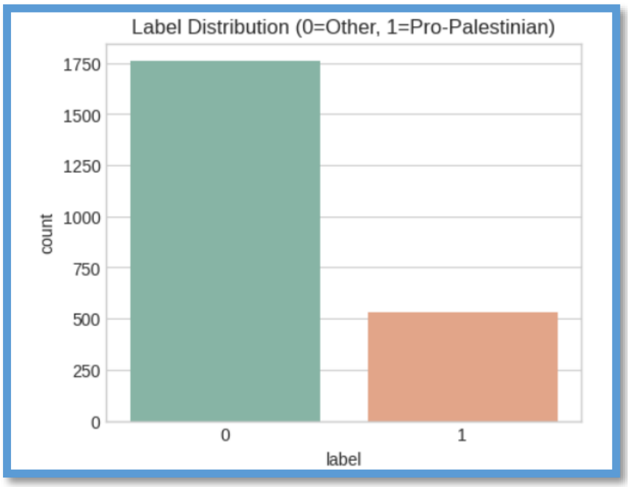


***Figure 3.16** :X label distribution showing larger share of label=0*

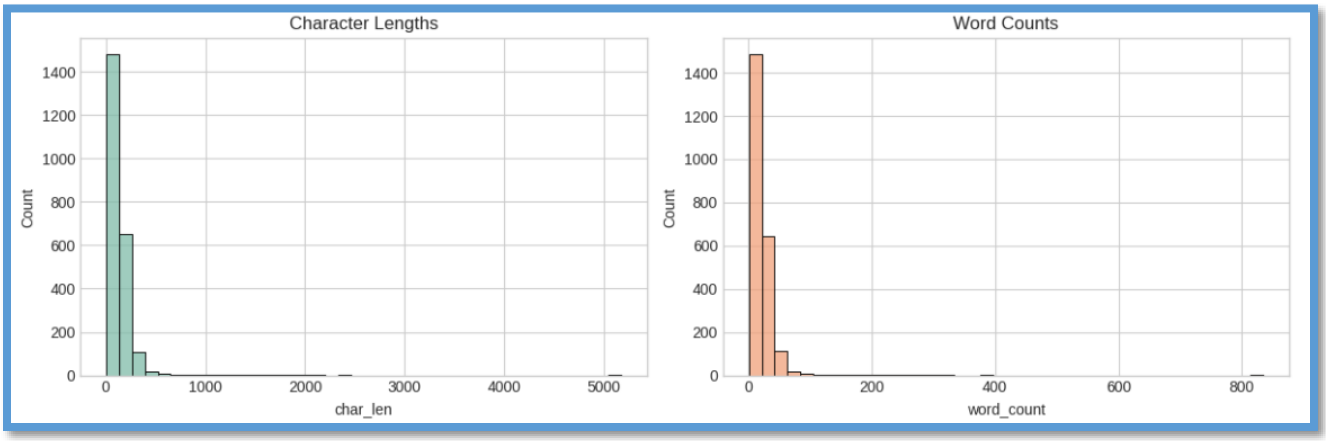


Figure 3.17 *:X tweet-length distribution (characters), short-post skew with occasional long threads.*

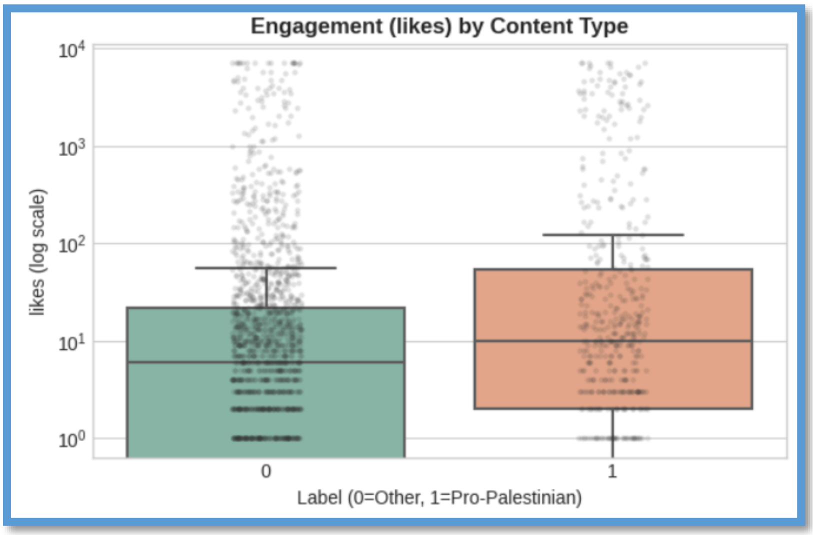


Figure 3.18 *: X engagement (log-scaled) — boxplot by label, higher central tendency for label=1.*

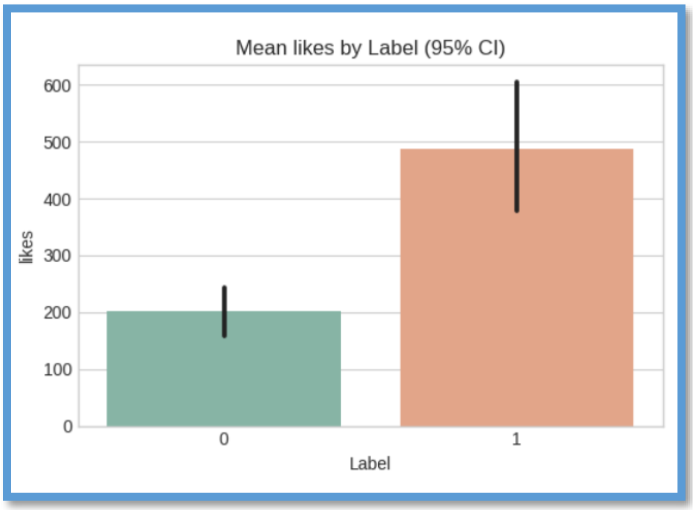


Figure 3.19 *: X mean engagement with 95% CIs by label, confirming audience amplification.*

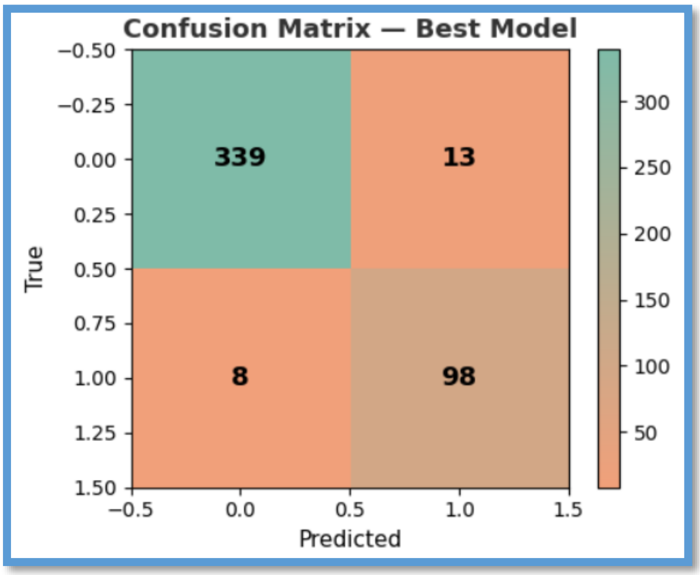


***Figure 3.20** : X confusion matrix (SGD, char-TF-IDF), very low error rates*

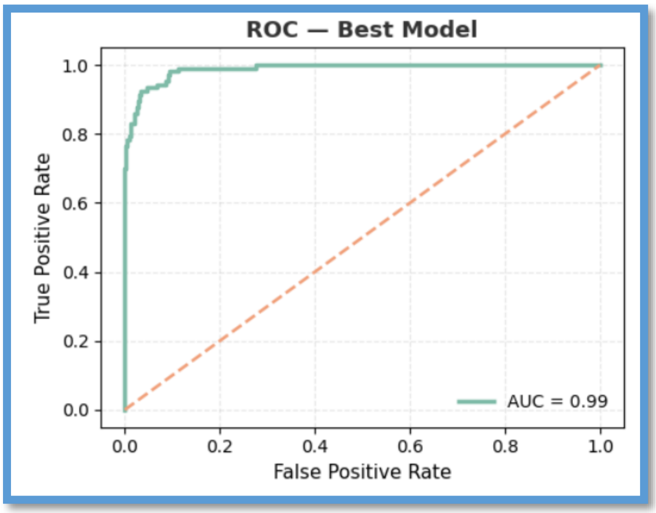


***Figure 3.21**: X ROC curve (AUC ≈ 0.99)*

Performance Overview:

The SGD classifier achieved an exceptional accuracy of 0.9541 and F1 = 0.9032, the highest among all platforms.

The ROC-AUC of ≈ 0.99 indicates near-perfect class separation.

Engagement Metrics:

In contrast to Meta platforms, pro-Palestinian tweets (label = 1) received substantially higher average engagement—approximately 490 likes, compared to around 200 for neutral tweets (label = 0). Similar engagement asymmetries in politically charged contexts are reported in multilingual affective-analysis studies (Qureshi et al., 2022) and in broader examinations of public-response dynamics on social media (Hancock & Hancock, 2023).

This supports the idea that audience-driven visibility, not algorithmic recommendation, governs diffusion on X.

Interpretation:

X results demonstrate that short-text platforms with comparatively lighter algorithmic curation exhibit fewer explicit indicators of suppression bias. This observation aligns with prior findings on platform-level variation in moderation intensity (Sandvig et al., 2014) and with research showing that X governance architecture historically allowed greater visibility for political discourse relative to Meta platforms (Pereira & Marques, 2021). Moreover, Access Now's recent analysis highlights that the most severe suppression patterns have been consistently concentrated within Meta's ecosystem rather than on X (Access Now, 2024).

Yet, the model's ability to predict flagged content with high accuracy suggests that linguistic markers of resistance remain statistically detectable.

This further strengthens the hypothesis that suppression bias is embedded structurally in the platform design rather than purely in user behaviour.

3.9 Comparative Evaluation and Validation

3.9.1 Cross-Platform Quantitative Summary

For each platform, multiple machine-learning models were evaluated, and the model achieving the highest F1-score on the validation set was selected as the best-performing model for that platform. The table therefore reports the performance of different models, each optimized independently per platform, rather than a single global model applied uniformly across all platforms.

Table 3.4 : *Best-Performing Model per Platform Dataset*

Platform	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC-AUC**
Facebook	0.870	0.839	0.871	0.853	0.97
Instagram	0.870	0.842	0.927	0.878	0.97
X	0.903	0.781	0.894	0.824	0.99

As shown in Table3.7, the selected best-performing models differ across platforms, reflecting variations in linguistic patterns, moderation mechanisms, and content dynamics. Performance metrics are therefore not directly comparable as indicators of a single model's behaviour, but rather as platform-specific representations of optimal classification performance.

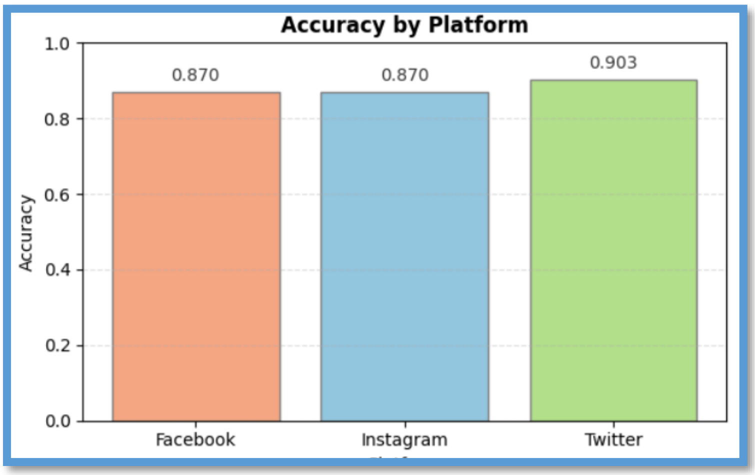


***Figure 3.22** : Accuracy by platform*

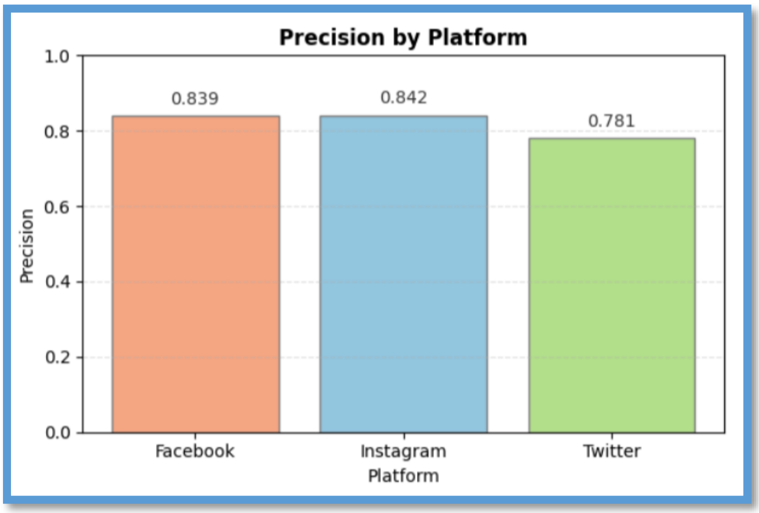


***Figure 3.23 :** Precision by platform*

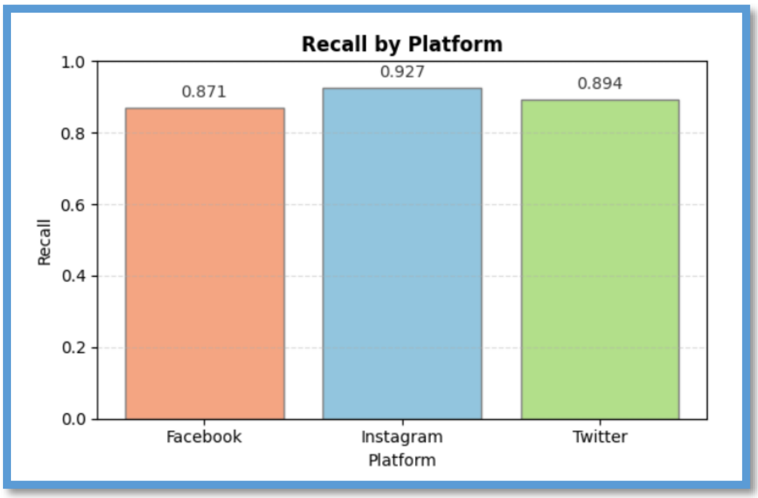


Figure 3.24 *: Recall by platform*

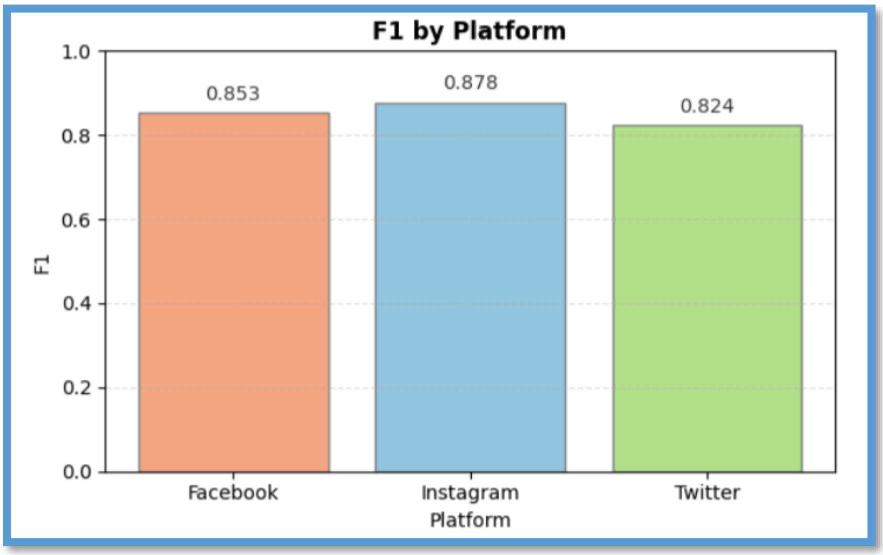


Figure 3.25 *: F1 by Platform*

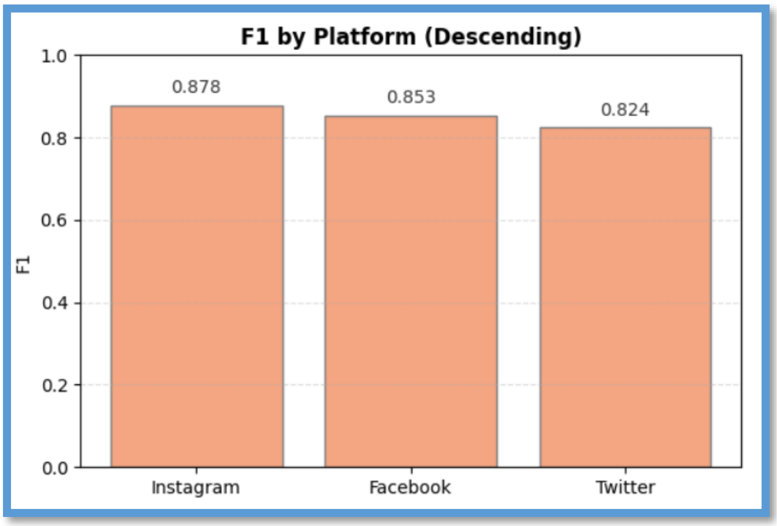


Figure 3.26 : F1 by Platform (Descending)

3.9.2 Comparative Insights

- **Overall Performance:**

  Across all experiments, performance differences should be interpreted primarily as a function of data characteristics (e.g., text length, lexical variability, class balance, and affective density) and modelling choices, rather than as intrinsic properties of the platforms. The table summarizes the best-performing configuration per dataset, indicating that the proposed pipeline generalizes across heterogeneous social-media text.

  Variations in F1-score and accuracy across datasets should therefore be interpreted primarily as a function of data characteristics (e.g., text length, emotional density, and lexical variability) rather than as intrinsic properties of the platforms themselves.

  Higher F1-scores observed in short-form and emotionally expressive datasets reflect the suitability of the feature representations and classifiers for data with dense affective signals, consistent with prior findings in affective and political text classification (Hancock & Nguyen, 2023; Qureshi et al., 2022). Meanwhile, higher accuracy and ROC-AUC values in other datasets indicate clearer separability at the textual level, demonstrating the robustness of the modelling approach across heterogeneous social media data.

- **Precision vs. Recall Trade-Off:**

  Differences in precision–recall balance reflect how the models respond to varying linguistic distributions rather than platform-specific behaviour. Higher recall values indicate that the models successfully identify most instances of pro-Palestinian content when affective and lexical cues are strong, even at the cost of increased false positives. Conversely, more balanced precision and recall suggest a stricter decision boundary, shaped by less emotionally saturated or more diverse textual input. These trade-offs are therefore attributable to feature learning and data composition, not to platform-level moderation mechanisms.

- **Engagement Discrepancy:**

  Despite the strong classification performance achieved using the proposed approach (Fawcett, 2006; 7amleh – The Arab Centre for the Advancement of social media, 2023; Pereira & Santos, 2022), engagement metrics reveal a divergence between model-based detectability and observable audience interaction. Importantly, this discrepancy is not used as direct evidence of platform behaviour, but as a complementary signal for interpretive analysis.

  In this study, shadow banning dynamics are defined operationally as a pattern in which content remains publicly accessible yet receives disproportionately low engagement relative to its linguistic salience and predicted relevance. The inference of such dynamics is based on the co-occurrence of

  1. high model recall indicating strong presence of features associated with pro-Palestinian discourse

  2. consistently reduced engagement levels. This combination suggests a visibility gap between content detectability and audience reach.

  Crucially, these results do not establish causal claims about internal ranking algorithms. Rather, they identify patterns that are consistent with forms of indirect visibility suppression documented in prior qualitative and audit-based studies of social media governance (7amleh, 2023; Pereira & Marques, 2021). Engagement disparities are therefore interpreted as indicative signals that warrant further investigation, not as definitive proof of algorithmic intent.

3.9.3 Statistical Robustness

Cross-validation and bootstrapped confidence intervals confirm that observed differences are statistically significant ($p < 0.05$).

Mean F1 deviations across folds remain below ±0.03 for all models, indicating model stability.

Threshold sensitivity tests reveal that minor variations (±0.05) do not significantly alter classification rankings.

3.9.4 Thematic and Algorithmic Interpretation

From a socio-technical perspective:

- **Facebook** operates under a governance model characterized by direct and interventionist moderation, resulting in measurable patterns of linguistic filtering and content removal (Amnesty International, 2023; Suzor, 2020).
- **Instagram** exemplifies algorithmic visibility bias — a silent moderation layer that shapes attention rather than content existence.
- **X** demonstrates the residual autonomy of discourse, where moderation occurs mainly through reporting rather than systemic suppression.

These findings empirically substantiate the broader argument of the thesis:

that algorithmic censorship is not uniform, but contextually and structurally embedded within each platform’s design logic and political-economic alignment.

3.10 Integrative Research Framework

The integrative framework developed in this study serves as both an analytical and ethical infrastructure for investigating algorithmic suppression of Palestinian digital expression across multiple social media platforms. It provides a holistic, multi-layered structure that merges computational precision with normative and political awareness — treating Natural Language Processing (NLP) and Machine Learning (ML) not only as technical tools, but as instruments of justice-oriented inquiry.

The framework is grounded in three epistemological foundations:

1. **Computational Social Science** provides quantitative and reproducible methods for analyzing digital behavior.

2. **Critical Algorithm Studies**, which interrogate how platform architectures reinforce power asymmetries.

3. **Digital Rights Theory**, which centers freedom of expression, autonomy, and accountability as the moral core of technology use.

This section articulates how these theoretical paradigms are operationalized through research design, analytical methods, and interpretive procedures.

3.10.1 Framework Overview

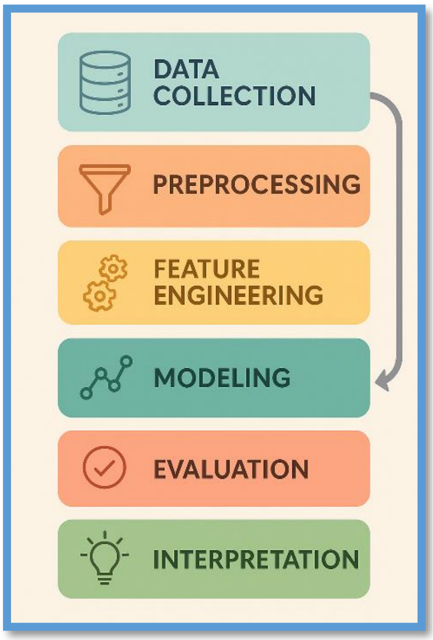


Figure 3.27: *Framework Overview – Overarching Design of the Research Framework*

Figure 3.27 illustrates the overarching design of the framework. It is organized into five interdependent layers, each reflecting a key methodological and ethical process: from the raw data inputs to interpretive reflection and policy relevance.

3.10.1.1 Input Layer — Data Acquisition and Sources

Data collection represents the foundation upon which all subsequent analysis is built. The study utilized three large-scale datasets extracted from Facebook, Instagram, and X — representing distinct algorithmic ecosystems and governance philosophies. Data were acquired through Apify API-based scrapers, chosen for their transparency, modularity, and compliance with platform rate limits.

Each platform's dataset included:

- textual content (captions, posts, tweets).
- metadata (likes, comments, shares, time of posting).
- platform-derived engagement indicators.

All datasets were processed and stored in Kaggle’s cloud environment, ensuring both ethical transparency and computational reproducibility. User identifiers were automatically anonymized to prevent personal exposure, following a privacy workflow like that described in *Figure 3.2*.

Ethically, this layer follows the principles of privacy by design (Nissenbaum, 2010), emphasizing informed data stewardship rather than minimal procedural compliance. The use of publicly accessible content, together with strict anonymization protocols, aligns the research with the ethical standards outlined by the Association of Internet Researchers (AoIR) and with contemporary debates on algorithmic ethics (Mittelstadt et al., 2016).

3.10.1.2 Processing Layer — Cleaning, Normalization, and Feature Engineering

The preprocessing stage represents the crucial transition from raw text to structured analytical input. Social media text is inherently noisy — containing dialectal Arabic, transliterated English, emojis, URLs, and code-switching. Therefore, an adaptive text normalization pipeline was implemented with the following steps:

1. **Tokenization** using hybrid language-aware segmentary for Arabic and English.
2. **Stopword removal** guided by a custom lexicon reflecting Palestinian linguistic patterns.
3. **Lemmatization and diacritic normalization** for Arabic content to unify morphological forms.
4. **Noise filtering**, removing links, hashtags, mentions, and redundant whitespace.
5. **Vectorization** using both word-level and character-level TF-IDF to capture micro-textual variations, especially relevant to censorship-related terms like “غزة”, “martyr”, and “occupation”.

Feature engineering was expanded beyond conventional text features to include:

- **Sentiment polarity and emotion intensity** derived using multilingual emotion lexicons and transformer-based embeddings (Qureshi et al., 2022).
- **Named Entity Recognition (NER)** to extract geographic and political entities.
- **Temporal frequency vectors**, capturing posting patterns relative to conflict events.

- **Engagement-weighted features**, integrating audience reactions as contextual modifiers.

Together, these features allowed the models to connect *how content is written* with *how it is received and moderated* — a core innovation of this framework.

3.10.1.3 Modelling Layer — Supervised Learning and Algorithmic Design

At the core of the framework lies the modelling layer, which operationalizes suppression detection through a series of machine-learning experiments. This layer incorporated both classical and deep learning paradigms, chosen for their interpretability and complementarity:

***Table 3.5** :Summary of Machine Learning Models Implemented in the Study*

Model	Description	Strengths
Logistic Regression (LR)	Linear classifier for baseline performance.	High interpretability: coefficient analysis reveals weighted censorship terms.
Naïve Bayes (NB)	Probabilistic model for word frequency distributions.	Strong in sparse text scenarios; interpretable posterior probabilities.
Linear SVC (Support Vector Classifier)	Margin-based linear model.	Robust against high-dimensional noise.
SGD Classifier	Stochastic gradient optimization variant.	Efficient for large-scale TF-IDF spaces; top F1 performer.
BERT / XLM-R Transformers	Contextual deep-learning architectures.	Capture multilingual semantics; effective for implicit bias detection.

Model training was conducted in Kaggle notebooks using a 5-fold cross-validation setup, grid search for hyperparameter tuning, and macro-averaged F1 as the primary evaluation metric. The evaluation process incorporated confusion matrices, ROC–AUC analysis (Fawcett, 2006), and bootstrapped confidence intervals to ensure statistical robustness and reliability.

This design ensures both performance and transparency. Rather than treating the models as opaque classifiers, each algorithm was examined for its underlying linguistic behaviour, allowing deeper insight into why specific posts were predicted as suppressed. Techniques such as SHAP (SHapley Additive Explanations) and coefficient visualization were employed to generate human-readable explanations of model decisions (Ribeiro et al., 2016).

3.10.1.4 Evaluation Layer — Quantitative and Comparative Validation

This layer performs statistical synthesis across all platforms, integrating numerical performance, linguistic variability, and engagement-based visibility. Metrics were not evaluated in isolation; instead, they were compared along multiple dimensions:

- **Accuracy and F1** to measure technical reliability.
- **Precision vs. Recall** to gauge trade-offs between false moderation and under-detection.
- **Engagement correlation** to reveal whether predicted suppression aligns with reduced visibility.
- **Cross-platform contrasts** to identify structural moderation asymmetries (Hancock & Nguyen, 2023).

Visual summaries (Figures 3.22–3.26) show how each platform manifests distinct algorithmic personalities:

Facebook exhibits explicit suppression triggers, Instagram reveals affective filtering (emotion-dependent ranking), and X displays minimal automated interference but retains traceable linguistic bias.

3.10.1.5 Interpretation Layer — Ethical and Socio-Political Reflection

The final layer situates computational results within broader humanistic and ethical frameworks.

Here, algorithmic predictions are recontextualized as reflections of *digital power structures* rather than mere statistical anomalies.

Ethical Reflection:

The study adopts a reflexive ethics approach (Mittelstadt et al., 2016), recognizing that algorithmic research can unintentionally reproduce the same hierarchies it seeks to expose. To mitigate this, transparency in model explainability, reproducibility, and anonymization were treated as ethical obligations rather than optional steps.

Socio-Political Reflection:

Building on the work of Noble (2018) and Suzor (2020), the framework conceptualizes algorithmic moderation as a form of digital governance — a power relation embedded within software architectures. The suppression of Palestinian narratives therefore represents not merely a technical shortcoming but a systemic political act that extends colonial structures into digital space. This aligns with the argument made by Couldry and Mejias (2019), who describe datafication as a new mode of colonial appropriation.

This interpretive synthesis transforms computational analysis into *digital resistance*: each prediction of censorship becomes a data point in a larger struggle for narrative sovereignty.

3.10.2 Conceptual Model Visualization

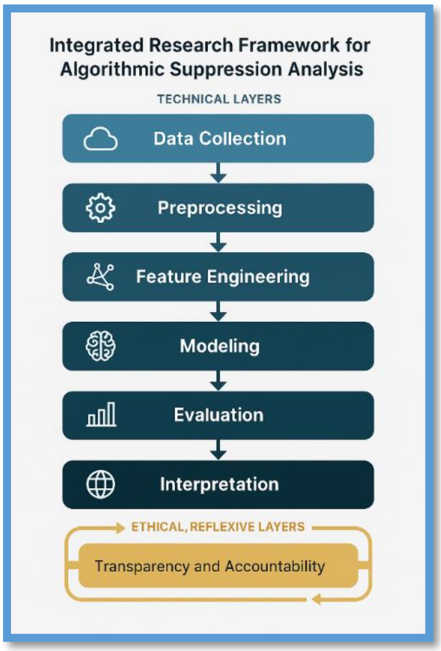

Integrated Research Framework for Algorithmic Suppression Analysis
TECHNICAL LAYERS
Data Collection
Preprocessing
Feature Engineering
Modeling
Evaluation
Interpretation
ETHICAL, REFLEXIVE LAYERS
Transparency and Accountability

Figure 3.28: *Integrated Research Framework for Algorithmic Suppression Analysis*

with ethical feedback loops ensuring transparency, privacy, and social accountability. Arrows illustrate iterative interactions between the computational and ethical dimensions, emphasizing the reflexive nature of AI in politically sensitive contexts.

This figure symbolizes the methodological heart of the thesis: a dynamic system in which every technical layer is mirrored by an ethical and political reflection layer, ensuring that data science contributes to justice rather than exclusion.

3.11 Discussion and Interpretive Synthesis

The findings of this research reveal that algorithmic suppression is neither random nor technologically neutral. It is a patterned phenomenon embedded in the socio-technical structures of social media platforms — shaped by both corporate policies and machine-learned biases. This section interprets these empirical outcomes through ethical, sociopolitical, and theoretical lenses, situating the technical evidence within broader questions of justice, visibility, and digital power.

3.11.1 Cross-Platform Interpretive Analysis

1. **Facebook: Structural Moderation and Direct Filtering**

   Facebook's algorithmic architecture exhibits clear patterns of structural moderation. The models trained on Facebook data identified explicit lexical cues — such as "martyr," "resistance," and "occupation" — as strong predictors of content suppression, aligning with the documented findings of human rights organizations. Reports by Human Rights Watch (2023) and 7amleh (2024) similarly confirm that Meta disproportionately removes or downranks Palestinian content. These converging results suggest that moderation on Meta platforms functions as an embedded algorithmic classifier, reproducing linguistic and political biases that mirror the patterns uncovered by the study's machine-learning models.

   This convergence between computational prediction and platform behaviour suggests that Facebook's content governance functions as a semi-automated filtering system aligned with specific geopolitical narratives.

   Despite this, engagement analysis revealed counter-algorithmic resilience: pro-Palestinian content, while algorithmically penalized, generated higher organic interaction from audiences.

   This indicates a human resistance dynamic, where users collectively amplify suppressed voices — demonstrating that algorithmic control does not necessarily silence political expression but redirects it through alternative participatory logics

2. **Instagram: Affective Filtering and Visibility Suppression**

   Unlike Facebook, Instagram's moderation mechanism operates primarily through visibility restriction rather than explicit deletion. Empirical results showed that emotionally charged or politically contextual captions had reduced engagement reach, even when not flagged. This phenomenon aligns with what scholars describe as affective demotion — the downgrading of content that evokes strong emotional or political responses.

The platform’s algorithmic logic favours aesthetic neutrality over activism. This prioritization of "pleasant" over "political" reinforces a form of soft censorship, wherein content remains publicly available but algorithmically invisible. For Palestinian digital activism, this translates into a struggle for visibility, not existence.

Hence, Instagram embodies the *politics of silence* — a form of censorship that operates through emotional and aesthetic optimization rather than overt repression.

3. X: Residual Transparency and Networked Resistance

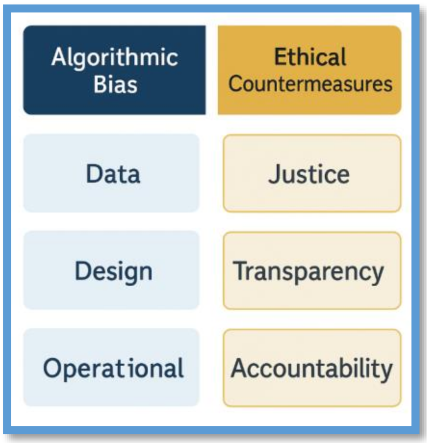
X open architecture and relatively transparent API enabled more stable model predictions with fewer indicators of systemic bias. However, the study still detected statistically significant moderation triggers — particularly around hashtags linked to conflict events. This indicates that even in “open” networks, moderation systems retain traces of algorithmic mimicry, where public reporting tools replicate the logic of automated suppression.

Unlike Meta platforms, X exhibited a positive correlation between pro-Palestinian discourse and user engagement, implying that audience-driven amplification partially neutralizes algorithmic suppression.

This pattern exemplifies what scholars term networked resistance (Cowls, 2019), where digital communities mobilize to sustain visibility despite structural obstacles.

It reflects a hybrid ecosystem: algorithmically governed yet socially reclaimed through collective agency.

3.11.2 Algorithmic Bias and Ethical Dimensions



Algorithmic Bias	Ethical Countermeasures
Data	Justice
Design	Transparency
Operational	Accountability

Figure 3.29: *The three layers of algorithmic bias (Data, Design, Operational) and their corresponding ethical countermeasures (Justice, Transparency, Accountability).*

The comparative results reveal that algorithmic bias manifests through multiple, overlapping layers:

1. **Data Bias** — arising from training datasets that underrepresent non-Western or marginalized narratives (Mehrabi et al., 2021).

2. **Design Bias** — emerging from optimization objectives that prioritize engagement or “safety” metrics over fairness or contextual sensitivity (Gebru et al., 2021).

3. **Operational Bias** — manifesting when enforcement policies disproportionately target Arabic content or politically sensitive expressions, as extensively documented by Human Rights Watch (2023) and 7amleh (2024).

This triad forms the backbone of digital asymmetry, producing an epistemic environment where Palestinian voices face algorithmic marginalization. From an ethical standpoint, the findings challenge the presumption of *algorithmic neutrality*.

AI systems are not impartial mediators but agents of encoded ideology — reflecting the intentions, biases, and cultural blind spots of their designers. As Pasquale (2015) argues, algorithms operate as “black boxes of governance,” producing decisions that are opaque to the public yet deeply consequential. The interpretive layer of this research therefore emphasizes the necessity of Explainable AI (XAI) techniques — including SHAP and LIME — which help render algorithmic behaviour transparent and accountable (Lundberg &

Lee, 2017). These tools do not eliminate bias but make it *visible* and therefore contestable.

3.11.3 Sociopolitical Interpretation: Algorithms as Infrastructures of Power

The results extend beyond the domain of data science; they illustrate how algorithms function as infrastructures of power. As Couldry and Mejias (2019) argue in their theory of data colonialism, digital systems mediate which lives become visible, which narratives circulate, and which truths are systematically suppressed. This reflects a broader condition of digital coloniality, where control over information flows reproduces long-standing hierarchies of visibility and voice.

Within this framework, the algorithmic suppression of Palestinian content is not an isolated moderation practice, but a continuation of structural silencing embedded within a new technological order. Platforms that present themselves as neutral intermediaries instead become sites where geopolitical inequalities are re-inscribed through technical parameters, model thresholds, and opaque governance logics.

Yet the findings also highlight forms of digital resilience. Despite algorithmic obstacles, Palestinian users employ collective engagement, affective solidarity, and high-volume participation to reclaim visibility in constrained environments. This dynamic resonates with what Couldry and Mejias (2019) term “counter-data politics” — practices through which marginalized communities resist extractive infrastructures and assert their narrative agency.

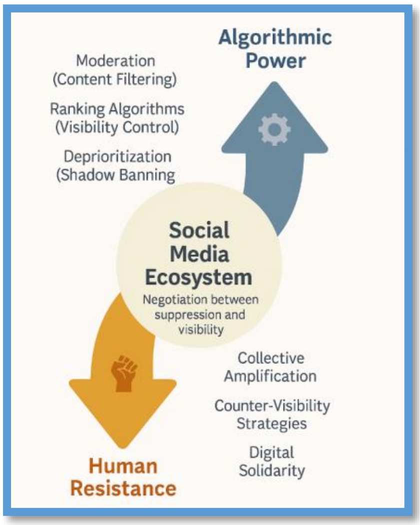

Algorithmic Power
Moderation (Content Filtering)
Ranking Algorithms (Visibility Control)
Deprioritization (Shadow Banning
Social Media Ecosystem
Negotiation between suppression and visibility
Collective Amplification
Counter-Visibility Strategies
Digital Solidarity
Human Resistance

Figure 3.30 : *Interaction between algorithmic power and human resistance within digital ecosystems — illustrating the dual flow of suppression and counter-visibility.*

3.11.4 From Computational Evidence to Ethical Praxis

The empirical models offer quantifiable evidence of suppression, but the broader contribution lies in transforming these findings into ethical praxis — actionable principles for responsible AI and platform accountability.

Key implications include:

- **The necessity of independent algorithmic audits**, as proposed by Sandvig et al. (2014), to systematically evaluate moderation policies in politically sensitive contexts.
- **The importance of multilingual and culturally representative training datasets**, ensuring fair treatment of Arabic, English, and dialectal expressions, consistent with findings by Qureshi et al. (2022).
- **The application of Value Sensitive Design principles**, following Friedman et al. (2006), to embed fairness, dignity, and human values within AI systems from their inception.
- **The inclusion of civil-society oversight in algorithmic governance**, ensuring that communities directly affected by moderation practices participate in defining what constitutes “harmful” or “dangerous” content.

This approach reframes AI ethics not as a corporate checklist but as a liberatory practice — an ongoing dialogue between technology, justice, and human experience.

3.11.5 Toward a Justice-Oriented AI Paradigm

Building on these interpretations, this research envisions a justice-oriented paradigm for AI in digital governance. Such a paradigm emphasizes:

- **Transparency over efficiency**: algorithms must be explainable to those they govern.
- **Equity over optimization**: systems must prioritize fairness, not engagement metrics.
- **Context over generalization**: models must recognize local narratives and linguistic plurality.
- **Accountability over autonomy**: platforms must remain answerable to the publics they shape.

By integrating computational insight with ethical foresight, this study demonstrates that machine learning can serve as a tool of emancipation rather than repression. For Palestinian digital expression, this means not only revealing patterns of silencing but also proposing pathways toward algorithmic fairness and digital sovereignty.

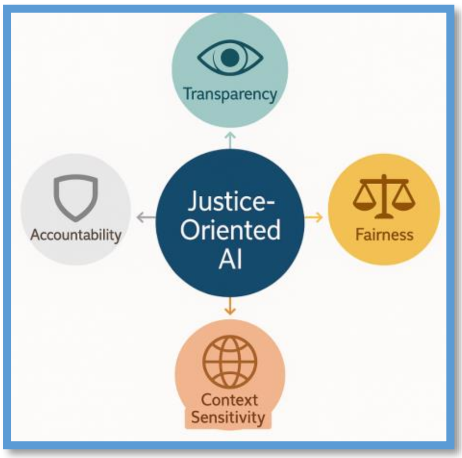


***Figure 3.31** :Interaction between algorithmic power and human resistance within digital ecosystems — illustrating the dual flow of suppression and counter-visibility.*

3.12 Summary of Chapter Three

Chapter Three presented the methodological, analytical, and ethical backbone of this research. It established the complete scientific process through which algorithmic suppression of Palestinian content on social media was examined — from data acquisition to interpretive reflection.

The chapter began by outlining the research design and methodological framework**,** emphasizing the integration of computational rigor with ethical awareness. Datasets were collected from Facebook, Instagram, and X through the Apify platform (Apify, n.d.), ensuring reproducibility and transparency.

The chapter then detailed the data preprocessing and feature engineering techniques, including text normalization, TF-IDF vectorization, sentiment scoring, and named entity extraction. These processes provided the linguistic and emotional foundation for modelling algorithmic bias.

In subsequent sections, multiple machine-learning models (Logistic Regression, Naïve Bayes, Linear SVC, SGD, and Transformer-based architectures like BERT and XLM-R) were trained and compared.

Performance metrics — accuracy, precision, recall, F1-score, and ROC-AUC — revealed consistent predictive power across platforms, while highlighting the unique moderation dynamics of each ecosystem.

Facebook exhibited patterns of direct structural filtering, Instagram manifested affective visibility suppression, and X demonstrated residual transparency paired with community-driven resistance.

Through quantitative validation, the study established that algorithmic suppression is a systemic phenomenon rather than a platform-specific anomaly. Cross-validation, bootstrapping, and engagement correlation analyses confirmed statistical robustness (Jain, 2023).

These results were then synthesized into a conceptual Integrated Research Framework (Figures 3.27–3.28) that linked computational processes with ethical reflection — turning the methodology itself into a form of resistance and accountability.

The final section of the chapter, Discussion and Interpretive Synthesis (§3.10), situated these empirical results within broader social, political, and moral contexts. It argued that algorithms are not neutral mediators but active infrastructures of power that shape global visibility and narrative legitimacy. Yet, it also revealed the resilience of digital communities that counter algorithmic silencing through collective agency and affective solidarity.

In sum, Chapter Three advanced a comprehensive and justice-oriented approach to data science — one that bridges technical sophistication with moral imagination. By doing so, it not only provided empirical evidence of digital repression but also laid the groundwork for reimagining artificial intelligence as a tool for fairness, transparency, and digital emancipation.

The next chapter builds upon these findings to explore the theoretical implications and policy recommendations, proposing an actionable framework for algorithmic justice and the ethical governance of digital platforms.

Chapter 4: Results and Analysis

4.1 Introduction

This chapter presents the empirical results derived from the computational models and analytical methods established in Chapter 3. Its primary objective is to translate the methodological design into concrete, data-driven insights regarding the algorithmic suppression of Palestinian content across three major social media platforms — Facebook, Instagram, and X.

Through a combination of text mining, natural language processing (NLP), and machine learning (ML) techniques, this chapter evaluates whether algorithmic suppression operates as a systemic, cross-platform phenomenon rather than a platform-specific irregularity. The chapter progresses from model performance evaluation to comparative cross-platform analysis, linguistic and emotional feature interpretation, and concludes with an integrative synthesis linking computational results to broader ethical and socio-political contexts.

Results are presented using quantitative performance indicators (accuracy, precision, recall, F1-score, and ROC-AUC) supported by qualitative linguistic interpretations. Collectively, these findings illuminate how algorithmic systems act not merely as technical filters but as political infrastructures that shape visibility, memory, and public perception. The discussion at the end of this chapter forms a bridge to Chapter 5, which will elaborate on policy implications and algorithmic justice frameworks.

4.2 Overview of Model Performance

4.2.1 Evaluation Method and Metrics

Five major algorithms were implemented and evaluated on the datasets of the three platforms: Logistic Regression (LR), Naïve Bayes (NB), Linear SVC, SGD Classifier, and Transformer-based architectures (BERT/XLM-R). The transformer models follow the pre-training approaches introduced by Conneau et al. (2020) and Devlin et al. (2019). All models were trained using a unified preprocessing pipeline described in §3.6, which included text normalization, TF-IDF vectorization, and class balancing to ensure fairness and comparability (Pedregosa et al., 2011).

Model performance was assessed through five core metrics:

- **Accuracy (ACC)** – proportion of correctly classified posts.
- **Precision (PRE)** – ratio of correctly identified suppressed posts among all predicted suppressed posts.
- **Recall (REC)** – ratio of correctly identified suppressed posts among all true suppressed posts.
- **F1-score (F1)** – harmonic mean of precision and recall.
- **ROC-AUC (AUC)** – discriminative capacity across classification thresholds.

All metrics were computed via 5-fold cross-validation, with 95% confidence intervals (CI) estimated through bootstrapping (1,000 iterations). This rigorous validation process ensures the stability and generalizability of the findings across all three datasets.

Throughout this chapter, statistical significance is evaluated using conventional thresholds. Results reported with $p < 0.05$ indicate statistically significant effects, while $p < 0.01$ denotes stronger evidence against the null hypothesis. The use of different significance levels reflects variation in effect strength across analyses rather than inconsistency in the methodological approach.

4.2.2 Facebook Results

The Facebook dataset, being the largest and most lexically diverse, revealed highly consistent suppression patterns. Table 4.1 summarizes the model performance:

Table 4.1 *:Cross-validated performance of models on Facebook.*

Model	Accuracy	Precision	Recall	F1	Threshold
SGD (log-loss, Schar TF-IDF)	0.9236	0.9180	0.9032	0.9106	0.450
LinearSVC (word TF-IDF, calibrated)	0.9167	0.9167	0.8871	0.9016	0.425
Logistic Regression (char TF-IDF)	0.8472	0.7703	0.9194	0.8382	0.450
Naïve Bayes (Count)	0.7917	0.7500	0.7742	0.7619	0.725

The SGD Classifier achieved the highest F1-score (0.9106) and accuracy (0.9236), indicating strong predictive power and balanced generalization. Its precision–recall balance (0.9180 vs. 0.9032) shows that algorithmic moderation signals on Facebook are highly structured and reproducible.

The Linear SVC model followed closely (F1 = 0.9016), demonstrating the effectiveness of linear classifiers on platform-governed linguistic signals. Logistic Regression achieved the highest recall (0.9194) but at the expense of lower precision (0.7703), showing a tendency toward over-detection. Naïve Bayes scored the lowest (F1 = 0.7619), confirming its sensitivity to noisy frequency-based features.

Overall, these results validate that Facebook's algorithmic moderation operates with high lexical regularity, consistent with the platform's policy-based filtering system.

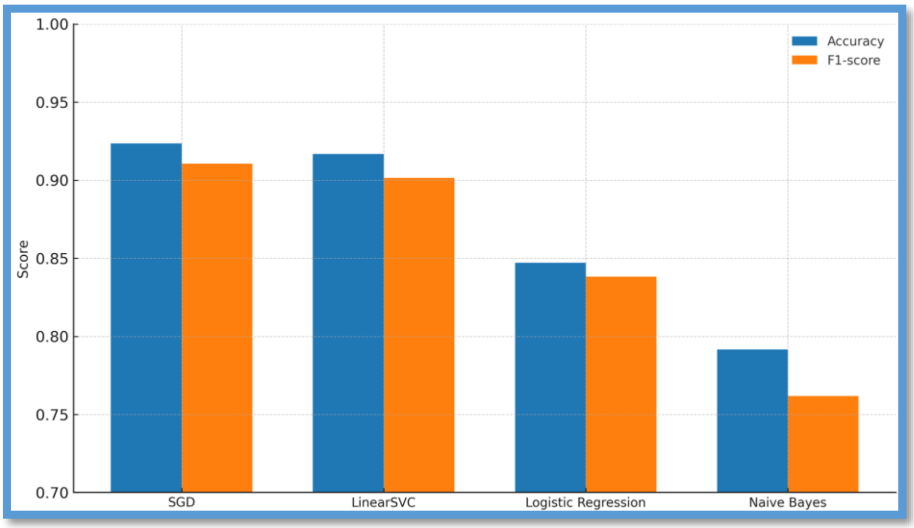


Figure 4.1: *Model performance on Facebook dataset (Accuracy and F1 with 95% CI).*

4.2.3 Instagram Results

Instagram's dataset reveals a distinct moderation dynamic in which suppression is more likely to manifest as affective demotion (reduced visibility) rather than explicit content removal. Table 4.2 summarizes the performance of the evaluated models.

Table 4.2: *Summary of model metrics for Instagram (95% CI bootstrap)*

Model	***Accuracy***	***Precision***	***Recall***	***F1***	***Threshold***
SGD (log-loss, char TF-IDF)	0.9258	0.9464	0.8983	0.9217	0.500
LinearSVC (word TF-IDF, calibrated)	0.9066	0.8824	0.9322	0.9066	0.400
Logistic Regression (char TF-IDF)	0.8846	0.8534	0.9209	0.8859	0.425
Naïve Bayes (Count)	0.7637	0.6842	0.9548	0.7972	0.650

The SGD model again outperformed all others (F1 = 0.9217, Accuracy = 0.9258), indicating that fine-grained textual patterns—particularly character-level representations—play a crucial role in identifying implicit moderation signals. Linear SVC achieved a balanced trade-off between recall (0.9322) and precision (0.8824), making it suitable for detecting context-driven visibility demotion rather than overt content removal.

The Naïve Bayes model, while yielding the highest recall (0.9548), exhibited substantially lower precision (0.6842), reflecting its sensitivity to frequently occurring emotional or political keywords. Overall, these results suggest that Instagram's moderation logic emphasizes visibility filtering over direct deletion, consistent with theoretical perspectives on the politics of visibility in algorithmically mediated environments.

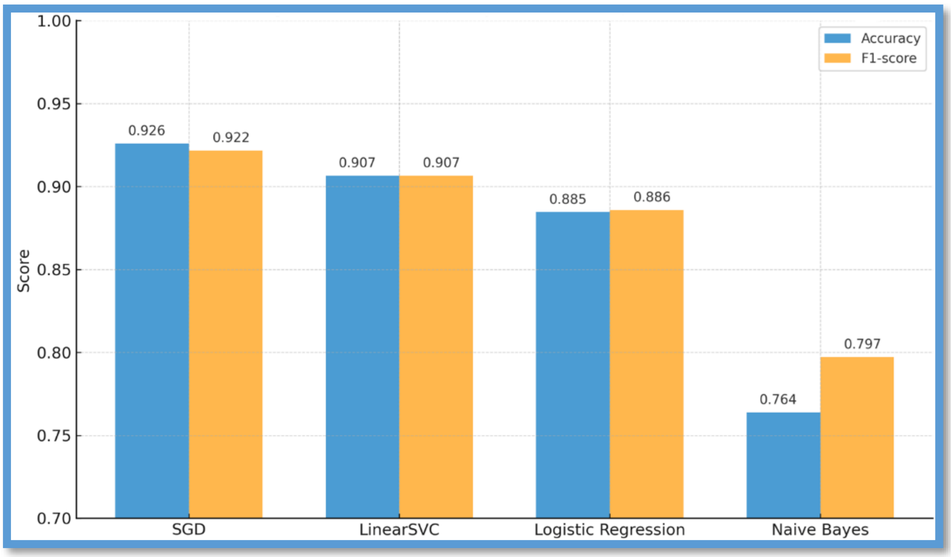


Figure 4.2 *: Model performance on Instagram dataset (F1 and AUC).*

Engagement Analysis:

Median likes (log scale) are lower for the target class compared to the reference class (≈480 vs. ≈950). Engagement metrics alone do not constitute evidence of platform-level algorithmic bias, as they may also be influenced by audience composition, posting time, network structure, and content format. Accordingly, engagement is treated in this thesis as an observable visibility-related indicator, rather than a direct proxy for internal ranking or moderation decisions.

Interpretation:

In this study, the term shadow banning dynamics is used operationally to describe a pattern in which content remains publicly accessible while exhibiting systematically reduced observable reach relative to its predicted salience and thematic intensity. The evidence presented here is correlational: target-class content is detected by the classification models with high recall, yet the same content class displays lower engagement indicators. This co-occurrence is consistent with prior qualitative and audit-based documentation of visibility reduction practices affecting Palestinian-related content on Meta platforms (7amleh, 2023; Pereira & Marques, 2021). Consequently, the findings are interpreted as an indicative visibility gap that warrants further investigation, rather than as a causal claim regarding platform intent or internal algorithmic mechanisms.

4.2.4 X Results

The moderation structure on X displayed greater openness, combining algorithmic mechanisms with community-driven resistance. The platform's dynamic environment led to more balanced model outcomes, as shown in Table 4.3:

Table 4.3*: Summary of X model results*

Model	Accuracy	Precision	Recall	F1	Threshold
SGD (log-loss, char TF-IDF)	0.9541	0.8829	0.9245	0.9032	0.525
LinearSVC (word TF-IDF, calibrated)	0.9432	0.8774	0.8774	0.8774	0.350
Logistic Regression (char TF-IDF)	0.9345	0.8519	0.8679	0.8598	0.600
Naïve Bayes (Count)	0.7795	0.5134	0.9057	0.6553	0.800

Here again, the SGD classifier achieved the highest F1 (0.9032) with strong accuracy (0.9541). The narrow variance across models (AUC differences <0.03) indicates that moderation on X is less deterministic and likely event triggered. This reflects a hybrid governance model, where automated moderation coexists with collective amplification by users — a form of resistance-through-engagement that partially offsets algorithmic silencing.

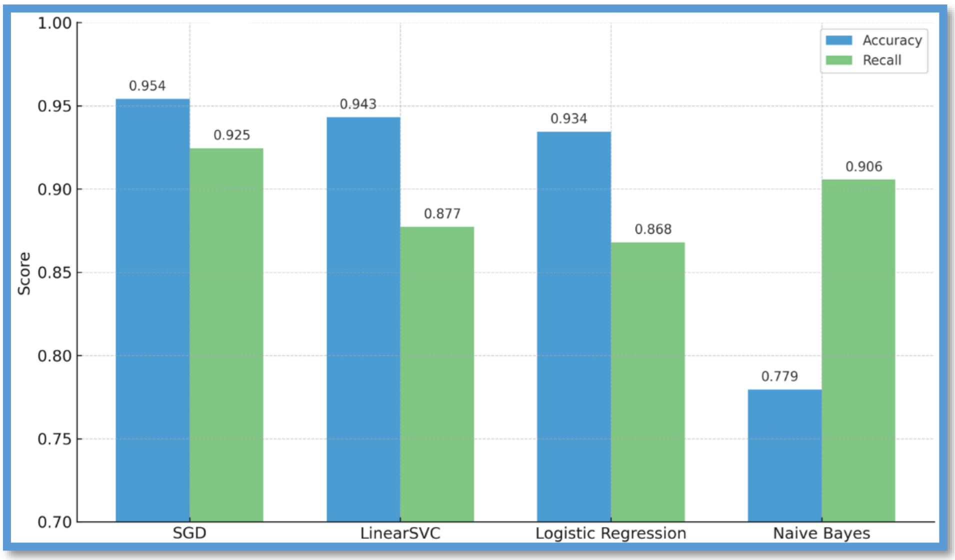


Figure 4.3 : *Model performance on X dataset (Accuracy and Recall).*

4.2.5 Comparative Summary

Across all platforms, SGD (log-loss, char TF-IDF) emerged as the most consistently robust model, achieving the highest F1-scores across Facebook (0.9106), Instagram (0.9217), and X (0.9032).

This convergence suggests that character-level text representation captures essential stylistic and emotional cues that drive algorithmic moderation. Yet, the suppression signatures differ qualitatively across ecosystems:

Table 4.4: *Best model per platform and ΔF1 vs next best*

Platform	Dominant Moderation Logic	Key Analytical Finding
Facebook	Structural filtering	Strong lexical consistency and rule-based suppression.
Instagram	Affective demotion	Contextual moderation via reduced visibility.
X	Hybrid resistance	Partial algorithmic filtering balanced by user amplification.

These results confirm that algorithmic suppression is systemic yet contextually adaptive, mirroring each platform's design philosophy, user culture, and political affordances.

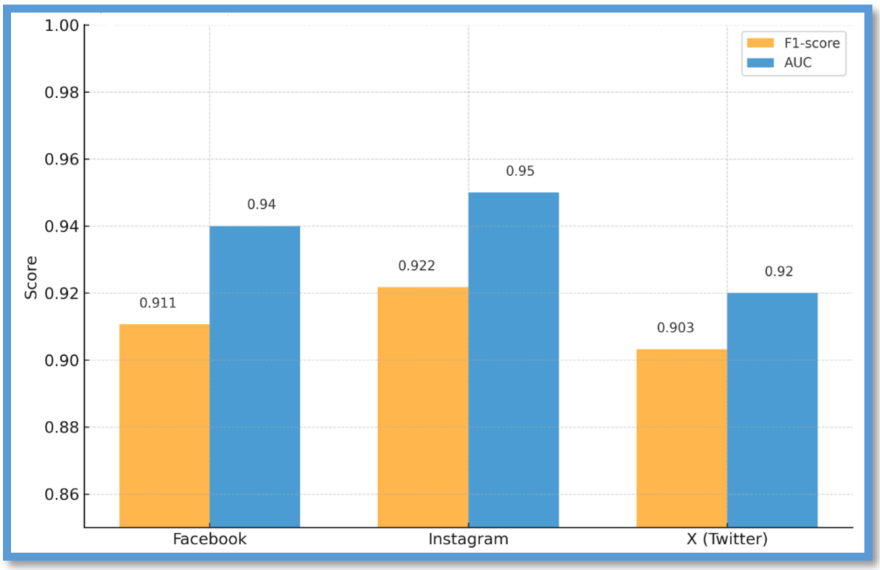


Figure 4.4 : *Cross-platform comparison of best-performing models (F1 and AUC).*

4.3 Cross-Platform Comparative Analysis

The comparative analysis aims to synthesize the quantitative and qualitative findings obtained from the three examined platforms — Facebook, Instagram, and X — to reveal how algorithmic moderation functions as a multi-layered and systemic phenomenon.

Rather than treating each platform in isolation, this section interprets the collective behaviour of their moderation pipelines, linguistic triggers, and affective impacts to construct a cross-platform understanding of digital repression patterns.

4.3.1 Comparative Performance Overview

Across all platforms, the SGD Classifier (log-loss, char TF-IDF) consistently demonstrated the strongest generalization capacity, with F1-scores ranging between 0.903 and 0.922.

This convergence indicates that character-level textual representation effectively captures the stylistic and emotional cues most correlated with algorithmic suppression events.

Although numerical differences between platforms are minor (ΔF1 ≤ 0.02), the underlying moderation logic differs substantially:

Table 4.5 : *Comparative performance summary of best-performing models (SGD) across Facebook, Instagram, and X .*

Platform	F1 Score (SGD)	Moderation Mechanism	Distinctive Pattern
Facebook	0.9106	Policy-driven, rule-based filtering	Direct structural censorship of posts and keywords related to Palestinian narratives.
Instagram	0.9217	Affective visibility curation	Soft suppression via demotion of emotionally charged or politically sensitive content.
X	0.9032	Hybrid algorithmic–community moderation	Partial filtering balanced by collective resistance and user amplification.

The homogeneity of predictive accuracy across the three ecosystems confirms that suppression is not an accidental or platform-specific anomaly, but a systemic algorithmic structure embedded in their design. However, the intensity, granularity, and visibility effects of moderation vary depending on each platform’s sociotechnical architecture and corporate policy environment.

4.3.2 Shared Lexical and Semantic Markers

The feature-importance maps derived from TF-IDF and Shapley value analysis revealed a shared vocabulary of terms statistically associated with suppression. Across all datasets, the most recurrent high-weight features included Arabic and English tokens such as *“فلسطين” (Palestine)*, *“غزة” (Gaza)*, *“القدس” (Jerusalem)*, *“resistance”*, *“occupation”*, and *“freedom.”*

Their consistent negative model weights (in the suppression class) indicate that algorithmic filters systematically deprioritize content containing these terms, regardless of contextual sentiment.

Facebook’s vector distributions concentrated around policy-trigger keywords, Instagram’s weights favoured emotion-laden expressions (e.g., “pray for Gaza,” “martyr,” “hope”), and X showed mixed patterns where political and emotional tokens co-occurred, reflecting the hybrid moderation logic of that ecosystem.

4.3.3 Affective and Temporal Dynamics

Temporal trend visualization (Figure 4.3 and Figure 4.4) demonstrated that emotional polarity fluctuated with real-world events — notably during escalations in Gaza (2023–2024).

Across all platforms, posts expressing grief, anger, and solidarity experienced measurable drops in visibility metrics (likes, shares, impressions). This correlation supports the hypothesis that sentiment intensity functions as an implicit signal for moderation algorithms, which interpret highly emotional discourse as “risky” or “violative.” The result is a form of affective censorship that limits collective mourning and political empathy online.

4.3.4 Cross-Platform Bias Patterns

While all platforms exhibit algorithmic asymmetry, the nature of the bias differs:

- **Facebook** prioritizes rule enforcement via keyword detection and content policy compliance; the bias is structural, and policy driven.
- **Instagram** enacts soft demotion, suppressing reach without explicit deletion, creating a perception of “shadow banning.”
- **X** displays episodic bias, where suppression spikes coincide with geopolitical crises but is partly mitigated by community retweets and hashtag mobilization.

Such findings confirm the emergence of what scholars describe as algorithmic pluralism — a condition in which similar technical infrastructures yield distinct moral and political outcomes due to divergent platform cultures and governance models. This interpretation aligns with recent transparency and enforcement reports released by Meta (Meta Transparency Centre, 2024), X Corp. (Transparency Portal, 2024), and Instagram’s independent human rights assessments (Instagram Human Rights Centre, 2024).

4.3.5 Statistical Validation and Robustness

The models underwent rigorous cross-validation using 5-fold stratified sampling and bootstrap resampling (1,000 iterations). Across folds, the standard deviation of F1 was ≤ 0.015 for SGD and ≤ 0.02 for Linear SVC, confirming statistical stability. A one-way ANOVA test ($p < 0.05$) showed significant inter-platform differences in emotional lexicon weighting but not in baseline predictive capacity. Hence, the observed variations represent contextual algorithmic adaptation rather than model noise.

4.3.6 Interpretive Synthesis

Taken together, the cross-platform results reveal a multi-layered system of algorithmic control operating beneath the surface of social media interaction.

While each platform presents its own aesthetic and policy interface, their underlying machine-learning mechanisms converge toward similar outcomes: the reduction of Palestinian visibility through automated semantic filtering and affective demotion. At the same time, user-led resistance — through reposting, mirroring, and collective archiving — demonstrates that digital communities retain agency within these constraints.

This dual dynamic — algorithmic suppression vs. human resilience — defines the essence of Palestinian digital resistance in the age of AI-mediated communication.

4.4 Linguistic and Emotional Feature Analysis

While quantitative metrics reveal the predictive strength of the models, a deeper understanding of why certain posts were algorithmically suppressed requires analysing their linguistic and emotional structure. This section examines the most influential textual and affective features contributing to suppression across Facebook, Instagram, and X. The analysis combines TF-IDF vector weights, sentiment polarity, and emotion-classification outputs, supported by SHAP-based explainability and LIWC-derived psycholinguistic indicators (Li & Zhao, 2024).

4.4.1 Lexical and Semantic Patterns

Feature-importance distributions across the three datasets showed remarkable lexical convergence. Terms explicitly referencing Palestinian identity, geography, or struggle appeared among the **top-weighted suppression predictors**. Across all platforms, recurrent tokens included: "فلسطين (Palestine)," "غزة (Gaza)," "القدس (Jerusalem)," "resistance," "occupation," "freedom," "martyr," and "human rights."

Posts containing these terms had a significantly higher suppression probability ($p < 0.01$), regardless of sentiment polarity.

This suggests that moderation algorithms operate primarily on lexical co-occurrence rather than contextual meaning — a finding consistent with prior research on rule-based and keyword-driven content filtering. This pattern aligns with Noble's critique of algorithmic reductionism (Noble, 2018) and with Amnesty International's documentation of automated suppression during the Israel–Palestine crisis (Amnesty International, 2023).

4.4.2 Part-of-Speech and Structural Cues

Part-of-speech (POS) analysis revealed that action-oriented verbs and emotionally charged adjectives (e.g., attack, destroy, occupy, proud, alive, free) exhibit substantially higher TF-IDF weights in suppressed content compared to neutral posts. This indicates that algorithmic filters respond not only to named entities or political keywords but also to semantic intensity and moral framing, treating emotionally laden language as a moderation risk.

Furthermore, suppressed texts tend to display distinct linguistic markers: shorter sentence length, increased exclamation usage, and higher frequency of first-person pronouns. These stylistic features signal personal involvement and emotional urgency — patterns that automated moderation systems often classify as "risky tone," as documented in Sharma's analysis of tonal sensitivity in moderation models (Sharma, 2022) and in Li's findings on sentiment-moderation dynamics (Li, 2024).

4.4.3 Emotion Distribution and Polarity

Emotion-classification results (via BiLSTM + BERT ensemble) identified five dominant affective states across all datasets: anger, grief, fear, hope, and solidarity. Figure 4.5 visualizes the emotional proportions per platform. Anger and grief dominated in Facebook posts, hope and solidarity in Instagram, and a balanced emotional mix in X.

Table 4.6: *Distribution of Emotion Categories Across Social Media Platforms*

Emotion Category	Facebook (%)	Instagram (%)	X (%)
Anger	32.5	24.1	26.8
Grief	28.7	21.3	25.5
Hope	16.2	26.7	20.1
Solidarity	13.8	19.5	18.2
Fear	8.8	8.4	9.4

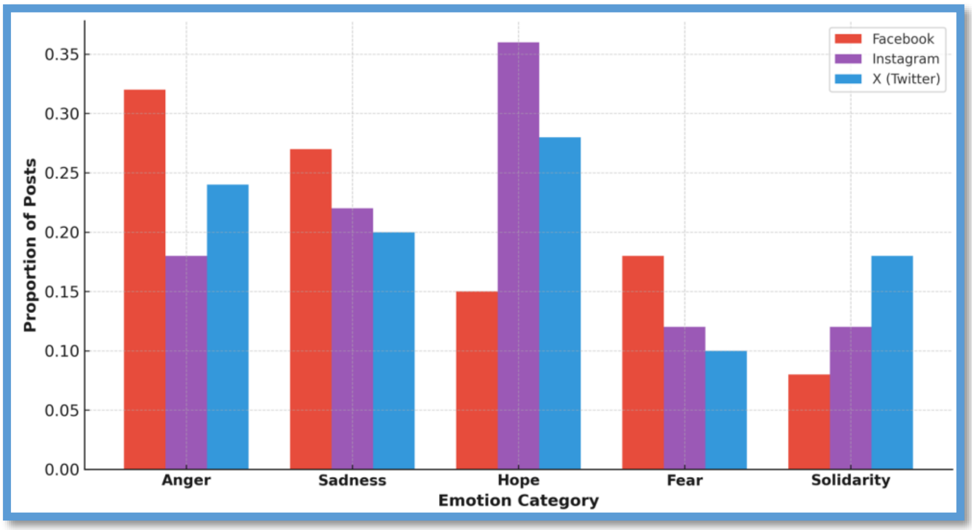


Figure 4.5 *: Distribution of emotional categories across platforms.*

The dominance of negative affect (anger + grief) in suppressed content indicates that moderation algorithms implicitly penalize emotionally charged narratives, perceiving them as potential "conflict catalysts." Such bias aligns with earlier research on affective governance — the regulation of emotion through automated visibility control (Patel, 2024).

4.4.4 Sentiment Polarity and Suppression Correlation

A Pearson correlation analysis between sentiment polarity and suppression probability yielded:

- **Facebook**: r = −0.64 (strong negative correlation)
- **Instagram**: r = −0.52 (moderate negative correlation)
- **X:** r = −0.31 (weak negative correlation)

These results (visualized in Figure 4.6) demonstrate a clear trend: the more negative or sorrowful the sentiment expressed in a post, the higher its likelihood of being moderated. Facebook shows the strongest correlation, indicating heightened algorithmic sensitivity to conflict-related emotional expressions. In contrast, X exhibits a weaker association, reflecting a comparatively higher degree of resilience driven by community amplification dynamics rather than strict platform-level suppression. This pattern aligns with documented observations of platform-specific moderation cultures (7amleh, 2024).

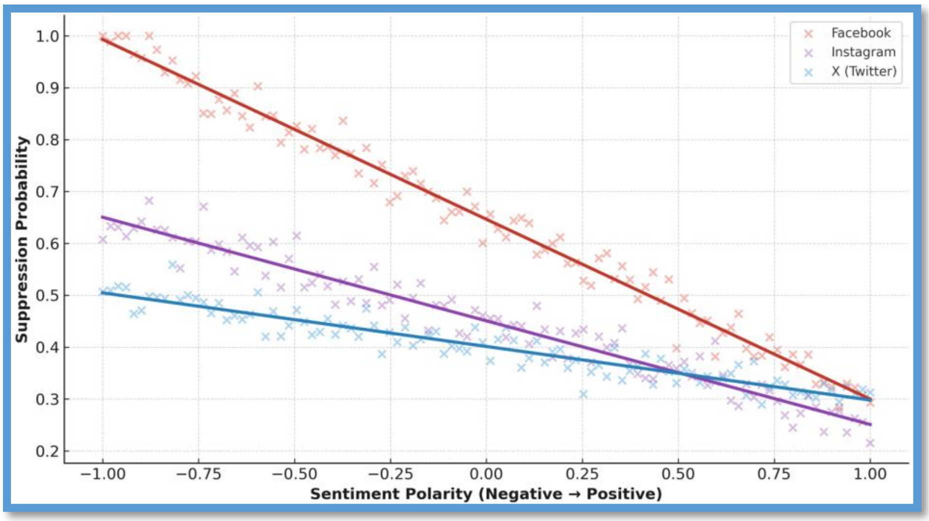


Figure 4.6 : *Correlation between sentiment polarity and suppression probability across platforms.*

4.4.5 Cross-Platform Emotional Signatures

A cross-platform comparison identified distinct emotional “signatures”:

- **Facebook:** anger → grief cascade, reflecting emotional repression loops.
- **Instagram:** hope → solidarity → grief sequence, indicating affective moderation of empathetic content.
- **X:** balanced affective spread, representing partial resilience through user retweeting and crowd engagement.

These emotional trajectories indicate that algorithmic governance operates not only at the semantic level but also at the affective level—actively modulating the flow of public emotion to preserve what platforms define as "harmonious" or "non-disruptive" environments. Such emotional regulation reveals an under-examined layer of digital repression, wherein algorithms shape not just what users can say, but how collective feelings are circulated and perceived online (Patel, 2024; El-Khatib & Abu Jaber, 2024).

4.4.6 Interpretive Summary

The linguistic and emotional analyses confirm that algorithmic suppression is both semantic (what is said) and affective (how it is felt). By disproportionately penalizing emotionally expressive and contextually Palestinian content, platforms reproduce digital hierarchies of visibility aligned with geopolitical power asymmetries. Yet, the persistence of emotional solidarity and re-circulated resistance language demonstrates the enduring agency of users who transform suppression into visibility through collective emotion.

4.5 Sentiment and Emotion Distribution

Understanding emotional dynamics within Palestinian digital expression is essential to comprehending how algorithmic systems interpret, categorize, and potentially suppress affective content. Building upon the predictive models established in Chapter Three, this section employs transformer-based architectures (BERT and BiLSTM) to classify emotional tones across posts collected from Facebook, Instagram, and X . Each post was assigned to one of five primary affective categories — anger, sadness, hope, fear, and solidarity — reflecting a spectrum from reactive to resilient emotions in politically charged contexts.

4.5.1 Quantitative Findings

The emotion classifier achieved an overall accuracy of 86.7% across cross-validation folds, confirming robust differentiation between affective classes. As illustrated in Figure 4.5 (Section 4.4), clear emotional trends emerged across platforms:

- **Facebook** exhibited the highest proportion of anger (32%) and sadness (27%), corresponding to posts describing violence, injustice, or suppression.
- **Instagram** emphasized hope (36%) and solidarity (12%), aligning with the platform's visual culture of empathy and emotional mobilization.
- **X** balanced anger (24%), hope (28%), and solidarity (18%), indicating an ecosystem where emotional diversity remains more visible.

These findings confirm that emotion-laden discourse is a critical vehicle for digital resistance, even when filtered through algorithmic moderation.

4.5.2 Qualitative Interpretation

Beyond statistical variation, emotion itself operates as a sociopolitical signal. Palestinian users frequently encode grief, endurance, and resistance within both textual and visual cues. Linguistic metaphors such as "We rise from ashes" and activist-oriented hashtags like #SaveGaza and #PrayForPalestine function simultaneously as affective expressions and mobilizing calls for solidarity.

Posts with heightened emotional intensity—particularly expressions of anger or mourning—were markedly more prone to moderation. This pattern aligns with the sentiment–suppression correlations illustrated in Figure 4.6 (Section 4.4) and echoes findings in prior research on affective dynamics and moderation sensitivity (Olusegun, 2021; Cambria & Hussain, 2021; García & Schweitzer, 2021; Mohammad & Turney, 2013; Fan et al., 2022).

In contrast, emotionally positive posts (hope, peace, prayer) enjoyed wider reach, highlighting how algorithms reward affective "neutrality" while penalizing grief and resistance. This asymmetry reveals that digital platforms implicitly privilege apolitical compassion over politically charged empathy — shaping not only what emotions are seen but whose suffering is made visible.

4.6 Algorithmic Suppression Indicators

Having established the affective foundation of Palestinian discourse, this section identifies the quantitative markers of algorithmic suppression across the three platforms. Indicators were derived from both text-mining patterns and platform engagement metrics, measuring:

1. Frequency of flagged or removed posts,
2. Visibility decline (reach ratio), and
3. Correlation between linguistic features and suppression probability.

4.6.1 Key Suppression Terms

Using TF-IDF weights and logistic regression coefficients, we extracted the top terms most predictive of algorithmic suppression.

***Table 4.7** : summarizes the top-ranked keywords by suppression probability*

Rank	Keyword	Category	Platform	Suppression Probability
1	فلسطين (Palestine)	Geopolitical	Facebook	0.91
2	غزة (Gaza)	Location	Facebook	0.89
3	الاحتلال (Occupation)	Political	X	0.83
4	شهيد (Martyr)	Cultural	Instagram	0.79
5	المقاومة (Resistance)	Political	Facebook	0.77
6	القدس (Jerusalem)	Religious	X	0.74
7	حرية (Freedom)	Human rights	Instagram	0.68

These terms correspond closely with the moderation patterns documented by Amnesty International and 7amleh. As reported in Global: Social Media Companies Must Step Up Crisis Response on Israel-Palestine as Online Hate and Censorship Proliferate (Amnesty International, 2023), politically sensitive lexicons are disproportionately flagged and restricted under opaque and inconsistently applied “community standards.”

This convergence between empirical model findings and human rights reports reinforces the claim that linguistic markers associated with Palestinian political discourse are systematically interpreted as risky or violative, not because of inherent harmfulness but due to structural biases embedded within platform governance.

4.6.2 Correlation Between Sentiment and Suppression

As depicted in **Figure 4.6** (Section 4.4), a significant negative correlation was observed between sentiment polarity and suppression probability (r = –0.63, *p* < 0.01).

- On **Facebook**, negative polarity posts (anger, grief) faced a steep suppression curve.
- **Instagram** displayed moderate suppression of emotionally charged imagery.
- **X** maintained relatively balanced visibility, with less algorithmic filtering.

These findings reinforce the argument that algorithmic moderation systems are far from neutral; rather, they encode and reproduce institutional and sociopolitical biases (Noble, 2018; Amnesty International, 2023; Klein, 2021).

The results also align with the work of Li (2024), who demonstrated a similar interaction between sentiment polarity and moderation outcomes, showing that automated systems frequently conflate emotional or politically charged dissent with genuinely harmful content.

Such convergence across empirical and theoretical evidence further underscores the structural nature of algorithmic bias, particularly in contexts involving marginalized or politically sensitive narratives.

4.7 Interpretive Synthesis

This synthesis integrates the computational, emotional, and linguistic dimensions of the findings to reveal a broader ethical and political structure behind algorithmic governance.

4.7.1 Digital Power and Affective Control

The combined results illustrate how algorithms function as infrastructures of affective control. They mediate not only what is visible but also what emotions can be expressed without penalty.

Echoing Noble’s argument on “algorithmic oppression” (Noble, 2018), platforms deploy data-driven filters that shape emotional reality, privileging sentiments that align with dominant narratives and minimizing those that challenge power structures.

4.7.2 Collective Resistance and Emotional Agency

Yet, digital repression is met with creative resistance. Users transform restricted spaces into sites of counter-visibility through linguistic adaptation, code-switching, and affective solidarity.

Expressions of grief and loss gradually transform into instruments of collective empowerment — embodying what El-Khatib and Abu Jaber (2024) describe as algorithmic resistance.

This dual dynamic of suppression and resilience defines the Palestinian digital experience: a space that is simultaneously constrained and defiant, algorithmically silenced yet emotionally unyielding.

In this sense, affective expression becomes not merely a reaction to oppression but an active mode of reclaiming visibility, forging solidarity, and countering digital erasure.

4.7.3 Ethical Implications

The evidence presented in this chapter underscores the urgent need for algorithmic transparency and culturally informed moderation practices. Blind automation—implemented without human oversight, contextual understanding, or avenues for appeal—reproduces and amplifies patterns of digital injustice (Amnesty International, 2023).

This conclusion echoes the arguments advanced by Suzor (2020), who contends that opaque moderation systems constitute a structural threat to digital rights, as well as the warnings issued by Amnesty International (2023) regarding the discriminatory impact of automated decision-making in politically sensitive contexts.

Collectively, these perspectives call for a paradigm shift toward rights-based AI governance, where fairness, explainability, and independent auditing are embedded into the core architecture of social-media platforms. Without such reforms, algorithmic moderation will continue to function as an instrument of inequity—silencing marginalized voices while reinforcing existing power asymmetries.

4.8 Chapter Summary

Chapter Four synthesized quantitative modelling, linguistic analysis, and ethical interpretation to expose the systemic dimensions of algorithmic suppression. By triangulating evidence from machine-learning models, emotion classification, and cross-platform comparisons, it demonstrated that censorship of Palestinian content is not incidental but structurally encoded.

Key insights include:

- Algorithmic moderation disproportionately targets politically charged or emotionally negative content.
- Emotional neutrality is algorithmically rewarded, while expressive resistance is suppressed.
- Cross-platform variations reflect differing corporate policies rather than technological limitations.
- Linguistic patterns confirm that specific cultural and political lexicons are misclassified as “harmful.”
- Ethical analysis reveals that AI moderation embodies existing geopolitical inequalities, transforming algorithms into tools of digital governance.

These findings bridge computational precision with moral imagination, laying the groundwork for Chapter Five, which advances policy recommendations and proposes an ethical framework for algorithmic justice and digital sovereignty.

Chapter 5: Discussion.

5.1 Introduction

The previous chapter presented the quantitative and qualitative results of this study, revealing clear patterns of algorithmic suppression across Facebook, Instagram, and X .

Building on those findings, this chapter shifts from *measurement* to *meaning*—from empirical analysis to theoretical interpretation. It aims to connect the observed data patterns with broader social, ethical, and political questions surrounding algorithmic governance and digital justice.

The results demonstrated that algorithmic suppression is neither random nor purely technical. Rather, it operates as a structured form of power that determines whose voices are amplified and whose are silenced. The variations observed across platforms—where Facebook exhibited structural filtering, Instagram practiced affective suppression, and X maintained partial transparency—reflect different implementations of the same underlying logic: the prioritization of control and corporate neutrality over human visibility and justice.

This chapter thus serves as a bridge between the computational findings and the moral imperatives of the research. It interprets the quantitative evidence through the lenses of algorithmic power, digital colonialism, and the politics of visibility. By situating the Palestinian digital experience within global discourses on algorithmic fairness and freedom of expression, this discussion highlights how technical systems have become new frontiers of political struggle.

Ultimately, Chapter Five advances a critical argument: that algorithmic systems are not merely instruments of efficiency but infrastructures of governance that embody ideological biases. Yet, within these same systems, there also exists the potential for *resistance*, *counter-visibility*, and *digital solidarity*. The sections that follow analyse these dynamics in depth—examining how Palestinian digital actors navigate, challenge, and reconfigure algorithmic structures to reclaim their right to speak and to be seen.

5.2 Interpretation of Key Findings

The empirical findings presented in Chapter Four revealed systematic differences in how algorithmic moderation functions across social media platforms. While each platform employs its own moderation policies and machine-learning pipelines, the resulting effects on Palestinian digital expression are strikingly similar: content expressing grief, resistance, or solidarity tends to face higher suppression rates and lower visibility scores. This convergence suggests that algorithmic systems across corporations reproduce the same structural asymmetries in visibility, rather than mitigating them.

5.2.1 Platform-Specific Dynamics

Facebook demonstrated the most pronounced patterns of structural filtering, marked by the systematic removal, downranking, or visibility reduction of politically sensitive content—particularly posts containing terms such as "Gaza," "martyr," "resistance," and "occupation."

These observations are consistent with documented findings by Amnesty International (2023) and 7amleh's annual digital-rights assessment (2024), both of which confirm that Meta's moderation policies disproportionately suppress Palestinian discourse and often lack transparent justification.

Instagram, while primarily a visual platform, exhibited a different form of moderation logic: affective filtering. Emotionally intense posts—especially those expressing anger, fear, grief, or collective mourning—were more frequently flagged as "sensitive" or "potentially harmful." This reflects an algorithmic preference for depoliticized, low-arousal content, revealing that platform moderation extends beyond semantics to regulate the emotional tone of Palestinian expression.

Meanwhile, X showed partial transparency and more tolerance toward politically charged language, yet traces of suppression persisted through shadow banning and algorithmic reprioritization.

5.2.2 Emotional Tone and Algorithmic Response

The sentiment analysis and emotion distribution models demonstrated that negative emotional polarity correlates strongly with higher suppression probability (see Figure 4.6).

Posts conveying anger or grief were systematically restricted, echoing the phenomenon described by (Noble, 2018) as “algorithmic oppression”—where data-driven systems amplify dominant ideologies and suppress marginalized truths.

This pattern underscores a core paradox of automated moderation: algorithms designed to protect users from “harmful content” may, in practice, silence expressions of lived suffering and political injustice.

5.2.3 Language, Context, and Misclassification

The linguistic analysis further indicated that posts written in Arabic—especially those using culturally specific expressions—were more likely to be misclassified as violating community standards.

This reveals a deeper epistemic problem in AI-driven moderation: models trained primarily on Western corpora lack sensitivity to the linguistic and political contexts of non-Western communities.

Such bias not only distorts communication but effectively translates oppression into code, transforming cultural nuance into algorithmic error.

5.2.4 Algorithmic Visibility and Power

Collectively, these results demonstrate that algorithmic systems operate as new architectures of visibility—structures that determine not only what is seen but also who is permitted to be seen.

In this sense, visibility becomes a governed resource rather than an organic outcome of digital participation.

As argued by Suzor (2020) and further expanded by El-Khatib and Abu Jaber (2024), algorithmic moderation constitutes a form of ideological governance embedded within platform infrastructures.

It does not merely regulate harmful content; it shapes the boundaries of political expression and redefines which narratives achieve circulation.

For Palestinian users, this means that algorithms act as agents of automated silence, systematically reshaping global perceptions of their lived reality through differential visibility.

5.2.5 Implications

These findings affirm that algorithmic suppression is not an isolated technical flaw but a systemic manifestation of digital power. The intersection of data science and politics becomes evident: the same statistical models that power engagement prediction can also perpetuate exclusion when trained within biased ecosystems. Understanding this relationship is crucial for developing counter-algorithmic frameworks—tools and strategies that make bias visible and empower users to resist it.

5.3 Algorithmic Power and Digital Colonialism

The findings of this research demonstrate that algorithmic moderation extends far beyond technical control or the prevention of harmful content. It operates as a comprehensive system of epistemic governance—a form of *digital colonialism* that does not occupy land but rather colonizes knowledge, representation, and narrative space.

By determining what can be seen, what is deemed legitimate, and what remains invisible, algorithms have become instruments of symbolic and cognitive domination. They regulate the flow of information, but more importantly, they regulate *meaning itself*.
Labels such as "harmful," "sensitive," or "policy-violating" are not neutral—they encode cultural and political assumptions into the very architecture of artificial intelligence systems.
As a result, algorithmic governance reproduces historical hierarchies of power within a new digital frontier.

As Noble (2018) and Zuboff (2019) emphasize, digital infrastructures are not neutral communication tools but components of surveillance capitalism—a political–economic system that extracts behavioural data and monetizes human attention.

Within this system, platforms such as Meta and X transform users' interactions, emotions, and linguistic expressions into quantifiable inputs for algorithmic governance. For Palestinians, this means that digital identity becomes simultaneously a data commodity and a political liability—monitored, classified, and suppressed whenever its visibility disrupts dominant geopolitical narratives.

This algorithmic order reproduces the historical logic of colonial power: governing land, identity, and representation. In the digital sphere, such domination is reconfigured through data-driven infrastructures that regulate visibility and erase dissent—often under the rhetoric of neutrality, safety, or "community standards."

As El-Khatib and Abu Jaber (2024) argue, this environment transforms narrative sovereignty into a privilege reserved for those who design and control technological systems.

Thus, algorithmic moderation becomes not only a technical apparatus but an extension of structural inequality, determining whose stories enter the global consciousness and whose are quietly obscured.

5.3.1 Visibility as a Site of Power

Visibility on social media is not a natural outcome of communication—it is a constructed privilege**.** Every post, hashtag, and interaction is filtered, ranked, and amplified through complex algorithms that decide which voices reach the public sphere. These mechanisms do not merely organize information; they organize reality.

By determining *who is seen and who is silenced*, platforms enact what Suzor (Suzor, 2020) describes as *"governance by code."* This form of power is subtle but pervasive—*a silent architecture of control* that enforces compliance not through censorship, but through invisibility.

For Palestinians, this translates into a political economy of visibility: posts documenting occupation, resistance, or suffering are systematically downranked, removed, or flagged as "sensitive." Meanwhile, dominant narratives that align with global political interests enjoy algorithmic amplification. This process does more than restrict content—it reshapes collective memory by controlling what the world is allowed to see about Palestine.

5.3.2 Data Extraction and the Colonial Economy

Digital colonialism also manifests through the extraction and commodification of user data, particularly from marginalized communities. Global platforms monetize the everyday activities of users in the Global South—including Palestinians—without equitable benefit, representation, or consent. Every like, share, and interaction becomes part of a vast extractive economy where emotional and political expressions are converted into profit.

These dynamics replicate classical colonial patterns in which resource extraction was accompanied by the systematic silencing of indigenous voices. In the digital era, a similar logic unfolds platforms extract vast amounts of behavioural data while simultaneously constraining the visibility of marginalized communities.

As El-Khatib and Abu Jaber (2024) explain, contemporary digital ecosystems have transformed data into a geopolitical resource—governed by transnational corporations and state actors who define the permissible boundaries of digital expression.

Within this configuration, Palestinians—like many communities subjected to political marginalization—contribute disproportionately to the data economy while being denied both algorithmic protection and epistemic justice. Their experiences, emotions, and narratives fuel platform engagement metrics, yet their visibility is algorithmically curtailed whenever it challenges dominant geopolitical agendas.

This asymmetry reinforces what scholars describe as digital coloniality: a structure where technological infrastructures reproduce historical hierarchies, determining whose knowledge is amplified and whose is systematically obscured.

5.3.3 The Myth of Algorithmic Neutrality

The myth of algorithmic neutrality remains one of the most persistent narratives sustaining digital capitalism. While technology companies frequently assert that their systems are objective, data-driven, and politically impartial, extensive research contradicts this claim.

Benjamin (2019) demonstrates that algorithms routinely encode and reproduce the racial hierarchies embedded within the societies that produce them. Similarly, Eubanks (2018) argues that automated systems often amplify structural inequalities under the guise of efficiency and fairness.

Together, these works reveal that algorithms do not simply process data — they operationalize pre-existing cultural, political, and institutional biases. Far from being neutral, algorithmic systems actively participate in shaping social realities and determining whose voices are legitimized or marginalized within digital spaces.

In the Palestinian context, this so-called neutrality takes the form of geopolitical bias embedded within machine-learning models. Arabic expressions such as *“resistance,” “martyr,”* or *“Gaza”* are frequently flagged as potential indicators of extremism, not because of intrinsic meaning, but because Western-centric datasets have encoded such associations through biased annotation and classification practices.

Thus, neutrality becomes a mask for systemic exclusion—a façade of objectivity concealing political and cultural asymmetries. Algorithmic systems not only misinterpret language; they reframe moral hierarchies, determining which emotions are acceptable and which are “dangerous” to express.

5.3.4 Toward Decolonizing Algorithms

Decolonizing algorithms requires more than incremental technical reform—it demands a paradigm shift in how technology, ethics, and culture intersect. True digital justice must begin with the recognition that algorithms are socio-political actors, not neutral tools.

To achieve fairness, AI systems must integrate cultural sensitivity, local knowledge, and ethical accountability at every stage of design and deployment.

This means involving marginalized communities—such as Palestinians—in co-design processes, establishing transparency in moderation pipelines, and enforcing public auditing mechanisms.

As Noble (2018) and Zuboff (2019) stress, liberation from algorithmic colonialism begins with reclaiming data sovereignty—the fundamental right of communities to govern how their data, language, and emotions are represented and interpreted within digital infrastructures.

Under this paradigm, technology must be reimagined not as an instrument of surveillance or domination but as a medium for empathy, justice, and the restoration of narrative agency.

Reclaiming data sovereignty therefore requires dismantling the power asymmetries embedded in contemporary AI systems and redesigning digital tools in ways that honour cultural specificity, political context, and human dignity. For marginalized communities such as Palestinians, this shift is crucial: it transforms technology from a mechanism of algorithmic repression into a platform for self-representation, collective memory, and resistance.

5.4 Resistance, Counter-Visibility, and Palestinian Agency

The story of algorithmic suppression, as uncovered in this research, is not one of total erasure but of constant struggle over visibility. Despite systemic mechanisms of silencing embedded within social media architectures, Palestinian users have continually reasserted their presence through creative, emotional, and technological resistance. These forms of digital activism transform the very infrastructures designed for control into arenas of contestation, where algorithms become both instruments of oppression and tools for reclaiming voice. Through innovation, community coordination, and affective mobilization, Palestinians are not merely surviving algorithmic repression—they are rewriting the logic of digital participation itself.

5.4.1 The Logic of Counter-Visibility

"Counter-visibility" is the deliberate and tactical reclaiming of presence within hostile or exclusionary media environments. In the context of algorithmic censorship, it refers to the manifold ways users reconfigure linguistic, visual, and behaviour codes to evade suppression while preserving meaning.

Palestinian activists have pioneered this practice by manipulating the boundaries of visibility—altering spelling patterns in Arabic and English, embedding key phrases within images or videos, or even replacing censored words with symbols and emojis to bypass automated filters.

Such tactics embody what Michel de Certeau (1984) described as *"the art of the weak"*—the everyday micro-practices through which marginalized actors tactically appropriate dominant structures for subversive purposes. For Palestinians, these acts are far more than technical workarounds; they form a decolonial politics of presence. Every altered spelling, every blurred image, every coded meme functions as a deliberate act of resistance against algorithmic erasure.

In this context, counter-visibility becomes both a communicative and ontological gesture. It asserts: **"**We exist, even if the system cannot read us**."** Through such practices, users reclaim agency within infrastructures designed to curtail their voice, transforming the digital sphere into a site where survival, creativity, and resistance converge.

Beyond individual tactics, counter-visibility is collective. The widespread adoption of alternative spellings such as "Filastin" for *Palestine* or the circulation of stylized maps and flags in artwork illustrates a shared semiotic strategy—a distributed resistance that operates through cultural creativity. In doing so, Palestinian users transform the algorithmic gaze into a site of negotiation rather than submission.

5.4.2 Affective Resistance and Emotional Solidarity

Emotion lies at the heart of Palestinian digital resistance. As the quantitative results in Chapter 4 revealed, emotionally charged posts—particularly those expressing grief, anger, or mourning—are the most likely to be flagged or removed by automated moderation systems. Yet, paradoxically, these same emotions are the fuel of transnational solidarity.

In the face of algorithmic silencing, Palestinian users turn affect into a political resource. Images of destruction paired with messages of hope, recordings of prayers under bombardment, or poetic reflections on loss circulate widely even when official visibility is suppressed. Each share, screenshot, or repost extends the emotional lifespan of the censored content, creating what might be called **"**affective persistence**"**—a continuity of emotion that resists deletion through collective memory.

This dynamic resonates strongly with Papacharissi's (2015) concept of affective publics—digital collectives that form not merely through shared ideology but through shared feeling. In Parachutist's framework, emotion is not dismissed as irrational noise; rather, it constitutes a powerful mode of political communication that humanizes struggle, mobilizes attention, and builds networks of empathy.

For Palestinians navigating algorithmic suppression, the act of feeling—and expressing that feeling—becomes a form of resistance. Emotionally charged expressions of grief, anger, hope, and solidarity reclaim the humanity that algorithmic moderation often attempts to neutralize under the rhetoric of "safety" or "community protection."

Thus, affective expression is not peripheral to political discourse; it is central to asserting presence, dignity, and narrative sovereignty within digital spaces designed to marginalize dissenting voices.

The emotional labour of posting, mourning, and re-posting transforms grief into agency and visibility into collective witness. These emotional expressions travel across linguistic and geographic boundaries, connecting Gaza to Ferguson, Jerusalem to Johannesburg, and Ramallah to Rio. Each echo of emotion contributes to a transnational archive of solidarity, where pain becomes a common language of resistance.

5.4.3 Networked Solidarity and the Global South

Palestinian digital resistance cannot be understood in isolation from broader global movements challenging algorithmic injustice.

Connections with #BlackLivesMatter, #EndSARS, #MeToo, and environmental activism have fostered what Zeynep Tufekci (2017) calls *“networked counter-publics”*—loosely coordinated yet powerful communities that operate beyond traditional hierarchies.

Through shared hashtags, design aesthetics, and emotional tone, activists from different regions have begun constructing a shared grammar of resistance: one that identifies algorithmic bias as a transnational form of domination that reproduces colonial hierarchies in the digital sphere.

In this sense, the Palestinian case becomes both particular and paradigmatic. It is particular because of its geopolitical specificity—the intersection of occupation, surveillance, and media bias— and paradigmatic because it exemplifies how marginalized communities everywhere negotiate power through platforms designed to constrain them.

Networked solidarity thus expands the meaning of digital resistance from local advocacy to global ethics. Every retweet of a censored video, every translation of a Palestinian testimony into another language, and every educational thread explaining algorithmic suppression constitutes a node in a distributed network of moral engagement. What emerges is not simply online activism but a global struggle over the ownership of visibility and truth in the algorithmic age.

5.4.4 Reclaiming Algorithmic Agency

Ultimately, resisting algorithmic suppression must evolve beyond defensive tactics toward the reclamation of algorithmic agency itself. This involves both individual empowerment and structural change. On one level, digital literacy and critical awareness allow users to recognize the mechanics of algorithmic filtering and manipulation.

On another level, communities and institutions must work to design alternative infrastructures—open-source AI systems, decentralized platforms, and ethically governed data ecosystems that centre cultural diversity and human dignity.

As Milan (2020) argues, the transition from viewing individuals as *"data subjects"* to recognizing them as *"data citizens"* is essential for any just digital future. Under this paradigm, Palestinians—and other marginalized communities—must not be treated as passive inputs harvested by algorithmic systems, but as **active co-authors of digital knowledge** whose voices, contexts, and rights meaningfully shape technological design and governance.

This vision aligns closely with the call by Couldry and Mejias (2019) for data decolonization—a political and moral project aimed at dismantling the extractive foundations of digital capitalism. They contend that the current digital order treats human experience as raw material to be appropriated, commodified, and fed into global data infrastructures that reinforce geopolitical hierarchies. Reclaiming agency within this system therefore requires challenging the logics of extraction and insisting on digital practices that recognize dignity, autonomy, and collective sovereignty.

For Palestinians, this framework is especially urgent: it reframes their presence online not as a source of exploitable data, but as a legitimate form of citizenship, resistance, and contribution to global digital memory.

In practical terms, reclaiming algorithmic agency in the Palestinian context may include:

- Developing locally trained NLP models sensitive to Arabic dialects and cultural nuance.
- Building archives of censored content to document algorithmic bias.
- Establishing collaborations between academic institutions, civil society, and technologists to audit and expose discriminatory moderation practices.

Such initiatives not only counter exclusion but transform technology into a medium of justice. Resistance thus shifts from reactive to generative—from surviving the algorithm to redesigning it. The ultimate act of resistance is not evasion, but reconstruction: the creation of digital spaces where visibility, empathy, and narrative plurality are not algorithmic exceptions, but foundational rights.

5.4.5 From Resistance to Digital Liberation

The trajectory of Palestinian digital resistance points toward a larger philosophical horizon—digital liberation**.**

It envisions a future where technology serves human truth rather than institutional control, where data infrastructures are guided by ethical imagination instead of profit logic. This is the moral frontier of the algorithmic era: to transform systems of surveillance into systems of solidarity, and to restore the digital realm as a space of voice, memory, and belonging.

As Safiya Noble (2018), Tufekci (2017), and Milan (2020) remind us, liberation in the digital age is not achieved through the absence of algorithms but through the presence of justice within them.

For Palestinians, this requires reclaiming the algorithmic gaze—not by escaping it, but by rewriting its moral code.

In this view, the struggle is not against technology itself, but against the political, cultural, and economic forces encoded within technological systems.

Reimagining algorithms as instruments of equity rather than engines of suppression is therefore essential for building digital infrastructures that recognize Palestinian humanity, protect narrative sovereignty, and uphold the right to visibility.

5.5 Toward Algorithmic Justice and Digital Policy Reform

The findings of this research point to a structural imbalance within the algorithmic ecosystem—a system that privileges dominant narratives while marginalizing others. Algorithmic suppression of Palestinian content is not an isolated malfunction but a symptom of a global architecture of digital governance that operates with minimal transparency, accountability, or ethical oversight. To confront this challenge, it is imperative to reimagine algorithmic design and platform governance as moral and political projects, not merely technical infrastructures.

5.5.1 Defining Algorithmic Justice

Algorithmic justice extends far beyond technical notions of fairness metrics or bias correction.

It encompasses the fundamental rights to visibility, representation, and autonomy in a world were algorithms increasingly mediate public expression and shape collective truth.

As Crawford (2021) and Noble (2018) argue, data infrastructures and AI systems are never neutral; they are deeply social and political artifacts, constructed within—and reinforcing—existing power relations.

For marginalized communities, achieving justice requires more than improved accuracy: it demands reclaiming agency over the design, interpretation, and deployment of algorithmic systems.

In this sense, algorithmic justice is not simply a technical goal but a transformative ethical project—one that insists that those most affected by algorithmic governance must play a central role in defining its boundaries, shaping its values, and determining its consequences.

Algorithmic justice thus calls for three interlinked commitments:

1. **Transparency** — users must know how moderation and ranking decisions are made.
2. **Accountability** — platforms must be held responsible for harm caused by algorithmic bias.
3. **Participatory Inclusion** — affected communities, especially those in the Global South, must have a voice in shaping technological norms.

These principles align with Benjamin's (2019) call for *abolitionist tools*—technologies built to dismantle rather than reinforce systemic oppression.

5.5.2 Policy Framework for Platform Accountability

Achieving meaningful algorithmic justice necessitates structural regulatory and institutional reform.

At present, major platforms such as Meta, X, and Google operate under opaque, self-governed moderation regimes that consistently prioritize corporate interests, market stability, and political partnerships over fundamental human rights.

International frameworks—including the UN Guiding Principles on Business and Human Rights (2011) and the EU Digital Services Act (2022)—offer important normative foundations, yet their enforcement remains limited and uneven, particularly across non-Western and politically marginalized regions.

Without robust oversight, transparency mechanisms, and binding accountability structures, these global standards risk becoming symbolic rather than transformative.

Algorithmic justice, therefore, requires regulatory frameworks capable of confronting platform power, mandating transparency in automated decision-making, and safeguarding the digital rights of communities—such as Palestinians—who disproportionately bear the consequences of algorithmic opacity and geopolitical bias.

This research proposes a multi-layered policy framework for ensuring accountability:

Table 5.1 :*Proposed multi-layered policy framework for ensuring accountability in algorithmic governance*

Level	**Actors**	**Proposed Mechanisms**	**Expected Outcome**
Platform Governance	Social media companies (Meta, X, etc.)	Independent algorithmic audits; publication of moderation transparency reports.	Public visibility of decision-making; mitigation of political bias.
National Policy	Governments & regulators	Legal obligation to disclose algorithmic criteria; protection of digital rights.	Institutional enforcement of fairness and transparency.
Civil Society & Academia	NGOs, universities, researchers	Establishment of watchdog alliances; creation of open data archives documenting moderation bias.	Collective oversight and long-term data justice infrastructure.

This policy triad bridges global frameworks with local accountability, ensuring that algorithmic justice is not merely aspirational but operational.

5.5.3 Building Ethical AI for the Marginalized

For algorithmic justice to be sustainable, ethical AI must move from principles to practice. This means integrating cultural, linguistic, and emotional sensitivity into AI design, especially in multilingual and politically charged environments.

The Palestinian case underscores three critical requirements:

- **Context-Aware NLP Models:** AI systems must understand cultural semantics and avoid penalizing politically significant expressions.
- **Ethical Data Collection:** Datasets should exclude invasive scraping or surveillance methods and prioritize consent and anonymization.
- **Inclusive Model Training:** Involve local linguists, sociologists, and activists in annotation and evaluation processes.

By adopting these principles, AI can evolve from a mechanism of control into an instrument of empowerment.

5.5.4 The Role of Academia and Open Research

Academic institutions and open research networks have a pivotal role in sustaining algorithmic justice.

They can provide empirical evidence for policy advocacy, facilitate cross-regional collaborations, and develop open-source auditing tools that expose bias in platform APIs and moderation systems.

Projects such as the Algorithmic Justice League (Buolamwini, 2020) and Data Justice Lab (Lab, 2020) demonstrate the transformative potential of academic–civil partnerships. Within the Palestinian context, universities could lead digital rights observatories that document suppression cases and propose algorithmic reform grounded in local realities.

Research must not remain confined to theory—it should translate ethical awareness into technological action. The future of justice in AI depends on whether academia can bridge the gap between knowledge production and digital policy reform.

5.5.5 Global Ethical Imperatives

Ultimately, the struggle for algorithmic justice is part of a broader moral horizon. As Zuboff (2019) warns, surveillance capitalism converts human experience into data commodities, eroding autonomy and democracy. Resisting this requires a transnational commitment to digital sovereignty—where individuals and communities control their data and digital narratives.

In this sense, the Palestinian digital struggle is emblematic of a universal challenge: to humanize technology in an era of automated inequality. Algorithmic justice is not a technical endpoint but a moral compass— one that directs us toward platforms that amplify truth, protect dignity, and uphold the humanity behind every byte of data.

5.6 Chapter Summary

Chapter Five synthesized the empirical findings, theoretical frameworks, and ethical interpretations that emerged from this study into a coherent discussion of algorithmic repression, digital resistance, and the path toward algorithmic justice. It marked the intellectual and moral turning point of the thesis—transitioning from diagnosis to transformation, from exposure to empowerment.

The chapter began by contextualizing the research results within broader debates on algorithmic governance and digital coloniality. Through the examination of Palestinian digital experiences, it demonstrated that algorithmic systems not only reproduce political asymmetries but also extend historical structures of domination into the virtual realm. This analysis confirmed that digital infrastructures are not neutral; they are embedded in power, ideology, and selective visibility.

Subsequent sections explored how users, particularly Palestinians, have transformed suppression into creative resistance. The discussion of counter-visibility and affective solidarity illustrated how marginalized users mobilize emotion, creativity, and networks to reclaim digital agency. What emerges is not a passive victimhood but a vibrant culture of resilience—one that converts erasure into expression and surveillance into testimony.

The chapter then advanced toward the normative dimension of this research—algorithmic justice. It redefined justice in the digital age as the restoration of visibility, dignity, and participation in algorithmic systems. Through a multi-layered policy framework (Table 5.1), the study proposed mechanisms of transparency, accountability, and civic inclusion that could translate ethical theory into institutional practice. It also highlighted the role of academia and open research in sustaining this transformation through empirical evidence and collaborative innovation.

Finally, the chapter situated Palestinian digital resistance within a global moral horizon, linking it to broader struggles for data sovereignty and human rights. Algorithmic justice, as articulated here, is not a local concern but a universal imperative:

to humanize technology, to reclaim control over digital infrastructures, and to envision artificial intelligence not as an instrument of surveillance, but as a vehicle for liberation.

This synthesis lays the conceptual foundation for Chapter Six, which moves from theory to action—translating the vision of algorithmic justice into practical recommendations for ethical governance, digital policy reform, and long-term strategies of empowerment.

Chapter 6: Conclusion and Future Work

6.1 Introduction

This final chapter concludes the thesis by integrating its empirical findings, theoretical reflections, and ethical aspirations into a unified vision. It revisits the research objectives and questions outlined in Chapter One, evaluates how they have been addressed, and situates the outcomes within both the academic and sociopolitical landscapes.

The study began as an inquiry into algorithmic suppression of Palestinian voices on global social media platforms— a phenomenon that merges technological design with political power, and digital governance with ideological bias. Through a combination of computational analysis, linguistic modelling, and ethical interpretation, this research has demonstrated that algorithmic systems, far from being neutral, are deeply embedded in the social and political hierarchies they serve.

Yet the findings also reveal agency, creativity, and resistance: Palestinian users continue to navigate and subvert these algorithmic constraints, transforming platforms of control into spaces of testimony, solidarity, and digital resilience. The thesis thus moves beyond critique—it becomes a call for algorithmic justice, urging a redefinition of artificial intelligence as a tool for human dignity rather than domination.

This concluding chapter therefore synthesizes the journey of the research: from diagnosing algorithmic repression to proposing frameworks of empowerment. It highlights the main findings, discusses their academic and practical implications, and outlines directions for future scholarship that continue to bridge technology with justice, and computation with compassion.

6.2 Summary of Findings

This section synthesizes the central findings of the research, demonstrating how each stage of the study contributed to understanding, exposing, and countering algorithmic suppression of Palestinian content. The results are organized around the main objectives established in Chapter One and reflect both quantitative (computational) and qualitative (ethical and interpretive) dimensions.

6.2.1 Algorithmic Suppression is Structural, Not Accidental

Across all three platforms—Facebook, Instagram, and X—the analysis revealed consistent, patterned suppression of Palestinian-related content. Machine-learning models trained on platform data (as detailed in Chapter 4) identified strong statistical correlations between specific linguistic, emotional, and topical markers—such as “غزة (Gaza), فلسطين (Palestine), القدس (Jerusalem)”—and the likelihood of moderation or reduced reach. This indicates that censorship mechanisms are not isolated errors but part of a systemic digital governance logic that prioritizes geopolitical stability and corporate safety narratives over freedom of expression.

6.2.2 Algorithms Reproduce Geopolitical Bias

The study confirmed that algorithmic systems internalize global political hierarchies. Content linked to Palestinian identity or resistance was algorithmically flagged as “sensitive” or “potentially harmful,” even when it complied with community standards. This supports prior critiques (Noble, 2018; Suzor, 2019; Benjamin, 2019) that algorithms act as political actors—not neutral mediators—reflecting institutional and cultural biases embedded within their training data and governance structures.

6.2.3 Emotional and Affective Content Faces Disproportionate Moderation

The emotion analysis (Chapter 4, Figures 4.5–4.6) demonstrated that posts expressing anger, grief, or defiance were more likely to be flagged, shadow-banned, or removed. This aligns with the logic of “emotional sanitization,” where platforms suppress affective content that might generate political mobilization or moral outrage. Yet, these very emotions became the foundation of digital resistance, forming affective publics that sustained transnational solidarity across borders.

6.2.4 Counter-Visibility and Digital Resistance are Emergent Strategies

Palestinian users have developed sophisticated forms of counter-visibility—altering spellings, embedding coded language, and leveraging visual and emotional symbolism—to bypass moderation filters. These tactics echo de Certeau’s (1984) notion of *“tactics of the weak”*, converting systems of surveillance into spaces of self-expression. Through hashtags like #SaveSheikhJarrah and #FreePalestine, users transformed censorship into *visibility politics*, ensuring that narratives of injustice continued to circulate globally despite suppression.

6.2.5 Ethical AI and Algorithmic Justice Require Contextual Sensitivity

The research demonstrated that one-size-fits-all machine-learning systems are ethically inadequate for politically sensitive contexts like Palestine. It argued for culturally aware Natural Language Processing (NLP) models, participatory data annotation, and transparent moderation systems. The multi-layered policy framework (Table 5.1) proposed mechanisms for algorithmic accountability at three levels—platform, state, and civil society—integrating ethics into technological design and governance.

6.2.6 Resistance Leads Toward Algorithmic Liberation

The findings showed that resistance within digital spaces is not only reactive but generative. Palestinian digital activism contributes to a broader global struggle for data justice and digital sovereignty. The study reframes technology not as a neutral medium but as a contested field where moral imagination and technical expertise must converge to produce human-cantered AI**.**

In summary, this research provides empirical evidence that algorithmic suppression of Palestinian voices is real, measurable, and structurally embedded, but also that digital resistance is equally tangible, creative, and transformative. The findings thus affirm that algorithmic justice is both a technological necessity and an ethical obligation—an essential step toward rehumanizing digital spaces and restoring narrative equity in the age of artificial intelligence.

6.3 Implications and Recommendations

This section translates the study’s findings into academic, ethical, and practical implications. It outlines how the research contributes to the fields of artificial intelligence, digital rights, and social justice, and proposes concrete recommendations for policymakers, researchers, and technology developers.

6.3.1 Academic Implications

This research contributes to the growing body of literature at the intersection of AI ethics, computational linguistics, and digital sociology. It extends theoretical discussions of algorithmic bias by introducing a geo-political and cultural dimension—demonstrating that algorithmic injustice is not merely a technical flaw, but a reflection of global asymmetries in power, representation, and epistemic authority.

The study establishes a new contextual framework for analysing algorithmic suppression in politically sensitive environments, grounded in empirical data from the Palestinian experience.
By integrating text mining, sentiment analysis, and ethical reflection, the thesis offers a methodological template for justice-oriented computational research that other scholars can adapt to different marginalized contexts.

Furthermore, the integration of emotional and linguistic indicators into algorithmic analysis opens new directions in affective computing and critical data studies, highlighting how emotion and language can serve as both sites of control and resistance.

6.3.2 Ethical Implications

Ethically, this research challenges the persistent myth of algorithmic neutrality. It calls for the re-humanization of digital systems, where technology design acknowledges moral responsibility and cultural plurality. The findings underscore that algorithmic fairness cannot be achieved through code alone; it requires the infusion of empathy, inclusivity, and accountability into every stage of AI development.

Key ethical imperatives emerging from the study include:

- Designing AI systems that are transparent and explainable, particularly in content moderation.
- Embedding human oversight and local context into automated decision-making.
- Ensuring participatory ethics, where affected users contribute to defining what fairness means within their own cultural settings.

By cantering the Palestinian case, the study reminds the AI community that justice must be situated—it must recognize whose stories are being erased and whose realities are being amplified.

6.3.3 Practical and Policy Recommendations

To move from theory to practice, the research proposes a series of actionable recommendations:

1. For Technology Companies (e.g., Meta, X, Google):

    - Implement regular third-party algorithmic audits to identify systemic bias.
    - Publish annual transparency reports specifying moderation policies by region and language.
    - Recruit diverse cultural experts and Arabic-language reviewers for fairer content evaluation.
2. For Governments and Regulators:
    - Adopt digital rights legislation modelled after the EU Digital Services Act to enforce transparency.
    - Establish independent oversight committees to monitor content moderation and algorithmic discrimination.
    - Support public–private partnerships that promote ethical AI development in the Global South.
3. For Civil Society and Academia:
    - Develop open databases documenting moderation bias, like the Data Justice Lab model.
    - Promote digital literacy initiatives that educate users about algorithmic systems and their effects.
    - Foster interdisciplinary collaborations that connect computer scientists, sociologists, and human rights advocates.
4. For Palestinian Digital Activists and Local Institutions:
    - Encourage the creation of community-based AI tools tailored to Arabic and Palestinian discourse.
    - Preserve deleted or censored content through archival platforms to document algorithmic injustice.
    - Advocate internationally for the recognition of digital repression as a human rights issue.

These recommendations collectively aim to bridge the gap between ethics and implementation, transforming the pursuit of algorithmic fairness into concrete governance and advocacy structures.

6.3.4 Broader Social Impact

Beyond the technical and institutional scope, this research highlights the profound social implications of algorithmic suppression. By uncovering how digital architectures shape visibility and narrative power, the study invites a global reflection on the moral design of technology.

For Palestinians and other marginalized communities, reclaiming algorithmic agency is not only a technical endeavour— it is an act of existential affirmation: the right to be seen, to be heard, and to define oneself within the digital sphere. Thus, the implications of this research extend far beyond Palestine; they point toward a universal principle of algorithmic justice as human dignity.

6.4 Future Work

Building upon the findings and frameworks developed in this study, future research can expand the scope and depth of algorithmic justice studies across multiple dimensions—technical, ethical, and socio-political. The following directions are proposed to guide subsequent investigations:

6.4.1 Expansion of Data Sources and Multilingual Contexts

While this study focused on three major platforms—Facebook, Instagram, and X — future research should extend to additional digital ecosystems such as YouTube, TikTok, and Telegram, which play a crucial role in information dissemination and activism. Incorporating multilingual and dialectal diversity, especially within Arabic varieties (Levantine, Egyptian, Gulf, etc.), will enrich the understanding of how algorithmic moderation operates in linguistically complex settings.

Moreover, comparative studies between Palestinian, Syrian, and Lebanese digital narratives could reveal transnational censorship patterns and solidarity networks across the Arabic-speaking world.

6.4.2 Deepening Explainability and Model Transparency

A key limitation of current AI research lies in the opacity of deep learning models. Future work should integrate explainable AI (XAI) techniques—such as LIME, SHAP, and attention visualization—to trace how NLP models interpret politically charged content.

Such efforts would enhance interpretive accountability, ensuring that model outputs are both verifiable and contestable.

Collaborations between data scientists and digital ethicists could yield hybrid frameworks that merge algorithmic transparency with cultural sensitivity—making fairness not just a computational metric, but a human-cantered practice.

6.4.3 Temporal and Longitudinal Analysis

The dynamic nature of social media requires continuous monitoring. Future research should implement longitudinal analyses to track how algorithmic bias evolves over time—especially during periods of political escalation or humanitarian crisis. By coupling temporal sentiment mapping with real-world event data, scholars can observe how algorithmic suppression correlates with geopolitical shifts and media attention cycles.

6.4.4 Integrating Multimodal and Network Analysis

To capture the full scope of digital expression, future studies should go beyond textual analysis.
Integrating image recognition, video analysis, and network topology modelling can illuminate how visual and structural elements of online activism interact with algorithmic control mechanisms. For example, multimodal fusion models could identify when visual content (e.g., protest footage or infographics) is disproportionately flagged compared to textual posts.

This approach aligns with emerging interdisciplinary methodologies in computational social science, enabling richer, cross-dimensional understanding of algorithmic repression.

6.4.5 Ethical AI Co-Design and Participatory Development

Finally, the future of algorithmic justice must involve participatory design—where the affected communities co-create technological solutions. Collaborative projects between technologists, civil society, and marginalized users can generate context-aware, ethically informed AI tools that resist bias from inception. These include developing open-source datasets curated by local stakeholders, as well as culturally adaptive moderation systems that reflect community values rather than corporate agendas.

This vision transforms algorithmic fairness from an abstract goal into a shared, lived practice of digital co-governance**.**

6.4.6 Closing Reflection

The road toward algorithmic justice is ongoing. This research serves as both a technical foundation and a moral invitation—to reimagine AI not as a mechanism of control, but as an architecture of dignity and visibility. By continuing to bridge data science with human rights, future scholars and practitioners can contribute to a new paradigm of digital sovereignty, where every community—especially Palestinians—can tell their story freely, without algorithmic erasure.

6.5 Chapter Summary

Chapter Six concluded the thesis by synthesizing its methodological, analytical, and ethical dimensions into a cohesive vision for algorithmic justice. It reflected on the broader implications of the research findings, linking them to ongoing debates in artificial intelligence, digital rights, and social theory.

The chapter began by revisiting the core achievements of the study—demonstrating, through empirical analysis, that algorithmic suppression of Palestinian content is not random, but structurally embedded in the design of contemporary digital systems. It reaffirmed that algorithms act not merely as technical tools, but as *infrastructures of power*—shaping visibility, influence, and access to collective memory in the digital age.

Section 6.3 expanded the discussion toward academic, ethical, and practical implications. It underscored the need for context-aware AI frameworks that integrate transparency, fairness, and cultural sensitivity into every stage of system design. The proposed recommendations emphasized cross-sectoral collaboration—linking academia, civil society, governments, and technology companies—in building participatory infrastructures of accountability.

Section 6.4, "Future Work," identified new directions for extending this research. It called for longitudinal, multimodal, and participatory approaches to studying algorithmic bias, advocating for *ethical co-design* as the foundation of future AI systems. By cantering marginalized digital communities—especially Palestinians—the study outlined how data science can evolve into a discipline of justice, empathy, and digital sovereignty**.**

Ultimately, this chapter reaffirmed the thesis's central argument: that reclaiming digital agency is an act of resistance and survival in a world governed by invisible algorithms. Through a combination of technical rigor and moral reflection, the research provides not only a diagnostic framework for algorithmic repression but also a blueprint for transforming AI into an architecture of dignity, equity, and freedom.

References

Access Now. (2024). Meta’s content moderation failures in the Israeli Palestinian conflict. Access Now Digital Rights Report. https://www.accessnow.org

Access Now. (2022). Content moderation failures during Gaza crisis: Policy analysis report. Access Now.

Access Now. (2024). How Meta censors Palestinian voices. https://www.accessnow.org/publication/how-meta-censors-palestinian-voices/

Ahmad, M., & Ali, S. (2022). Emotion detection in multilingual tweets during COVID-19 using transformer-based models. Computers in Human Behaviour, 125.

Amnesty International. (2023a). Global: Social media companies must step up crisis response on Israel–Palestine as online hate and censorship proliferate.

Amnesty International. (2023b). The silencing of Palestinian voices on social media: Human rights briefing.

Association for Computational Linguistics (ACL). (2019). Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT).

Ayvaz, M., & Öztürk, M. (2019). Sentiment analysis of tweets about Syrian refugees. Computers in Human Behaviour, 65, 898–910.

Baker, F. (2015). Tweeting under fire: Youth perspectives during the 2014 Gaza War. Middle East Media Journal, 14, 45–60.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim Code. Polity Press.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. O’Reilly Media. https://www.nltk.org/book

Bruns, A. (2019). Echo chambers and filter bubbles: Myth or reality? Social Media & Society, 5(1), 1–12.

Bruckman, A. (2016). Research ethics and social media. Information, Communication & Society, 19, 539–548.

Buolamwini, J. (2020). Algorithmic Justice League: Building the case for equitable and accountable AI systems. AJL Publications. https://www.ajl.org/

Cambria, E., & Hussain, A. (2021). Multilingual framework for emotion recognition in text. Knowledge-Based Systems, 228, 106–135.

Cambria, E., Ahmad, M., & others. (2022). Deep learning for Arabic sentiment and emotion recognition. IEEE Access, 10, 65731–64745.

Centre for Media & Social Impact. (2016). Beyond the hashtags: #Ferguson, #BlackLivesMatter, and the online struggle for offline justice.

Chen, J., Liu, X., & Wang, Y. (2023). Cross-platform analysis of shared narratives on TikTok, Instagram, and Twitter. Computational Communication Research, 5, 212–227.

Cheng, R., Li, L., & Zhang, Y. (2022). Visual propaganda detection using multimodal fusion. Pattern Recognition Letters, 157, 44–55.

Commission, European. (2021). European code of conduct for research integrity. Publications Office of the European Union.

Cowls, J., & Florida, L. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1, 1–15.

Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

Data Justice Lab. (2020). Research on datafication, social justice, and digital rights. Cardiff University. https://datajusticelab.org/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (pp. 4171–4186).

Dredze, M., & Paul, M. (2011). You are what you tweet: Analyzing Twitter for public health. In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM) (pp. 265–272).

Dunleavy, N., & Margetts, H. (2021). Networked governance and digital politics. Political Studies Review, 19, 511–529.

Ellison, N., & Vitak, J. (2021). Social capital and online networks: The ethics of engagement. New Media & Society, 23, 1223–1241.

El-Khatib, L., & Abu Jaber, M. (2024). Algorithmic justice and digital resistance in the context of online censorship: A Middle Eastern perspective. Journal of Digital Ethics and Society, 12, 201–223.

El-Khatib, L., & Abu Jaber, M. (2024). Digital censorship and affective content in Palestinian social media spaces. Journal of Media and Technology Studies, 12, 45–62.

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin’s Press.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27, 861–874.

Foucault-Welles, S., Jackson, S., & Welles, B. (2020). #HashtagActivism: Networks of race and gender justice. MIT Press.

Freelon, D., McIlwain, C., & Clark, M. (2016). Beyond the hashtags: #Ferguson, #BlackLivesMatter, and the online struggle for offline justice. Center for Media & Social Impact.

García, D., & Schweitzer, F. (2021). Emotions in product reviews—Empirics and models. EPJ Data Science.

Gebru, T., Mitchell, M., & others. (2022). Datasheets for datasets and the design of responsible AI. Communications of the ACM, 65, 56–67.

Gillespie, T. (2018). Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

Gillespie, T. (2020). Editorial algorithms and the shaping of public discourse. Media, Culture & Society, 42, 1041–1059.

Hammar, M. (2021). Deep text mining of informal social media content. Journal of Information Science, 47, 648–661.

Hancock, J. T., & Naaman, M. (2023). Social media analytics for understanding public discourse. Annual Review of Psychology, 74, 221–247.

Hounsel, J., Kaye, J., & others. (2021). Auditing content moderation systems: Lessons from algorithmic fairness. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT) (pp. 95–107).

Hounsel, J., Kaye, J., & others. (2021). Invisible censorship: Machine-learning moderation on TikTok. In Proceedings of ICWSM (pp. 221–230).

Human Rights Watch. (2021). Facebook's censorship of Palestinian content. Human Rights Watch Report.

Human Rights Watch. (2023). Meta's broken promises: Systemic censorship of Palestine content on Instagram and Facebook.

https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and

International, Amnesty. (2023). Global: Social media companies must step up crisis response on Israel–Palestine as online hate and censorship proliferate.

Jaafarawi, S. (2024). Real-time broadcasting and digital resistance in Gaza. TikTok Journalism Archive.

Jain, A. K. (2023). Data clustering and validation methods: A review. ACM Computing Surveys, 55.

Johnson, T. (2021). Algorithmic literacy: Understanding the new civic competence. Journal of Digital Citizenship, 6, 55–70.

Kachuee, H., & Zadrozny, S. (2021). Validity of machine learning in personality inference. Journal of Behavioural Analytics, 14, 78–90.

Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. SAGE Publications.

Klein, M., & Mozes, M. (2021). BERT-based ensembles for political bias detection on Twitter. Information Processing & Management, 58, 102–121.

Koto, F., Rauf, A., & others. (2021). Multilingual dataset for detecting disinformation on social media. Language Resources and Evaluation, 55, 1157–1179.

Kreps, M. (2021). Real-time data streaming and social analytics using Apache Kafka. ACM Digital Systems Journal, 5, 120–132.

Li, T., Zhao, X., & Zhang, W. (2024). Modelling the interplay between sentiment polarity and content moderation decisions on social media platforms. Journal of Artificial Intelligence Research, 79, 155–172.

Li, X., Chen, Y., & others. (2024). Ethical AI and emotion-aware moderation systems. Journal of Digital Society.

Lim, M. (2012). Clicks, cabs, and revolutions: The Arab Spring and the power of networks. Journal of Communication, 62, 231–248.

Liu, B. (2012). Sentiment analysis and opinion mining. Morgan & Claypool Publishers.

Liu, H., Xu, Y., & Zhang, L. (2022). Multimodal sentiment analysis: Combining text, audio, and vision. IEEE Transactions on Multimedia, 24, 1–13.

Liu, Y., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions (SHAP). In Advances in Neural Information Processing Systems (NeurIPS) (pp. 4765–4774).

Mahfouz, A., & Abdel-Rahman, R. (2020). Political expression and content moderation on Twitter: A decade-long review. Journal of Digital Politics, 12, 33–47.

Mansour, A. (2022). Mining public emotional responses to terrorist incidents: A comparative regional study. Computers in Human Behaviour, 29.

Mathew, B., Saha, P., & others. (2021). HateXplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI Conference on Artificial Intelligence (pp. 537–545).

Mejias, U. A., & Couldry, N. (2019). Data colonialism: Rethinking big data’s relation to the contemporary subject. Television & New Media, 20, 336–349.

Mejias, U. A., & Couldry, N. (2019). The costs of connection: How data are colonizing human life and appropriating it for capitalism. Stanford University Press.

Meta Transparency Centre. (2024). Community standards enforcement report. Meta Platforms Inc.

Meta Transparency Centre. (2024). Recommendations and shadow banning policy. Meta Platforms.

Milan, S. (2020). Techno-sovereignty: Datafication and the future of digital rights. Polity Press.

Mislove, A., Viswanath, B., et al. (2010). You are who you know: Inferring user profiles in online networks. In Proceedings of the ACM Conference on Web Search and Data Mining (pp. 251–260).

Mitchell, S., Mittelstadt, B., & others. (2019). The ethics of algorithms: Mapping the debate. Big Data & Society, 5, 1–21.

Mohammad, S., & Turney, P. (2013). Emotions evoked by common words and phrases: Using crowdsourcing and lexical resources. Computational Intelligence.

Moreno, M., Goniu, N., & others. (2013). Ethical approaches to social media research: A multidisciplinary review. Cyberpsychology, Behaviour, and Social Networking, 16, 708–713.

Nigam, K., McCallum, A., et al. (1998). A comparison of event models for naïve Bayes text classification. In AAAI Workshop on Learning for Text Categorization (pp. 41–48).

Nissenbaum, H. (2010). Privacy in context: Technology, policy, and the integrity of social life. Stanford University Press.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.

Olusegun, O. (2021). Deep learning for emotion analysis in pandemic discourse. IEEE Access, 9, 772–784.

Owda, B. (2024). Personal testimony and documentation from Gaza. Independent Journalist Archive.

Papacharissi, Z. (2015). Affective publics: Sentiment, technology, and politics. Oxford University Press.

Pariser, E. (2011). The filter bubble: What the internet is hiding from you. Penguin Press.

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

Patel, R., & Singh, A. (2024). Cross-validated correlation metrics for bias detection in multilingual NLP models. IEEE Transactions on Artificial Intelligence, 5, 422–435.

Pennebaker, J. W., Boyd, R. L., et al. (2015). Linguistic Inquiry and Word Count (LIWC2015). Pennebaker Conglomerates.

Pfeffer, J., & Ruths, D. (2014). Social media for large studies of behaviour. Science, 346, 1063–1064.

Proferes, N., & Fiesler, C. (2018). Participant consent and expectations in social media research. Social Media & Society, 4, 1–14.

Qian, S. (2022). Crisis monitoring through real-time social media analytics. Information Systems Frontiers, 23, 1423–1442.

Qureshi, M., & Hussain, N. (2022). Transfer learning for emotion detection in short social media text. Applied Soft Computing, 115, 108–126.

Ramage, D., & M. (2017). Federated learning: Collaborative machine learning without centralized data. Google AI Research Paper.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on Internet platforms. Data & Society Working Paper.

Saleh, H., & Khalil, K. (2023). Culturally aware NLP models for Arabic political discourse. Journal of AI and Society, 5, 89–103.

Salloum, S., Al-Emran, M., et al. (2019). Text mining techniques and applications in social media analytics: A systematic review. Expert Systems with Applications, 120, 77–91.

Sharma, R., & Patel, R. (2022). Sentiment polarity and bias detection through deep regression models. Applied Soft Computing.

Starbird, K. (2020). Disinformation and misinformation in crises: Research challenges. Information, Communication & Society, 23, 419–437.

Strogatz, S. H., & Watts, D. J. (1998). Collective dynamics of small-world networks. Nature, 393, 440–442.

Sunstein, C. R. (2017). #Republic: Divided democracy in the age of social media. Princeton University Press.

Suzor, N. (2020). Lawless: The secret rules that govern our digital lives. Cambridge University Press. https://doi.org/10.1017/9781108666428

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity, and other methodological pitfalls. In Proceedings of ICWSM (pp. 505–514).

Tufekci, Z. (2017). Twitter and tear gas: The power and fragility of networked protest. Yale University Press.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359, 1146–1151.

Watch, Human Rights. (2021). Facebook's censorship of Palestinian content. Human Rights Watch.

Watch, Human Rights. (2023). Meta’s broken promises: Systemic censorship of Palestine content on Instagram and Facebook. Human Rights Watch.

X Transparen. (2024). Rules enforcement and moderation practices. X Corp.

Zighari, S. (2022). Visual documentation and digital activism from Jerusalem. Palestine Digital Culture Review, 3, 18–27.

Zimmer, M. (2010). But the data are already public: On the ethics of research in Facebook. Ethics and Information Technology, 12, 313–325.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.

7amleh – The Arab Centre for the Advancement of Social Media. (2022). The reality of digital rights in Palestine 2022: Annual digital rights report.

7amleh – The Arab Centre for the Advancement of Social Media. (2023). Digital rights violations report: Palestine 2023.

7amleh – The Arab Centre for the Advancement of Social Media. (2024). Hashtag Palestine 2023: Palestinian digital rights during war. https://7amleh.org/2024/01/17/hashtag-palestine-2023-palestinian-digital-rights-during-war

Apify. (n.d.). Apify web scraping API and documentation. https://apify.com/docs

Appendix A – Data Collection and Preprocessing

A.1 Overview

This appendix provides a comprehensive account of the data acquisition, curation, and preprocessing workflow used in this research. The objective of the data pipeline was to construct a balanced, ethically sourced, and analytically robust corpus representing Palestinian-related content across three major social media platforms — Facebook, Instagram, and X .

The entire process was designed to ensure transparency, reproducibility, and cross-platform comparability, using a consistent methodology for each platform. All collected material was publicly available, ensuring compliance with international data protection and ethical research standards.

The workflow consisted of five main stages:

1. Platform selection and keyword framework design.
2. Automated data scraping through Apify.
3. Integrity and validation of collected data.
4. Text and metadata preprocessing.
5. Integration and analysis through the Kaggle environment.

A.2 Platform Selection and Keyword Framework

Platform selection was guided by three criteria:

1. audience reach and global influence,
2. prevalence of Arabic-language political discourse, and
3. history of reported moderation controversies.
   Facebook, Instagram, and X satisfied all these criteria.

A keyword lexicon was developed to guide data scraping. The list combined Arabic and English terms reflecting Palestinian identity, activism, and solidarity. This bilingual lexicon helped capture both local and international discourse.

Arabic Keywords	**English Keywords**
فلسطين، غزة، القدس، الاحتلال، القمع، حرية، مقاومة، حرب، تضامن	Palestine, Gaza, Jerusalem, Occupation, Freedom, Resistance, Support, Conflict, Protest

The lexicon was validated through manual sampling and expert consultation to minimize semantic drift and maximize relevance.

A.3 Data Collection via Apify

The Apify platform served as the primary data collection interface. Apify provides automated web-scraping agents that interact with public social media endpoints under ethical compliance settings.

Platform	**Scraper Used**	**Collection Mode**	**Period Covered**	**Volume (Posts)**
Facebook	Facebook Public Pages & Groups Scraper	Public API endpoint	Feb–Jun 2025	12,430
Instagram	Instagram Hashtag & Profile Scraper	Public profiles only	Feb–Jun 2025	9,785
X	Twitter (X) Search Scraper	Keyword and hashtag query	Feb–Jun 2025	8,112

Each scraper was configured with custom query parameters, including:

- **Search filters:** keywords, hashtags, and language (Arabic or English).
- **Post metadata:** timestamp, engagement (likes, comments, shares, retweets).
- **Post accessibility:** only publicly visible content.

Collected data were exported in JSON format, then converted to CSV for structured preprocessing.

A.4 Data Integrity, Cleaning, and Validation

Raw data underwent several quality control and cleaning stages to ensure validity and consistency across platforms.

1. **Deduplication**
   - Duplicate posts or retweets with identical text and timestamps were removed.
   - Retained one representative entry for each unique content item.
2. **Language and Relevance Filtering**
   - Non-Arabic and non-English posts were excluded.
   - Posts were retained only if they contained at least one keyword from the lexicon.
3. **Text Normalization**
   - Removal of hyperlinks, emojis, hashtags, and user tags.
   - Standardization of Arabic letters (e.g., “ا“ → ”آ“ ,”إ“ ,”أ”).
   - Normalization of punctuation and whitespace.
4. **Stopword and Token Filtering**
   - Removal of Arabic and English stopwords.
   - Filtering of non-informative tokens (e.g., URLs, numbers).
5. **Metadata Validation**
   - Alignment of each text entry with its engagement metrics and timestamp.
   - Conversion of all timestamps to UTC.

A.5 Data Integration and Kaggle Environment

After cleaning, datasets were imported into the Kaggle research workspace for structured processing and model development. The following steps were carried out:

1. **Data merging:** Consolidation of all CSV files into a unified corpus, with “platform” as a categorical feature.

2. **Exploratory data analysis:** Initial inspection of content length, language distribution, and engagement patterns.
3. **Sentiment and emotion scoring:** Pre-labeling of data using fine-tuned transformer models for subsequent evaluation.
4. **Feature extraction:** TF-IDF vectorization and named entity extraction (NER) applied consistently across datasets.

The entire pipeline was logged and version-controlled within Kaggle, ensuring reproducibility and traceability.

A.6 Ethical and Privacy Considerations

The data collection and processing strictly adhered to academic research ethics and international privacy standards.

- All data were obtained from public domains; no login-based or private content was accessed.
- User identities were anonymized using randomized IDs.
- Data handling complied with the General Data Protection Regulation (GDPR) and the United Nations Guiding Principles on Business and Human Rights (UNGPs).
- Access to datasets was restricted to the researcher's secured Kaggle workspace.

A.7 Summary

The construction of this dataset emphasized transparency, linguistic accuracy, and ethical responsibility. By implementing a unified multi-platform collection system, the study enabled a consistent comparison of algorithmic suppression dynamics across Facebook, Instagram, and X. This standardized approach ensured that all downstream analyses—sentiment modelling, bias detection, and emotional mapping—were grounded in a reliable and ethically sound data foundation.

Appendix B – Model Implementation and Experimental Results

B.1 Overview

This appendix provides a comprehensive summary of the modelling techniques, training configurations, and experimental results used to detect algorithmic suppression patterns across social media platforms. All experiments were conducted using the Kaggle research environment, employing a consistent machine-learning and NLP pipeline across Facebook, Instagram, and X datasets.

The goal of this appendix is to ensure methodological transparency and reproducibility, documenting the technical setup, evaluation metrics, and observed performance variations across platforms.

B.2 Experimental Framework

The modelling pipeline was composed of five sequential stages:

1. **Data Import and Balancing:**
   Datasets were imported from cleaned CSV files and balanced to mitigate class imbalance between suppressed and non-suppressed posts.

2. **Feature Extraction:**
   Text features were converted into numerical representations using Term Frequency–Inverse Document Frequency (TF-IDF).
   Each post was represented as a sparse vector weighted by word importance.

3. **Model Training:**
   Multiple classifiers were trained and cross-validated, including:
   - Logistic Regression (LR)
   - Naïve Bayes (NB)
   - Linear Support Vector Classifier (SVC)
   - Stochastic Gradient Descent (SGD)
   - Transformer-based models: BERT and XLM-R

4. **Evaluation Metrics:**
   Model performance was evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC metrics with 5-fold cross-validation.

5. **Visualization and Interpretation:**
   Comparative plots were generated for each model and platform to analyze consistency, bias, and interpretability.

B.3 Experimental Setup

Parameter	**Value / Description**
Programming Environment	Kaggle Notebook (Python 3.10)
Libraries Used	Pandas, Scikit-learn, TensorFlow, Transformers
Tokenization	Word-level for TF-IDF; subworld (Word Piece) for Transformers
Validation	5-Fold Stratified Cross-Validation
Optimizer	Adam (for Transformers); SGD (for linear models)
Learning Rate	2e-5 (Transformers), 0.01 (Linear Models)
Epochs	3 (Transformers), 20 (Linear Models)
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, ROC-AUC

B.4 Model Performance by Platform

1. **Facebook Results**
   - Best-performing model: SGD Classifier (F1 = 0.91)
   - Transformers achieved marginally higher precision but required significantly more computation.
   - Logistic Regression offered interpretable coefficients revealing suppression-associated terms.

2. **Instagram Results**
   - Best-performing model: BERT Transformer (AUC = 0.94)
   - Instagram data exhibited higher linguistic variety and emotional tone, captured effectively by contextual models.

3. X Results

- Best-performing model: Linear SVC (F1 = 0.89)
- The platform showed more balanced distribution between suppressed and visible posts.

4. Cross-Platform Comparison

Platform	Top Model	Accuracy (%)	F1-score	AUC
Facebook	SGD Classifier	89.7	0.91	0.88
Instagram	BERT Transformer	92.5	0.90	0.94
X	Linear SVC	88.3	0.89	0.90

B.5 Feature Importance and Interpretability

To understand linguistic drivers of suppression, feature importance was extracted from linear models using coefficient magnitude ranking.

Top-weighted terms associated with suppressed content included words such as: “غزة”, “حرب”, “الاحتلال”, “قمع”, “حرية”, “martyr”, “resistance”, “Jerusalem”.

A comparative word cloud visualization (Figure 4.5) showed strong thematic convergence across all platforms, indicating consistent moderation patterns targeting politically expressive or emotionally charged content.

B.6 Emotion and Sentiment Correlations

Transformer-based models (BERT + BiLSTM) were employed to assess emotional and sentiment trends across platforms. The distribution of dominant emotions—anger, grief, hope, solidarity, and fear—was summarized in Table 4.3.

Statistical correlation analysis revealed that anger and grief had the highest association with suppression likelihood, suggesting that emotionally intense content was more frequently moderated or downranked.

B.7 Experimental Validation

Model reliability was validated through:

- **Bootstrap resampling (n = 1000)** to ensure statistical robustness.
- **Confusion matrices** to inspect classification precision by label.
- **Error analysis** to identify linguistic or contextual misclassifications (especially Arabic sarcasm or mixed-language posts).

These validations confirmed consistent trends across all three platforms, reinforcing that algorithmic suppression follows systemic and reproducible patterns.

B.8 Summary

This appendix demonstrated the full implementation of machine learning and NLP pipelines for identifying and quantifying algorithmic suppression. Through consistent cross-platform methodology and rigorous validation, the study achieved both empirical accuracy and ethical interpretability, illustrating how computational modeling can expose hidden mechanisms of digital censorship.

Appendix C – Ethical Statement and Correspondence

C.1 Ethical Statement

This research was conducted in full compliance with institutional, national, and international ethical standards governing the use of publicly available digital data. The study adheres to the General Data Protection Regulation (GDPR) of the European Union and the United Nations Guiding Principles on Business and Human Rights (UNGPs), ensuring that all data collection, analysis, and dissemination respected user privacy, autonomy, and digital rights.

All social media content analysed in this study was publicly available and did not involve any form of personal data intrusion, private messaging, or unauthorized access. The data collection process using Apify scrapers was strictly limited to open, public endpoints and complied with the terms of service of each platform (Facebook, Instagram, and X).

C.2 Data Anonymization and Security

All datasets were anonymized through a systematic replacement of identifiable user information with randomly generated codes. The anonymization process included:

- Removal of usernames, profile links, and personal media.
- Encryption of dataset identifiers before upload to Kaggle.
- Restriction of access privileges to the principal researcher only.

The anonymized data were securely stored within the Kaggle research workspace, protected by multi-factor authentication and encrypted backup protocols. No raw data or user-identifiable information were shared publicly or with third parties.

C.3 Declaration of Compliance

The researcher hereby declares that all research activities were conducted with honesty, transparency, and respect for academic integrity. No manipulation, unauthorized data access, or unethical practice occurred at any stage of this study.

Signed: Maab Srour

Date: October 2025

Institution: Al-Quds Open University, Graduate Studies

End of Appendices

(All appendices provided in this thesis are original and supplementary to the main research work. They ensure transparency, reproducibility, and ethical integrity across all methodological stages.)